Integrating network motifs into a genetic network: a case of study based on the Phop/PhoQ two-component system

Oscar Harari [1], Coral del Val[1], Igor Zwir [1,2]
[1]Department Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[2]Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri, USA
E-mail: zwir@borcim.wustl.edu

ABSTRACT

Genetic and genomic approaches have been successfully employed to assign genes to distinct regulatory networks. The strength of the connections in these networks must be specified to define the kinetics of a group of genes, but the uncertainty concerning the connections between genes, the ambiguity inherent to the biological processes, and the impossibility of experimentally determining the underlying biological properties only allow a rough prediction of gene interaction. Here we describe a framework that examines promoter sequences and identifies those *cis*-acting features that define transcriptional network motifs. Then, we employ an iterative process, based on Ordinary Differential Equations, to learn a network architecture that appropriately integrates these motifs into a full structure. The application of this method to the two component systems PhoP/PhoQ and the PmrA/PmrB in *Salmonella* enterica uncovered novel mechanisms that enable the inter-connection of these networks. The predictions were experimentally verified.

1. INTRODUCTION

Gene expression is determined by interactions among regulatory proteins, called transcription factors, and RNA polymerase(s), as well as the interactions of the transacting factors and RNA Polymerase with *cis*-acting DNA sequences in the promoters of regulated genes [1]. Computational tools that look for these *cis*-elements in genome sequences and genome-wide gene expression patterns (usually in the form of microarray data) provide the raw material for the characterization and understanding of transcription regulation of target genes. Recurrent patterns of interactions among these features define network motifs, which are elementary building blocks [2]. Few works have been devoted to integrate those blocks into a more complex genetic network [3-5].

In this work, we present a framework to infer gene network topologies based on genome sequences, and previous knowledge obtained by experimental assays. First, we enhance the discovery of the network motifs by providing a computational approach to improve the sensitivity while detecting *cis*-acting elements in promoter regions, including the analysis of transcription factor binding sites and RNA polymerase binding sites. Second, we connect network motifs and incorporate knowledge form the literature to conform complete network kinetic models. These allow the automatically test of hypotheses about the network motif integration, and select the most probable one [6]. We also consider different qualitative measures such as the *realism* of the topologies, the *flexibility* to reproduce the distinct behaviors under the distinct stimuli and their *robustness* to preserve functional characteristics when their parameters change (*e.g.,* initial concentrations, degradation rates).

We applied our method to analyze the expression of genes controlled by the PhoP/PhoQ regulatory system of *Salmonella* enterica serovar Typhimurium [7]. This specie has a cross-link between PhoP/PhoQ and PmrA/PmrB two component systems, enabling the simulation of a variety of network motifs. Measurements of time-dependent mRNA levels validated that our predictions could describe distinct kinetics even within a same network motif.

1.1. PhoP/PhoQ and PmrA/PmrB two component systems

The PhoP/PhoQ two-component system constitutes a master regulator in *Salmonella enterica*, regulating the transcription of more than 3% of the genes in response of a low extra cellular Mg^{2+}. Some of the genes regulated by PhoP/PhoQ two-component system are also regulated by the PmrA/PmrB two-component system, which is related to the polymyxin B antibiotic inducted resistance; and resistance to cell death mediated by Fe^{3+} among others. Thus these target genes respond independently to two signals: high level of extra cellular Fe^{3+}, sensed by the PmrB protein; and low levels of Mg^{2+}, sensed by the PhoQ protein.

1.2. Modeling genetic networks

We employ continuous models that determine the level of gene expression and relationships among them. These allow capturing biological properties that can be experimentally observed. Ordinary Differential Equations (ODE's) are good approximations that calculate the difference of concentration of species (i.e. RNA, proteins) along the time. Statical ODE's [6] model the systems when they reach their steady state (i.e. the system has reached an equilibrium in which the difference of concentrations of species in function of time is equal to zero). Dynamic models [5] enable the simulation of the gene expression behavior before their reach their equilibrium.

2. Results

2.1. *Cis*-regulatory features devoted to infer PhoP regulatory network

We investigated four types of *cis*-acting promoter features by extracting the maximal amount of useful information from datasets and then creating models, which in turn allowed the inference of the PhoP regulon. We incorporated into our analysis and learning process the PhoP regulated genes of *Escherichia coli,* bacterium specie closely related to *Salmonella,* which has been widely studied. It has been showed that *phoP* gene could complement a *Salmonella* phoP mutant [7].

