Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study

Carito Guziolowski[1] & Philippe Veber & Michel Le Borgne[1] & Ovidiu Radulescu[1,2] & Anne Siegel[1]

[1]IRISA (Inria, UMR CNRS, Université de Rennes 1), Campus de Beaulieu, 35042 Rennes Cedex
[2]IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex

Abstract

We showed in previous papers how to define and to check consistency between experimental measurements and a graphical regulatory model. The purpose of the present work is to validate this methodology on a real-size setting. More precisely, we show 1. that the algorithms we proposed are able to handle models with thousands of genes and reactions, 2. that our methodology is an effective strategy to extract biologically relevant information from gene expression data.

1 Introduction

There exists a wide range of techniques for the analysis of gene expression data. Following a review by Slonim [13], we may classify them according to the particular output they compute: 1. list of significantly over/under-expressed genes under a particular condition, 2. dimension reduction of expression profiles for visualisation, 3. clustering of co-expressed genes, 4. classification algorithms for protein function, tissue categorization, disease outcom, 5. inferred regulatory networks. The last category may be extended to all model-based approaches, where experimental measurements are used to build, verify or refine a model of the system under study.

Following this line of research, we showed in previous papers (see [9], [12] and [14]) how to define and to check consistency between experimental measurements and a graphical regulatory model formalized as an interaction graph. The purpose of the present work is to validate this methodology on a real-size setting. More precisely, we show 1. that the algorithms we proposed in [14] are able to handle models with thousands of genes and reactions, 2. that our methodology is an effective strategy to extract biologically relevant information from gene expression data.

For this we built an interaction graph for the regulatory network of *E. coli* K12, mainly relying on the highly accurate database RegulonDB [10], [11]. Then we compared the predictions of our model with three independant microarray experiments. Incompatibilities between experimental data and our model revealed:

- either expression data which are not consistent with what is described in the literature – *i.e.* there exists publications which contradict the experimental measurement,
- either missing interactions in the model

We are, of course, not the first to address this issue. We gave a special attention to the work of Gutierrez-Rios and co-workers [6]. The authors designed on-purpose microarray experiments in order to measure gene expression profiles in *E. coli* K12 under different conditions. They evaluate the consistency of their experimental results first with those reported in the literature, second with a rule-based formalism they propose. Our main contribution is the use of algorithmic tools that allow inference/prediction of gene expression, and diagnosis in case of inconsistency between a model and expression data.

2 Mathematical framework

2.1 Introductory example

Let us have a look at Fig. 1. This is a common representation for biochemical systems where arrows show activation or inhibition. Basically, an arrow between A and B means that an increase of A tends to increase or decrease B depending on the shape of the arrow. Pushing this intuitive reasoning forward, it could be said that an increase of allolactose (node A on Fig 1) should result in a decrease of $LacI$ protein. However, if both $LacI$ and $cAMP - CRP$ increase, then nothing can be said about the variation of $LacY$.

The aim of this section is first, to provide a formal interpretation for the graphical notation used in Fig. 1; second, to derive constraints on experimental measurements, which mimic and justify our intuitive reasoning. For this, we resort to qualitative modeling ([8]), which may be seen as a principled way to derive a discrete system from a continuous one.

2.2 Equilibrium shift of a differential system

Let us consider a network of interacting cellular constituents (mRNA, protein, metabolite). We denote by $X(i)$ the concentration of the ith species. We assume that the system can be adequately described by a system of differential equations of the form:

$$\frac{dX}{dt} = F(X, P)$$

where P denotes a set of control parameters (inputs to the system). A *steady state* of the system is a solution of the algebraic equation $F(X, P) = 0$ for fixed P. The particular form of F is unknown in general, but will not be needed in the following. Indeed, the only information we need about F is the sign of its partial derivatives $\frac{\partial F_i}{\partial X(j)}$. We call *interaction graph* the graph whose nodes are the constituents $\{1, \ldots, n\}$, and where there is an edge $i \to j$ iff $\frac{\partial F_i}{\partial X(j)} \neq 0$. As soon as F is non linear, $\frac{\partial F_i}{\partial X(j)}$ may depend on the actual state X. In the following, we will assume that the *sign* of $\frac{\partial F_i}{\partial X(j)}$ is constant, that is, the interaction graph is independant of the state. Thus an arrow $i \to j$ means that the rate of production of i depends on $X(j)$.

