

Curating a Large-scale Regulatory Network by Evaluating its Consistency with Expression Datasets

Carito Guziolowski¹, Jeremy Gruel², Ovidiu Radulescu³, and Anne Siegel⁴

¹ Centre INRIA Rennes Bretagne Atlantique, IRISA, Campus de Beaulieu, 35042
Rennes, France,

`carito.guziolowski@irisa.fr`,

WWW home page: http://www.irisa.fr/symbiose/carito_guziolowski

² INSERM U620, 2 Av Pr. L. Bernard, 35043 Rennes, France

³ Université de Rennes 1, IRMAR Campus de Beaulieu, 35042 Rennes, France

⁴ CNRS, UMR 6074, IRISA, Campus de Beaulieu, 35042 Rennes, France

Abstract. The analysis of large-scale regulatory models using data issued from genome-scale high-throughput experimental techniques is an actual challenge in the systems biology field. This kind of analysis faces three common problems: the size of the model, the uncertainty in the expression datasets, and the heterogeneity of the data. On that account, we propose a method that analyses large-scale networks with small – but reliable – expression datasets. Our method relates regulatory knowledge with heterogeneous expression datasets using a simple *consistency rule*. If a global consistency is found, we predict the changes in gene expression or protein activity of some components of the network. When the whole model is inconsistent, we highlight regions in the network where the regulatory knowledge is incomplete. Confronting our predictions with mRNA expression experiments allows us to determine the missing post-transcriptional interactions of our model. We tested this approach with the transcriptional network of *E. coli*.

1 Introduction

Reconciling gene expression data with large-scale regulatory network structures is a subject of particular interest due to the urgent need of curating regulatory models that are likely to be incomplete and may contain errors. Based on the large amount of genome-scale data yielded by high-throughput experimental techniques, many data-driven approaches for reconstructing regulatory network structures have been proposed [1–3]. Meanwhile, for some well studied organisms there are already large-scale models of regulations derived from literature curations or from computational predictions; as for example the RegulonDB database [4] for the bacteria *E. coli*. On account of this, large-scale models are presented as a compilation of interactions deduced from different methods or from experiments applied under different conditions. The accumulation of diverse sources in the construction of regulatory models may cause errors. A challenging solution

is therefore to design automatic methods that integrate experimental datasets to known regulatory structures enabling biologists to conciliate heterogeneous data types, find inconsistencies, and refine and diagnose a regulatory model [5–7].

On the basis of these arguments, we propose an iterative method to identify and fix regions in a regulatory network model where proposed regulations are logically inconsistent with that of microarray expression data. Using our method we can: (i) detect the regions in the network inconsistent with the experimental data, and (ii) extend the initial set of expression data in order to generate computational predictions. We applied our method to the transcriptional regulatory network of *E.coli K12* (1763 products and 4491 interactions) extracted from the RegulonDB database.

By using fast and performant algorithms we are able to reason over the large system of constraints representing a large-scale regulatory network. Thus, we detect global inconsistencies involving contradictions among different regions in the network and go one step further than other methods [8] that detect only local contradictions involving only network modules formed by direct predecessors of a node in the graph. Regarding the *E. coli* model, we concluded that its transcriptional interactions did not adequately explain the initial experimental data composed by 45 differentially expressed genes under the exponential-stationary growth shift. On that account, we detected specific post-transcriptional interactions in order to obtain a globally explained model by the experimental data.

Once a global consistency is obtained, our method generates gene-expression computational predictions, which are the result of a consistency test performed previously to the entire network. As opposed to other methods [9, 10], we do not simulate the response of our model to an artificial perturbation; we describe the (possibly huge) set of models that are consistent with the initial dataset, then we look for invariants in this set. In this way, our computational predictions correspond to the nodes whose expression change fluctuates in the same direction in all the models consistent with the initial dataset. Regarding the *E. coli* regulatory model, we calculated 502 gene-expression predictions that correspond to nearly 30% of the network products predicted to change considerably under the analysed condition. These predictions were validated with microarray outputs, obtaining an agreement of 80%. This percentage can be comparable to the one obtained by other methods working on *E. coli* data [11–13], and considerable, since we used only a transcriptional model without including metabolic regulations.

