An ensemble method for querying gene expression compendia with experimental lists

Riet De Smet, Kathleen Marchal
Department of Microbial and Molecular systems
Katholieke Universiteit Leuven, Belgium
E-mail: riet.desmet@esat.kuleuven.be, kathleen.marchal@biw.kuleuven.be

***Abstract*—Query-based biclustering can be used to explore public gene expression data for genes coexpressed with genes of interest to a certain researcher (the query). These methods, however, fail when faced with a list of query-genes with diverse expression profiles. In addition, a threshold on the minimal coexpression with the query-genes needs to be defined in advance. To deal with these problems we introduce an ensemble approach for 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 coexpression thresholds are merged in a statistically robust way. Graph clustering is used to obtain non-redundant consensus biclusters from this matrix. We tested out different ensemble construction schemes and illustrate the effectiveness of this approach.**

***Keywords*-biclustering; ensemble method; gene expression compendia; query-based**

I. Introduction

Interpreting results of experimental assays in light of the increasing amount of publicly available gene expression data can help revealing inconsistencies between own experiments and available data. In addition, such a comparison allows viewing experimental results in the more global context of the cell unveiled by the global expression data. Query-based search strategies, such as prioritization methods [1]–[3] and query-based biclustering methods [4], [5], have been developed to query gene expression compendia for genes tightly coexpressed with gene(s) of interest and therefore can be potentially used to this purpose. Applying query-based methods to search for gene coexpressed with genes from an experimental list, leads however to tedious post-processing and difficult interpretation of the results.

First, experimentally derived query-sets are often heterogeneous in their expression profiles. As existing query-based tools generally query the expression compendium with the average expression profile of the query-set, query-based biclustering must be applied to each gene from the query-list separately to avoid the query-profile to be deteriorated by query-genes that are not mutually coexpressed. Some of the query-genes might, however, still be coexpressed since they are derived from the same experimental assay. Therefore this procedure often results in at least partially redundant biclustering solutions.

A second issue concerns defining a threshold on the minimal level of coexpression of additionally recruited genes with the query. Indeed, as it is often not *a priori* known how tightly a set of genes should be coexpressed to be biologically meaningful, several query-based biclustering methods [4], [5] avoid hard thresholding and incorporate a resolution sweep approach in which a whole range of possible threshold values is scanned in one algorithmic run. This, however, requires the user to select the most relevant solution a posteriori.

Consequently, when existing query-based strategies are applied to explore expression compendia for genes coexpressed with experimentally derived gene sets, these problems require the user to scan lots of potentially redundant query-based biclustering results obtained for many possible thresholds on coexpression. Here we introduce an ensemble strategy [6], [7] to merge multiple query-based biclustering results obtained for a whole range of different thresholds into a few non-redundant consensus biclusters, which allows for easy interpretation of the query-list within the context of the gene expression compendium.

II. Materials and methods

A. Datasets

An *Escherichia coli* cross-platform gene expression compendium containing 4557 genes and 870 conditions derived from publicly available gene expression data was used [8]. A query-list was obtained from an *E. coli* FNR ChIP-chip experiment [9]. In this experiment 63 genomic regions bound by FNR were identified, which were mapped to 90 genes. Gene functional GO-categories were taken from EcoCyc [10].

B. Query-driven biclustering

The strategy proposed in this paper can be used in conjunction with any query-based strategy. For illustrative purposes we use here query-driven biclustering (QDB) [5]. Briefly, QDB takes as input one or multiple query-genes and a gene expression data set and outputs a bicluster solution centered on the average expression profile of the query-genes. The algorithm uses a resolution sweep approach to evaluate in a single run of the algorithm all possible solutions

