Algorithmic Operations Research Vol.3 (2008) 147–164

Discovering Trends in Gene Expression Data Using a Hybrid Evolutionary Algorithm

Stefan Bleuler

Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland

Philip Zimmermann

Institute of Plant Science, ETH Zurich, Switzerland

Markus Friberg

Institute of Computational Science, ETH Zurich, Switzerland

Eckart Zitzler

Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland

Abstract

High-throughput technology has enabled molecular biologists to study genes and gene products of living organisms on a systems level: nowadays, it is possible to measure the activity of thousands of genes in a single experiment. With this type of measurement, one aims at revealing the structure and the dynamics of the underlying genetic regulatory network. In particular, one is interested in identifying groups of genes with shared functions or shared regulatory mechanisms which leads to various challenging optimization problems.

Here, we consider the problem of finding multiple, diverse modules of genes that exhibit similar trends regarding one or several gene expression data sets. We present a hybrid evolutionary algorithm for this task that distinguishes itself from previous approaches in three aspects: (i) a set of diverse modules can be found in a single optimization run, (ii) multiple data sets can be considered simultaneously without mixing the corresponding data, and (iii) the trade-off between available runtime and quality of the generated solution can be set by the user.

Key words: gene expression data, clustering, biclustering, evolutionary computation

1. Introduction

A fundamental goal in systems biology is to understand how genes control cellular processes. Colloquially speaking, genes are the blueprints for proteins, and mRNA is the first intermediate during the production of a protein from the genetically encoded information. The *expression level* of a gene denotes the concentration of the corresponding mRNA molecule and is an indicator for gene activity, or more precisely: for the amount of protein that is currently being produced. Nowadays, the expression levels of thousands of genes, possibly all genes in an organism, can be measured simultaneously in a single experiment using microarrays. By performing series of microarray measurements under different conditions and treatments, one obtains a matrix of gene expression values. Related experiments are usually pooled in one data set, while measurements stemming from diverse environmental settings or different technology platforms, laboratories, etc. are summarized in terms of separate data sets.

Def. 1 *A **gene expression data set** is a real-valued $m \times n$ matrix $E := (e_{i,j})_{m \times n}$ where the element $e_{i,j}$ represents the gene expression value of gene i under experimental condition j. A **collection** $\mathbf{E}$ of l gene expression data sets is a vector $\mathbf{E} = (E^1, E^2, \ldots, E^l)$ where $E^k := (e^k_{i,j})_{m \times n_k}$. The **combined gene expression data set** of a collection $\mathbf{E}$ is the matrix $E^{\mathbf{E}} := (e^{\mathbf{E}}_{i,j})_{m \times n^{\mathbf{E}}}$*

Email: Stefan Bleuler [e-mailxxxx], Philip Zimmermann [e-mailXXXX], Markus Friberg [e-mailXXXX], Eckart Zitzler [e-mailxxxx].

with $n^{\mathbf{E}} := \sum_{1 \leq k \leq l} n_k$ *where the* i*th row is defined as* $(e^1_{i,1}, \ldots, e^1_{i,n_1}, e^2_{i,1}, \ldots, e^2_{i,n_2}, \ldots, e^l_{i,1}, \ldots, e^l_{i,n_l})$.

The identification of groups of genes that participate in similar cellular processes is one of the key issues in analyzing genome wide gene expression measurements. To this end, biologists are looking for a subset of genes which are similarly expressed over a subset of conditions—under the assumption that genes with shared functions or regulatory mechanisms exhibit, in specific situations, similar expression levels. Such a combination of selected genes G and conditions C is denoted as *bicluster*; usually, not a single bicluster, but a diverse set of ideally non-overlapping biclusters, a *biclustering*, is sought.

Def. 2 *Let* $\mathbf{E}$ *be a collection of* l *gene expression data sets. A* ***bicluster*** $\mathbf{B}$ *is a vector* $\mathbf{B} = (G, C_1, C_2, \ldots, C_l)$ *where* G *is a subset of genes* $G \subseteq \{1, \ldots, m\}$ *and* C_k *a subset of conditions in data set* k $C_k \subseteq \{1, \ldots, n_k\}$ *for* $1 \leq k \leq l$*; the set of all possible biclusters is denoted as* $\mathcal{B}$*. A* ***biclustering*** D *is a multi-set of biclusters; the set of all possible biclusterings is denoted as* $\mathcal{D}$*.*

The task of finding biclusters in a collection of gene expression data sets can be formalized as an optimization problem in different ways—depending on the specific biological question and scenario. Here, we consider the discovery of gene groups that represent *trends*: only the order of the expression values matters, but the absolute differences between expression values are not taken into account. Trends are a useful concept especially with the analysis of time course data sets where each matrix column corresponds to a specific point in time under the same environmental conditions and for the same organism. For instance, suppose the expression values of three genes are constantly increasing over the course of time, i.e., they follow the same trend. In this case, the order of the expression levels is the same for all three genes: the first condition represent rank 1, while the last condition stands for the highest rank. Nevertheless, the absolute expression values among the three genes can differ strongly: the values for one gene may drastically go up, while another gene leads to small changes in expression only. That means the similarity of the three genes may be low regarding the absolute expression levels, but perfect with respect to the order of their expression values.

In this paper, we propose an evolutionary algorithm in combination with a greedy heuristic in order to determine a biclustering that is based on the order of the expression values. In contrast to existing approaches, the proposed module identification method

- is capable of identifying a set of diverse biclusters, each following a trend, in a single optimization run;
- allows to operate on multiple data sets simultaneously without the need of mixing data from different experiments;
- enables the user to individually set the trade-off between run-time and solution quality.

This analysis of multiple input data sets leads to problem formulations which are clearly different from existing biclustering approaches based on evolutionary algorithms which all work on a single input matrix only [4,1,8,23,22,2,10,21]. The present work is in part based on a preliminary study reported in [5]. While the basic algorithm is similar, it has been adapted to identify also biclusters which do not extend over all columns and it uses a new variant of the environmental selection. Additionally, the present paper provides an extensive algorithmic Al comparison to two alternative techniques, validates the biological relevance by means of a promoter motif analysis which also allows to determine the effects of mixing multiple data sets and it provides a detailed discussion of exemplary biclusters.

2. Related Work

In the literature, a variety of methods has been proposed and employed for gene module identification. Classical clustering methods such as hierarchical clustering [27] and k-means clustering [17] partition the set of genes into disjoint groups according to the similarity of their expression patterns over *all* conditions of a single data set; that means every gene is contained in exactly one (bi)cluster:

$$\forall i \in \{1, \ldots, m\} : |\{(G, C)\,;\, (G, C) \in D \wedge i \in G\}| = 1$$

and every (bi)cluster includes all conditions:

$$\forall (G, C) \in D : C = \{1, \ldots, n\}$$

Although these approaches have been successfully applied to gene expression data analysis [11], some underlying assumptions do not reflect biological reality: (i) genes may have several functions and therefore may be contained in multiple biclusters, and (ii) certain processes may be active only over some but not all conditions.

In contrast, the concept of biclustering, which goes back to the work of Hartigan [16], overcomes these limitations and focuses on local subpatterns in an arbitrary

data matrix: a bicluster is defined as a subset of the rows and a subset of the columns. Cheng and Church [9] were the first to transfer this concept to the analysis of gene expression data sets, and meanwhile various biclustering methods have been presented in this context, e.g., [18,3,29]. An extensive comparative study demonstrated on both synthetic and real data that popular biclustering approaches do find gene modules that could not be identified by a hierarchical clustering algorithm [24]. Biclustering problems in general are highly complex optimization problems and most approaches use heuristics for identifying good biclusters. Following the first EA for biclustering in [4], a number of studies have demonstrated the usefulness of EAs for solving such problems [1,8,23,22,2,10,21].

Despite the recent advances, there are several open issues. Firstly, most biclustering methods are based on greedy strategies that can be considered as local search methods which are fast but often yield suboptimal results. However, the computation resources needed are often less critical than the quality of the outcome: in comparison to the amount of lab work required to perform the measurements, run-times of several minutes up to a couple of hours may be still acceptable if it can be justified by a substantial improvement in quality. Secondly, many biclustering techniques are designed to find a single bicluster, and they need to be applied iteratively in order to obtain a biclustering. For instance, in [9] found biclusters are simply replaced by randomly chosen values which hinders overlapping biclusters to be identified. Thirdly, existing clustering and biclustering algorithms can only operate on a single data matrix, which implies that multiple data sets need to be mixed ($E^{\mathbf{E}}$) for a combined analysis [11,13,26]. However, often the measured values can be compared more reliably within one data set than between data sets; significant differences have been found in measurements performed by different labs or with different technologies [19,15]. An alternative approach would be to perform a cluster analysis on each data set separately, thus avoiding the problem of mixing. Nevertheless, it is unclear how to combine such results to find groups of genes that are similarly expressed over all data sets since looking for the intersections of the clustering results is usually too restrictive.

3. Optimization Framework

In the following, the proposed approach for discovering trends is detailed. We start with a formal description of the underlying optimization model, before dealing with the algorithms.