2 Approach

2.1 Global approach

We analysed regulatory networks formed by the interactions among certain molecules. As molecules we refer to genes, sigma factors, active proteins, or protein complexes; and as interactions to Protein-DNA, Sigma-gene, and complex formation. Even so, any kind of interactions may be studied as long as we

can map them as *influence* relations, i.e. *A influences B if increasing or decreasing A's concentration causes a change in B's concentration*. Molecules that hold influence relations form an *influence network*, the central object of this study (Fig. 1).

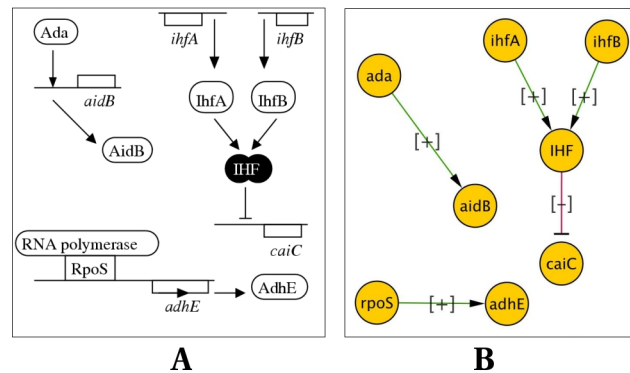


Fig. 1. A regulatory network (**A**) mapped into an influence network (**B**). Influences among molecules create an influence network; the arrows in the influence network represent a positive (+) or negative (-) influence.

The interactions of the influence network must be described qualitatively as +, -, and ?, where: +, represents a positive influence (e.g. activation of gene transcription, recognition of a gene promoter region, or formation of proteins complexes); -, a negative influence (e.g. inhibition of gene transcription, inactivation of a protein); and ?, a dual or complex regulation. Differential data issued from perturbation steady state experiments should be also provided in the form of qualitative (+, -) *concentration changes* of some network molecules. One type of concentration changes may be statistically significant mRNA-expression responses described qualitatively as: + up-regulation, - down-regulation.

An influence network is analysed using the following consistency rule: "*The variation of the concentration level of one molecule in the network must be explained by an influence received from at least one of its predecessors, different from itself, in the network*". Checking the validity of this rule for an influence network alone, or for an influence network plus a set of reliable concentration changes, will be referred from now on as the *consistency check process*. The mathematical bases of the consistency rule are formally proven in [14, 15]. In these studies the authors assumed that the concentration changes must correspond to changes between two stable conditions in the cell.

Let us intuitively explain the logic of the consistency rule in the simple influence network presented in Fig. 2. *A* and *B* may represent two proteins that activate the transcription of gene *C*. The consistency rule states that if *A* and *B* are both up-regulated (+) under certain condition, then *C* must be up-regulated,

i.e. a + prediction will be assigned to C (Fig. 2A). Similarly, the concentration change of C will be predicted as – if both, A and B , were down-regulated. When A is up-regulated (+) and B is down-regulated (–) the expression level of C cannot be *predicted*, as both expression levels (up/down regulated) are possible for C and do not contradict the consistency rule (Fig. 2B). Only if C was a protein complex formed by proteins A and B , we may *predict* the expression level of C depending on the change in expression of the limiting former protein. To state this, we have extended the theory in [14, 15] in a context where *the weakest takes it all* concluding in a rule called the *protein complex behaviour rule*. This rule applies when the values of the concentrations of A and B are well separated (one of them is much smaller than the other). The details of this rule are provided in the Supplementary material.