978-1-4244-8305-1/10/$26.00 ©2010 IEEE

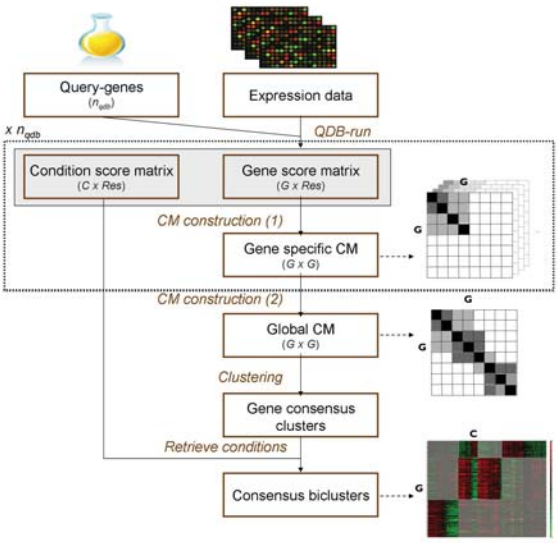

Figure 1. Overview of the computational framework. n_{qdb} targets from a ChIP-chip analysis were each taken as input of a query-driven biclustering algorithm (QDB) [5]. Each QDB-run outputs the results for a single ChIP-chip target and consists of a gene score matrix and condition score matrix. G refers to the dataset's genes, C to the conditions and Res to the resolution parameters. The global consensus matrix (CM) construction is a two-step procedure in which first a gene-specific consensus matrix is constructed for each query, resolving the resolution issue. In a second step the gene-specific consensus matrices for all n_{qdb} genes in the ChIP-chip list are merged into a single global consensus matrix, to remove redundancy amongst QDB-outcomes (shades of grey are representative of the magnitude of the consensus scores). Finally, this global consensus matrix is clustered and the corresponding conditions are retrieved from the condition score matrices, resulting in consensus biclusters.

that correspond to different degrees of coexpression (*i.e.* the resolution), first outputting bicluster solutions with only a limited number of genes that are all tightly coexpressed with the query-profile and then gradually adding genes that are less tightly coexpressed with the query. In this work, a single run of the algorithm outputs the results for a specific query-gene and consists of two matrices: the gene score matrix and condition score matrix (Figure 1). These matrices contain for each resolution (Res) the loglikelihood ratio of the genes (gene scores Gs_i) and conditions (condition scores Cs_i) belonging to the bicluster (hence for each resolution a gene score vector and condition score vector is obtained). We refer to the output of such a single run as the "QDB-outcome". Here, as is motivated in Introduction, we run QDB on all n_{qdb} genes from the query-list separately, generating an ensemble of QDB-outcomes.

C. An ensemble strategy for query-based biclustering

The computational framework to derive a consensus solution for the ensemble of QDB-outcomes is outlined in Figure 1. We first focus on generating a consensus in the gene direction before retrieving the relevant conditions.

1) Global consensus matrix construction: The global consensus matrix is a gene-by-gene matrix containing consensus scores C_{ij}. These scores approximate the similarity between two genes by calculating the frequency of co-occurrence of these genes across all n_{qdb} QDB-runs. We follow a two-step procedure to construct the consensus matrix.

In the first step we resolve the resolution issue. Specifically, we construct a *gene-specific consensus matrix* which merges the results for separate QDB-runs, each obtained for a single query-gene and for n_{res} different values of the resolution parameter. Matrix entries reflect the average gene pair-to-bicluster membership across all resolutions tested: $C_{ij}^{qdb}=\frac{\sum Gs_i^{n_{res}}.Gs_j^{n_{res}}}{n_{res}}$. The rationale behind this is that genes that co-occur in both fine-grained and coarser-grained biclusters (corresponding to a decreasing tightness of coexpression) get a higher consensus score than genes that only co-occur in coarser-grained biclusters.