3.1. *Model*

The problem of finding one or several biclusters is inherently multi-objective. On the one hand, one is interested in finding large biclusters, i.e., modules containing many genes exhibiting a similar behavior over many experimental conditions. On the other hand, the similarity among the chosen matrix elements is to be maximized, i.e., the biclusters should be *homogeneous*. These criteria are naturally conflicting as the latter is usually the higher the less genes and conditions are involved. We will first formally define these criteria before discussing how to combine them.

As to the size criterion, we simply consider the number of matrix cells associated with a given bicluster.

Def. 3 *Given a data set collection* $\mathbf{E}$*, the* ***size score*** f_{size} *of a bicluster* $\mathbf{B}$ *is defined as the number of contained matrix elements, i.e.,*

$$f_{size}(\mathbf{B}) := |G| \cdot \sum_{1 \le k \le l} |C_k|$$

Since the focus is on trends, the homogeneity criterion is based on the order of the expression values for the selected conditions. In a first step, the expression values are transformed into ranks—for each data set and gene separately. In principle, the rank of a value corresponds to its position in the sorted list of all values for the conditions under consideration; however, here we normalize the ranks to the interval $[0, 1]$ to make data sets with different number of columns comparable.

Def. 4 *Let* $\mathbf{E}$ *be a collection of data sets. The rank of gene* i *at condition* j *for data set* k *and the selection* C *of conditions is given by*

$$rk(i, j, k, C) := 1 + s_< + (s_= + 1)/2$$

with $s_< := |\{e^k_{i,j'}\,;\, e^k_{i,j'} < e^k_{i,j} \ \wedge \ j' \in C\}|$ *and* $s_= := |\{e^k_{i,j'}\,;\, e^k_{i,j'} = e^k_{i,j} \ \wedge \ j' \in C\}|$*. The normalized rank is defined as*

$$nrk(i, j, k, C) := \frac{rk(i, j, k, C) - 1}{n_k - 1}$$

To quantify differences in the order of the expression values for a given bicluster, the rank variance over the selected genes is computed for each selected condition in a second step. Finally, the average rank variance over the conditions gives the *rank-homogeneity score*.

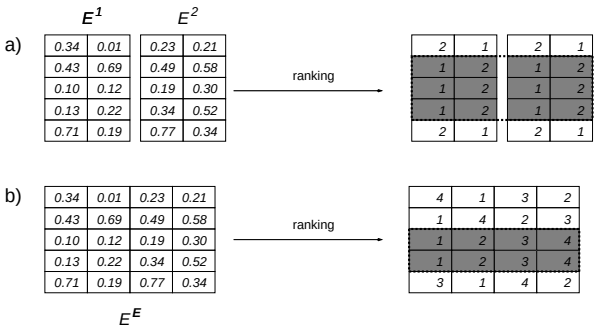

Fig. 1. a) On the left hand side, a collection $\mathbf{E} = (E^1, E^2)$ of two gene expression data sets is shown, on the right hand side, the corresponding expression levels are replaced by their (unnormalized) ranks within each row; the shaded area marks the largest bicluster with $f^1_{hom} = f^2_{hom} = 0$. b) The same is shown for the combined gene expression data set of the collection $\mathbf{E}$; here, the resulting largest bicluster contains fewer genes as an effect of mixing E^1 and E^2.

Def. 5 *The **rank-homogeneity score** f^k_{hom} of a bicluster $\mathbf{B}$ for the kth data set in a collection $\mathbf{E}$ is defined as*

$$f^k_{hom}(\mathbf{B}) := \frac{1}{|C_k|} \sum_{j \in C_k} \left(\frac{1}{|G|} \sum_{i \in G} rd(i,j,k,G,C_k)^2 \right)$$

where the rank deviation rd is

$$rd(i,j,k,G,C) := nrk(i,j,k,C) - \frac{\sum_{i' \in G} nrk(i',j,k,C)}{|G|}$$

It can be easily seen that $f^k_{hom} = 0$ if and only if the ranks for each selected conditions are the same for all selected genes, i.e., $rk(i_1,j,k,C) = rk(i_2,j,k,C)$ for any two genes i_1, i_2 in the bicluster; this is illustrated in Fig. 1 for a collection of two data sets. Furthermore, the score is related to the scoring schemes proposed in [9] and [3]. On the one side, it equals the mean residue score [9] when the ranks and not the absolute expression values are considered. On the other side, it represents a relaxation of the strict order preserving criterion [3] where all genes are required to induce the same order on the expression values over the selected conditions. In [3], it was also shown that the decision problem of whether a data matrix contains a bicluster of given size with rank-homogeneity score of 0 is NP-complete.

Similarly to [9,3,18], the size f_{size} is taken as an objective function and the homogeneity f^k_{hom} is transformed into a constraint in order to resolve the conflicts between the two criteria. In the case of a collection of data sets, for each data set k a separate threshold δ^k can be specified, but due to the normalization of the ranks the same threshold can be used for all k. In addition, we consider a constraint on the number of contained conditions per data set; the reason is that it is usually much harder to find biclusters with a large number of conditions and a few genes only in comparison to biclusters with many genes but only a few conditions.

Def. 6 *The **width** f^k_{width} of a bicluster $\mathbf{B}$ gives the portion of conditions that $\mathbf{B}$ comprises for each distinct data set E^k in a collection $\mathbf{E}$:*

$$f^k_{width}(\mathbf{B}) := \frac{|C_k|}{n_k}$$

Thus, when focusing on a single bicluster, the problem is to find a bicluster $\mathbf{B} \in \mathcal{B}$ such that $f_{size}(\mathbf{B})$ is maximum while $f^k_{hom}(\mathbf{B}) \leq \delta^k$ and $f^k_{width}(\mathbf{B}) \geq \gamma^k$ for $k \in \{1, \ldots, l\}$ with $0 \leq \delta^k, \gamma^k \leq 1$. To extend this model to multiple biclusters, a further criterion comes into play that quantifies the distribution of the biclusters found. This is important because largely overlapping biclusters only provide little information, while a diverse set of biclusters is biologically more interesting. Ideally, a biclustering covers a wide range of genes which can be formalized in terms of a *coverage score*.

Def. 7 *The **coverage score** f_{cov} of a biclustering D denotes the overall number of different matrix cells covered by the union of the biclusters contained in D, formally:*

$$f_{cov}(D) := |\{\ (i,j,k)\,; \exists (G, C_1, \ldots, C_l) \in D\ \wedge\ 1 \leq k' \leq l : i \in G\ \wedge\ j \in C_{k'}\ \wedge\ k = k'\ \}|$$

Now, given the two objectives of bicluster size and biclustering coverage, we formulate the overall optimization problem as finding a biclustering that maximizes coverage and the average size of the biclusters included. Again, these objectives are conflicting as the largest average bicluster size is achieved by filling the multiset D with copies of the largest bicluster which in turn leads to low coverage. This conflict is resolved in the following by ranking the objectives: first, f_{cov} is to be maximized, and then f_{size} is considered.

Def. 8 *Let d be the maximum number of biclusters to be found and γ^k the minimum portion of conditions that each bicluster should comprise with regard to data set k, and δ^k the corresponding homogeneity threshold;*

Algorithm 1 Multiple Gene Deletion

▷ **Input:** $\mathbf{B}, \mathbf{E}, \delta^k, t_G, \alpha$
▷ **Output:** $\mathbf{B}$
▷ Iterate the following as long as $|G|$ is large and the homogeneity threshold is not reached.
while $|G| > t_G$ **and** $f^k_{hom}(\mathbf{B}) > \delta^k \; \forall\, 1 \le k \le l$ **do**
  $r \leftarrow$ FALSE ▷ Gene removed?
  ▷ Remove each gene with highly dissimilar pattern, i. e., $p^k_i > \alpha f^k_{hom}(\mathbf{B})$.
  for all $i \in G$ **do**
    $p^k_i \leftarrow \frac{1}{|C|} \sum_{j \in C} (e^k_{ij} - e^k_{Gj})^2 \quad \forall\, 1 \le k \le l$
    if $p^k_i > \alpha f^k_{hom}(\mathbf{B})$ **then**
      $G \leftarrow G \setminus \{i\}$ ▷ Remove gene.
      $r \leftarrow$ TRUE
    end if
  end for
  if $r =$ FALSE **then**
    switch to Single Node Deletion
  end if
end while

*then, the **rank-based biclustering problem** is defined as follows:*

$$
\begin{array}{ll}
\textit{lex max} & (f_1, f_2) \\
\textit{with} & f_1 = f_{cov}(D) \\
& f_2 = \sum_{\mathbf{B} \in D} f_{size}(\mathbf{B}) \\
\textit{subject to} & \forall \mathbf{B} \in D : \forall 1 \le k \le l : f^k_{hom}(\mathbf{B}) \le \delta^k \\
& \forall \mathbf{B} \in D : \forall 1 \le k \le l : f^k_{width}(\mathbf{B}) \ge \gamma^k \\
& D \in \mathcal{D} \\
& |D| \le d
\end{array}
$$

Note that this model assumes that the number of biclusters sought is small compared to the number of measurements, i.e., $d \ll m \cdot n^{\mathbf{E}}$; otherwise, the coverage and size objectives need to be combined differently.

