BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 14 2011, pages 1948–1956
doi:10.1093/bioinformatics/btr307

Gene expression

Advance Access publication May 18, 2011

# An ensemble biclustering approach for querying gene expression compendia with experimental lists

Riet De Smet[1,2] and Kathleen Marchal[3,*]

[1]Department of Plant Systems Biology, VIB, [2]Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, Ghent and [3]Department of Microbial and Molecular systems, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium

Associate Editor: John Quackenbush

## ABSTRACT

**Motivation:** Query-based biclustering techniques allow interrogating a gene expression compendium with a given gene or gene list. They do so by searching for genes in the compendium that have a profile close to the average expression profile of the genes in this query-list. As it can often not be guaranteed that the genes in a long query-list will all be mutually coexpressed, it is advisable to use each gene separately as a query. This approach, however, leaves the user with a tedious post-processing of partially redundant biclustering results. The fact that for each query-gene multiple parameter settings need to be tested in order to detect the 'most optimal bicluster size' adds to the redundancy problem.

**Results:** To aid with this post-processing, we developed an ensemble approach to be used in combination with query-based biclustering. The method relies on a specifically designed consensus matrix in which the biclustering outcomes for multiple query-genes and for different possible parameter settings are merged in a statistically robust way. Clustering of this matrix results in distinct, non-redundant consensus biclusters that maximally reflect the information contained within the original query-based biclustering results. The usefulness of the developed approach is illustrated on a biological case study in *Escherichia coli*.

**Availability and implementation:** Compiled Matlab code is available from http://homes.esat.kuleuven.be/~kmarchal/Supplementary_ Information_DeSmet_2011/.

**Contact:** kathleen.marchal@biw.kuleuven.be

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on November 16, 2010; revised on April 15, 2011; accepted on May 11, 2011

## 1 INTRODUCTION

With the large body of publicly available gene expression data, compendia are being compiled that assess gene expression in a plethora of conditions. Comparing one's own experimental data with these large scale gene expression compendia allows own findings to be viewed in a more global cellular context, and inconsistencies between public data and own experiments to be pinpointed. Query-based search approaches such as prioritization-based methods (Adler *et al*., 2009; Hibbs *et al*., 2007; Owen *et al*., 2003) and query-based biclustering techniques (Dhollander *et al*., 2007; Ihmels *et al*., 2002; Zhao *et al*., 2011) have been developed to query a gene expression compendium for genes that are coexpressed with a given gene or gene list (the query). These approaches generally combine gene with condition selection to identify genes that are coexpressed with the query in a subset of the compendium conditions.

Query-based methods usually work well when the query-list contains one gene only or a set of genes that are mutually tightly coexpressed, since they query the expression compendium with the average expression profile of the query-set. However, when query-lists are compiled from the output of experimental assays this list will often contain genes with diverse expression profiles. For instance, a query-list derived from a ChIP-chip experiment might partition into different coexpressed groups depending on which other trancription factors (TFs) the ChIP-assayed TF is interacting with. Hence, when faced with a query-set that is heterogeneous in its expression, the query-profile will be deteriorated, and query-based methods will fail to output meaningful clustering results. Running query-based methods on each gene from the query-list separately circumvents this problem, but will inevitably result in at least partially redundant bicluster solutions as mutually coexpressed genes within the query will output similar biclusters.

A second issue when using query-based biclustering relates to the definition of a threshold on the minimal level by which the bicluster genes should be coexpressed with the query-gene. Indeed, it is often not a priori known how tightly a set of genes should be coexpressed to be biologically meaningful. In addition, this level of coexpression might depend on the biological process the query-genes are involved in (some processes are more tightly coexpressed than others). To allow for a maximum flexibility, some query-based biclustering methods offer the possibility to use a resolution sweep in which a whole range of possible threshold values is scanned. The most relevant solutions can then be selected *a posteriori*, either based on the intuition of the user or by using other *ad hoc* defined selection criteria (such as functional over-representation).

The combined effect of having to run query-based biclustering on each of the genes from the query-list separately with the fact that for each of these single runs also an optional parameter sweep can be performed will result in a whole set of highly redundant biclustering results.

In this work, we developed an ensemble approach to merge such multiple query-based biclustering results into a few non-redundant

*To whom correspondence should be addressed.

