


| | | | | |
|--|----------------|----------------|----------------------------|-------------------------|
|  aptara <small>The Content Transformation Company</small> | NYAS | nyas_1465000 | Dispatch: 11-9-2008 | CE: N/A |
| | Journal | MSP No. | No. of pages: 7 | PE: Amanda/Carey |

The Condition-Dependent Transcriptional Network in *Escherichia coli*

Karen Lemmens,^a Tijn De Bie,^{b,e} Thomas Dhollander,^a Pieter Monsieurs,^c Bart De Moor,^a Julio Collado-Vides,^d Kristof Engelen,^c and Kathleen Marchal^c

^a*Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium*

^b*Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom*

^c*Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Leuven, Belgium*

^d*Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, México*

^e*OKP Research Group, Katholieke Universiteit Leuven, Leuven, Belgium*

Thanks to the availability of high-throughput omics data, bioinformatics approaches are able to hypothesize thus-far undocumented genetic interactions. However, due to the amount of noise in these data, inferences based on a single data source are often unreliable. A popular approach to overcome this problem is to integrate different data sources. In this study, we describe DISTILLER, a novel framework for data integration that simultaneously analyzes microarray and motif information to find modules that consist of genes that are co-expressed in a subset of conditions, and their corresponding regulators. By applying our method on publicly available data, we evaluated the condition-specific transcriptional network of *Escherichia coli*. DISTILLER confirmed 62% of 736 interactions described in RegulonDB, and 278 novel interactions were predicted.

Key words: transcriptional modules; frequent itemset mining; DISTILLER

Introduction

The transcriptional network of *E. coli* is among the best characterized networks; it has been estimated that 25% of its interactions have been resolved,¹ and for about 70 regulators the regulatory binding sites are known.² Thanks to the availability of high-throughput omics data, we can use computational approaches to predict novel interactions. As such, methods that infer transcriptional interactions from microarray only have been successfully applied to further extend the *E. coli* tran-

scriptional network.³ However, high noise levels often make predictions from single data sources unreliable. It is well known that data integration may alleviate this problem, by exploiting complementarities in different data sources.

In this study we describe DISTILLER (Data Integration System To Identify Links in Expression Regulation), a data integration framework that simultaneously analyzes microarray and motif information to find modules of genes that are co-expressed in a subset of conditions, together with their corresponding regulators. We applied DISTILLER to a cross-platform array compendium and motif data to further complete the RegulonDB network and to study its condition dependency.

Address for correspondence: Kathleen Marchal, Kasteelpark Arenberg 20, B-3001 Heverlee (Leuven), Belgium. Voice: +3-216-329-685; fax: +3-216-321-966. kathleen.marchal@biw.kuleuven.be

Results and Discussion

Distiller

The method introduced in this paper builds upon a previous regulatory module detection tool,^{4,5} but has been significantly redesigned to increase scalability and reduce the number of parameters required. More importantly, DISTILLER includes a condition selection strategy: co-expression of genes is required in only a significant subset of the complete condition set (bicluster strategy). By including this condition selection we can apply the algorithm to large heterogeneous expression compendia.

Our methodology consists of three steps: 1) the identification of seed modules; 2) the reduction of the set of all seed modules to a manageable set of nonredundant (small overlap) and significant (association in the data cannot be explained by chance) seed modules; and 3) the extension of the seed modules thus obtained with additional genes.

Although our approach can be extended to any number of input matrices or data sources, we used two input matrices to generate the results in this paper:

- An expression data compendium \mathbf{A} (see Materials and Methods) with dimensions $\mathcal{N}_G \times \mathcal{N}_C$, where \mathcal{N}_G indicates the total number of genes in the compendium and \mathcal{N}_C the total number of conditions.
- A binary regulatory motif matrix \mathbf{R} (input interaction matrix) with dimensions $\mathcal{N}_G \times \mathcal{N}_R$, where \mathcal{N}_R is the total number of regulators for which motif data is available. Each element r_{ip} of this matrix indicates whether the upstream region of a specific gene i contains a statistically significant motif instance of the known regulatory motif model of that specific regulator p .