3.2. *Multi-Matrix Greedy Algorithm*

As a first step towards solving the rank-based biclustering problem (Def. 8), this section describes a greedy strategy for finding one bicluster that satisfies the homogeneity constraint $f^k_{hom}(\mathbf{B}) \le \delta^k$. The method consists of three parts described in Algorithm 1–3 which are based on the general procedure proposed in [9]: Starting with a given matrix, the rows and columns that contribute most to the inhomogeneity of the bicluster are iteratively removed until the constraint is satisfied (cf. Algorithm 2). Then, the algorithm adds all rows and columns that can be included without increasing the inhomogeneity (cf. Algorithm 3). For large input matrices removing rows or columns one by one and recalculating the current inhomogeneity is computationally expensive. To reduce the running time, multiple rows or columns can be removed in one iteration as long as the number of rows and columns is still comparatively high with respect to the target size. Algorithm 1 details this procedure for the removal of multiple genes as this is the more common case than the removal of multiple conditions which is done correspondingly. The intensity of multiple gene deletion is controlled by the two thresholding parameters α and t_G (cf. Algorithm 1). Note that all three algorithms calculate an intermediate measure p^k_i which determines how well a row fits into

Algorithm 2 Single Node Deletion

▷ **Input:** $\mathbf{B}, \mathbf{E}, \delta^k, \gamma^k$
▷ **Output:** $\mathbf{B}$
▷ While any homogeneity constraints are violated do the following.
while $\exists f^k_{hom}(\mathbf{B}) > \delta_k$ for any data set k **do**
  ▷ Find the gene which fits worst.
  for all $i \in G$ **do**
    $p^k_i \leftarrow \frac{1}{|C_k|} \sum_{j \in C_k} (e_{ij} - e_{Gj})^2 \quad \forall 1 \le k \le l$
    $s_i \leftarrow \frac{1}{l} \sum_k p^k_i$
  end for
  $i_{max} \leftarrow \arg\max(s_i)$
  ▷ Find the condition which fits worst.
  for $k \leftarrow 1$ to l **do**
    if $\frac{|C_k|-1}{n_k} \ge \gamma^k$ **then**
      for all $j \in C_k$ **do**
        $C^{*j}_k \leftarrow C_k \setminus \{j\}$
        $\mathbf{B}^* \leftarrow \{G, C_1, C_2, \ldots, C^{*j}_k, \ldots, C_l\}$
        $q^k_j \leftarrow f^k_{hom}(\mathbf{B}) - f^k_{hom}(\mathbf{B}^*)$
      end for
    else
      $q^k_j \leftarrow -\infty \quad \forall j \in C_k$
    end if
  end for
  ▷ Decide whether to remove a gene or a condition.
  $(k_{max}, j_{max}) \leftarrow \arg\max(q^k_j)$
  if $\max(s_i) \ge \max(q^k_j)$ **then**
    $G \leftarrow G \setminus \{i_{max}\}$ ▷ Remove gene.
  else
    $C_{k_{max}} \leftarrow C^{*j_{max}}_{k_{max}}$ ▷ Remove condition.
  end if
end while
continue with Node Addition

Algorithm 3 Node Addition

1: ▷ **Input:** $\mathbf{B}, \mathbf{E}, \delta^k$
2: ▷ **Output:** $\mathbf{B}$
3: **repeat**
4: $a \leftarrow$ FALSE ▷ Gene or condition added?
5: ▷ Add any condition which can be added without increasing f^k_{hom}.
6: **for** $k \leftarrow 1$ to l **do**
7: **for all** $j \notin C_k,\ 1 \leq j \leq n_k$ **do**
8: $C^*_k \leftarrow C_k \cup \{j\}$
9: $\mathbf{B}^* \leftarrow \{G, C_1, C_2, \ldots, C^*_k, \ldots, C_l\}$
10: **if** $f^k_{hom}(\mathbf{B}^*) < \delta^k$ **then**
11: $\mathbf{B} \leftarrow \mathbf{B}^*$ ▷ Add condition j.
12: $a \leftarrow$ TRUE
13: **end if**
14: **end for**
15: **end for**
16: ▷ Add any gene which can be added without increasing f^k_{hom}.
17: **for all** $i \notin G,\ 1 \leq i \leq m$ **do**
18: $p^k_i \leftarrow \frac{1}{|C_k|} \sum_{j \in C_k} (e_{ij} - e_{Gj})^2 \quad \forall 1 \leq k \leq l$
19: **if** $p^k_i < f^k_{hom}(\mathbf{B}) \quad \forall\, 1 \leq k \leq l$ **then**
20: $G \leftarrow G \cup \{i\}$ ▷ Add gene i.
21: $a \leftarrow$ TRUE
22: **end if**
23: **end for**
24: **until** $a =$ FALSE

the bicluster.

The proposed procedure differs from the original algorithm in two central aspects: First, the adaptation to the rank-based problem formulation requires to calculate the exact inhomogeneity for each candidate when removing conditions. As opposed to removing genes, this requires a re-ranking of the expression values (compare Step 14 to Step 5 in Algorithm 2). Second, the algorithms were extended work on collections of gene expression data sets. Additionally, we have limited the removal of columns to enforce the constraint $f^k_{width}(\mathbf{B}) \geq \gamma^k$ given that the input bicluster satisfied the same constraint as well. Note that the proposed procedure also guarantees that the resulting bicluster satisfies the homogeneity constraint $f^k_{hom}(\mathbf{B}) \leq \delta^k$ for any $\delta^k \geq 0$ as it is always possible to reduce the bicluster to one gene and thereby reducing $f^k_{hom}(\mathbf{B})$ to zero.

3.3. *Hybrid Evolutionary Algorithm*

The heuristics described in the previous section identifies a single feasible bicluster. In order to optimize a whole biclustering we employ a combination of a global search method, namely an evolutionary algorithm (EA), with the greedy strategy (Algorithms 1–3). The latter is used to improve each bicluster generated by the global search while the EA optimizes a whole population of biclusters simultaneously. This is done by applying the greedy strategy to each individual before its evaluation. Thanks to this combination, not only the size of each individual bicluster is improved, thereby optimizing f_2, but the population is also distributed over the space of possible biclusters in order to optimize f_1. To this end, we propose a special kind of environmental selection which favors diverse biclusters. In essence, the algorithm switches between optimizing f_2 in the mating selection based on the objective function and optimizing f_1 in the environmental selection. The greedy strategy consisting of Algorithms 1–3 guarantees that the homogeneity constraint is met for each bicluster while bicluster size is increased as much as possible.

Besides the optimization of a whole biclustering, the EA can also improve the performance of the greedy strategy with respect to the optimization of a single cluster. The greedy strategy is likely to get stuck in local optima and can thus profit from the global search which chooses suitable biclusters for the greedy method to start with. In the following, we discuss the details of the hybrid evolutionary algorithm.

3.3.0.1 Representation Each individual represents one bicluster. For reasons of simplicity we have chosen to use a binary representation with two bit strings: one of length m for the genes and a second one of length $n^{\mathbf{E}}$ for the conditions. A bit is set to 1 if the corresponding gene or condition is included in the bicluster.

3.3.0.2 Initialization The initial population should be generated such that a high diversity of biclusters is attained. A simple strategy for example which sets each bit to 1 with a probability of 0.5 produces a set of biclusters containing different genes and conditions but all biclusters will have similar sizes as shown in Figure 2. To avoid this problem, the proposed procedure deterministically chooses the number of genes and conditions to include in each bicluster such that the biclusters are uniformly distributed in the plane spanned by the number of genes and the number of conditions contained in a bicluster (see Figure 3). Which of the

genes or conditions are included is then randomly chosen. This strategy also assures that the full matrix is always part of the initial population.

3.3.0.3 Variation The mutation operator flips each bit in both bit strings with probability p_{mut} and we apply uniform crossover which for each bit picks the value of either of the parents with equal probability.

3.3.0.4 Selection For mating selection, a tournament selection is used, i. e., τ individuals are chosen from the population with replacement and the fittest one is copied to the pool of parents. In choosing the value of τ the selection pressure can be influenced: A higher τ results in more pressure towards fit solutions.

As described, we introduce a specific environmental selection to maintain diversity in the population. While the objective function and the constraints relate to the bicluster size and homogeneity, respectively, the goal of the environmental search is to maximize the coverage. The general idea of the algorithm is to select those biclusters which add most to the coverage of the matrix. This iterative process works as follows: First the biggest bicluster is selected and the elements which are contained in this bicluster are marked. In each following step the algorithm selects the bicluster which contains the largest number of unmarked cells. These steps are iterated until enough individuals have been selected (cf. Algorithm 4). If even more diverse biclusters are sought, a variant of this algorithm can be applied which minimizes the overlap instead of minimizing the number of remaining cells. This modification is achieved easily by replacing line 24 in Algorithm 4 with "**if** $\mathrm{level}_r^{\mathbf{B}} < \mathrm{level}_r^{\mathrm{best}}$ **then**". In the results section we will refer to the former variant as EA R for and the latter as EA O.