A typical experiment consists in pertubating a system in a given initial condition, and see what changes. Especially with DNA chips, it is difficult to measure absolute concentrations; rather, it is more often obtained the list of genes that are significantly up or down-regulated. In our setting, this could be interpreted in the following way: initially the system is at steady state X^1_{eq}; following a change in the parameters, the system reaches another steady state X^2_{eq}, if we wait long enough. An experimental measurement corresponds to determining the sign of $X^1_{eq} - X^2_{eq}$.

2.3 Qualitative constraints

In the following, we introduce an equation that relates the sign of variation of a species to that of its predecessors in the interaction graph. To state this result with enough clarity, we need to introduce the following algebra on signs.

We call sign algebra the set $\{+, -, ?\}$ where **?** stands for indeterminate sign. It is provided with addition, multiplication and qualitative equality, defined as:

$+ + - = ?$ $+ + + = +$ $- + - = -$ $+ \times - = -$ $+ \times + = +$ $- \times - = +$
$? + - = ?$ $? + + = ?$ $? + ? = ?$ $? \times - = ?$ $? \times + = ?$ $? \times ? = ?$

$\approx$	+	−	?
+	T	F	T
−	F	T	T
?	T	T	T

Some peculiarities deserve to be mentionned:

- the sum of + and − is indeterminate, as is the sum of anything with indeterminate
- qualitative equality is reflexive, symetric but not transitive, because **?** is qualitatively equal to anything

To summarize, we consider experiments that can be modelled as an equilibrium shift of a differential system under a change of its control parameters. In this setting, DNA chips provide the sign of variation in concentration of many (but not necessarily all) species in the network. Put in another way, a measure is a sign $s(X^2_{eq}(i) - X^1_{eq}(j))$ of the variation of some species i between initial state X^1_{eq} and final state X^2_{eq}. Both states are stationary and unkown.

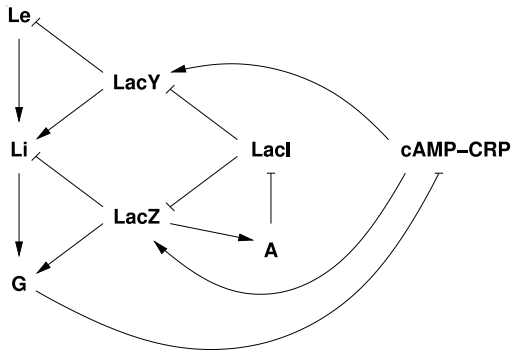
In [9], we proved that under some reasonable assumptions, in particular if the sign of $\frac{\partial F_i}{\partial X(j)}$ is constant, then the following relation holds in sign algebra for all species i:

$$s(X^2_{eq}(i) - X^1_{eq}(i)) \approx \sum_{j \in pred(i)} s(\frac{\partial F_i}{\partial X_j}) s(X^2_{eq}(j) - X^1_{eq}(j)) \quad (1)$$

where $s : \mathbb{R} \to \{+, -\}$ is the sign function, and where $pred(i)$ stands for the set of predecessors of species i in the interaction graph.

2.4 Analyzing a network

Let us now describe a practical use of these results. Given an interaction graph, say for instance the graph illustrated in Fig. 1, we use equation 1 at each node of the graph to build a qualitative system of constraints. The qualitative system associated to our lactose operon model is proposed in the right side of Fig. 1. The variables of this model are the signs of variation for each species. If the model is correct, then any experimental measurement should satisfy the constraints. More precisely, we define a *solution* of the qualitative system as a valuation of its variables, which does not contain any "**?**" (otherwise, the constraints would have a trivial solution with all variables set to "**?**").



$$\begin{cases} LacI \approx -A & (1) \\ A \approx LacZ & (2) \\ LacZ \approx cAMP - LacI & (3) \\ Li \approx Le + LacY - LacZ & (4) \\ G \approx Li + LacZ & (5) \\ cAMP \approx -G & (6) \\ LacY \approx cAMP - LacI & (7) \end{cases}$$

Figure 1: Interaction graph for the lactose operon and its associated qualitative system

Once the system is built, *self-consistency* has to be checked, that is, find if the qualitative system has at least one solution. If this is the case, then it is possible to determine if the model predicts some variations. Namely, it happens that a given variable has the same value in all solutions of the system. Such variable is called a *hard component*. Predictions of the model exactly correspond to the set of hard components of the associated qualitative system.