A third situation may occur when all the molecules are observed, let us say, A is up-regulated, B is up-regulated, and C is down-regulated. The consistency rule states that C should be up-regulated; the experiment, however, shows the contrary. Thus, we arrive to a *contradiction* between the network and the experiment, also called *inconsistent* model (Fig. 2C). No predictions may be generated from an inconsistent model, yet, a region in the network is identified together with the expression data that created the conflict.

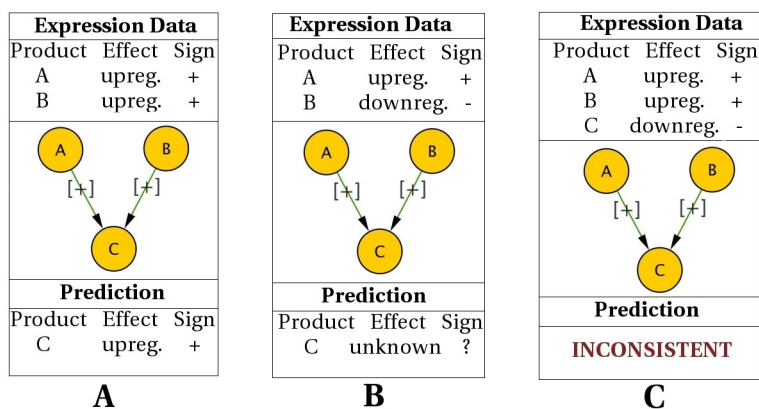


Fig. 2. Examples explaining the consistency check process. **A.** Expression predictions when a set of consistent expression data is provided. **B.** Consistent expression data may not generate a new prediction. **C.** Expression data provided resulted inconsistent with the influence network.

An influence network consistent with expression data is a network where all the expression data is explained by the (consensual or not) fluctuation of all the nodes in the network. In Fig. 2 we illustrated that depending on the expression data, we may obtain up to three different results: consistency and prediction, just consistency, or inconsistency. In order to evaluate the consistency of a large-scale network, we represent it as a system of qualitative constraints in $\{+, -, ?\}$,

where each constraint relates a node with its direct predecessors. All constraints in the system should satisfy the consistency rule, and should not contradict any other. Computing the satisfiability of a large system of qualitative constraints is an NP-complete problem for even linear qualitative systems [16]. As classic methods of resolution do not allow to solve this kind of problems [17], we used an efficient representation based on binary decision diagrams for the set of solutions of a qualitative system proposed in [18]. The consistency of a large system of qualitative constraints is thus computed using *Pyquali*, a Python library devoted to solve huge systems of constraints in three values (+, -, ?) that is based on the mentioned representation. As a result it is possible to decide in a matter of seconds if a large-scale regulatory model is consistent with an initial dataset and to detect inconsistent regions in the model when needed.

In Fig. 3 we illustrate a flow-chart of the complete consistency check process. As input data we receive a qualitative influence network and a set of qualitative concentration changes. We build from this initial data a system of mathematical constraints, its consistency will be analysed afterwards using *Pyquali*. If the system is consistent we predict new concentration changes that after being compared with other experiments may generate new inputs to the initial set of concentration sets, however, it is also possible that thanks to this comparison new nodes and edges in the influence network will be added. If the system is inconsistent, the influence network must be corrected either by searching in the literature or by experimental results.

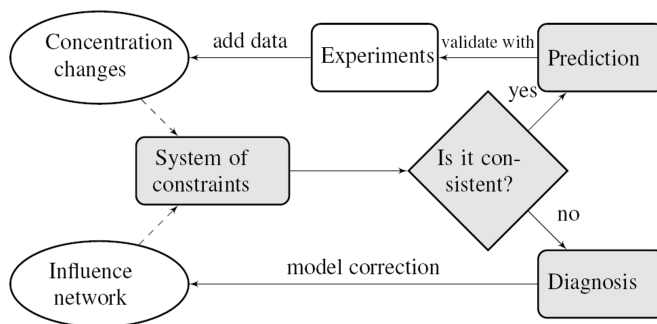


Fig. 3. Consistency check process. (1) We build a system of constraints from an influence network with a set of concentration changes, (2) we check the consistency of the system, (3) if it is consistent and an initial dataset of concentration changes was provided, we may predict new concentration changes of the molecules in the network that after compared with real measurements may question the original dataset and model. If it is not consistent, we report the inconsistent region in order to correct the network or initial dataset. The shaded blocks represent the automatically calculated steps.