3.3.0.5 Fitness Assignment Before the evaluation of an individual it is subjected to greedy heuristic described in the previous section. An individual is evaluated based on the size of the resulting bicluster, i. e., the fitness is calculated as the inverse of its size

$$F(i) = \frac{1}{f_{size}(\mathbf{B})}$$

which leads to a minimization problem. The result of the greedy strategy can either replace the original individual or be just used to determine the fitness of the original individual while the latter one remains unchanged. In this study we use the second strategy called Baldwinian evolution, since it is able to generate a more diverse set of solutions.

Algorithm 4 Environmental Selection

1: ▷ **Input:**
2: P: Population of biclusters.
3: n_{sel}: number of individuals to select ($n_{sel} < |P|$).
4: m, n_k, l: dimensions of the input data set.
5: ▷ **Output:**
6: S: Set of selected individuals.
7: $\mathrm{taken}_{i,j}^{k} \leftarrow 0 \quad \forall\,(i,j,k), 1 \le i \le m, 1 \le j \le n_k, 1 \le k \le l$
8: $S \leftarrow \arg\max_P f_{size}(\mathbf{B})$ ▷ Select largest bicluster.
9: **while** $|S| < n_{sel}$ **do**
10: **for all** $\mathbf{B} \in P$ **do**
11: ▷ Count how many elements of each bicluster are already covered by 1, 2, 3,... selected biclusters.
12: $\mathrm{level}_r^{\mathbf{B}} \leftarrow 0 \quad \forall\, 0 \le r \le n_{sel}$
13: **for all** $i \in G, j \in C_k, 1 \le k \le l$ **do**
14: $\mathrm{temp} \leftarrow \mathrm{taken}_{i,j}^{k}$
15: $\mathrm{level}_{\mathrm{temp}}^{\mathbf{B}} \leftarrow \mathrm{level}_{\mathrm{temp}}^{\mathbf{B}} + 1$
16: **end for**
17: **end for**
18: $T \leftarrow P \setminus S$ ▷ Biclusters not yet selected.
19: $\mathrm{best} \leftarrow$ first element of T
20: ▷ Find the bicluster with the highest number of uncovered or lightly covered elements.
21: **for all** $\mathbf{B} \in T$ **do**
22: $r \leftarrow 0$
23: ▷ As long as both biclusters are equal go to the next level.
24: **while** $\mathrm{level}_r^{\mathbf{B}} = \mathrm{level}_r^{\mathrm{best}}$ **and** $r \le n_{sel}$ **do**
25: $r \leftarrow r + 1$
26: **end while**
27: **if** $f_{size}(\mathbf{B}) - \mathrm{level}_r^{\mathbf{B}} > f_{size}(\mathrm{best}) - \mathrm{level}_r^{\mathrm{best}}$ **then**
28: $best \leftarrow \mathbf{B}$
29: **end if**
30: **end for**
31: $S \leftarrow S \cup \{\mathrm{best}\}$ ▷ Select the best bicluster.
32: $\mathrm{taken}_{i,j}^{k} \leftarrow \mathrm{taken}_{i,j}^{k} + 1 \quad \forall\,(i,j,k), i \in G^{\mathrm{best}}, j \in C_k^{\mathrm{best}}, 1 \le k \le l$
33: **end while**

4. Experimental Results

The experimental validation serves two main goals: (i) to assess the performance of the hybrid evolutionary

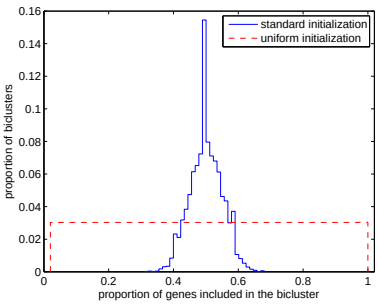

Fig. 2. Histograms of the number of genes in the biclusters of the initial population with standard initialization (setting each bit to 1 with probability 0.5) and with uniform initialization.

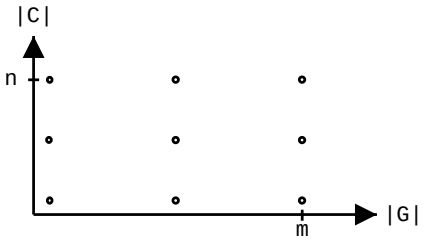

Fig. 3. Schematic drawing of the distribution of an initial population of nine biclusters. m and n are the total number of genes and conditions, respectively. $|G|$ and $|C|$ are the numbers of genes and conditions included in the bicluster.

algorithm by comparing it to two alternative methods and (ii) to compare our strategy for the analysis of a collection of gene expression data sets to the standard approach of combining multiple data sets. Based on these results, a detailed discussion of some exemplary biclusters demonstrates that the proposed method can extract interesting biological information. Additionally, we highlight the flexibility of the optimization framework by solving a related problem where biclusters are sought that exhibit co-expression in one data set but differential expression in other data sets.

4.1. *Experimental Setup*

4.1.1. *Alternative Algorithms included in the Empirical Comparison*

4.1.1.1 OPSM The goal of the strategy proposed in [3] is to identify the largest order preserving submatrix (OPSM) containing a given number of columns. Note that an OPSM is equivalent to a bicluster with a rank homogeneity score of zero ($f_{hom}(\mathbf{B}) = 0$). The algorithm in [3] is run iteratively to search for OPSMs with increasing number of columns. As this approach does not allow to relax the strict order preserving criterion, one can only compare it to our method for the cases where we search for biclusters with $f_{hom}(\mathbf{B}) = 0$. Additionally, it is not targeted to identify a well distributed set of biclusters. Thus, we will compare it to the other methods with respect to the goal of finding one maximal bicluster with perfect ordering.

4.1.1.2 Adapted Cheng and Church Method In [9] Cheng and Church proposed a method for finding a biclustering with multiple diverse biclusters by identifying single biclusters iteratively and removing them from the data set by replacing the corresponding expression values with random data. As mentioned, the multi-matrix greedy algorithm is an adaptation of the greedy strategy used in [9]. We adapted the Cheng and Church method to the case of multiple input matrices by applying the multi-matrix greedy strategy iteratively and after each run replacing the expression values within the identified bicluster with random values.

$$e^k_{i,j} \leftarrow \text{uniform}(\min(e^k), \max(e^k)) \quad \forall\, i,j \in \mathbf{B}$$

This replacement ensures that in consecutive runs of the greedy algorithm different biclusters are identified.

4.1.2. *Algorithm Parameters*

The EA parameter settings used in the following simulations are described in Table 2. α influences the amount of multiple gene deletion. It should be above 1 and values closer to one lead to more multiple gene deletions. The crossover rate refers to the percentage of parents involved in crossover. The mutation rate is the probability for bit flips in the independent bit mutation. The tournament size can be used to influence the importance of f_2: higher values lead to more pressure towards large biclusters but can affect the diversity negatively. Unless stated otherwise, 11 replicates with different random number generator seeds were performed for each run of the evolutionary algorithm and the adapted Cheng and Church method.

The OPSM algorithm takes a parameter l describing how many candidate solutions should be further investigated during the greedy search for OPSMs, see [3] for the details. Consistent with the value used in [3] we set l to 100.

Table 1

α	1.2
probability of 1 in initialization	0.001
mutation rate	0.001
crossover rate	0.1
tournament size	3
population size	100
number of generations	100

Default parameter settings for this study.

4.1.3. *Data Set Preparation*

The simulation runs were performed on gene expression data from a small plant named *Arabidopsis thaliana*. Two collections of genes expression data sets are used: a first collection in which all data sets stem from similar experimental setups and a second one which is more diverse. All data sets measure gene expression in time course experiments using the Affymetrix GeneChip platform.

4.1.3.1 Homogeneous Data Sets The first collection investigates the response of Arabidopsis to different kinds of stresses (cold, salt, osmotic, drought). For each stress experiment gene expression was measured in leaves and roots. The data was provided by the AtGenExpress consortium [1] and consists of 8 time series with 6 time points each. The total expression matrix thus contains 22746 genes and 48 conditions. This data set represents a case where the expression values are well comparable across the different time courses; the experimental setup was identical for all different kind of stresses, all measurements were performed by the same laboratory using the same microarray technology and the expression values were normalized with RMA [6] a state-of-the-art method and logratios were calculated using measurements from an untreated control plant. This reference time course was the same for all stress experiments.

4.1.3.2 Diverse Data Sets The second collection contains time courses of Arabidopsis that are much more diverse than those in the first data set. Like the first data set it consists of 8 time courses with a total of 48 conditions but the number of time points varies. The experiments include different type of treatments such as heat stress, infection with Pseudomonas syringae, and measurements of diurnal changes. These experiments were performed by different labs using different organs such as roots, leaves, and cell cultures. All measurements were performed using Affymetrix GeneChips and all expression values were normalized using RMA. In contrast to the first data set absolute expression values are used.

[1] See http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm

4.2. *Comparison to Alternative Algorithms*

The following simulation runs determine the relative performance of the proposed hybrid EA, the OPSM method and adapted Cheng and Church algorithm and additionally they compare the two variants of the environmental selection. We evaluate the algorithms with respect to finding one bicluster and with respect to the problem of identifying diverse sets of biclusters. Since all algorithms are subject to the same homogeneity threshold we measure performance in the former case by comparing the size of the largest bicluster and in the latter scenario by comparing both the average bicluster size and the coverage of the input matrix.