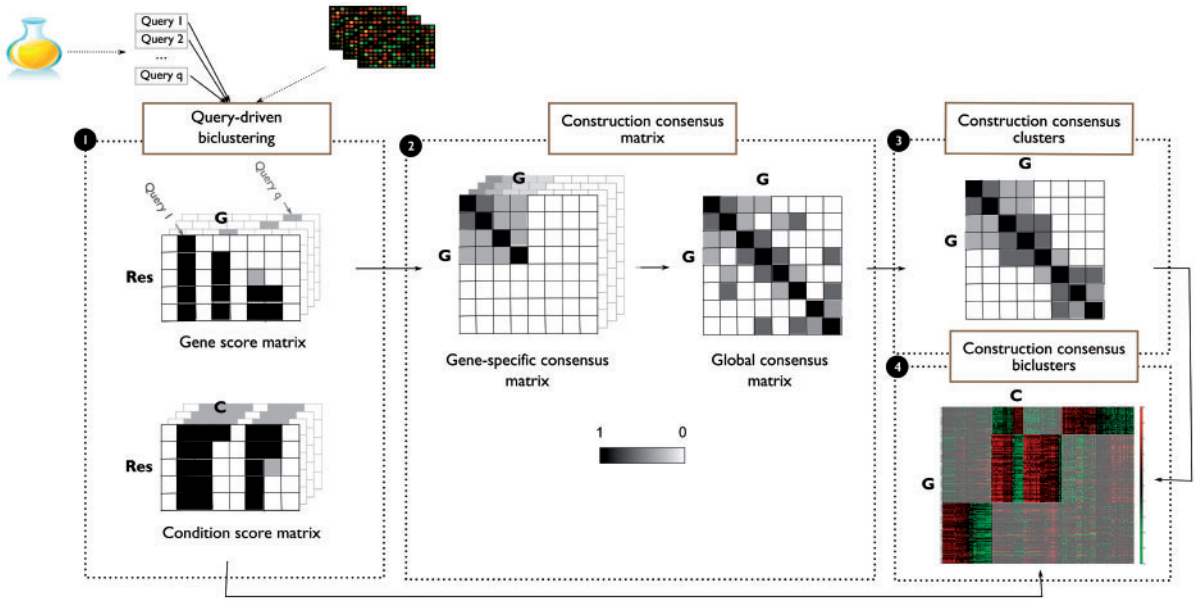


**Fig. 1.** Overview of the split-and-merge ensemble biclustering approach. Step 1: genes from an input-list are each taken separately as a query for QDB (Dhollander *et al*., 2007). For each query-gene QDB results in a gene score and condition score matrix (*G* and *C* refer respectively to the gene and condition dimension of the matrices), containing for each value of the resolution parameter (indicated with *Res*) a score that reflects to what extent the respective gene or condition belongs to the bicluster. Shades of grey are representative for the magnitude of the gene scores and condition scores. Step 2: constructing the consensus matrix proceeds in two steps in which first for each query a gene-specific consensus matrix is constructed from its gene score matrix. This gene-specific consensus matrix summarizes for each query-gene the biclustering-solution obtained at different values of the resolution parameter. In the second step, the gene-specific consensus matrices for all genes in the query-list are merged into a single consensus matrix, representing the frequency of co-occurrence (again indicated by shades of grey) of two genes in the different biclustering solutions that contain at least one of the genes of the pair. Step 3: next, by applying graph clustering, the consensus matrix is partitioned into consensus clusters. Step 4: eventually, consensus biclusters can be obtained by retrieving for each consensus cluster the corresponding conditions from the original QDB-solutions.

consensus biclusters. The usefulness of the developed approach is illustrated on a biological case study.

## 2 OVERVIEW OF THE APPROACH

In our work, the goal is to use query-based biclustering to interrogate gene expression compendia for query-lists heterogeneous in their expression profiles. To this end we introduce a 'split-and-merge strategy' in which query-based biclustering is applied to each gene of the query-list separately (split step) (Fig. 1, Step 1) and redundant results are subsequently summarized using an ensemble strategy (merge step) (Fig. 1, Steps 2–4).

As a query-based biclustering algorithm, we used query-driven biclustering (QDB) (Dhollander *et al*., 2007) as this one incorporates a resolution sweep approach. In this resolution sweep, in a single run of the algorithm the parameter that determines the level of coexpression of the genes within a bicluster is varied. Hence, QDB results for each query-gene in a gene and condition score matrix, which contain for each setting of the resolution parameter respectively the gene and condition scores, i.e. the probability of a gene/condition to belong to the bicluster (Fig. 1, Step 1).

In the ensemble strategy, we separate the task of merging the gene sets obtained for the query-genes from that of merging the condition sets. We first merge different biclustering results in the gene direction. To this end, a consensus matrix (Fig. 1, Step 2) is built that represents the evidence for co-clustering of genes within a certain gene pair by assessing how often these genes co-occur in multiple biclustering results of the data (Monti *et al*., 2003).

Construction of this consensus matrix runs over two phases. In a first phase, we summarize in a gene-specific consensus matrix for each query-gene the multiple biclustering results obtained by applying the resolution sweep approach of QDB. We reason that genes that co-occur in both fine-grained and coarser-grained biclusters, corresponding to a decreasing tightness of coexpression, are more likely to be truly functionally related than genes that only co-occur in coarser-grained biclusters. Therefore, genes that frequently co-occur over the results obtained with the varying resolution parameter will obtain higher consensus scores (i.e. have a higher weight of belonging to the same consensus bicluster) than those that only sporadically co-occur. This first step is only needed when applying a query-based biclustering algorithm that uses a resolution sweep (e.g. Dhollander *et al*., 2007; Ihmels *et al*., 2002).

In a second phase, the gene-specific consensus matrices of the different query-genes are merged into a single consensus matrix, which summarizes the outcomes of all query-based biclustering runs obtained for all the query-genes in the list. Query-genes with a similar expression profile result in redundant biclustering results, while query-genes with mutually very different expression profiles are expected to result in different non-redundant bicluster solutions. To summarize these results we designed a consensus strategy that both reduces redundancy, by combining genes that co-cluster consistently across different runs, but also retains biclustering

outcomes specific to a certain query-gene. As such we retain as much information as possible contained in the original query-based biclustering runs.

In a last step (Fig. 1, Step 3) consensus clusters are extracted from the consensus matrix by using graph clustering. Each resulting consensus cluster consists of genes that repeatedly co-occur over different query-based biclustering runs. Finally, for each of the obtained consensus clusters the corresponding conditions are retrieved from the original biclustering-outputs (Fig. 1, Step 4).

# 3 METHODS

## 3.1 QDB

The strategy proposed in this article can be used in conjunction with any query-based strategy. For illustrative purposes we use here the QDB algorithm (Dhollander *et al*., 2007). As QDB incorporates a resolution sweep approach, one run of the algorithm on a single query-gene (the 'QDB-run') outputs multiple biclustering solutions, each corresponding to a different value of the resolution parameter. We further refer to the output of one such run as the 'QDB-solution'.

## 3.2 Ensemble approach for QDB

*3.2.1 Consensus matrix construction* In a first step, all results for a single QDB-run of a single query-gene from the input list, obtained for $n_{\text{res}}$ different values of the resolution parameter, are merged into a *gene-specific consensus matrix* according to the same principle as Monti *et al*. (2003). Matrix entries $C_{ij}^{(r)}$ reflect the average genepair-to-bicluster membership as estimated over all results obtained at different settings of the resolution parameter (i.e. the coexpression threshold):

$$C_{ij}^{(r)} = \frac{\sum_{t=1}^{n_{\text{res}}} Gs_{i,t}.Gs_{j,t}}{n_{\text{res}}} \quad (1)$$

Here, $Gs_{i,t}$ represents the gene score for gene $i$ for the $t$-th value of in total $n_{\text{res}}$ possible values for the resolution parameter. We obtain in total $n_{\text{qdb}}$ gene-specific consensus matrices, one for every query-gene: $C_{ij}^{(1)}, C_{ij}^{(2)}, \ldots, C_{ij}^{(n_{\text{qdb}})}$.