2.2 Small example

To understand how a consistent model may generate computational predictions, let us analyse in detail the consistency check process applied to a small region of the *E. coli* influence network. The influence network presented in Fig. 4 is analysed during the exponential-stationary growth shift. An initial dataset of concentration changes (obtained from the literature) was initially provided for *fic*, *ihfA*, and *ihfB* [19, 20]. From this initial data the analysis proceeds as follows. As *fic* is up-regulated, and it only receives an influence from the sigma factor RpoS, then RpoS is fixed to be up-regulated. The protein complex IHF is fixed as down-regulated after applying the *protein complex behaviour rule* deduced from observing the concentration of its former proteins and the metabolically stable behaviour of IHF in its dimeric form (see the Supplementary material). The gene transcribing one of its former proteins, *ihfB*, appears down-regulated; hence, it must receive a negative influence from one of its three predecessors in the network: IHF's influence is positive as it is down-regulated but it inhibits *ihfB*'s expression; RpoS's influence is also positive as it was predicted to be up-regulated; only RpoD's value can be fixed to down-regulated (-), causing a negative influence over *ihfB*. The genes that receive a unique influence from RpoD, *crp* and *ompA*, are fixed as down-regulated because of the negative influence coming from RpoD. No other alternatives nor contradictions appear in the fixed values of IHF, RpoS, RpoD, *crp*, and *ompA*; consequently, they are the computational predictions.

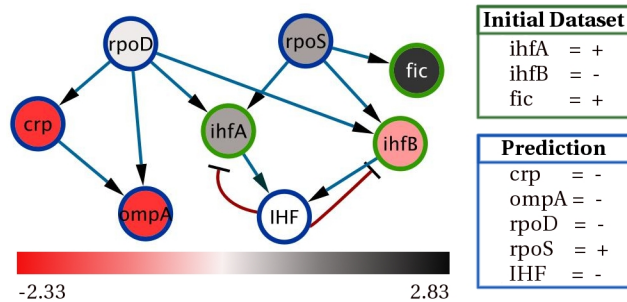


Fig. 4. Consistency check process for 8 products of the *E. coli* regulatory network under the exponential-stationary growth shift. All transcriptional influences that each product receives appear in the network. The grey-red intensity of each product reflects the experimentally observed change in mRNA expression (\log_2 ratio) during the studied condition. Products with a green border refer to those present in the initial dataset obtained from the literature, whereas products with a blue border refer to our computational predictions. Experimentally observed mRNA-expression changes agree with the computational predictions except for RpoD, where the predicted changes correspond to variations on the *active protein* and cannot be observed on mRNA expression levels.

3 Results

3.1 Construction of the network

We constructed the influence network from the regulatory model of *E. coli* available in the RegulonDB database. The influences in our network represented transcriptional, complex formation, and Sigma-gene regulations. The products of the network were genes, active proteins, and protein complexes. The information to reconstruct this model was extracted from the database on June 2007. In Fig. 5A we show the complete and reduced network.