Seed Modules

Let $\mathbf{g}^{(m)}$ be the set of gene indices that correspond to genes in module m . DISTILLER ini-

tially identifies seed modules that satisfy three constraints:

(C_G) The module should contain a minimum number G of genes (that is, a gene content threshold), or $|\mathbf{g}^{(m)}| \geq G$.

(C_R) All genes $\mathbf{g}^{(m)}$ in the module should contain motif instances for a sufficient number R of common, a priori, unspecified regulators. Let $\mathbf{1}$ represent an all-ones vector of appropriate dimensions. Then, this constraint can be formulated mathematically as $|\{p \mid \mathbf{1}^T \mathbf{R}_{\mathbf{g}^{(m)}, p} = \mathbf{g}^{(m)}\}| \geq R$.

(C_C) All genes $\mathbf{g}^{(m)}$ in the module should be significantly co-expressed in a sufficiently large, a priori unspecified set of experimental conditions. We indicate this required number of conditions by C . In order to calculate the maximal valid condition set $\mathbf{c}^{(m)}$ given a certain gene set $\mathbf{g}^{(m)}$, we first compute the difference between the largest and smallest expression levels in the gene set for each of the conditions: $BW_j = \max(\mathbf{A}_{\mathbf{g}^{(m)}, j}) - \min(\mathbf{A}_{\mathbf{g}^{(m)}, j})$. We call BW_j the bandwidth for condition j . Subsequently, these bandwidths are sorted in increasing order to obtain a sorted bandwidth sequence $BW_j^{(s)}$ that satisfies $BW_{j-1}^{(s)} \leq BW_j^{(s)} \leq BW_{j+1}^{(s)}$ (see Fig. 1A). The sequence $BW_j^{(s)}$ is then compared with a prespecified threshold bandwidth sequence $BW_j^{(th)}$ that is sorted as well: $BW_{j-1}^{(th)} \leq BW_j^{(th)} \leq BW_{j+1}^{(th)}$. Gene set $\mathbf{g}^{(m)}$ is said to be co-expressed in C_{max} conditions if the sorted bandwidth sequence $BW_j^{(s)}$ is completely within the threshold bandwidth sequence $BW_j^{(th)}$ for $1 \leq j \leq C_{max}$, that is, if $BW_j^{(s)} \leq BW_j^{(th)}$ for $1 \leq j \leq C_{max}$. Constraint C_C is satisfied if this property holds for $C_{max} \geq C$.

A naive exhaustive search for valid modules as defined above would require checking all possible combinations of genes, motif instances, and experimental conditions. This is unfeasible for data sets of any reasonable size. Itemset mining algorithms are well suited to solve this problem. In our previous work we therefore have adopted a breadth-first strategy that resembles the Apriori algorithm.^{4,5} In the current work we use a depth-first search more similar to CHARM.⁶ The depth-first strategy guarantees a better scalability, especially on non-sparse data sets (such as the expression compendium, which was necessary to obtain the results in the current paper). Even though the