2.1.1. PhoP binding site patterns

We decomposed set of binding site sequences corresponding to PhoP into four patterns and then combined them to increased the sensitivity to weak sites without losing specificity (Fig. 1A) [8]. This allowed the recovery of promoters, such as that corresponding to the *E. coli hdeA* gene or the *Salmonella pmrD*, that had not been detected by the single position weight matrix model [9] despite being footprinted by the PhoP protein [10].

2.1.2. RNA polymerase binding site patterns and location

We identified six patterns among PhoP-regulated promoters of *E. coli* and *Salmonella* that combine promoter class and distance between the PhoP box and the RNA polymerase site [8]. For example, *ugtL* and *pagC* promoters share the PhoP box but differ in the distance of the PhoP box to the RNA polymerase binding site [11] (Fig. 1B). The RNA polymerase site feature was evaluated and obtained an 82% sensitivity and 95% specificity for detecting RNA polymerase sites [8].

2.1.3. Activated/ repressed promoters

We determined that the location of binding sites functioning in activation is different from that corresponding to sites functioning in repression [8] (Fig. 1C). For example, we identified a PhoP binding site at a relative distance to the RNA polymerase consistent with repression in the promoter region of the *hilA* gene, which encodes a master regulator of *Salmonella* invasion and had been known to be under transcriptional repression by the PhoP/PhoQ system [7]. Several promoters, including those of the *Salmonella pipD* and *nmpC* genes, were classified as candidates for being both activated and repressed, because the distance between the predicted transcription start site and the PhoP box is consistent with either activation or repression.

2.1.4. Binding sites for other transcription factors

We analyzed the intergenic regions of the *E. coli* and *Salmonella* genomes for the presence of binding sites for 54 transcription factors [8,12]. We then investigated the co-occurrence of 24 sites with the binding site of the PhoP protein in an effort to uncover different types of independent or orchestrated co-regulation of PhoP and other TFs (Fig. 1D). By analyzing both the binding site quality and the location of transcription factor binding sites, we increase the chances of identifying co-regulated promoters.). For example, the *Salmonella pmrD*, *ugd* and *yrbL* promoters and the *E. coli yrbL* promoter harbor PhoP- and PmrA-binding sites, consistent with the experimentally verified regulation by both the PhoP and PmrA proteins [13].

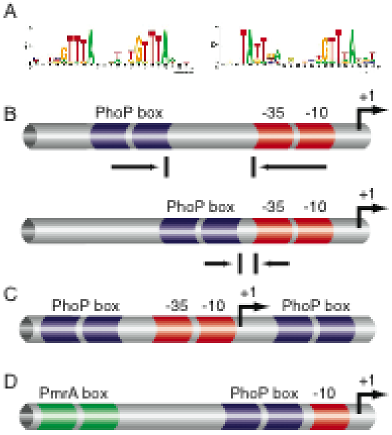

Figure 1. *Cis*-features identified in pomoter regions of PhoP regulated genes. **(A)** The PhoP protein recognizes a binding site motif consisting of a hexameric direct repeat separated by 5 bp, but distinguishes between different patterns with different specificities (i.e. *phoP* on the left and *pmrD* on the right). **(B)** PhoP-regulated promoters differ in the RNA polymerase sites. The PhoP-activated *ugtL* and *pagC* genes exhibit a class I sigma 70 promoter, but differ in the distance between the PhoP box and the RNA polymerase site. **(C)** The *mgtC* promoter harbors a PhoP binding box upstream of the RNA Polymerase binding site, positioned in a typical activation location. It also harbors a PhoP binding box downstream of the RNA Polymerase binding site, in a relative distance usual employed by PhoP to repress expression. **(D)** The *Salmonella pmrD* promoter harbors experimentally verified PhoP- and PmrA-binding sites.

2.2. PhoP network motifs uncovered by the *cis*-regulatory features

Complex biological systems are often modeled as networks. Network motifs are understandable patterns of connections that occur significantly more often than pure chance or random networks [2], and alleviate the complexity of the study of transcriptional regulatory networks. One of the most elemental network motif is the autoregulation, whereas a gene is controlled by its own protein product (Fig. 2A). We identified a PhoP binding site upstream and close to a RNAP binding site indicating a positive autoregulation. Indeed. PhoP autoregulation is critical for *Salmonella* virulence [14]. Mice inoculated with wild-type *Salmonella* died, where mice inoculated with mutant *phoP* promoter survived. [14]. Other network motifs were identified in the same fashion (Fig. 2).