In the second step, a *global consensus matrix*, which aggregates all the gene-specific consensus matrices for the different query-genes, is constructed. Here, we remove the redundancy amongst the QDB-outcomes by assuming that genes that co-occur repeatedly across different QDB-runs form a single grouping. However, as not all query-genes give rise to similar QDB-outcomes, we do not only aim to reduce the redundancy amongst the gene sets but also to preserve the QDB-outcomes that were not repeatedly retrieved for different query-genes (*i.e* the non-redundant outcomes). Therefore, we introduce a *distributed consensus matrix construction approach*. Here, the frequency of co-occurrence for a gene-pair (gene consensus score) is calculated as its sum over all gene-specific consensus matrices, accounting for the number of times a certain gene pair co-occurred in the specific consensus matrices: $C_{ij}^{global}=\frac{\sum_{n_{qdb}} C_{ij}^{n_{qdb}}}{\sum_{n_{qdb}} O(C_i^{n_{qdb}}, C_j^{n_{qdb}})}$, with $O(C_i^{n_{qdb}}, C_j^{n_{qdb}})$ representing the co-occurrence function which is 1 if both genes belong to the same specific consensus matrix and otherwise 0. Simply averaging the gene-specific matrices across all QDB-solutions would erroneously downweigh those genepairs specific to a certain QDB-run and reward genepairs retrieved by all QDB-runs.

We also tested two consensus matrix transformations to see whether we could further improve upon the ensemble solution. The first transformation concerns the Topological Overlap Matrix (TOM) [11] which does not only account for pairwise gene-gene co-occurrence but also for similarity in the other genes with which both genes co-occur in the QDB-solutions. For the second transformation we tested whether pruning the consensus matrix could improve the outcome. To this end we used the disparity filter [12] to put statistically insignificant consensus scores to zero. We

choose our significance level for the disparity filter such that 90% of the total consensus score (*i.e.* the sum of all consensus scores) was retained as to not eliminate too many elements with large consensus scores from the matrix.

2) Extracting consensus clusters from the consensus matrix: To obtain non non-redundant gene consensus clusters we cluster the global consensus matrix. In particular, as the global consensus matrix can be considered as a weighted graph, with weights corresponding to the consensus scores, we compared several graph clustering methods that can be applied for this purpose. These methods include the Newman spectral modularity algorithm [13], affinity propagation (AP) [14], Markov clustering (MCL) [15] hierarchical clustering and a recently published fuzzy spectral graph clustering method [16].

The Newman spectral modularity algorithm and the fuzzy spectral method select automatically the number of clusters. For AP we use the default parameters, for MCL the efficiency measure [15] and for hierarchical clustering the silhouette coefficient to select the optimal number of clusters.

We refer to the whole set of consensus clusters obtained from the consensus matrix as the "consensus solution". For each of the graph clustering methods, gene 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) Constructing consensus biclusters by retrieving the relevant conditions: To map the conditions to the gene consensus clusters, we trace back the obtained gene consensus clusters to the original QDB-outcomes from which they were derived. Specifically, we use the geometric coefficient [17] to quantify for a gene consensus cluster its overlap in its genes with each of the the QDB-outcomes: $Overlap = \frac{|G_{cons} \cap G_{qdb}|}{\sqrt{|G_{cons}||G_{qdb}|}}$, with G_{cons} representing the genes in the consensus cluster and G_{qdb} the genes in the QDB-bicluster. Since, each QDB-outcome corresponds to different gene sets retrieved for different values of the resolution parameter, the overlap is calculated 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 gene consensus cluster are calculated as the weighted mean of all condition score vectors (one per QDB-outcome) retained for this consensus cluster. The weight is chosen equal to the geometric coefficient, hence giving higher weight to condition score vectors belonging to QDB-outcomes better reflected by the gene consensus clusters. Finally conditions with a consensus score exceeding 0.75 (conditions occur in at least 75% of the condition score vectors) are retained.