For the special case of OPSMs that extend over all columns, i. e., $\delta = 0$ and $\gamma^k = 1 \quad \forall\, 1 \leq k \leq l$, the problem of finding the largest OPSM becomes tractable with a time complexity of $O(m^2)$. Thus, the results of the EA and the OPSM algorithm can be compared to the true optimum. In the first experiment we ran both algorithms on each of the eight time courses of the homogeneous data sets (cf. Section 4.1.3.) separately. The largest bicluster found by both the EA and the OPSM algorithm equaled the optimal one in all cases [2]. Often this optimal bicluster was found by the EA after only a few generations.

In a second set of experiments, we reduced the minimum number of columns in a bicluster, now searching for "real" biclusters. The data set used consisted of the concatenation of the two "cold stress" time courses resulting in a matrix with 12 conditions. The largest bicluster found by the EA equaled the size of those found by the OPSM algorithm for all tested settings (cf. Table 3) and they were substantially better than results of the adapted Cheng and Church methods. Figures 4 and 5 summarize the quality of the biclusterings. The first

[2] For seven of the eight data sets the EA found the optimal bicluster in all of 30 replicate runs. For the "osmotic roots" data set 5 of the 11 EA runs identified only the second largest bicluster.

Table 2

δ	γ	OPSM	CC	EA R	EA O
0	6/12	3888	2142 (1.8)	3888 (0)	3888 (0)
0	7/12	1295	861 (0)	1295 (0)	1295 (8.5)
0	8/12	512	352 (0)	512 (0)	512 (0)
0	9/12	216	162 (0)	216 (0)	216 (0)

Size of the largest bicluster ($\max(f_{size}(\mathbf{B}))$) found in the combined "cold stress" data set by the OPSM, the adapted Cheng and Church method (CC) and the EA using the two variants of the environmental selection. For the randomized methods, values denote median and (standard deviations).

version of the environmental selection which maximizes coverage is better suited to achieve high values for the average size of the biclusters as well as high coverage than the version focusing on small overlaps. The former clearly outperforms the alternative methods while the latter one is in some cases inferior to the adapted Cheng and Church method.

A more difficult problem setting consists in searching the concatenation of all eight homogeneous data sets resulting in a matrix with 48 conditions. When requiring perfectly ordered biclusters ($\delta = 0$) the EA variants were in some cases able to identify larger biclusters than the OPSM method while in general the performance was similar (cf. Table 4). The comparison of the biclusterings identified by the two EA variants and the adapted Cheng and Church method reveals a similar situation as for the smaller data set: the overlap version of the environmental search performs similarly as the adapted Cheng and Church method while the coverage version clearly outperforms both other methods (cf. Figures 6 and 7).

So far, we have only considered perfectly ordered biclusters ($\delta = 0$). In two experiments on the same data set, the restrictions on perfect order were removed by setting δ to 0.001 and 0.005, respectively. As expected, the size of the biclusters increases when increasing δ from 0 to 0.001 for the same value of γ. However, the relation between the performance of the different methods basically remains the same (cf. Table 4 and Figures 6 and 7). A typical bicluster is shown in Figure 14 where it can be seen that the expression patterns all follow similar trends in all data sets.

A considerable advantage of the EA optimization framework is that it can analyze multiple data sets simultaneously. We compared the adapted Cheng and Church method to the two EA variants on four pairs of expression data sets by searching for perfectly ordered biclusters ($\delta = 0$) which extend over all six columns of each data set ($\gamma^k = 1$). The results are summarized in Figures 8 and 9. As for the single data sets, the EA results show substantially larger average sizes and coverages than the adapted Cheng and Church method. However, for this setup both variants of the environmental search lead to similar results.

The comparison of the results for the two variants of the environmental selection in Tables 3, 4 and Figures 4–9 show that the version which minimizes the remaining uncovered area (EA R) almost always performs better than the version which minimizes overlap (EA O) both with respect to the bicluster size as well as the coverage. Correspondingly, it is recommended to use the former unless the resulting biclusterings clearly lack diversity. In such a case, the latter method can help to better distribute the biclusters.

As an additional advantage, the iterative scheme of the EA allows to explicitly choose the trade-off between running time and solution quality while most alternative methods have fixed running times. Figures 10 and 11 show how the size of the largest bicluster and the coverage evolves over a typical run. The user can stop the algorithm when the desired quality is achieved or after a given amount of time.

4.3. *Effects of Combining Data Sets*

As discussed in Section 2. it is often not desirable or not possible to concatenate several data sets into one expression matrix. However, existing clustering and biclustering algorithms require this and thereby the information about which measurements belonged to the same experiment and which did not is lost. In the following we investigate the effects of mixing different data sets in the context of the rank-based biclustering problem. A first part of the following analysis is based on the assumption that a difference between two expression values stemming from two different time data sets need not be relevant and contrarily differences between values within one data set always are meaningful. A second part does not use this assumption but investigates the biological significance of the biclusters for both the combined and the separate data sets.

In a first set of analyses we searched for perfect order preserving biclusters ($\delta = 0$) which extend over all columns in the matrix. The EA was run on 4 pairs of time courses: first with the data combined into one matrix and then with keeping the two time courses separately. As expected (cf. Table 5) the resulting biclusters are much larger when the time courses are kept sep-

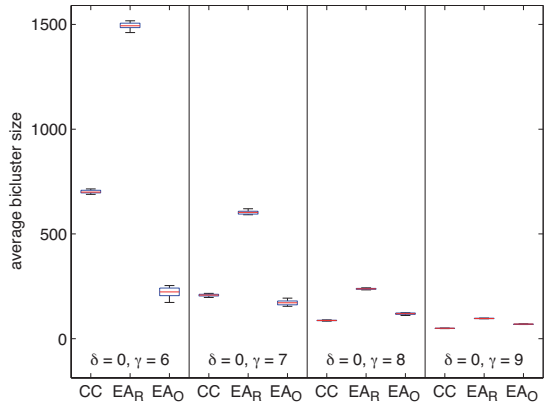

Fig. 4. Analysis of the combination of the two "cold" data sets. Average size of the biclusters for the adapted Cheng and Church method (CC), the EA using the two variants of the environmental selection..

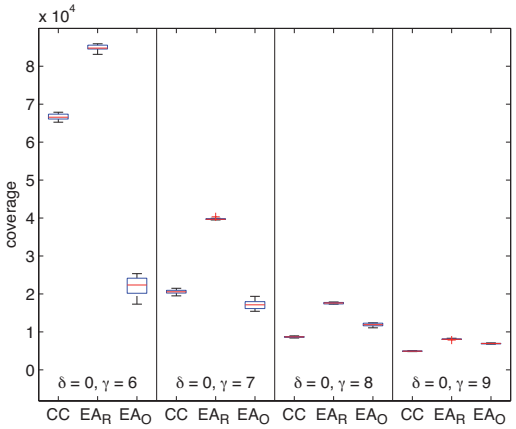

Fig. 5. Analysis of the combination of the two "cold" data sets. Coverage of the biclusters for the adapted Cheng and Church method (CC), the EA using the two variants of the environmental selection.

Table 3

δ	γ	OPSM	CC	EA R	EA O
0	8/48	1992	1528 (183)	2144 (2.4)	1984 (82.7)
0	10/48	520	400 (0)	630 (22.1)	550 (52.0)
0.001	10/48	-	670 (42.0)	1140 (71.0)	920 (91.4)
0.005	20/48	-	4980 (0)	4980 (18.1)	4980 (36.2)

Size of the largest bicluster ($\max(f_{size}(\mathbf{B}))$) found in the combined homogeneous data set by the OPSM, the adapted Cheng and Church method (CC) and the EA using the two variants of the environmental selection. For the randomized methods, values denote median and (standard deviations).

arate. Often it is not possible to find a bicluster with more than a minimal number of genes when mixing the time courses but keeping them separate results in useful biclusters. A characteristic example is the pair of the two "cold stress" experiments where the largest bicluster for the concatenated matrix consists of 2 genes and 32 genes for the simultaneous biclustering of the two data sets.

The same comparison can be made for relaxed constraints on the ordering ($\delta > 0$). However, setting a certain value for δ is not equally restrictive for two time courses with ranks 1–6 as for one combined data set with ranks 1–12. To ensure a fair comparison, we transformed the analysis of the two separate data sets into the analysis of one data set by ranking the expression values in the first time course experiment with ranks 1–6 and those in the second experiment with ranks 7–12. This corresponds a version of the simultaneous biclustering where the constraint is put on the sum of the δ values for both data sets. For the same pair of time courses ("osmotic") and $\delta = 1$ the number genes in the largest bicluster was 21 on average for the concatenated data sets and 474 for the separate time courses. This

Table 4

stress	combined	separate	overlap	independent
cold	2	32		20
osmotic	6	118		65
salt	4	12		3
drought	2	6		0

Number of genes in the largest biclusters for two time courses with $\delta = 0$ which corresponds to searching for OPSMs. Results for mixing of the data sets, joint analysis, and separate analysis with intersection of the best biclusters found.