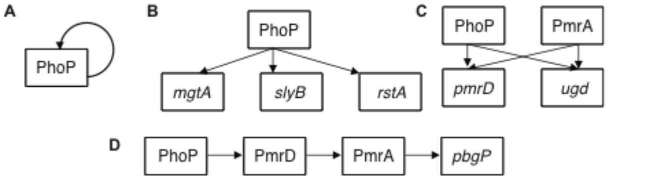

Figure 2. **The PhoP/PhoQ system employs a variety of network motifs to regulate gene transcription**. **(A)** Autoregulation: the *phoP* gene is regulated by its own product the PhoP protein. **(B)** In the simple-input module, PhoP as a single transcription factor regulates a set of genes. **(C)** In the bi-fan module, a set of genes (i.e. *pmrD* and *ugd*) are each regulated by a combination of transcription factors (i.e. PhoP and PmrA). **(D)** In the chained motif, genes are regulated in an ordered cascade.

Chained modules, allows transducing a signal (Fig. 2D). The cross-link between PhoP/PhoQ and PmrA/PmrB two-component systems is mediated by the *pmrD* gene, which resulting protein PmrD can bind the PmrA protein probably in a posttranscriptional or posttranslational fashion. As a result some of the genes governed by PmrA, including *pbgP*, are expressed even without the presence of the inducting signal of this two-component system. Curiously, the repression of *pmrD* by PmrA results in a negative feedback that closes the regulatory loop (Fig. 3). Although, this system has been widely studied [13], the exact mechanisms that defines the system dynamics is still controversial.

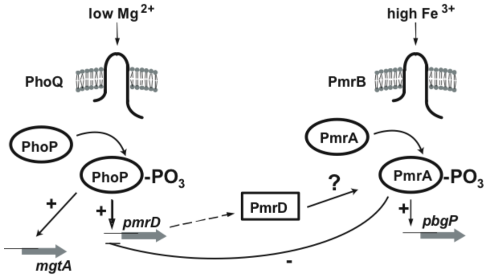

Figure 3: The PhoP/PhoQ-PmrA/PmrB functional scheme in *Salmonella enterica serovar Typhimurium*. The PhoQ protein senses low Mg^{2+} and the PmrB protein high Fe^{3+} concentrations from the environment and both proteins phosphorylate their cognate response regulators PhoP and PmrA, respectively. Although each of these proteins control the expressions of their target genes in response to their own signal, an alternative cross-talk suggest that some genes regulated by the PmrA protein can be regulated by PhoP in low Mg^{2+} conditions via the PmrD protein. Indeed, a transcriptional negative feedback has been detected in the *pmrD* gene.

2.3. Integrating network motifs into a genetic network

The initial model included the activation of *phoP, mgtA* and *pmrD* genes by phosporylated-PhoP, representing the positive autoregulation and simple network motifs respectively. Identically, we modeled the activation of *pmrA* and *pbgP* genes by phosporylated-PmrA. We translated the architecture into a system of ODE's, by employing the Ingeneue library [5], which allows simulating the dynamic behavior of the network architecture. The constraints imposed to this model only reflected the expression of *mgtA and pmrD* under low Mg^{2+} and *pbgP* on high Fe^{3+}. As expected, we observed the model could satisfy every input/output pattern with high probabilities, and could obtain high probability values (p=0.83) to satisfy all (*i.e.*, AND operation) of the constraints simultaneously.

The next proposed architecture incorporated the "backward" connection between PhoP/PhoQ and PmrA/PmrB systems as a first attempt to connect both network motifs in the whole network. Binding sites evidence and ChIP experiments showed that PmrA represses the expression of *prmD* gene [10]. Consequently, we specified that *pmrD* should be repressed under low Mg^{2+} and high Fe^{3+} environment concentrations (p=0.81).