D. Performance evaluation

To evaluate the performance of the obtained consensus biclusters for the different consensus matrix transformations and graph clustering methods we define the following evaluation metrics which assess different aspects of the ensemble framework:

1) The *overlap measure* evaluates the agreement with original QDB-outcomes by calculating for each consensus cluster its maximal overlap with the original QDB-outcome as measured by the geometric coefficient (see above).
2) The *redundancy measure* evaluates the extent to which redundancy is removed from the origingal QDB-outcomes. Ideally query-genes with similar (*i.e.* redundant) QDB-outcomes should end up in the same consensus cluster. Therefore we compare a clustering of the query-genes based on their similarity in QDB-outcomes (overlap in gene content as assessed by geometric coefficient is used as similarity measure) with the clustering of the query-genes in the consensus solution. We use the Normalized Mutual Information (NMI) [7] to quantify the extent to which both clusterings of the query-genes (*i.e.* based on QDB-outcome and based on consensus solution) are the same.
3) The *functional coherency* evaluates the biological relevance of the consensus clusters. We calculate for each consensus cluster the *functional coherence* using the hypergeometric test ($p < 0.01$, Bonferroni-corrected for multiple testing). We use the clustering score function [18] to aggregate all p-values into one score obtained for a certain consensus clustering: 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 $f_c = 1 - \frac{\sum_{k=1}^{n_s} min(p_k) + (n_i \cdot c)}{(n_s + n_i) \cdot c}$
4) The *modularity* [19] assesses the statistical quality of the consensus clusters by comparing the fraction of the edges that fall within a given cluster minus the expected fraction if edges were distributed at random. The higher the modularity the better the cluster separation.

All metrics have a maximal value of 1, which makes their interpretation straightforward.

III. Results

Here we apply our ensemble framework for query-based biclustering to a query-list of 90 FNR ChIP-chip targets in *E. coli* [9]. We first generate an ensemble of query-based biclustering solutions by running QDB [5] for each gene from the query-list separately on an *E. coli* gene expression compendium [8]. This results in 44 query-driven biclustering solutions (not for all query-genes a solution could be obtained) that at least partially overlap: 23/44 QDB-outcomes show an overlap of at least 50% (geometric coefficient) with at least one other QDB-outcome. To remove

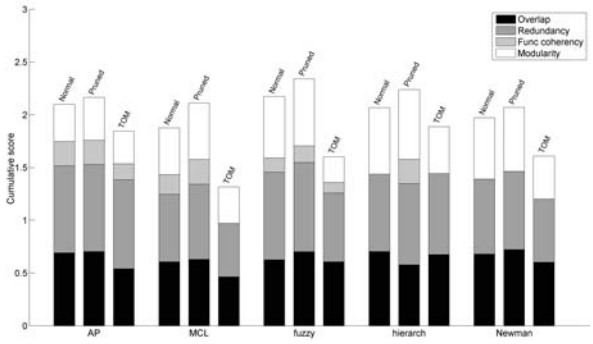

Figure 2. Comparison of different consensus constructions. Each bar corresponds to a different consensus matrix transformation ('Normal' refers to no transformation). Bars are grouped per clustering method (x-axis).

Table I
COMPARISON OF CONSENSUS CONSTRUCTIONS FOR DIFFERENT GRAPH CLUSTERING METHODS, ON THE PRUNED CONSENSUS MATRIX

	AP	MCL	Fuzzy	Hierarch	Newman
Overlap	0.70	0.63	0.70	0.58	0.72
Redundancy	0.82	0.71	0.84	0.77	0.74
Func coherency	0.23	0.24	0.16	0.23	0
Modularity	0.41	0.54	0.64	0.66	0.61

the redundancy amongst the QDB-outcomes we apply the proposed ensemble approach. As such the 44 bicluster solutions are merged into 4 to 21 consensus biclusters, depending on the combination of consensus matrix transformation and the graph clustering method used.

A. Validation ensemble approach

In Figure 2 we evaluate the different consensus matrix constructions and graph clustering methods using the evaluation metrics introduced in Materials and Methods. For each combination of consensus matrix transformation and graph clustering we define a cumulative score as the sum of the 4 performance measures introduced above. These cumulative scores are represented as stacked bar graphs in Figure 2 and allow for the comparison of different consensus constructs. As each evaluation metric has a maximal value of 1, the maximal value for the cumulative score is 4.