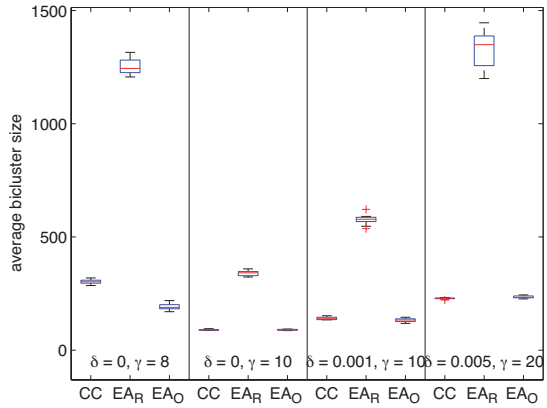

Fig. 6. Analysis of the combination of all eight homogeneous data sets. Average size of the biclusters for the adapted Cheng and Church method (CC) and the EA using the two variants of the environmental selection.

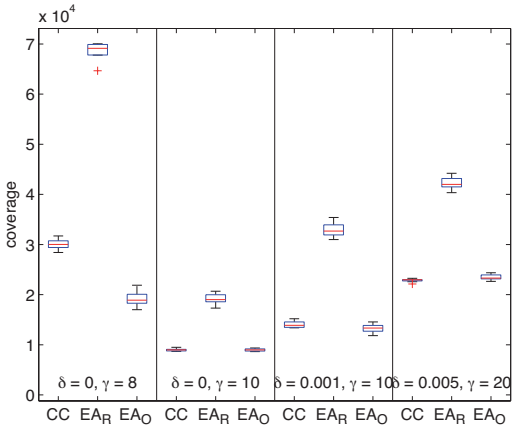

Fig. 7. Analysis of the combination of all eight homogeneous data sets. Coverage of the biclusters for the adapted Cheng and Church method (CC) and Church method (CC) and the EA using the two variants of the environmental selection.

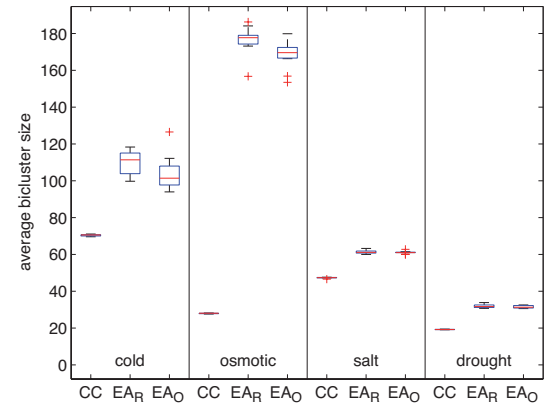

Fig. 8. Analysis of four pairs of data sets. Average size of the biclusters for the adapted Cheng and Church method (CC) and the EA using the two variants of the environmental selection.

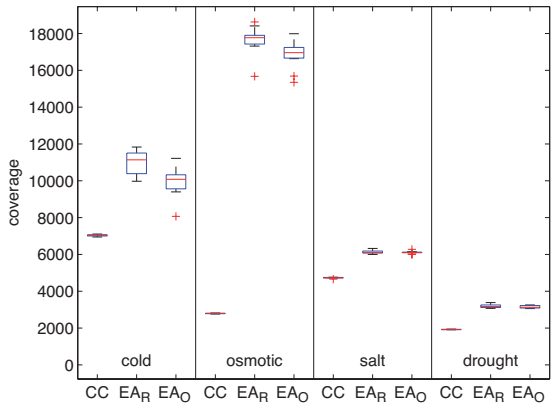

Fig. 9. Analysis of four pairs of data sets. Coverage of the biclusters for the adapted Cheng and Church method (CC) and the EA using the two variants of the environmental selection.

demonstrates that mixing the time courses results in an unnecessarily restrictive optimization problem and most large biclusters are missed.

An alternative strategy to mixing multiple data sets is to perform the bicluster analysis separately on each data set and then combine the results by looking for overlaps. The third column of Table 5 shows the size of the overlap of the optimal biclusters in each data set. For none of the four pairs of data sets the best bicluster from the joint analysis could be recovered by this procedure. For the special case of $\delta = 0$ the largest bicluster could theoretically be recovered by determining the set of all biclusters for each data set and then calculating the intersection of all combination of biclusters. However, this is only practical in the case of $\gamma = 1$. Correspondingly, a separate bicluster analysis combined with the search for overlaps is not a valid alternative to avoid mixing of data sets.

So far we have investigated the effects of mixing data sets on the level of the bicluster size and homogeneity score. This analysis was based on the assumption that comparing measurement values across different data sets is not meaningful. We now drop this assumption and compare the two strategies with respect

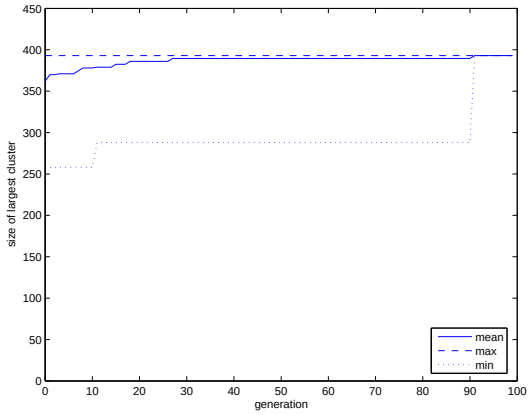

Fig. 10. Size of the largest cluster during the optimization run. (Data from 30 runs)

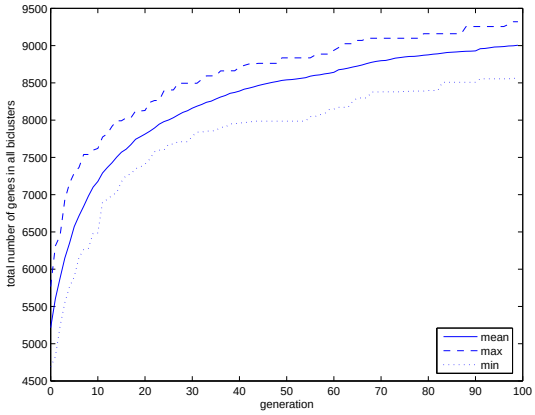

Fig. 11. Total number of genes covered by clusters during the optimization run. (Data from 30 runs)

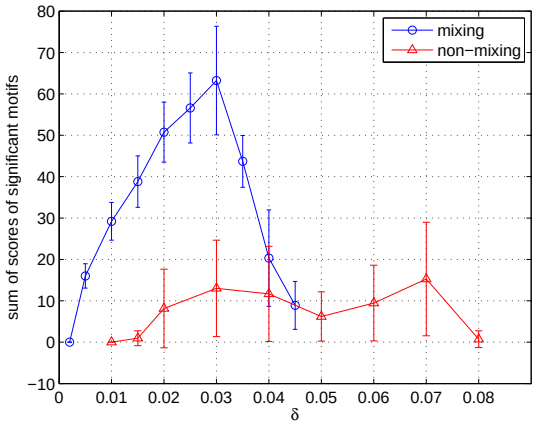

Fig. 12. Number of biclusters found in the homogeneous stress data set with promoter motifs that have a score $s > 3$ for different similarity thresholds delta. Biclusters for low values of δ) are too small to contain significant motifs while biclusters for high values of δ are too big and too diverse. Data from 5 runs per setting. The line represents the mean and the error-bars have a length of 2 standard deviations.

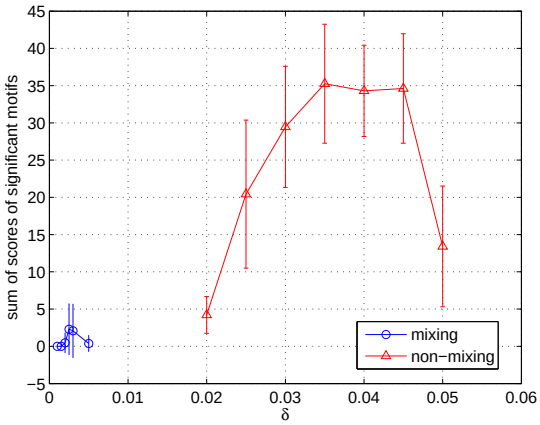

Fig. 13. Number of biclusters found in the diverse data set with promoter motifs that have a score $s > 3$. Biclusters for low values of δ) are too small to contain significant motifs while biclusters for high values of δ are too big and too diverse. Data from 5 runs per setting. The line represents the mean and the error-bars have a length of 2 standard deviations.

to the biological relevance of the resulting biclusters. To this end we searched for 100 biclusters that extend over all eight time courses for a range of different δ values. In each module we then searched for new promoter motifs using the method described in [12]. A highly significant motif is an indicator of a functional relationship between the genes in the bicluster. Many highly significant promoter motifs were discovered in the resulting biclusters. Figures 12 and 13 show the sum of motif scores of the modules with a score [3] above 3. For the homogeneous data sets (Figure 12) both mixing of the time courses and keeping the time courses separate in the analysis result in biclusters with highly significant motifs. Mixing leads to slightly more biclusters

[3] The score is calculated as the distance (measured in standard deviations) from the mean of the distribution of randomly chosen biclusters.

with significant motifs. For the diverse data sets (Figure 13) mixing of the time course prevents the detection of more then few motifs that have scores just above the threshold. However, keeping the time courses separate leads to the identification of many modules with highly significant motifs. In the case of combined analysis of diverse data sets, it is thus detrimental to mix data sets into one matrix while in the case of highly homogeneous data sets mixing of the time courses has a slightly positive effect on the results. However, for many biological studies it is desirable or even necessary to include data from different experiments, different labs or even different technologies.