Several frustrating attempts to integrate the network motifs into a whole network model derived the search towards relaxing all the constraints imposed to the system (i.e., values 1 in the tables). After an exhaustive search, we found that relaxing the expected output for the *pmrD* gene in the first constraint under low Mg^{2+} and high Fe^{3+} (Fig. 4; Table 1) allowed us to find a set of solutions with high probability that satisfy all constraints simultaneously (p=0.83). Moreover, this is the only architecture that could obtain such probabilities. The analysis of these results suggests that the *pmrD* gene can alternate between activated and repressing states according to the concentrations of its activation by the PhoP protein and repression by the PmrA protein. In addition, the state of this gene depends on the time when of occurrence of these events. We conclude that the uncertainty about the state of the *pmrD* gene does not constraint the connection of both network motifs when low Mg^{2+} and low Fe^{3+} occur (Table 1, constraint 2) and that, in that case, the PmrD protein protects the phosporylated-PmrA form the dephosphorylating activity of PmrB allowing the activation of the *pbgP* gene. Overall, the finding of sets of parameters that concurrently satisfy all constraints of the last architecture with high probability permits the interaction of all network motifs, including simple, autoregulated, chained, and bi-fan, into a model of the whole regulatory network.

Table 1: Patterns of input/output and constraints for the final PhoP/PhoQ-PrmA/PrmB architecture[*]

Input		Output			
Mg^{2+}	Fe^{3+}	*mgtA*	*pmrD*	*pbgP*	Probability
1	1	1	-[*]	1	0,87
1	0	1	1	**1**	0,90
0	1	0	0	1	0,89
0	0	0	0	0	0,98
AND					0.83

[*] both values are acceptable. AND indicates probability of solutions satisfying all the constraints simultaneously

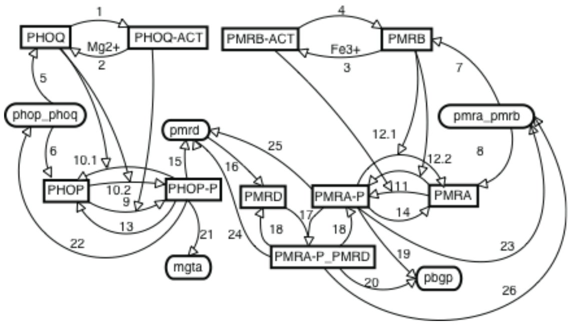

Figure 4: PhoP/PhoQ-PmrA/PmrB architecture. The species interact as follows: 1/2- Low/High Mg^{2+} level favors the PHOQ-ACT(ivated)/PHOQ state in equilibrium. 3/4- High/Low Fe^{3+} level favors PMRB-ACT(ivated)/PMRB state in equilibrium. 5/6- phop_phoq is translated into PHOQ/PHOP proteins. 7/8- pmra_pmrb is translated into PMRB/PMRA proteins. 9- PHOP is phosphorilated (PHOP-P) by PHOQ-ACT kinase activity. 10.1- PHOP-P is desphosphorilated to PHOP by PHOQ phosphatase activity. 10.2- PHOP is phosphorilated to PHOP-P by PHOQ kinase activity. 11- PMRA is phosphorilated to PMRA-P by PMRB-ACT kinase activity. 12.1- PMRA-P is desphosphorilated to PMRA by PMRB phosphatase activity. 12.2- PMRA is phosphorilated to PMRA-P by PMRB kinase activity. 13/14- PHOP-P/PMRA-P is spontaneous desphosphorilated to PHOP/PMRA. 15- PHOP-P activates the *pmrD* transcription. 16- *pmrD* is translated into PMRD. 17- PMRD binds PMRA-P (constituting PMRD_PMRA-P) which activates *pbgP* and represses *pmrD* genes, but it is not affected by the phosphatase activity of PMRB-ACT. 18- PMRA-P_PMRD unbinds into PMRD and PMRA-P. 19/20- PMRA-P/ PMRA-P_PMRD activates the transcription of *pbgP* gene. 21/22- PHOP-P activates the transcription of *mgta*/*phoP_phoQ*. 23- PMRA-P activates the transcription of *pmrA_pmrB*. 24/25- PMRA-P_PMRD/PMRA-P represses the transcription of *pmrD*. 26- PMRA-P_PMRD activates the transcription of *pmrA_pmrB*.