Now, *checking consistency* between experimental measurements and an interaction graph boils down to instantiating the variables which are measured with their experimental value, and see if the resulting system still has a solution, and if so, compute the set of its hard components.

Whenever the system has no solution, a simple strategy to *diagnose the problem* is to isolate a minimal set of inconsistent equations. In our experiments, a greedy approach was enough to solve all inconsistencies (see next section). Note that isolating a subset of the equations is equivalent in our setting to isolating a subgraph of the interaction graph. This property is very useful for visualisation, when refining a model.

Finally, let us mention that we provided in [14] an efficient representation of qualitative systems, leading to effective algorithms, some of them could be used to get further insights into the model under study. We shall see in the next section that they are enough to deal with large scale networks.

3 Results

3.1 Construction of the Escherichia coli regulatory network

For building E.coli network we used as source the interaction network of E.coli release appeared in RegulonDB [10], [11] on March 2006. From the file *TF − gene interactions* we have built the regulatory network of E.coli as a set of regulations of the form $A \rightarrow B\ sign$ where $sign$ is the value of the interaction: $+, -, ?$ (expressed, repressed, undetermined) and A and B can be considered as genes or proteins, depending on the following situations:

- The relation $genA \rightarrow genB$ was created when among the information provided by RegulonDB, we found protein A, synthesized by $genA$, as the transcriptional factor that regulates $genB$. See Figure 2 A.
- Relation $TF \rightarrow genB$ was created when we found TF as an heterodimer protein (protein-complex formed by the union of 2 proteins) that regulates $genB$. See Figure 2 B. In E.coli transcriptional network we have found 4 protein-complexes which are: IHF, HU, RcsB, and GatR.
- Relation $genA \rightarrow TF$ was created when we found the transcriptional factor TF as an heterodimer protein and $genA$ synthesizes one of the proteins that form TF. See Figure 2 B.

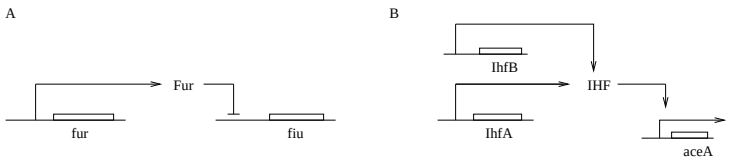


Figure 2: Representation of genetical interactions. **(A)** Negative regulation (repression) of gene fiu by the transcription factor Fur represented as $fur \rightarrow fiu -$. **(B)** Biological interaction of genes $ihfA$ and $ihfB$ forming the protein-complex IHF represented as $ihfA \rightarrow IHF +$ and $ihfB \rightarrow IHF +$, positive regulation of gene $aceA$ by the protein complex IHF represented by $IHF \rightarrow aceA +$

3.2 Adding $sigmafactors$ to obtain self-consistency

Using methods and algorithms described with detail in [14] we built a qualitative system of equations for the interaction graph obtained from *E.coli* network. This system was not found to be self-consistent because it had an incomplete set of interactions that caused an inconsistent system. Thus, interactions among $sigmafactors$ were added to complete the network and in this manner obtaining a network of 3885 interactions and 1529 components (genes, protein-complexes, and sigma-factors). This final network (global network) was found to be self-consistent.

3.3 Compatibility of a network with a set of observations

Once we count with a compatible network, it can be tested with different sets of observations of different conditions: hot, hunger, lack of air, etc. An observation is a pair of values of the form $gene = sign$ where $sign$ can be + or − indicating that the gene is expressed or repressed respectively under certain condition. To test the global network, we have chosen a set of 40 observations for the stationary phase condition of *E. coli* provided by RegulonDB (Table 1).