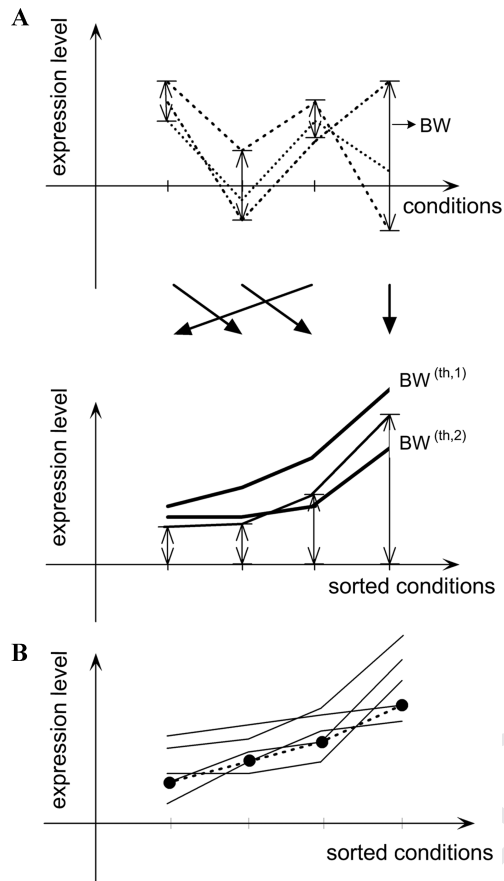


Figure 1. (A) Definition of the bandwidth and its use for condition selection. The top half shows hypothetical expression profiles for three different genes. For each of the four conditions, the bandwidth for this set of genes is indicated with a vertical bidirectional arrow. The lower half shows the bandwidth sequence for these expression data, obtained by sorting the bandwidths in increasing order. Two threshold bandwidth sequences ($BW^{(th,1)}$ and $BW^{(th,2)}$) are shown as well (bold lines). If we were to use $BW^{(th,1)}$, this set of three genes would qualify as a module for the expression data, since the bandwidth sequences lie entirely below $BW^{(th,1)}$. However, if we were using $BW^{(th,2)}$, this set of genes would not qualify as a module. (B) Selection of the bandwidth threshold based on random data. To illustrate how a threshold bandwidth sequence is obtained from the data, we show five bandwidth sequences corresponding to randomly sampled sets of genes. The dashed line is a candidate threshold bandwidth sequence, obtained by connecting all second smallest values for all four sorted conditions. The threshold bandwidth sequence that is actually used is the one that qualifies a fixed fraction of randomly sampled gene sets as a module (box P -value).

exploration of all gene sets has become feasible in this way, the number of modules may still be impractically large. DISTILLER solves this problem by reporting only closed modules. These are modules that cannot be further extended by any other gene without reducing the number of motifs that all of its genes share, or the number of conditions in which its genes are expressed similarly. DISTILLER applies techniques used in CHARM to efficiently ignore nonclosed modules while selecting all the closed ones. We report only closed modules with a number of genes larger than or equal to a threshold G (the “gene content threshold”) because only those modules containing a minimum number of genes are interesting.

Selecting Interesting Nonredundant Modules

Despite the massive reduction in the number of modules achieved by using CHARM, the output may still be too large to explore: small amounts of noise in the data may cause one module to appear as a large number of separate partially overlapping modules, all very similar in gene, regulator, and condition content. To address this problem we apply an iterative procedure that selects the most interesting modules one by one. It takes into account the significance of individual modules but at the same time penalizes overlap with modules that have already been reported.

Seed Module Extension

In a subsequent extension step, we recruit additional candidate module genes that did not pass the stringent seed discovery step but should be considered part of the module, such as downstream operon genes that do not contain a motif instance in their promoter regions but are subject to its regulatory influence. The relaxed criteria for adding additional genes to the module are the following: 1) the gene’s expression profile should have a correlation with the module’s mean expression profile of at least a fraction α of the module correlation, defined as the lowest correlation value between a seed gene’s

expression profile and the average expression profile for the modules conditions, and 2) the genes should have a motif instance with P -value below a threshold β . Both requirements have to be fulfilled unless a gene is part of an operon for which the first gene is present in the seed module. In this case only the first criteriom has to be satisfied.

Condition-Dependent Regulator-Target Interactions

We used our data integration framework, to infer the condition-dependence of modules in *E. coli* and their degree of combinatorial regulation. The input data consisted of an *E. coli* microarray compendium and information on the interaction between a regulator and its targets (see Materials and Methods). The microarray compendium is, to our knowledge the first cross-platform compendium combining 870 arrays, from 70 experiments, performed on four different platforms.