First, with respect to the consensus matrix transformation used, we observe that pruning the consensus matrix improves for all graph clustering methods the cumulative score of all evaluation metrics (Figure 2). The fact that a better cumulative score is obtained than for the non-transformed consensus matrix illustrates that putting non-significant consensus scores to zero before clustering the consensus matrix results in consensus biclusters that better represent the original QDB-outcomes (higher 'Overlap-score'). Applying TOM, on the other hand, seems to disturb the relation of the consensus solution to the original QDB-oucomes as indicated by the low overlap and low redundancy measures.

Secondly, we observe that for the pruned consensus matrix fuzzy clustering outperforms the other graph clustering methods with respect to agreement with the original QDB-solutions. Indeed, having on average an overlap of 70% with the original QDB-outcomes and a redundancy score of 0.84, the results obtained by fuzzy clustering on the pruned consensus matrix seem to be truthful to the original QDB-outcomes (Table I). AP performs similarly for these metrics but fuzzy clustering outperforms AP w.r.t. cluster coherency as assessed by the modularity measure.

Table II
INTERPRETATION CHIP-CHIP TARGETS IN TERMS OF CONSENSUS BICLUSTER CONTENT

		In interesting bicluster	
	In consensus	enrich	coverage
Novel target	37	7	8
Known target	24	17	20
Total	61	24	28

B. Biological interpretation consensus biclusters

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. 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 transcription factor (TF). Here we interpret the outcome of the ChIP-chip assay [9] in terms of the obtained consensus biclusters. We choose the results obtained with fuzzy clustering on the pruned consensus matrix as for this combination the best performance was observed (see above). Here the 44 QDB-outcomes were merged into 17 consensus biclusters that cover 61 of the 90 ChIP-chip targets, amongst which 24 known FNR targets [20] (Table II - 'In consensus').

To get a sense of reliability of the output of the ChIP-chip assay we assessed (1) whether ChIP-chip targets are mutually coexpressed (or co-cluster) and (2) whether they are coexpressed (co-cluster) with known FNR-targets. As for (1) we could identify 2 consensus biclusters with a significant enrichment for ChIP-chip targets ($p < 0.05$, hypergeometric test, Bonferroni-corrected). These 2 consensus biclusters contain in total 24 ChIP-chip targets of which 17 are documented to be regulated by FNR [20], whereas the 7 remaining targets are novel FNR targets identified by the ChIP-chip assay (Table II - 'enrich'). As these ChIP-chip targets are not only bound by the same TF but also coexpressed they are likely functional FNR-targets. Regarding (2), we observed that 6 consensus biclusters showed a high coverage for known FNR-targets (*i.e.* $> 33\%$ of the genes within these biclusters were known FNR-targets [20]), including the 2 biclusters that were significantly enriched in the ChIP-chip list. These 6 consensus biclusters contained in total 28 ChIP-chip targets, of which 20 are documented FNR-targets (Table II - 'coverage') and 8 are novel ones.

Through this analysis we can confirm 20 out of the 24

ChIP-chip targets that are known to be regulated by FNR [20]) and for which we could retrieve a consensus bicluster (Table II), illustrating that we identify true positives of the assay with a high sensitivity.

IV. Discussion and conclusion

In this paper we tackled the problems that query-based search methods are confronted with when faced with a list of experimentally derived genes by introducing an ensemble method for query-based biclustering.

This ensemble method deals with the biological resolution issue by stressing genes that are repeatedly retrieved for multiple resolutions. In addition it copes with the heterogeneity of the input list by employing a "split and merge strategy": we first derive query-based biclustering solutions for each query-gene separately and then merge the partially redundant solutions in a consensus solution, which retains the distinct solutions amongst the ensemble of query-based biclustering solutions. Whereas ensemble methods have traditionally been used to enforce robustness and increase accuracy of clustering results [6], [7], here we applied it in a novel way: we introduced a distributed consensus matrix approach to remove redundancy and to simultaneously retain as much information as possible of the original QDB-solutions.