Table 1: Table of the 40 variations of products observed under stationary growth phase condition. Source: RegulonDB March 2006

(a)

gene	effect
acnA	+
acrA	+
adhE	+
appB	+
appC	+
appY	+
blc	+
bolA	+

(b)

gene	effect
csiE	+
cspD	+
dnaN	+
dppA	+
fi c	+
gabP	+
gadA	+
gadB	+

(c)

gene	effect
gadC	+
hmp	+
hns	+
hyaA	+
ihfA	−
ihfB	−
lrp	+
mpl	+

(d)

gene	effect
osmB	+
osmE	+
osmY	+
otsA	+
otsB	+
polA	+
proP	+
proX	+

(e)

gene	effect
recF	+
rob	+
sdaA	−
sohB	−
treA	+
yeiL	+
yfi D	+
yihI	−

The set of 40 observations of the stationary phase was found to be inconsistent with the global network of *E. coli*. We found a direct inconsistency in the system of equations caused by the values fixed by the given observations to ihfA and ihfB: $\{ihfA = -, ihfB = -\}$, implying repression of these genes under stationary phase. This mathematical incompatibility agreed with the literature related to genes $ihfA$ and $ihfB$ expression under stationary growing phase. Studies [1],[2],[4],[15] agree that transcription of $ihfA$ and $ihfB$ increases during stationary phase. Supported by this information, we have modified the observations of $ihfA$ and $ihfB$ and the compatibility test of the global network of *E.coli* was successful.

3.4 Predictions over a compatible network from a set of observations

When a regulatory network represented as a system of qualitative equations results to be compatible with a given set of observations (fixed values), then the qualitative system has at least one solution. This means that all the variables of this qualitative system will be fixed to + or − in order to give a solution to the system. If a variable is fixed the same value (+ or −) in all the solutions found, then mathematically we are talking about a hard component of the system and we name this hard component as the *prediction or inference* for this set of observations.

Because of the topology of *E.coli* network, for the 40 observations of stationary phase the qualitative system resulted from the compatibility of the network had $2,66 \cdot 10^{16}$ solutions. In all these solutions, 401 variables of the system were hard components (see Fig. 3). In other words, we were able to predict the variation: expressed $(+)$ or repressed $(-)$ of 401 components of our network (26% of the products of the network). We provide a subset of these predictions in Table 2.

3.5 Validation of the predicted genes

We have showed previously that 401 variations of genes (expression or repression) of bacteria *E.Coli* were predicted from 40 observations of the stationary phase. To verify whether these predictions were valid, we have compared them with three sets of microarray data related to the expression of genes of *E.Coli* during stationary phase. The result obtained is showed in Table 3.

Table 2: Table of 43 products inferred under stationary phase condition.

(a)

gene	value
IHF	+
ada	+
agaR	+
alsR	+
araC	+
argP	+
argR	+
baeR	+
cadC	+

(b)

gene	value
cpxR	+
crp	+
cusR	+
cynR	+
cysB	+
cytR	+
dnaA	+
dsdC	+
evgA	+

(c)

gene	value
fucR	+
fur	+
galR	+
gcvA	+
glcC	+
gntR	+
ilvY	+
iscR	+
lexA	+

(d)

gene	value
lysR	+
melR	+
mngR	+
oxyR	+
phoB	+
prpR	+
rbsR	+
rhaR	+
rpoD	+

(e)