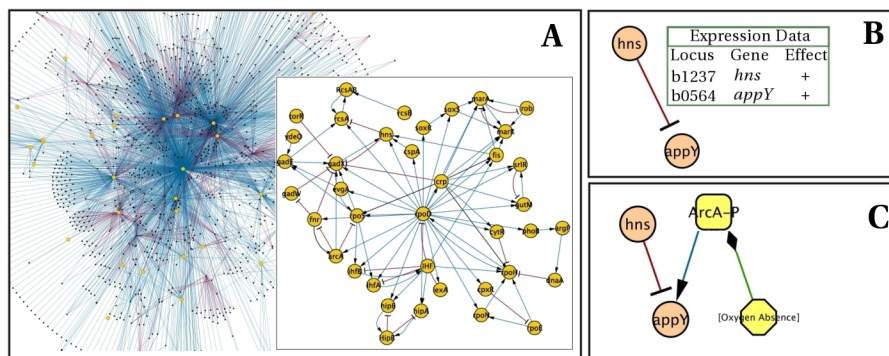


Fig. 5. A. *E. coli* influence network (1763 products and 4491 interactions). A small region of it is presented to the right (39 products and 94 interactions). The products forming this region control most of the components of the bigger network. **B.** The inhibition of gene *appY* by the *hns* product causes an inconsistency with the expression data related to the exponential-stationary growth shift. **C.** Correction of the inconsistency by adding a positive regulation from ArcA-P (phosphorylated protein ArcA) to *appY*; this regulation occurs in the absence of oxygen.

3.2 Consistency test and correction using one stress condition

To start with the consistency check process, two initial data were provided: (i) *E. coli* influence network, and (ii) a set of 45 differentially expressed genes in the transition from exponential to stationary growth phase (see the complete list of observed genes in the Supplementary material). This set of phenotypes was collected from the literature based on initial information provided in RegulonDB. The first result obtained was an inconsistency between the model and the experimental data. The inconsistent region, highlighted by the method, is the one shown in Fig. 5B. It represents the inhibition of transcription of the *appY* gene by the H-NS protein; no other transcriptional regulation of the *appY* gene was reported in the RegulonDB database. These products (*appY*, *hns*, and therefore H-NS) are however, shown to increase their levels in the exponential-stationary

growth shift of the cell [21, 22]. Hence, the source of the conflict may be in the model.

Searching in the primary literature, we found that the *appY* gene is induced during entry into stationary phase, and that during oxygen-limiting conditions the stationary-phase induction is partially dependent on ArcA [23]. The protein ArcA is activated via phosphorylation by the ArcB sensor under conditions of reduced respiration [24]; the signal which leads to the activation of ArcA during entry into stationary phase may be the deprivation of oxygen caused by an increase in cell density [23]. Based on these studies we corrected our influence network adding new interactions (see Fig. 5C), obtaining a model consistent with the experimental data.

3.3 Computational predictions and validation

Once our network was consistent with the expression data provided for the exponential-stationary growth shift we generated the computational predictions. From the 45 initial expression phenotypes, 526 changes in other components of the network during the same experimental condition were predicted. We characterized them into 12 functional groups using the *DAVIS* software [25]. To validate our computational predictions, we obtained from the *Many Microbe Microarray Database* [2] a dataset of differentially expressed genes after 720 minutes of growth (stationary phase) in a rich medium [26]. This dataset was compared to our predictions (Fig. 6). The extended view of this comparison is included in the Supplementary material.

Expression profiling of this microarray dataset identified 926 genes that change significantly (2-fold) in transcription in response to the growth shift from exponential to stationary phase. The 526 products, computationally predicted, could be classified into four categories: 131 agreed with significant expression changes; 32 had a predicted expression change in a direction opposite to that of the experimental data; 329 had a predicted expression change that was not found to be statistically significant in the experimental data; and for 34 products there was no expression data available (some products of the network were protein complexes and could not be compared to mRNA expression). Thus, of the 163 (=131+32) significant differentially expressed genes that could be compared between the computational predictions and the experiment, 131 (or 80% consensus) agreed. Only the 31% (coverage) of our predictions could be compared with 2-fold expression changes; therefore we performed the same analysis choosing different thresholds (1.5-fold, 0 fold). In this way, new consensus and coverage percentages were calculated showing that the higher the threshold is, the better is the consensus and the worse is the coverage of the predicted data (Fig. 7).