In a second step, these gene-specific consensus matrices are merged into a *final consensus matrix*, which summarizes the outcomes of the QDB-runs of all $n_{\text{qdb}}$ query-genes in the input list. For this purpose, we introduce a *distributed consensus matrix construction* approach. Here, the frequency of co-occurrence for a gene-pair $ij$ (i.e. the consensus score) is calculated as the element-wise sum of the entries $C_{ij}^{(r)}$ across $n_{\text{qdb}}$ gene-specific consensus matrices, normalized by the number of gene-specific consensus matrices in which a certain gene pair co-occurred:

$$C_{ij}^{\text{global}} = \frac{\sum_{r=1}^{n_{\text{qdb}}} C_{ij}^{(r)}}{\sum_{r=1}^{n_{\text{qdb}}} O(C_i^{(r)}, C_j^{(r)})} \quad (2)$$

with $O(C_i^{(r)}, C_j^{(r)})$ representing the co-occurrence function, which is 1 if both genes belong to the same gene-specific consensus matrix and otherwise 0. The reason for this altered normalization as compared to the gene-specific consensus matrix is that simply averaging the gene-specific matrices over all $n_{\text{qdb}}$ QDB-runs would erroneously down weigh those gene pairs specific to a certain QDB-run (i.e. the non-redundant biclustering solutions) and reward gene pairs that were retrieved by QDB-runs of multiple query-genes (i.e. those that occur in the biclustering result obtained for query-genes that are mutually coexpressed).

We also tested whether the following transformations of the consensus matrix could further improve the quality of the obtained ensemble solution:

- The Topological Overlap Matrix (TOM) (Zhang and Horvath, 2005), which replaces the consensus scores by the topological overlap. This topological overlap reflects for each gene pair not only its pairwise co-occurrence, but increases the score if both genes in a pair frequently co-occur with the same other genes in the output of the QDB-runs.
- Pruning the consensus matrix by setting statistically insignificant consensus scores (i.e. low consensus scores) to zero. Statistical relevance of consensus scores is assessed by the disparity filter (Serrano *et al*., 2009). This method compares for each gene $i$ the distribution of its consensus scores (i.e. the values $C_i^{\text{global}}$) to a null model and sets the least significant scores to zero. We choose our significance threshold such that the sum of the pruned consensus matrix did not fall below 90% of the sum of the non-pruned consensus matrix, this in order to avoid eliminating too many elements with large consensus scores from the matrix.

*3.2.2 Extracting consensus clusters from the consensus matrix* This step aims at obtaining non-redundant consensus clusters from the consensus matrix. This problem can be approached as the clustering of a weighted graph, with weighted edges representing the gene consensus scores and nodes representing the genes. We compared several graph clustering methods that can be applied to weighted graphs. These methods include the Newman spectral modularity algorithm (Newman, 2006), affinity propagation (AP) (Frey and Dueck, 2007), Markov clustering (MCL) (Van Dongen, 2000), hierarchical clustering and a recently published fuzzy spectral graph clustering method (Joshi *et al*., 2008).

The Newman spectral modularity algorithm and the fuzzy spectral method select automatically the number of clusters. To select the optimal number of clusters for AP, MCL and hierarchical clustering we used respectively the default parameters, the efficiency measure (Van Dongen, 2000) and the median split silhouette coefficient (Pollard and van der Laan, 2005).

Consensus clusters not containing any of the genes included in the query-list were discarded, as they were of no further relevance to the study.

*3.2.3 Obtaining consensus biclusters* To map the conditions to the consensus clusters, we trace back the obtained consensus clusters to the original QDB-solutions from which they were derived. To find the corresponding QDB-solutions we use the geometric coefficient (Goldberg and Roth, 2003) to quantify the overlap in the gene content between a consensus clusters and each of the original QDB-solutions:

$$\text{Overlap} = \frac{|G_{\text{cons}} \cap G_{\text{qdb}}|}{\sqrt{|G_{\text{cons}}||G_{\text{qdb}}|}} \quad (3)$$

with $G_{\text{cons}}$ representing the genes in the consensus cluster and $G_{\text{qdb}}$ the genes in the original QDB-solution. Since, each QDB-solution corresponds to different gene sets retrieved for different values of the resolution parameter, the overlap is calculated for the results obtained for each resolution separately. The condition score vector that corresponds to the resolution for which this overlap is maximized is then retained. Next, the condition consensus scores for a particular consensus cluster are calculated as the weighted mean of all condition score vectors retained for this consensus cluster. The weight is chosen equal to the geometric coefficient, hence giving higher weight to condition score vectors belonging to bicluster outcomes better reflected by the consensus clusters. Finally conditions with a consensus score exceeding 0.75 (conditions occur in at least 75% of the condition score vectors) are retained.

## 3.3 Applying the ensemble biclustering approach

As a proof of concept we applied the proposed ensemble biclustering approach to presumed targets of the fumarate nitrate reduction transcriptional regulator (FNR) obtained by ChIP-chip analysis (Grainger *et al*., 2007). In this experiment binding of FNR under anaerobic conditions was evaluated. The authors identified 63 genomic regions at which FNR binds. These 63 genomic regions were mapped to 90 genes as the authors assigned a bound region located in the promoter region of two divergently regulated genes to both genes (Grainger *et al*., 2007).

Each of the 90 potential FNR-targets was used separately as query in the QDB-algorithm (Dhollander *et al*., 2007). As a gene expression dataset an *Escherichia coli* gene expression compendium spanning 870 conditions

was used (Lemmens *et al*., 2009). For each of these query-genes 200 biclustering outcomes were obtained corresponding to 200 different values of the resolution parameter. For 44 out of the 90 query-genes QDB-solutions could be retrieved which contained all together 61 out of the 90 FNR ChIP-chip targets. For the remaining 29 genes no significant QDB-solutions were obtained, either because no additional genes were found to be coexpressed with the query-gene (26 cases) or because the number of conditions under which the genes were found to be coexpressed was not sufficient (here at least 10 conditions were required to be included in the bicluster).