Identifying Regulatory Modules and Regulator-Target Interactions

Motif data were integrated with our expression compendium to generate condition-dependent modules. The 150 statistically most significant modules recovered by DISTILLER represent 454 interactions corresponding to 62% of 736 interactions for 67 regulators with known binding sites described in RegulonDB.² Most modules are enriched for functions in which the regulator was known to be involved. For 37 of the 67 regulators at least part of their regulon could be confirmed. For the remaining 30 regulators no interaction was found; most likely either the number of genes in the corresponding modules fell below the gene content threshold, or the conditions needed to trigger these interactions are not present in our compendium.

In addition to identifying 454 previously described interactions, we predict 278 novel interactions that have not previously been described in RegulonDB. Supplementary file 1

gives an overview of the number of known interactions that were identified per regulator, as well as the number of new, predicted interactions. It shows how for many well-studied regulators the known part of the regulon could still be extended considerably. The large number of predictions for FNR, CRP, ArcA, Fis, and IHF confirms their role as global hubs in the network. For some regulators only the previously described interactions, and/or a few additional ones could be retrieved (for example, CueR, GlpR). Because we found these regulators active in conditions of the compendium but could not extend these regulons any further, we postulate that these regulators are nearly completely characterized and indeed target only a few genes (operons) triggering very specific pathways. As for most of the newly predicted interactions (hypothetical genes), no additional confirmation existed in literature, we assigned them a level of confidence based on the gene composition of the module in which the target was retrieved. If the module contained many previously confirmed targets, tightly co-expressed with the novel target, we can be more confident in its prediction.

Conditional Dependency of the Regulatory Network

When working with expression compendia containing many heterogeneous conditions, it can be expected that coregulated genes are no longer co-expressed over all conditions but only in a subset of conditions. To be able to exploit the continuously growing number of publicly available microarrays, DISTILLER uses a bicluster strategy that not only identifies sets of genes that are co-expressed but also selects the conditions under which these genes are co-expressed.

Figure 2 shows an example of two modules that were inferred by DISTILLER. The genes in both modules are regulated by the two-component regulatory ArcAB system. ArcAB functions as a major control system for the regulation of expression of genes encoding enzymes involved in both aerobic and anaerobic