The quality of our computational predictions is related to the experimental condition used to calculate them. Three important phases in the cell (exponential, early stationary-phase, late stationary-phase), appear in the exponential-stationary growth shift. During the shift among these phases, many molecules in the cell change their behaviour considerably. In our study we have chosen to compare the late stationary phase, when the number of cells does not change,

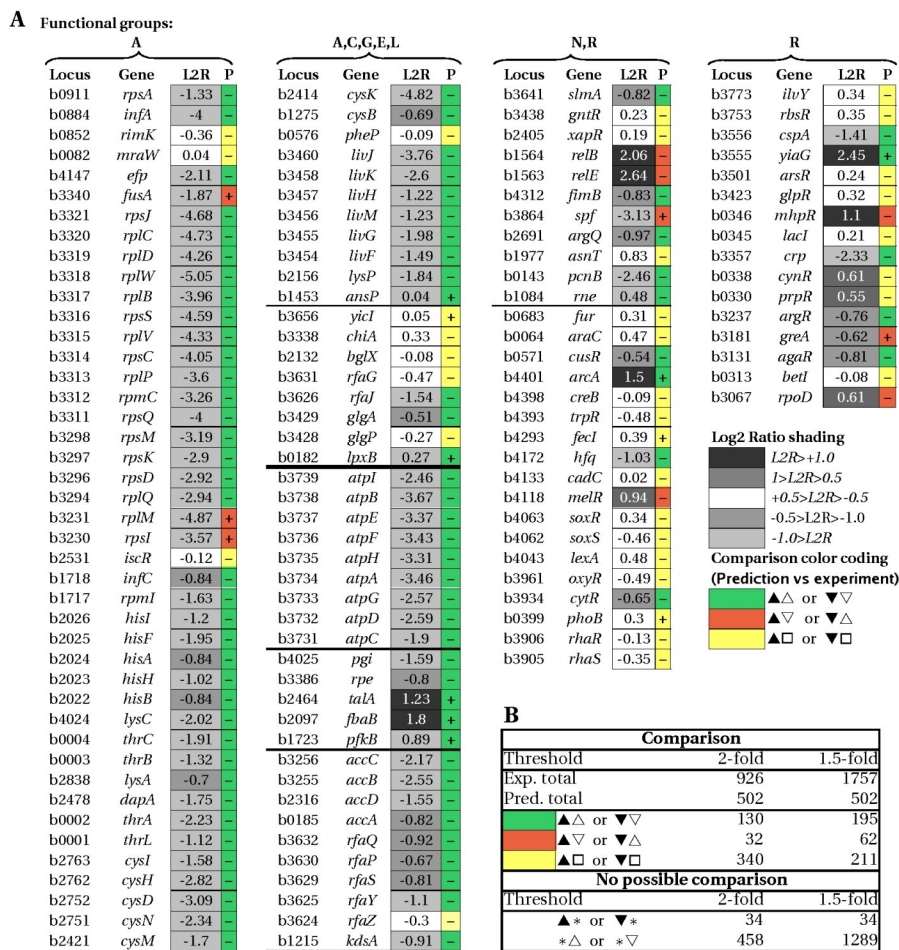


Fig. 6. Table of observed vs. predicted gene-expression responses in the *E.coli* network under the exponential-stationary growth shift condition. **A.** The locus numbers, gene names, and the \log_2 ratio (L2R) of gene expression (exponential to stationary) are shown for some of the 526 predicted expression changes (+,-). Genes were divided by functional groups into: A, amino acids metabolism and biosynthesis; C, carbohydrates metabolism and biosynthesis; E, energy metabolism; G, glucose catabolism; L, lipid metabolism and biosynthesis; N, nucleic acids metabolism; R, regulatory function; S, cell structure; SI, signal peptides; T, transport; V, vitamin metabolism and biosynthesis; and U, unassigned. The L2R is shaded depending on the magnitude of the expression shift. Filled and open symbols indicate computational predictions and experimental data, respectively; squares indicate no change in gene expression; triangles indicate a change in expression, as well as the direction of change (up-regulated or down-regulated). **B.** Comparison between all predicted and observed expression changes. An * symbol indicates either that our model did not predict a gene expression or that no expression data related to a gene in our model was found.