For each of these 44 query-genes a gene-specific consensus matrix was constructed to aggregate its 200 biclustering outcomes. QDB-solutions for each of these 44 query-genes were at least partially overlapping (Supplementary Fig. S1), therefore these 44 gene-specific consensus matrices are merged into one consensus matrix. Consensus biclusters are finally obtained by applying graph clustering to the (transformed) consensus matrix and by retrieving the matching condition set from the 44 QDB-solutions.

To analyse the gene content of these consensus biclusters gene functional GO-categories were taken from EcoCyc (Keseler *et al*., 2009). To verify the presence of known FNR-targets within the consensus biclusters the known *E. coli* regulatory network was taken from RegulonDB (Gama-Castro *et al*., 2008). Heatmap visualizations of the consensus biclusters were made by using ViTraM (Sun *et al*., 2009).

## 3.4 Performance evaluation

*3.4.1 Comparison of different ensemble designs* To assess the impact of using a certain combination of matrix transformation and graph clustering approach (hereto further referred as the 'ensemble design'), we introduce the following quality measures for the obtained consensus biclusters:

- *Degree to which the consensus biclusters recapitulate the information contained within the original QDB-solutions*: The *agreement* calculates for each consensus bicluster its maximal overlap in number of genes with the original QDB-solution, by the geometric coefficient [see formula (3)]. By averaging this measure for the different consensus biclusters obtained from a single consensus matrix, a global agreement measure for all consensus clusters was obtained. The agreement is asymmetric: different consensus biclusters might show maximal overlap with the same QDB-solution and hence the overlap might be high while the consensus solution is biased towards certain QDB-solutions only. Therefore we also calculate the query-gene coverage (*coverage measure*), which assesses whether the obtained consensus biclusters cover the information content of the QDB-solutions in its entirety. The query-gene coverage is calculated as the number of query-genes in the original QDB-solutions that belong to a non-trivial consensus bicluster (i.e. a consensus bicluster with more than 1 gene).
- *Degree to which consensus clusters remove redundancy*: The *redundancy measure* evaluates the extent to which the consensus biclusters are able to reduce the redundancy present in the original QDB-solutions. We assume that query-genes with largely overlapping (or highly redundant) QDB-solutions should belong to the same consensus bicluster. Consequently, we use Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002) to assess how well a clustering of the query-genes based on overlap in their QDB-solutions corresponds to the partitioning of the query-genes according to the consensus biclusters (see Supplementary Data for more information on this measure).
- *Biological relevance*: The *functional coherence measure* assesses the biological relevance of the set of consensus biclusters produced from the consensus matrix. For each consensus bicluster a $P$-value for functional enrichment is calculated using the hypergeometric test ($P<0.01$, Bonferroni-corrected for multiple testing). As from each consensus matrix multiple consensus biclusters are obtained, we use the clustering score function (Asur *et al*., 2007) to aggregate all $P$-values obtained for all consensus biclusters derived from the same consensus matrix into a single score. Let $n_s$ be the number of significantly enriched clusters and $n_i$ the number of insignificant clusters for a $P$-value cut-off $c$, then the functional coherence of a consensus solution is defined as follows:

$$f_c = 1 - \frac{\sum_{k=1}^{n_s} \min(p_k) + (n_i * c)}{(n_s + n_i) * c} \quad (4)$$

- *Statistical quality*: We also assessed the objective quality of the consensus biclusters by assessing whether consensus clusters derived from the consensus matrix have more intra-cluster edges than between-cluster edges as evaluated by the *modularity* function. Modularity Q (Newman, 2004) compares, given a clustering and corresponding consensus matrix $C_{ij}$, the fraction of the edges that falls within a given cluster minus the expected fraction if edges were distributed at random. The higher the modularity the better the cluster separation, with a maximum value of 1 for strong modular structures. Let $k_i$ be the degree of node $i$, $m$ the total number of edges and the $\delta$ function yields 1 if vertices $i$ and $j$ belong to the same cluster (otherwise the function is 0), then the modularity is given by:

$$Q = \frac{1}{2m} \sum_{ij} (C_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \quad (5)$$

Each of these metrics has a maximum value of 1, which makes their interpretation and comparison rather straightforward.

*3.4.2 Comparison with standard application of query-based biclustering* We also compared the here introduced ensemble approach with the standard application of QDB as described in (Dhollander *et al*., 2007):

- Takes one or multiple query-genes as input. In case of the latter, the average expression profile of the query-genes is used as query-profile.
- Uses a resolution sweep approach to scan multiple possible biclustering solutions. Selection of the most appropriate biclustering solution for a certain range of resolution parameter values is based on the Akaike Information Criterion (AIC) that comes with the algorithm (Dhollander *et al*., 2007).

For this comparison we used the *overlap measure* which assesses for a certain method the redundancy or overlap of its biclustering solutions. This measure corresponds to the proportion of the biclustering solutions for which the gene content overlaps at least 70% with that of another biclustering solution obtained by the same method. Overlap is calculated by the geometric coefficient [formula (3)].

*3.4.3 Comparison with state-of-the-art biclustering* We used the state-of-the-art biclustering methods SAMBA and ISA for comparison. SAMBA was applied through the Expander software package (Ulitsky *et al*., 2010), whereas for ISA the R-implementation was used (Csardi *et al*., 2010).

## 3.5 Code availability

The code for the ensemble framework was implemented in Matlab version 7.10.0 and the compiled code is freely available from http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_DeSmet_2011/. For the *E.coli* ChIP-chip results, running the ensemble code on the 44 biclustering outcomes took 30 min on a 2800 MHz Dual-Core AMD Opteron(tm).