gene	value
rpoH	+
rpoS	+
soxR	+
soxS	+
srlR	+
trpR	+
tyrR	+

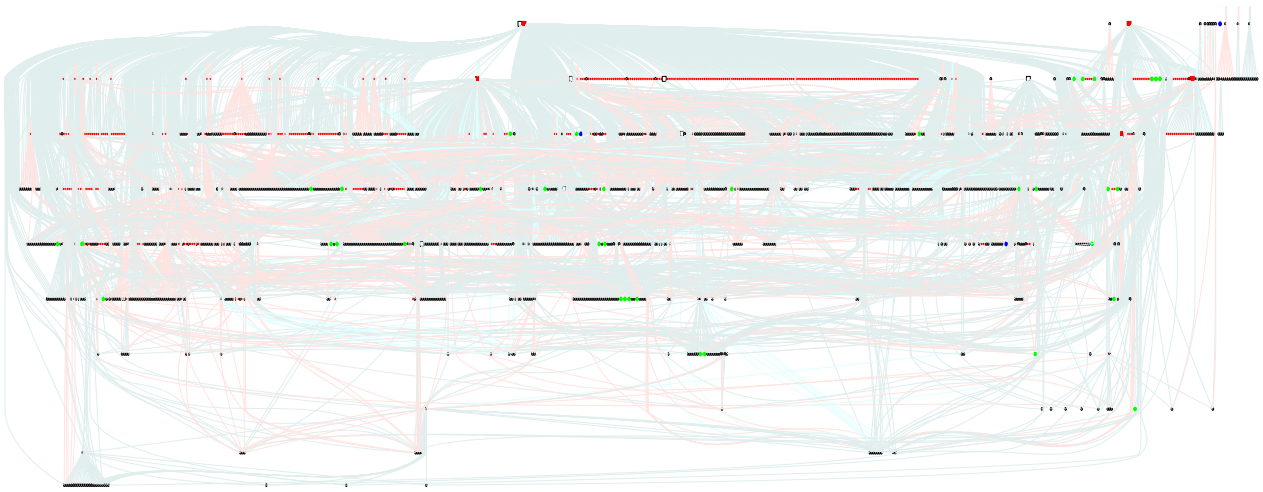


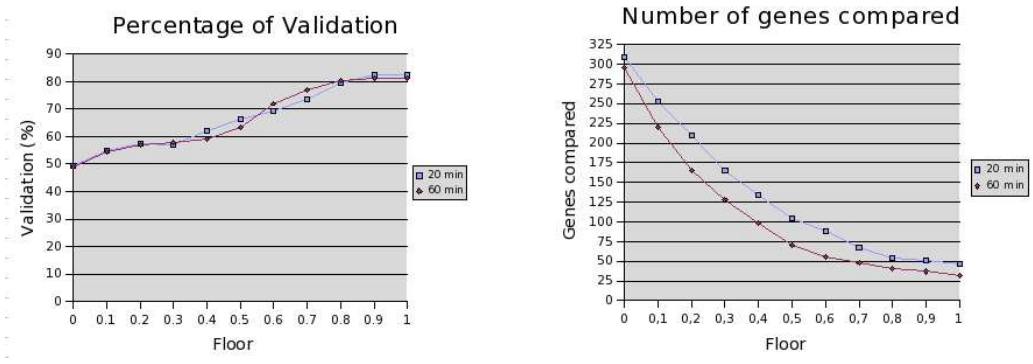
Figure 3: Global *E.coli* regulatory network with transcriptional and sigma-factors interactions (3885 interactions and 1529 products). Blue and red interactions represent activation or, respectively, repression. Green and blue nodes correspond to positive and negative observations (40). Red nodes (401) are the total inferred variations of products under stationary growth phase condition.

Table 3: Validation of the prediction with microarray data sets

Compared genes	Validated genes (%)	Source of microarray data
275	44%	[6], stationary phase
307	49.51%	Gene Expression Omnibus ([3],[5]), stationary phase after 20 minutes
294	49.1%	Gene Expression Omnibus ([3],[5]), stationary phase after 60 minutes

The number of compared genes corresponds to the common genes, the validated genes are those genes which variation in the the prediction is the same as in the microarray data set.

From the sets of microarray data provided by GEO (Gene Expression Omnibus) for stationary phase measured after 20 and 60 minutes, we have taken into account gene expressions which absolute value is above an specific floor and compare only these genes expression with the 401 predictions. The percentage of validation obtained for different values of floors is illustrated in Figure 4.



(a) (b)

Figure 4: (a) Percentage of validation of the 401 predicted variations of genes with microarray data sets from GEO (Gene Expression Omnibus) for stationary phase after 20 and 60 minutes. For both experiments we validate the 401 predictions with different sets of microarray observations considering only those genes which absolute value of expression is above certain value (fbor). (b) Number of genes considered for the validation for the different used fbors of both microarray data sets.

The percentage of our predictions that does not agree with the microarray results is in some cases:

- A mistake of the microarray result: for example in the microarray data set provided in [6] some genes as: $xthA$, cfa, $cpxA$, $cpxR$, gor known to be expressed under stationary phase, as references [7] and [16] can show, are predicted as expressed by our model and as repressed by the microarray data.
- Incompleteness of our network model: for example in [17] is shown that protein $IlvC$ decreases its level in stationary phase due to an interaction with $clpP$. Our prediction shows that gene $ilvC$ expresses, and in our network model there is not any interaction between $ilvC$ and $clpP$. Thus, this result highlights that some interactions are not taken into account in our model of E.Coli network.

4 Conclusions

Given an interaction graph of a thousand products, as *E.coli* regulatory network, we were able to find its compatibility using the mathematical framework explained before and the algorithms and methods described in [14]. Thus we found that: 1. *E.coli* transcriptional regulatory network, obtained from RegulonDB site [10],[11], was found incompatible. 2. *E.coli* transcriptional regulatory network plus sigma-factors interactions was found compatible.