Using different evaluation metrics and comparing different ways of constructing the ensemble it was illustrated that consensus biclusters can be obtained that are in good agreement with the original bicluster solutions.

Acknowledgment

RDS is a research assistant of IWT. This work is further supported by: 1) KUL: GOA AMBioRICS, GOA/08/011, CoE EF/05/007 SymBioSys, CREA/08/023; 2) IWT: SBO-BioFrame; 3) IUAP P6/25 (BioMaGNet); 4) FWO IOK-B9725-G.0329.09; 5) HFSP-RGY0079/2007C.

References

[1] M. Hibbs, D. Hess, C. Myers, C. Huttenhower, K. Li, and O. Troyanskaya, "Exploring the functional landscape of gene expression: directed search of large microarray compendia," *Bioinformatics*, vol. 23, no. 20, pp. 2692–2699, 2007.

[2] A. Owen, J. Stuart, K. Mach, A. Villeneuve, and S. Kim, "A gene recommender algorithm to identify coexpressed genes in C. elegans," *Genome Research*, vol. 13, no. 8, pp. 1828–1837, 2003.

[3] P. Adler, R. Kolde, M. Kull, A. Tkachenko, H. Peterson, J. Reimand, and J. Vilo, "Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods," *Genome Biology*, vol. 10, no. 12, p. R139, 2009.

[4] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing modular organization in the yeast transcriptional network," *Nature genetics*, vol. 31, no. 4, pp. 370–377, 2002.

[5] T. Dhollander, Q. Sheng, K. Lemmens, B. De Moor, K. Marchal, and Y. Moreau, "Query-driven module discovery in microarray data," *Bioinformatics*, vol. 23, no. 19, pp. 2573–2580, 2007.

[6] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[7] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[8] K. Lemmens, T. De Bie, T. Dhollander, S. De Keersmaecker, I. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. Collado-Vides *et al.*, "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli," *Genome Biology*, vol. 10, no. 3, p. R27, 2009.

[9] D. Grainger, H. Aiba, D. Hurd, D. Browning, and S. Busby, "Transcription factor distribution in Escherichia coli: studies with FNR protein," *Nucleic acids research*, vol. 35, no. 1, pp. 269–278, 2006.

[10] I. Keseler, C. Bonavides-Martinez, J. Collado-Vides, S. Gama-Castro, R. Gunsalus, D. Aaron Johnson, M. Krummenacker, L. Nolan, S. Paley, I. Paulsen *et al.*, "EcoCyc: a comprehensive view of Escherichia coli biology," *Nucleic acids research*, vol. 37, no. Database issue, pp. D46–D470, 2009.

[11] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, pp. 1128–1164, 2005.

[12] M. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[13] M. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[14] B. J. Frey and D. Dueck, "Clustering by passing messages between data points." *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[15] S. van Dongen, "Graph clustering by flow simulation," Ph.D. dissertation, U. of Utrecht, 2000.

[16] A. Joshi, Y. Van de Peer, and T. Michoel, "Analysis of a gibbs sampler method for model-based clustering of gene expression data." *Bioinformatics*, vol. 24, no. 2, pp. 176–183, 2008.

[17] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world." *Proceedings of the National Academy of Sciences*, vol. 100, no. 8, pp. 4372–4376, 2003.

[18] S. Asur, D. Ucar, and S. Parthasarathy, "An ensemble framework for clustering protein-protein interaction networks." *Bioinformatics*, vol. 23, no. 13, pp. i29–i40, 2007.

[19] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, pp. 056 131–1–056 131–9, 2004.

[20] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado *et al.*, "Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation." *Nucleic acids research*, vol. 36, no. Database issue, pp. D120–D124, 2008.