# 4 RESULTS

## 4.1 Comparison different ensemble designs

To develop an ensemble approach that was able to maximally remove redundancy by merging the outcome of redundant biclusters into a single consensus bicluster, while also retaining as much as possible the information contained within the original query-based biclustering results (the results obtained before applying

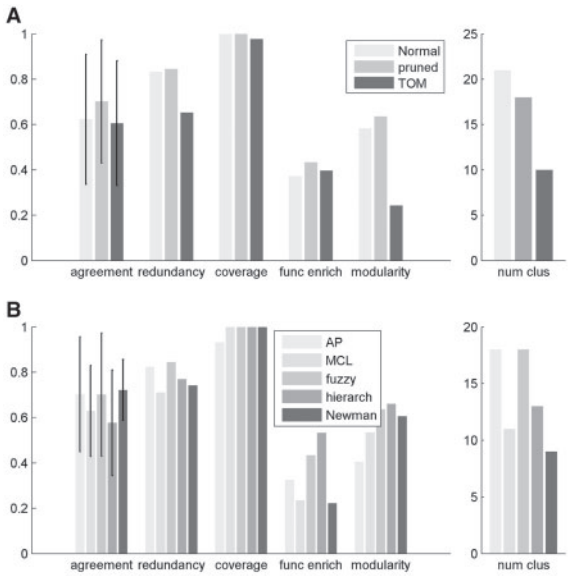


**Fig. 2.** Comparison of different ways to construct the consensus biclusters. (**A**) Compares the influence of using different consensus matrix transformations on the quality of the final consensus biclusters assessed by respectively their overlap with the original QDB-solutions ('agreement'), the extent to which redundancy among the QDB-solutions is removed ('redundancy'), their coverage for query-genes ('coverage'), their functional coherency ('func enrich'), the modularity of the obtained clustered consensus matrix ('modularity') and the number of consensus biclusters ('num clus') (*x*-axis). For illustrative purposes we show the assessment of the final consensus biclusters for different matrix transformations, each time used in combination with fuzzy clustering. (**B**) Comparison of the effect of using different graph clustering methods to extract from the consensus matrix the final consensus biclusters. Same assessment criteria as in panel A were used. For illustrative purposes only results obtained on the pruned consensus matrix are shown.

the ensemble approach), we tested (i) different transformations of the final consensus matrix (see Section 3) and (ii) different graph clustering methods (see Section 3) to extract the consensus biclusters from the consensus matrix. The final consensus matrix, before applying any of the matrix transformation methods, was obtained as described in Section 3.3. For this comparison we used the evaluation metrics defined in Section 3.4.1.

We applied all possible combinations of consensus matrix transformation and graph clustering methods. Their effect seemed not to be confounded as the effect of using a different consensus matrix transformation was the same irrespective of the used graph clustering method and vice versa (Supplementary Fig. S2). Therefore, we here only show the results of representative cases.

Regarding the effect of using different transformations (Fig. 2A, showing the results for fuzzy clustering), it seems that the best results were obtained with a pruned consensus matrix rather than with a non-transformed consensus matrix or a TOM. Compared to not using any transformation, pruning the consensus matrix resulted in consensus biclusters that better represented the original QDB-solutions ('agreement'), that were more biologically relevant ('functional coherence') and that were more densely connected ('modularity'). Pruning the consensus matrix seems to improve the outcome of the ensemble biclustering approach by excluding noise from the consensus matrix, as this filtering sets low consensus scores to zero. Theoretically, applying TOM is expected to increase the robustness of the consensus scores by not only taking into account pairwise co-occurrence of the genes within a biclustering solution, but by also accounting for their joint co-occurrence with other genes. This effect was, however, less obvious from our results. Rather, compared to using a non-transformed or pruned consensus matrix, applying TOM seems to lower the agreement between the consensus and the original QDB-solutions, as indicated by the lower agreement and redundancy measures.

The effect of using different graph clustering methods on the quality metrics is illustrated in Figure 2B (contains results for the pruned consensus matrix). We observe that both AP and fuzzy clustering show the best trade-off between removing redundancy while still agreeing largely with the original QDB-solutions (with a redundancy score of 0.84 and an average agreement of 70% with the original QDB-solutions). However, compared to AP, consensus clusters obtained with fuzzy clustering have a higher coverage for query-genes, a better functional coherence and a more pronounced modularity. Supplementary Figure S3 indeed shows that the clustered consensus matrix obtained with fuzzy clustering shows a more consistent block-diagonal structure than that obtained with other cluster algorithms.

As the combination of the pruning transformation with the fuzzy clustering outperformed the other methods for the used quality criteria, we used this combination in the subsequent application.

### 4.2 Comparison of the ensemble strategy with standard query-based biclustering

Here, we compared our 'split-and-merge ensemble strategy' ('QDB-ensemble') with alternative ways of applying query-based biclustering on the same input gene list. We compared with a first alternative strategy, referred to as 'QDB-split'. Here, QDB is also applied to each gene of the query-list separately, using a resolution sweep. However, instead of merging the different results obtained after the resolution sweep in a gene-specific consensus matrix, only one representative solution is withheld using the AIC (see Section 3). Resulting biclustering outcomes of single query-genes are not further merged as is done in 'QDB-ensemble'. As a second alternative strategy we used 'QDB-nosplit': here the average expression profile of all genes in the list is considered as the query-profile. In contrast to both previous strategies, the genes in the query-list are not treated separately. Here also, only one representative solution from the resolution sweep is chosen based on the AIC (see Section 3). As a last alternative strategy we used 'QDB-partialsplit': instead of relying on the average profile of all genes in the query list, we used here subsets of coexpressed query-genes from which the average query-profile was derived. These subsets were obtained by clustering the expression profiles of the query genes over all conditions (K-means clustering). For each query-profile, one representative solution from the resolution sweep is chosen based on AIC (see Section 3). Here also, resulting biclustering outcomes for the query-gene subsets are not merged.