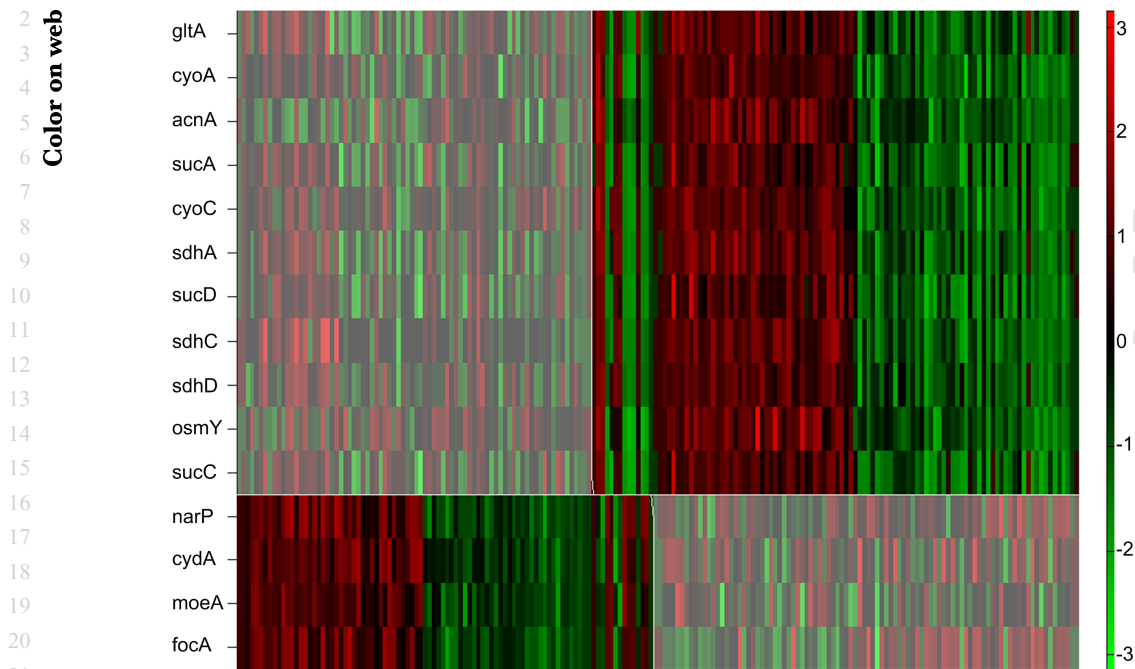


Figure 2. Two ArcA modules that were identified by DISTILLER. The first module consists of four genes: three known ArcA targets and one putative target (*narP*). The known genes are all activated by ArcA in anaerobic conditions. The second module consists of genes involved in aerobic metabolism. These genes are repressed by ArcA in an anaerobic environment. The expression behavior illustrates that, although the genes in both modules are all under the control of ArcA, they are regulated differently. This is also evident from the opposite expression behavior that can be observed in the common conditions between the modules.

catabolic pathways. The first module contains
 Q8 three known targets of ArcA (*cydA*, *moeA*, *focA*)
 that are activated by ArcA under anaerobic
 conditions. The fourth gene of the module
 is a putative ArcA target that encodes for a
 Q9 regulator NarP. NarP is responsible for the
 up-regulation of genes involved in the NO₃-
 respiration, a process known to take place in
 the absence of oxygen. The second module
 contains 11 ArcA targets, including one pu-
 Q10 tative target (*omp1*). The previously confirmed
 targets of ArcAB in this module are involved in
 aerobic metabolism and are repressed by Arc-
 AB in anaerobic conditions. Figure 2 clearly
 shows that the genes of both modules are co-
 expressed in different subsets of conditions. The
 modules also have a limited number of condi-
 tions in common. In these conditions, the genes
 involved in aerobic metabolism (first module)
 show an expression behavior (regulation) that

is opposite to the anaerobic targets of ArcAB
 (second module).

Perspectives and Future Work

DISTILLER combines bicluster functional-
 ity with data integration functionality. Since bi-
 clustering is a combinatorial and hence compu-
 tationally hard problem, we exploit advanced
 itemset mining algorithms. With the growing
 number of publicly available microarray data,
 further reduction in computational complexity
 may be achieved by imposing additional con-
 straints, for instance by grouping time series
 conditions.

Although the current work combines mi-
 croarray and motif data only, our framework
 makes it straightforward to include any number
 of additional data sources related to transcrip-
 tional interactions, including ChIP-chip data. Q11
 Novel state-of-the-art data sources such as deep

sequencing data are of particular interest, as they may obtain more information on the presence and amount of mRNA in a particular experimental condition, and hence a more detailed view on the transcriptional network.

Materials and Methods

Microarray Compendium

Our cross-platform compendium contains a large collection of publicly available microarrays. The data were collected from the three major microarray databases: Stanford Microarray Database (SMD),⁷ Gene Expression Omnibus (GEO),⁸ and ArrayExpress (AE).⁹ Additionally, we added four microarray experiments described in the literature that were available as supplementary information. After removing redundant information in the microarray databases, we obtained a total of 870 microarrays.

Input Interaction Data

The input interaction data were based on both experimentally verified and predicted regulatory binding sites. Known binding sites in the motif matrix were derived from RegulonDB. Whenever a motif instance in the promoter region of a gene was experimentally confirmed according to RegulonDB, its corresponding regulator-target interaction was set to “1” in the motif matrix.