Corrections over an incompatible model were proposed: incompatibility for *E.coli* transcriptional regulatory network was caused by lack of interactions. Corrections over an incompatible set of observations of an specific experiment were also proposed: incompatibility problem for *E.coli* transcriptional regulatory network plus sigma-factors interactions with Stationary Phase experimental data, was caused by a wrong (different from literature) set of experimental data.

Finally a step of inference/prediction was achieved being able to infer 401 new variations of products (26% of the total products of the network) from *E.coli* global network (transcriptional plus sigma-factors interactions).

We plan to use this approach to test different experimental conditions over *E.coli* network in order to complete its interaction network model. It should be also interesting to test it with different (signed and oriented) regulatory networks. A package with all the algorithms proposed is in preparation, and we are working on expanding its functionalities.

References

[1] T Ali Azam, A Iwata, A Nishimura, S Ueda, and A Ishihama. Growth Phase-Dependent Variation in Protein Composition of the Escherichia coli Nucleoid. *J Bacteriol*, 181(20):6361–6370, August 1991.

[2] M Aviv, H Giladi, G Schreiber, AB Oppenheim, and G Glaser. Expression of the genes coding for the Escherichia coli integration host factor are controlled by growth phase, rpoS, ppGpp and by autoregulation. *Mol Microbiol*, 14(5):1021–1031, Dec 1994.

[3] T Barrett, T O Suzek, D B Troup, S E Wilhite, W C Ngau, P Ledoux, D Rudnev, A E Lash, W Fujibuchi, and R Edgar. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res*, 33:D562–6, Jan 2005.

[4] M D Ditto, D Roberts, and R A Weisberg. Growth phase variation of integration host factor level in Escherichia coli. *J Bacteriol*, 176(12):3738–3748, June 1994.

[5] R Edgar, M Domrachev, and A E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, Jan 2002.

[6] Rosa Maria Gutierrez-Rios, David A Rosenblueth, Jose Antonio Loza, Araceli M Huerta, Jeremy D Glasner, Fred R Blattner, and Julio Collado-Vides. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res*, 13(11):2435–2443, Nov 2003. Evaluation Studies.

[7] A Ishihama. Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol*, 54:499–518, 2000.

[8] B Kuipers. *Qualitative reasoning.* MIT Press, 1994.

[9] Ovidiu Radulescu, Sandrine Lagarrigue, Anne Siegel, Philippe Veber, and Michel Le Borgne. Topology and static response of interaction networks in molecular biology. *J R Soc Interface*, 3(6):185–196, Feb 2006.

[10] H Salgado, S Gama-Castro, M Peralta-Gil, E Diaz-Peredo, F Sanchez-Solano, A Santos-Zavaleta, I Martinez-Flores, V Jimenez-Jacinto, C Bonavides-Martinez, J Segura-Salazar, A Martinez-Antonio, and J Collado-Vides. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 1(34):D394–7, Jan 2006.

[11] H Salgado, A Santos-Zavaleta, S Gama-Castro, M Peralta-Gil, MI Penaloza-Spinola, A Martinez-Antonio, P D Karp, and J Collado-Vides. The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinformatics*, 7(1):5, Jan 2006.

[12] A Siegel, O Radulescu, M Le Borgne, P Veber, J Ouy, and S Lagarrigue. Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *Biosystems*, 84(2):153–174, May 2006.

[13] Donna K Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet*, 32 Suppl:502–508, Dec 2002.

[14] Philippe Veber, Michel Le Borgne, Anne Siegel, Sandrine Lagarrigue, and Ovidiu Radulescu. Complex qualitative models in biology: A new approach. *Complexus*, 2(3-4):140–151, 2004.

[15] A Weglenska, B Jacob, and A Sirko. Transcriptional pattern of Escherichia coli ihfB (himD) gene expression. *Gene*, 181(1-2):85–8, Nov 1996.

[16] D Weichart, R Lange, N Henneberg, and R Hengge-Aronis. Identification and characterization of stationary phase-inducible genes in Escherichia coli. *Mol Microbiol*, 10(2):405–20, Oct 1993.

[17] D Weichart, Querfurth N, Dreger M, and Hengge-Aronis R. Global role for ClpP-containing proteases in stationary-phase adaptation of Escherichia coli. *J Bacteriol*, 185(1):115–125, Jan 2003.