We compared the outcome of these different QDB-strategies using 'functional coherence' and 'coverage' to assess the biological relevance of the obtained biclusters and the extent to which they contain genes from the input list. In addition, we assess the extent to which the biclusters obtained with a single strategy are

**Table 1.** Comparison of alternative ways of applying query-based biclustering

| | Biclusters (*n*) | Coverage | Overlap | Functional coherency |
|---|---|---|---|---|
| QDB-ensemble | 17 | 0.68 | 0 | 0.46 |
| QDB-split | 44 | 0.64 | 0.57 | 0.56 |
| QDB-nosplit | NA | NA | NA | NA |
| QDB-partial | 12 | 0.36 | 0.17 | 0.56 |

Coverage, Functional coherence and Overlap measures are defined similarly as in Section 3. NA designates that for the particular strategy no bicluster outputs were obtained.

mutually redundant ('overlap' measure). The results are summarized in Table 1.

These results illustrate that when dealing with input lists containing genes that are heterogeneous in their expression profiles, it is absolutely necessary to apply query-based biclustering to each gene of the query-list separately instead of using an average query-profile. Indeed, if the QDB-algorithm is initialized with an average profile of too many genes no output is obtained ('QDB-nosplit') (Table 1, 'Coverage'). Even when using as a query the average profile of genes that are similar in expression ('QDB-partialsplit'), results in terms of the coverage for query-genes were inferior to those obtained for 'QDB-ensemble' and 'QDB-split', suggesting that the obtained results largely depend on which genes were combined to calculate the average profile.

Among the strategies that use each gene of the query-list separately, 'QDB-ensemble' outperforms 'QDB-split': compared to the latter, the former significantly reduces the number of biclusters while increasing at least slightly the coverage of the query-genes. The lower redundancy is due to the ensemble strategy: the relatively larger number of biclustering results obtained by 'QDB-split' showed indeed to be partially redundant (Table 1, 'Overlap'). The higher coverage on the other hand probably relates to circumventing the thresholding of the resolution parameter based on the AIC criterion.

Circumventing this thresholding, however, comes at the cost of reducing the functional coherency of the biclusters (Table 1, 'Functional coherency'). Indeed, by retaining all solutions from the resolution sweep, 'QDB-ensemble' also retains the bicluster solutions that are not as tightly coexpressed (relaxation of the resolution parameter) and as such functional coherency values for the obtained biclusters are slightly lower than the ones obtained for 'QDB-split' and 'QDB-partialsplit'.

### 4.3 Comparison with state-of-the-art biclustering

The advantage of a query-based strategy over a global biclustering approach that searches for global patterns in the data is illustrated in Table 2. Here, we compared the results of our ensemble query-based biclustering strategy to the results of state-of-the-art biclustering algorithms SAMBA (Tanay *et al*., 2002) and ISA (Ihmels *et al*., 2004), which were applied to the same *E.coli* expression dataset. For the ISA-algorithm we took advantage of the 'smart seeding' option to initialize the algorithm with the genes from the query-list. For both methods default parameter settings were used.

These results illustrate that for global biclustering approaches, it is difficult to balance query-gene coverage and bicluster quality.

**Table 2.** Comparison with state-of-the-art biclustering

| | Biclusters (*n*) | Coverage | Bicluster with query (%) | Overlap | Functional coherency |
|---|---|---|---|---|---|
| QDB-ensemble | 17 | 0.68 | 1 | 0 | 0.46 |
| SAMBA | 408 | 1 | 0.77 | 0.25 | 0.13 |
| ISA | 33 | 0.48 | 0.76 | 0 | 0.36 |

Same evaluation metrics as in Table 1 were used. In addition to these evaluation metrics we also calculated the 'Biclusters with query (%)' to get an idea of the number of biclusters obtained that were not relevant for the query-list.

Indeed, as the biclusters containing the query-genes are most likely not all the most prominent ones in the dataset, increasing query-gene coverage requires lowering the stringency of the coexpression quality measures defined by the global strategies, resulting in many biclusters with a lower quality and functional enrichment (SAMBA). In our results it seems that ISA when using default settings searches for high-quality biclusters (high functional over-representation), but misses many query-genes. Whereas SAMBA under default settings works in a regime that allows for a higher coverage of query-genes, but with biclusters of much lower quality. By focusing its search strategy on the query-genes, query-based approaches find a better trade-off in obtaining a high coverage and quality (the latter assessed by functional enrichment).

### 4.4 A ChIP-chip case study

Chromatine immunoprecipation in combination with microarray technology (ChIP-chip) is increasingly being used to measure protein–DNA interactions *in vivo*. Being a high-throughput technology, ChIP-chip data inevitably gives rise to false positives. In addition, the technology fails to distinguish non-functional from functional binding (Wade *et al*., 2007). Hence ChIP-chip experiments need to be backed up by expression data that provide information on whether the identified target genes are indeed being regulated by the bound TF.

We applied the proposed workflow to presumed FNR-targets obtained by ChIP-chip analysis (see Section 3) (Grainger *et al*., 2007). In this experiment binding of FNR under anaerobic conditions was evaluated, yielding a list of 90 query-genes containing 26 known FNR-targets (Gama-Castro *et al*., 2008). For 44 out of the 90 query-genes, biclusters could be retrieved that were efficiently merged into 17 consensus biclusters (Supplementary Table S1). These 17 biclusters cover 61 of the 90 ChIP-chip targets, amongst which 24 known FNR-targets (Supplementary Table S2).

In what follows we use the results of these consensus biclusters to interpret the results of the FNR ChIP-chip experiment, i.e. to distinguish within the list of possible ChIP-chip targets the functional from the non-functional or false positive ones and to pinpoint likely false negative targets that were not recovered by the ChIP-chip experiment.