To predict novel binding site instances, motif weight matrices corresponding to the binding sites of 67 regulators were downloaded from the RegulonDB website (version 5.6). Upstream regions of all annotated *Escherichia coli* K12 (NC_000913) genes were screened with these motif models. For motif screening and P -value calculations for the identified motif instances, we used the method of Hertzberg.¹⁰ The P -values were used to construct the “motif matrix”, a binary matrix that assigns a motif instance to a gene whenever the gene’s upstream sequence contains at least one instance

of the motif, with a P -value below a threshold of 0.001.

Benchmarking with RegulonDB and Novel Interactions

For genes organized into operons, usually only the promoter region of the first operon gene contains a motif instance. Because in RegulonDB the direct interaction between a regulator and a target gene is derived from the presence of an experimentally verified motif instance, only the interaction between a regulator and the first operon gene is reported. RegulonDB contains information on 736 such interactions.² Therefore, when comparing the interactions inferred by DISTILLER with the known direct interactions in RegulonDB, we only consider those genes that have the motif instance in their promoter region. We consider all direct interactions inferred by DISTILLER that are not direct interactions in RegulonDB as novel interactions. Some of these interactions might have been reported in recent literature not yet covered by RegulonDB.

Running Parameters

We choose our parameter settings such that the seed module consists of at least four genes (the gene content threshold) that share at least one motif and 50 conditions. The threshold for the P -value of the motif instances was set to 0.001. In order to choose the box threshold, 100,000 randomizations were carried out and the box P -value threshold was set to 0.0001. In the post-processing steps, a subset of 150 minimally redundant modules was selected, and the threshold for the coefficient of variation was set to 0.6. For the seed module extension step, correlation parameter α was set to 90% and the relaxed P -value threshold β to 0.05.

Acknowledgments

TD is research assistant of the FWO-Vlaanderen. This work is supported by 1)

Research Council KUL: GOA AMBioRICS, GOA/08/011, CoE EF/05/007 SymBioSys; 2) IWT: SBO-BioFrame, TAD-BioScope-IT; 3) Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet); and 4) EU-RTD: FP6-NoE Biopattern.

Conflicts of Interest

The authors declare no conflicts of interest.

Supporting Information

The following Supporting Information is available for this article:

TABLE S1. Interactions recovered by DISTILLER

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

1. Resendis-Antonio, O. *et al.* 2005. Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genet.* **21**: 16–20.
2. Salgado, H. *et al.* 2006. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* **34**: D394–D397.
3. Faith, J.J. *et al.* 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**: e8.
4. De Bie, T. *et al.* 2005. Discovering transcriptional modules from motif, chip-chip and microarray data. In *Proceedings of the Pac. Symp. Biocomput.* 483–494.
5. Lemmens, K. *et al.* 2006. Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.* **7**: R37.
6. Zaki, M.J. & C. Hsiao. 2002. CHARM: An efficient algorithm for Closed Itemset Mining. In *Proceedings of the Second SIAM International Conference on Data Mining (SDM '02)*. R. Grossman, J. Han, V. Kumar, *et al.*, Eds.: 457–473.
7. Demeter, J. *et al.* 2007. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.* **35**: D766–D770.
8. Barrett, T. *et al.* 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.* **35**: D760–D765.
9. Parkinson, H. *et al.* 2007. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**: D747–D750.
10. Hertzberg, L. *et al.* 2005. Finding motifs in promoter regions. *J. Comput. Biol.* **12**: 314–330.

Queries

- Q1** Author: Is this correct or do you mean genomics and/or proteomics?
- Q2** Author: Should this be "a priori"?
- Q3** Author: Please define acronyms on first use.
- Q4** Author: Is this short title acceptable? If not please edit.
- Q5** Author: Please define these abbreviations.
- Q6** Author: Please define abbreviations.
- Q7** Author: Please define abbreviation.
- Q8** Author: Please define abbreviations.
- Q9** Author: Please define abbreviation.
- Q10** Author: Please define abbreviation.
- Q11** Author: Please define abbreviation.
- Q12** Author: Is this change correct?