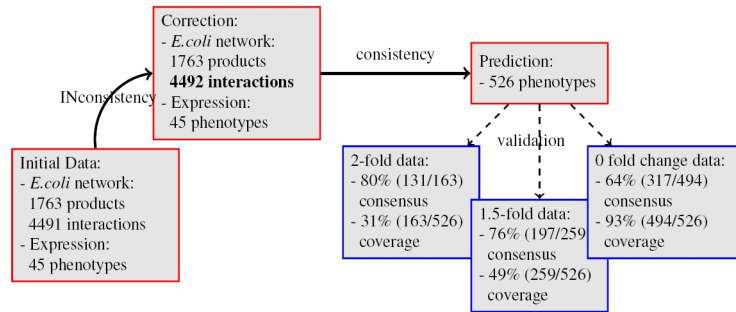


Fig. 7. Results of the consistency check process applied to the *E.coli* transcriptional network using 45 phenotypes related to the exponential-stationary growth shift. We validated our computational predictions using the observations in a microarray dataset filtered with three thresholds. Consensus refers to the percentage of validated model predictions; coverage indicates the percentage of compared predictions.

with the late exponential growth phase, when the cell continues to grow exponentially. We believe that these two conditions represent instants in the cell where the genes and proteins do not significantly change their concentration. However, the 45 initial expression data may correspond to slightly different time points of the exponential phase and thus induce divergences in our computational predictions. In spite of this, a high percentage of our predictions was validated when compared with the microarray data.

4 Discussion

We have introduced an iterative method that checks the consistency between a regulatory model and expression data. To validate our approach we chose the *E. coli* transcriptional regulatory network obtained from the RegulonDB database. By using an expression dataset of 45 *E. coli* genes significantly expressed under the exponential-stationary growth shift, we corrected the model and predicted the outputs of microarray experiments with an 80% of agreement. This percentage is comparable to the one obtained by other methods working on a more complex regulatory model for *E. coli* [11–13]. After the comparison of our computational predictions with microarray measurements, some disagreements were detected. A reasonable explanation for these divergences resides on our previously made assumption that some level of correlation exists between the transcription factor protein and the target gene expression without considering in detail the post-transcriptional effects; this problem was also reported in [8]. An example of this case of disagreement was illustrated in Fig. 4; nevertheless, we can still use this type of errors in our predictions to complete the regulatory model with post-transcriptional regulations. A second possible reason is the complexity of the exponential-stationary growth shift which induces heterogeneity in the observations obtained from different sources of the literature [27]. We il-

illustrated that our automatized framework checks in very short time the global consistency of a large-scale system of constraints, into which regulatory and expression knowledge can be represented. Computationally, the solution of this type of problems is not trivial, nevertheless we analysed the complete large-scale regulatory network of *E.coli* in less than one minute. Biologically speaking, we are able to integrate large-scale regulatory data and confront it to experimental observations in order to provide a diagnosis of regions in the network where the regulatory knowledge is contradictory with the expression data. Moreover, we compute a set of computational predictions that represent the invariant regions in the network that explain our expression data consistently. These results reflect important global configurations in a regulatory network that can be practically used to diagnose models and in a future to modulate their global behaviour. What we have shown for the *E.coli* network, can be applied to other networks and to different expression datasets.

Availability and Supplementary material

Sources and a working application of our method can be accessed on-line at: <http://www.irisa.fr/symbiose/bioquali/>. The supplementary material is available on-line at: <http://www.irisa.fr/symbiose/projects/networks>.

Acknowledgments

We would like to thank Michel Le Borgne and Philippe Veber for their invaluable work implementing Pyquali. The work of the first autor was financially supported by CONICYT (Comisión Nacional de Investigación Científica y Tecnológica), Chile. This work was supported by the ANR program ANR-06-BYOS-0004.