Figure 3 represents 4 biclusters chosen based on their enrichment for ChIP-chip targets (bicluster 5 and 12) and/or a high coverage for previously described FNR-targets (bicluster 10 and 16) (Supplementary Table S1). We chose these criteria as they suggest that the ChIP-chip targets within these biclusters are functional targets. We indeed expect that ChIP-chip targets within biclusters enriched for the ChIP-chip list constitute functional targets as they

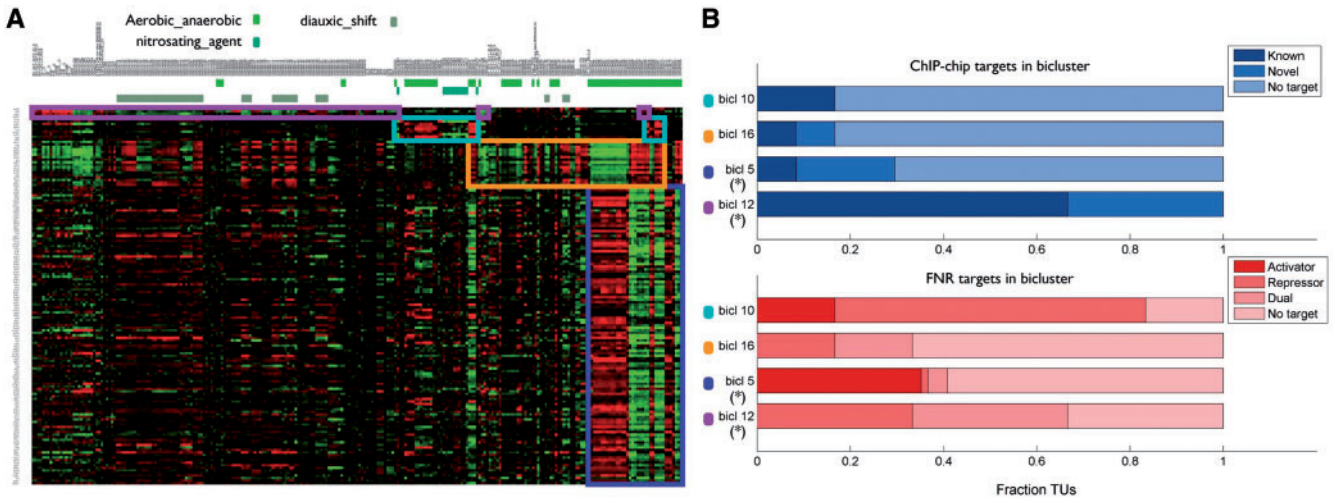


**Fig. 3.** Consensus biclusters obtained by interrogating an *E.coli* expression compendium with the target list of a FNR ChIP-chip experiment. (**A**). Heatmap representation of the consensus biclusters 5, 10, 12 and 16. Rows represent the genes, whereas columns represent the conditions. Different consensus biclusters are indicated by colored rectangles. At the top of the picture conditional categories present within the gene expression dataset are shown (Lemmens *et al*., 2009). A colored square on top of the heatmap indicates that a condition belongs to a particular conditional category. (**B**). Overview of the content of the 4 consensus biclusters in terms of the number of ChIP-chip targets (top) and previously described FNR-targets (bottom). Stacked bars represent the proportion of transcription units (TUs) in the consensus bicluster belonging to a certain category. In the top bar chart 'Known' represents ChIP-chip targets that are documented to be regulated by FNR according to RegulonDB (Gama-Castro *et al*., 2008), 'Novel' the ChIP-chip targets that are no documented regulators of FNR and 'No target' refers to the remainder of the TUs in the bicluster that were not identified by ChIP-chip. Consensus biclusters indicated with an asterisk are significantly enriched in ChIP-chip targets. In the bottom bar chart TUs are grouped according to the regulatory mode of FNR (repressor, activator or dual regulator) and 'No target' here refers to the TUs that are not documented to be regulated by FNR according to RegulonDB.

are not only bound by the same TF but also mutually coexpressed. As can be expected consensus biclusters 5 and 12, both of which are significantly enriched in the ChIP-chip targets, are indeed mainly composed of conditions that measure the effect of oxygen (Fig. 3A) and show a high coverage for known FNR-targets (Fig. 3B). In total, these consensus biclusters covered 24 FNR ChIP-chip targets of which 7 novel ones, not documented in RegulonDB (Gama-Castro *et al*., 2008). In addition they contained 14 previously described FNR-targets that were missed by the ChIP-chip analysis (false negatives).

The two other biclusters in Figure 3, consensus bicluster 10 and 16, are not enriched in the ChIP-chip targets, but show a high coverage of previously described FNR-targets (Supplementary Table S1). In addition, they are just like biclusters 5 and 12 enriched in oxygen-related conditions (Fig. 3). Interestingly, the expression pattern of the genes within bicluster 16 is anti-correlated to that of the genes in the ChIP-chip enriched consensus biclusters (Fig. 3A). This anti-correlated behaviour reflects the different mode of action of the dual regulator FNR, which acts as an activator on the targets of the ChIP-chip enriched consensus biclusters 5 and 12 and as a repressor on consensus bicluster 16 (Fig. 3B).

The distinct expression behaviour of the genes in consensus bicluster 10 can be explained by joint regulation of the genes within this consensus bicluster by NsrR and FNR (Gama-Castro *et al*., 2008), which also explains the presence of nitrosating conditions within this consensus bicluster (Fig. 3A). Similarly to consensus bicluster 16 the genes within this consensus bicluster not retrieved by the ChIP-chip experiment are also known to be repressed by FNR. Seemingly the conditions used in the set up of Grainger *et al*. (2007) were biased towards selecting positively regulated targets (bicluster 5 and 12), but missed most of the repressed targets (bicluster 10 and 16). Together these consensus biclusters 10 and 16 contained 3 ChIP-chip targets of which 1 novel one [i.e. not documented as an FNR-target in RegulonDB (Gama-Castro *et al*., 2008)] and 7 additional previously described FNR-targets not retrieved by the ChIP-chip experiment (Fig. 3B).

The remaining 33 ChIP-chip targets belong to biclusters not enriched with genes from the ChIP-chip experiment, nor having a high proportion of known FNR-targets. For these targets the results are less conclusive. Six of these ChIP-chip targets are known FNR-targets according to RegulonDB (Supplementary Table S2). Considering that many targets of FNR perform global cellular functions, it is indeed possible that due to pleiotropic functions of these genes some FNR-targets end up in biclusters not having a high coverage for known FNR-targets. However, we can expect a large proportion of these 33 genes to correspond to false positives or non-functional targets. Not only because of the ChIP-chip procedure itself, but also because of the way the ChIP bound regions were mapped to the genes: the presence of a ChIP bound region located in the intergenic region between two divergently transcribed genes does not automatically imply that both genes are transcriptionally regulated by the bound TF (Gao *et al*., 2004).