4.4. *Differential Coexpression*

As mentioned above, with the proposed framework one can not only search for co-expression but also look for differential co-expression, i. e., groups of genes that are similarly expressed in some data sets but show diverging expression patterns in others. This problem formulation is actually a special case of a joint analysis of separate data sets. The goal of finding differences in co-regulation is far less often pursued than looking for co-regulation but has some potentially interesting applications since it allows to investigate condition specific co-regulation. This type of analysis was first proposed in [20] in the context of cancer studies where a break-down in the co-regulation of specific genes can be observed in tumor tissue. While the method in [20] was specifically designed for the case of two data sets our approach can more generally be applied to multiple data sets.

Using this problem formulation, we identified groups of genes that are co-expressed in one type of stress but show inhomogeneous expression patterns in response to the other stresses. This was done by maximizing the dissimiliarities for some data sets (S_{inhom}) instead of the bicluster size while maintaining the homogeneity constraints for other data sets: $f_2 = \sum_{\mathbf{B}\in D} f^k_{hom}(B) \text{for} k \in S_{inhom}$.

A typical example is shown in Figure 15 where the cluster was conditioned on similarity in osmotic stress (third and fourth data set) and dissimilarities in all other treatments. All genes included in the bicluster exhibit perfectly ordered expression profiles for the two osmotic stress data sets while their profiles in the other data sets are much more diverse.

4.5. *Biological Content of Exemplary Biclusters*

The identification of significant promoter motifs is a good indicator for the general biological relevance of the clustering results. In order to further confirm the validity of the approach, we have analyzed two typical biclusters in more detail. For both, it was found that many of the genes included were known to be involved with in the same processes and for some genes a more detailed annotation could be suggested based on the clustering results. For details of this analysis the reader is referred to the Appendix.

5. Conclusions

Current clustering and biclustering algorithms generally operate on one data matrix. In contrast many studies of gene expression involve multiple sets of experiments between which measurements cannot be compared reliably, e. g., the measurements were performed in different laboratories or even using different microarray technologies. With respect to this discrepancy, this paper proposed a flexible biclustering framework that can jointly analyze multiple expression data sets without comparing measurement values between the different data sets and compared this approach to the standard method of mixing different data sets.

While the proposed framework is flexible with respect to the exact problem formulation in this study we have focused on a specific one, namely the rank based biclustering problem. To this end we have introduced a new scoring scheme that allows to arbitrarily scale the degree of orderedness required for a bicluster and integrated it into the biclustering framework which allows to address the aforementioned questions.

In general the framework provides the following main benefits:

- It allows a simultaneous bicluster analysis of separate data sets.
- Multiple well distributed biclusters can be found in a single optimization run.
- The method can be easily adapted to address more specific questions such as differential co-expression or different problem formulations in general.

In an empirical comparison on various data sets the proposed hybrid EA showed similar performance to the OPSM algorithm [3] when considering the largest bicluster. To verify the EAs ability to find diverse biclusterings, we have compared the coverage and the average bicluster size of the results for two variants of EA to an

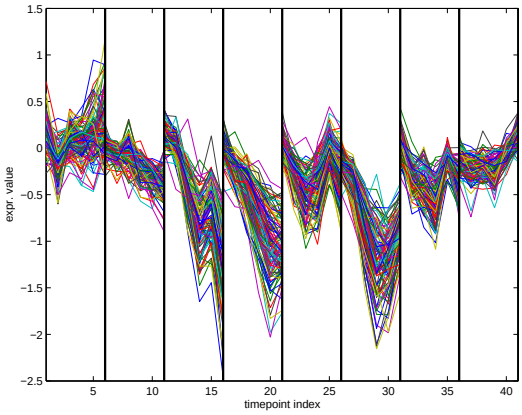

Fig. 14. Expression profiles of the 106 genes in cluster 1 in the cold, osmotic, salt and drought stress time courses. For each stress green tissue is displayed first and roots as second.

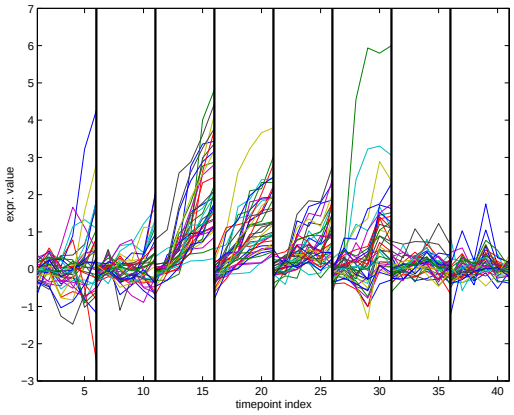

Fig. 15. Expression profiles of the 35 genes from cluster 2 in the cold, osmotic, salt, and drought stress time courses. For each stress green tissue is displayed first and roots as second. The cluster was conditioned on similarity in osmotic stress and dissimilarity in the other treatments.

adaptation of the Cheng and Church method [9]. With the first variant of the environmental search which optimizes for high coverage the EA clearly outperformed the adapted Cheng and Church method over a range of different problem setting. The alternative environmental search which minimizes overlap of the biclusters is able to produce even more diverse sets of biclusters. However, thereby also the average size and the coverage are reduced.

In a second set of experiments, we have investigated the effects of combining different time courses into one data matrix for bicluster analysis. To this end we have analyzed two different expression data sets for *Arabidopsis thaliana*, each one including 8 time course measurements. The biological relevance of the biclustering results has been assessed by an analysis of the promoter motifs common to the genes in the biclusters. This analysis showed that combining different data sets into one matrix is feasible or even advantageous in a setting where all time courses measurements are highly homogeneous but can be detrimental to the results when the data sets are more diverse. The proposed method of a combined analysis does not suffer from this problem.

The proposed strategy is not restricted to the analysis of multiple gene expression data sets. As interesting direction for further research, our method could be used to integrate other types of genome data with gene expression for a combined cluster analysis.

Appendix

In this section we look at two typical biclusters in more detail and discuss their biological implications.

The first bicluster was identified in the stress data set by mixing the time courses but similar biclusters containing the same promoter motifs were identified in the second data set when keeping the time courses separate.

The first module (cf. Figure 14) comprises 106 genes, of which 63 have been annotated as encoding 40S and 60S ribosomal proteins. 16 of the remaining genes are related to RNA metabolism, protein synthesis and protein folding (nascent polypeptide associated complex alpha chain protein, nuclear RNA-binding protein, eukaryotic translation initiation factor, phenylalanyl-tRNA synthetase, and chaperonins). This module comprises genes that are strongly downregulated in response to salt and osmotic stress. Results from Genevestigator [33,32] show that it is additionally downregulated in senescing cell culture, genotoxic stress, and cycloheximide. In contrast, it is consistently upregulated by isoxaben, lovastatin and norflurazon. A similar cluster with strong downregulation in response to stress was described in Eisen et al. (1998). Clusters enriched with ribosomal proteins have also previously been described in yeast and were generally associated with environmental stress responses [13,14]. An analysis of promoter sequences using the MAP scoring function [12] applied to the data

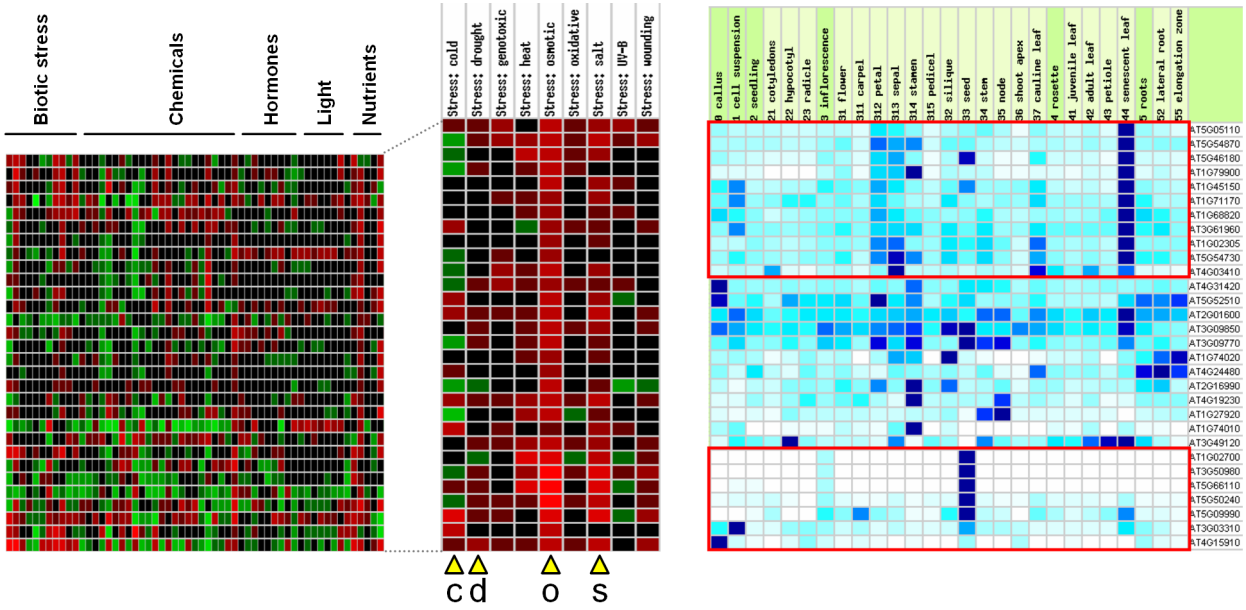

Fig. 16. Expression profiles of the cluster from Figure 15 in response to different stimuli and in several organs/tissues (data from Genevestigator). There is a conditional co-expression both at the response as well as the organ level. The heat-map in the center shows the same dataset as the one used for clustering. The treatments discussed in the paper are indicated (o, osmotic; s, salt; c, cold; and d, drought).

set of interest and to 100 random data sets (z-score), revealed a highly significant sequence motif (AAACCCT). [30] show that the ACCCTA motif (telo-box) is found in the majority of Arabidopsis genes encoding ribosomal proteins and is related to their expression. Additionally, this motif often appears together with a second motif TGGGCC or TGGGCT.