References

1. Joyce, A., Palsson, B.: The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell. Biol.* 7, 198–210 (2006)
2. Faith, J., Hayete, B., Thaden J., et al.: Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5, e8 (2007)
3. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and Di Bernardo, D.: How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78 (2007)
4. Salgado, H., Gama-Castro, S., Peralta-Gil, M., et al.: RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* 34, D394–7 (2006)
5. Ideker, T., Galitski, T., and Hood, L.: A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2, 343–72 (2001)
6. Palsson, B.: The challenges of in silico biology. *Nature Biotechnology* 18, 1147-1150 (2000)
7. Kitano, H.: Systems Biology: A Brief Overview. *Science* 295, 1662-1664 (2002)

8. Herrgard, J., Lee B., Portnoy, V., Palsson, B.: Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* 16, 627–35 (2003)
9. Ideker, T., Thorsson, V., Ranish, J.A., et al.: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–34 (2001)
10. Rawool, S.B., Venkatesh K.V.: Steady state approach to model gene regulatory networks—simulation of microarray experiments *Biosystems* 90, 636–55 (2007)
11. Covert, M., Knight E., Reed, J., et al.: Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92–6 (2004)
12. Covert, M., Palsson B.: Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. biol. chem.* 277, 28058–28064 (2002)
13. Edwards, J.S., Palsson, B.O.: The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *PNAS* 97, 5528–33 (2000)
14. Radulescu, O., Lagarrigue, S., Siegel, A., et al.: Topology and static response of interaction networks in molecular biology. *J. R. Soc. Interface* 3, 185–96 (2006)
15. Siegel, A., Radulescu, O., Le Borgne, M., et al.: Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *Biosystems* 84, 153–74 (2006)
16. Dormoy J.: Controlling qualitative resolution. In: 7th National Conference on Artificial Intelligence, pp. 319–323. Morgan Kaufman Press, Saint Paul, Min. (1988)
17. Travé-Massuyès, L., Dague, P.: Modèles et raisonnements qualitatifs. Hermes Sciences, Paris (2003)
18. Veber, P., Le Borgne, M., Siegel, A., et al.: Complex Qualitative Models in Biology: A new approach. *Complexus* 3–4 (2005)
19. Hiratsu, K., Shinagawa, H., Makino, K.: Mode of promoter recognition by the *Escherichia coli* RNA polymerase holoenzyme containing the sigma S subunit: identification of the recognition sequence of the fic promoter. *Mol. Microbiol.* 18, (1995)
20. Aviv, M., Giladi, H., Schreiber, G., et al.: 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, 1021–31 (1994)
21. Dersch, P., Schmidt, K., Bremer, E.: Synthesis of the *Escherichia coli* K-12 nucleoid-associated DNA-binding protein H-NS is subjected to growth-phase control and autoregulation. *Mol Microbiol.* 8, 875–89 (1993)
22. Atlung, T., Sund, S., Olesen, K., Brndsted; L.: The histone-like protein H-NS acts as a transcriptional repressor for expression of the anaerobic and growth phase activator AppY of *Escherichia coli*. *J. Bacteriol.* 178, 3418–3425 (1996)
23. Brndsted, L. and Atlung, T.: Effect of growth conditions on expression of the acid phosphatase (cyx-appA) operon and the appY gene, which encodes a transcriptional activator of *Escherichia coli*. *J. Bacteriol.* 178, (1996)
24. Iuchi, S., Chepuri, V. Fu, H., et al.: Requirement for terminal cytochromes in generation of the aerobic signal for the arc regulatory system in *Escherichia coli*: study utilizing deletions and lac fusions of cyo and cyd. *J. Bacteriol.* 172, 6020–6025 (1990)
25. Dennis, G., Sherman B., Hosack, D., et al.: VID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, (2003)
26. Allen, T., Herrgrd, M., Liu, M., et al.: Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.* 185, 6392–9 (2003)
27. Gutierrez-Rios, R., Rosenblueth, D., Loza, J., et al.: Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13, 2435–2443 (2003)