## 5 DISCUSSION

In this article, we developed an ensemble approach to be used in combination with query-based biclustering methods for the

interrogation of expression compendia with a list of experimentally derived genes.

The method exploits the possibility some query-based biclustering methods offer to explore a whole range of thresholds that influence the bicluster size. Instead of having to choose the 'best bicluster with the most optimal coexpression level' based on some user-defined *ad hoc* criteria, our ensemble approach merges the results of the multiple runs in a single consensus matrix, whereby genes that were repeatedly retrieved at multiple biological resolutions will receive a higher weight to belong to the same consensus cluster. The ensemble approach thus offers a statistically inspired way to merge the outcomes for different thresholds on coexpression.

The ensemble approach is also devised to cope with the 'split-and-merge strategy' that is needed when using a query-list containing genes with different expression behaviour as input. The ensemble strategy is used to merge the partially redundant biclustering-outcomes that were obtained by running query-based biclustering on each of the genes of the query-list separately. Query-genes with a very similar profile will result in highly redundant biclustering results, while query-genes with a profile very different from that of the other query-genes will result in unique but equally interesting biclusters. Therefore, the ensemble strategy is designed to reduce redundancy of the obtained solutions, while at same time maintaining to a maximal extent the distinct solutions that were present in the query-based biclustering outcomes of the individual query-genes. This application of an ensemble-based strategy is inherently different from its traditional use where it is mainly meant to increase accuracy of clustering results by searching for genes that were found frequently coexpressed over multiple runs (Monti *et al*., 2003; Strehl and Ghosh, 2002).

The ensemble approach was validated using different evaluation metrics that assess both the agreement with the original biclustering solutions as the quality of the consensus biclusters independent of these query-based biclustering outcomes. We tested the influence of using different transformations of the consensus matrix in combination with different graph clustering methods on the quality of the consensus biclusters. While all tested combinations of matrix transformations and graph clustering methods resulted in consensus biclusters that recapitulate the original query-based biclustering solutions and reduce redundancy, using fuzzy clustering to extract consensus clusters from a pruned consensus matrix gave the overall best results.

Further, we showed that whenever dealing with an input set of which at least some of the genes are expected to have a strongly different expression profile, the ensemble based 'split-and-merge strategy' clearly outperforms other strategies of using query-based biclustering in identifying biologically relevant biclusters. Biclustering results are in general also more focused, with a higher coverage in query-genes, when compared to those obtained with any other global biclustering approach.

To illustrate how query-based biclustering in combination with our ensemble approach can be used to interrogate a gene expression compendium with own experimental data, we applied it to an FNR ChIP-chip case study. By combining the ChIP-chip list with the public data we could obtain a view on its quality: not only could the analysis point out potential false positive ChIP-chip targets, but it also showed that most of the targets repressed by FNR were missing from the ChIP-chip list.

## ACKNOWLEDGEMENTS

All authors read and approved the final manuscript. We would like to thank the Associate Editor and the three anonymous reviewers whose comments substantially improved the manuscript. We thank Hong Sun and Karen Lemmens for their assistance with the ViTraM software and useful discussions.

*Funding*: At the time of writing R.D.S. was a research assistant of Flemmish government agency for Innovation by Science and Technology. This work is supported by Katholieke Universiteit Leuven (GOA AMBioRICS, GOA/08/011, CoE EF/05/007, SymBioSys and; REA/08/023); IWT (SBO-BioFrame); IUAP P6/25 (BioMaGNet); FWO IOK-B9725-G.0329.09; HFSP-RGY0079/2007C.

*Conflict of Interest*: none declared.

## REFERENCES

Adler,P. *et al*. (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.*, **10**, R139.

Asur,S. *et al*. (2007) An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics*, **23**, i29–i40.

Csardi,G. *et al*. (2010) Modular analysis of gene expression data with R. *Bioinformatics*, **26**, 1376–1377.

Dhollander,T. *et al*. (2007) Query-driven module discovery in microarray data. *Bioinformatics*, **23**, 2573–2580.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gama-Castro,S. *et al*. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.

Gao,F. *et al*. (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.

Grainger,D.C. *et al*. (2007) Transcription factor distribution in Escherichia coli: studies with FNR protein. *Nucleic Acids Res.*, **35**, 269–278.

Hibbs,M.A. *et al*. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.

Ihmels,J. *et al*. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Joshi,A. *et al*. (2008) Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*, **24**, 176–183.

Keseler,I.M. *et al*. (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.

Lemmens,K. *et al*. (2009) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome Biol.*, **10**, R27.

Monti,S. *et al*. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Lear.*, **52**, 91–118.

Newman,M.E. (2004) Analysis of weighted networks. *Phys. Rev. E*, **70**, 056131-1-056131-9.

Newman,M.E. (2006) Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA*, **103**, 8577–8582.

Owen,A.B. *et al*. (2003) A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res.*, **13**, 1828–1837.

Pollard,K. and van der Laan,M. (2005) Cluster analysis of genomic data. In Gentleman,R. *et al*. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, pp. 209–228.

Serrano,M.A. *et al*. (2009) Extracting the multiscale backbone of complex weighted networks. *Proc. Natl Acad. Sci. USA*, **106**, 6483–6488.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

Sun,H. *et al*. (2009) ViTraM: visualization of transcriptional modules. *Bioinformatics*, **25**, 2450–2451.

Tanay,A. *et al*. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl. 1), S136–S144.

Ulitsky,I. *et al*. (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.*, **5**, 303–322.

Van Dongen,S. (2000) Graph clustering by flow simulation. *PhD Thesis*, University of Utrecht, Utrecht, The Netherlands.

Wade,J.T. *et al*. (2007) Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization. *Mol. Microbiol.*, **65**, 21–26.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, 1–37.

Zhao,H. *et al*. (2011) Query-based biclustering of gene expression data using Probabilistic Relational Models. *BMC Bioinformatics*, **12** (Suppl. 1), S37.