As mentioned, the proposed framework is able not only to look for co-expression but also to look for differential co-expression, i. e., groups of genes that are similarly expressed in some time courses but show diverging expression patterns in others. Using this problem formulation, we identified groups of genes that are co-expressed in one type of stress but show inhomogeneous expression patterns in response to the other stresses.

The module shown in Figure 15 contains several genes that have previously been associated with osmotic, drought, and pathogen stress responses:

(1) microtubule associated protein (MAP65/ASE1) family protein [28]
(2) drought-responsive protein / drought-induced protein (Di21)
(3) dehydrin, putative similar to dehydrin Xero 1 [25]
(4) strictosidine synthase genes [31]

Osmotic stress is a common component of drought, salt and cold stress and coordinates cross-talk in the regulatory network between these stresses [7]. Both ABA-dependant and ABA-independent pathways have been associated with osmotic stress. In compliance with this model, most genes of this module, which was conditioned for co-expression in the osmotic stress treatment, are upregulated strongly in response to osmotic stress, but also (with lower intensity) in the salt stress and ABA treatments, as well as partially in the cold stress treatments cf. Figure 15. To further investigate the expression regulation of genes from this module, stimulus response and anatomy profiles were retrieved from Genevestigator [33] (see Figure 16). As obtained in the biclustering approach, genes were consistently upregulated in osmotic and salt stress, but also to nitrogen deficiency, treatment with ABA, with the elicitor syringolin, and with Pseudomonas syringae. The responses to other treatments were not similar for all genes, revealing that these genes are conditionally co-regulated and could only be identified using an approach that specifically searches for differences in co-expression. This differential pattern of expression is also seen at the organ-level: two larger modules appear, one with genes preferentially expressed in senescent leaves, and the other with seed-specific gene expression. The

remaining genes show strong expression in tissues with reduced or no photosynthetic activity (silique, seed, stamen, sepal, petal, roots). It is known that ABA signalling pathways, which are regulated in response to osmotic changes, are also particularly active in these responses and tissues, where they regulate several metabolic and developmental processes.

The cytoskeleton has previously been implicated in abiotic stress responses such as in osmotic regulation and is known to modulate the activity of ion channels. Additionally, both plant-pathogen and symbiotic interactions involve changes in cell polarity and cellular trafficking in plants and thus are intimately associated with the reorganization of the cytoskeleton. The dehydrin gene Xero2 from Arabidopsis has been shown to respond to ABA, wounding, cold and dehydration. Promoter-GUS studies revealed the presence of several motifs involved in theses responses [25].

Interestingly, although several genes within this module have been annotated as drought-related, the effects of the stresses considered on genes from this module are most intense in osmotic stress, followed by salt and cold stresses, whereas the effect of drought is minimal. This result suggests that these genes are controlled rather by osmotic stress, which is a subcomponent of drought stress, and less by drought-specific signaling pathways. The use of this clustering technique therefore allows to allocate genes much more precisely to subnetworks of signaling pathways, especially when cross-talk exists between those pathways.

References

[1] J. S. Aguilar-Ruiz and R. Divina. Evolutionary Biclustering of Microarray Data. In *Evo Workshops 2005*, LNCS, pages 1–10. Springer, 2005.

[2] H. Banka and S. Mitra. Evolutionary biclustering of gene expressions. *Ubiquity*, 7(42):1–12, 2006.

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem. In *International Conference on Computational Biology (ICCB 2002)*, pages 49–57, New York, NY, USA, 2002. ACM Press.

[4] S. Bleuler, A. Prelić, and E. Zitzler. An EA Framework for Biclustering of Gene Expression Data. In *Congress on Evolutionary Computation (CEC 2004)*, pages 166–173, Piscataway, NJ, 2004. IEEE.

[5] S. Bleuler and E. Zitzler. Order Preserving Clustering over Multiple Time Course Experiments. In *EvoWorkshops 2005*, volume 3449 of *LNCS*, pages 33–43. Springer, 2005.

[6] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A Comparison of Normalization Methods for High Density Oligonucleotide Array Based on Variance and Bias. *Bioinformatics*, 19(2):185–193, 2003.

[7] M. Boudsocq and C. Laurière. Osmotic Signaling in Plants: Multiple Pathways Mediated by Emerging Kinase Families. *Plant Physiology*, 183(3):1185–1194, 2005.

[8] A. Chakraborty and H. Maka. Biclustering of Gene Expression Data Using Genetic Algorithm. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '05)*, pages 1–8, 2005.

[9] Y. Cheng and G. M. Church. Biclustering of Gene Expression Data. In *ISMB 2000*, pages 93–103, 2000. http://cheng.ecescs.uc.edu/biclustering.

[10] F. Divina and J. S. Aguilar-Ruiz. Biclustering of Expression Data with Evolutionary Computation. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):590–602, 2006.

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868, 1998.

[12] M. Friberg, P. von Rohr, and G. Gonnet. Scoring functions for transcription factor binding site prediction. *BMC Bioinformatics*, 6(1):84, 2005.

[13] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.

[14] A. P. Gash and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11), 2002.

[15] D. R. Goldstein, M. Delorenzi, R. Luthi-Carter, and T. Sengstag. Comparison of meta-analysis to combined analysis of a replicated study. In R. Guerra and D. B. Allison, editors, *Meta-Analysis and Combining Information in Genetics*. CRC Press, 2006.

[16] J. A. Hartigan. Direct Clustering of a Data Matrix. *Journal of the Amercian Statistical Association*, 67(337):123–129, 1972.

[17] J. A. Hartigan and M. A. Wong. A k-means Clustering Algorithm. *Applied Statistics*, 28:100–108, 1979.

[18] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–377, 2002.

[19] R. A. Irizarry et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–350, 2005.

[20] D. Kostka and R. Spang. Finding Disease Specific Alterations in the Co-Expression of Genes. *Bioinformatics*, 20(Suppl. 1):i194–i199, 2004.

[21] F. Liu, H. Zhou, J. Liu, and G. He. Biclustering of Gene Expression Data Using EDA-GA Hybrid. In *Congress on Evolutionary Computation (CEC 2006)*, pages 1598–1602. IEEE, 2006.

[22] S. Mitra and H. Banka. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39:2464–2477, 2006.

[23] S. Mitra, H. Banka, and S. K. Pal. A MOE Framework for Biclustering of Microarray Data. In *International Conference on Pattern Recognition (ICPR 2006)*, pages 1154–1157, Washington, DC, USA, 2006. IEEE Computer Society.

[24] A. Prelić, 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, 2006.

[25] D. T. Rouse, R. Marotta, and R. W. Parish. Promoter and expression studies on an Arabidopsis thaliana dehydrin gene. *FEBS Lett*, 381(3):252–256, 1996.

[26] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

[27] R. R. Sokal and C. D. Michener. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

Received 13 March 2007; revised 1 August 2007; accepted 22 November 2007

[28] D. Takemoto and A. R. Hardham. The cytoskeleton as a regulator and target of biotic interactions in plants. *Plant Physiol*, 136(4):3864–3876, 2004.

[29] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(Suppl. 1):S136–S144, 2002.

[30] D. Tremousaygue, L. Garnier, C. Bardet, P. Dabos, C. Herve, and B. Lescure. Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in Arabidopsis thaliana cycling cells. *Plant J*, 33(6):957–966, 2003.

[31] L. van der Fits, H. Zhang, F. L. Menke, M. Deneka, and J. Memelink. A Catharanthus roseus BPF-1 homologue interacts with an elicitor-responsive region of the secondary metabolite biosynthetic gene Str and is induced by elicitor via a JA-independent signal transduction pathway. *Plant Mol Biol*, 44(5):675–685, 2000.

[32] P. Zimmermann, L. Hennig, and W. Gruissem. Gene-expression analysis and network discovery using Genevestigator. *Trends in Plant Science*, 9(10):407–409, 2005.

[33] P. Zimmermann, M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*, 136(1):2621–2632, 2004.