2.3.1. *Learning parameters of the model*

We tested two inference methods to estimate the parameters of the network by executing the native random walk already codified in the Ingeneue package [5], and compare the results obtained by a genetic algorithm (GA) implemented by ourselves. The solutions obtained were scored by a function that evaluates if the predicted concentrations of distinguished species match the expected ones (the lower, the better). The GA was executed using different configurations (i.e. population size, number of generations) and observed that both the population size and the maximum number of executions independently improve the quality of the results (Table 2). Furthermore, we compare the solutions obtained by the GA to the solutions obtained by the random walk approach, obtaining a score difference above 0.20 (Table 3).

Table 2: Evaluation of the performance of the GA.

Pop. size	Nr. Generations	Evaluations	Best score	Generation
50	100	5,000	0.1914	20
200	100	20,000	0.0522	9
50	250	12,500	0.0473	22

Table 3: Performance comparison (Random walk vs. GA)

Population size	Evaluations	Best score
Random Walk	100,000	>0.25
GA	1,100/12,500*	0.0473

* The GA obtained the best score after 1,100 evaluations. Heuristics like stall time can decrease the number of evaluations by indicating possible algorithm's stop condition.

2.3.2. *Validating the prediction of simulated species*

The mRNA levels, product of the transcription of our genes were experimentally measured with an interval of 15 minutes six times [14]. We interpolated the activity signal, and calculated the Pearson's coefficient of correlation (*c*) to the predictions of the model. We observed that PhoP, the distinguished specie that represents the positive autoregulation network motif, showed a high correlation (*c*=0.97), and exhibited a high level of expression. Similarly, the remaining network motifs showed a correlation of: *c*=0.95 for *mgtA* gene which represents the simple module; *c*=0.88 for *pmrD*, which exemplifies the bi-fan module; and *c*=0.92 for *pbgP* (chained module). These results reflect a highly correlated kinetics between our predictions and the experimentally observed values (Fig. 5).

2.3.3. *Robustness of the model*

Our analysis of the robustness of the network architecture for the PhoP/PhoQ-PrmA/PrmB system shows a tolerance of different magnitude order for distinct set of parameters. Indeed, we analyzed a random feasible solution for the final network architecture and found that some parameters (*e.g.,* Hill coefficient for *mgtA* - nu_phop_mgta) could take the entire biological meaningful range, and that only 3 of the 68 parameters could accept less than 25% of their entire biological significant range. We repeated this analysis for 10 additional random feasible solutions, grouped the parameters according to their types (*e.g*., Phosphorylation, Dimerization, etc.) and calculated the average of the percentage of range acceptance (Fig. 6). We found high average values what showed the robustness quality of our final architecture

2.3.4. *Predicting simple input network motif regulated genes by scanning ranges of feasible solutions*

Genes embedded in a same network motif can show differential expression. Thus, we hypothesize about the different kinetic behavior that genes co-regulated by PhoP might exhibit by scanning the parameters related to the *mgtA* specie.

We simulated the previously learned system, ranging the Hill coefficient for *mgtA* (*i.e.*, nu_mgtA took the values of 1, 5 , 10) while also changing the half-maximum rate of activation (i.e., k_mgtA was assigned with 0.01, 0.025, 0.05, 0.075, and 0.1); the remaining parameters were not modified (i.e, H_PhoP_P=10; H_mgtA=20). We obtained 15 dynamics and observed that these could produce different kynetics. We clustered the patterns by applying the hierarchical method (Fig. 7), and found three groups that exhibit distinct kinetics. Cluster 1 groups patterns that show smooth peaks of expression that tends to decay fast. Cluster 2 characterizes patterns that also exhibit smooth peaks, but in contrast to the previous one, these transcriptions decay slowly. Finally, patterns grouped by cluster 3 show a high rate of transcription after minute 20.

We employed experimentally measured mRNA levels to evaluate the kinetics of PhoP regulated genes [14], and calculated the correlation (*c*) between these observation and the patterns predicted (Fig. 8). Our analysis showed that pattern 7 (cluster 3) predicted the dynamics of genes with early rise time and high level of transcription (*i.e. mig-14* *c*=0.76); pattern 5 (cluster 2) correlates to genes with a late rise time and low level of expression (*i.e. pagC* *c*=0.71)); and finally that pattern 8 (cluster 1) correlates to kinetic behavior of *pagD* gene (*i.e., c*=0.73).

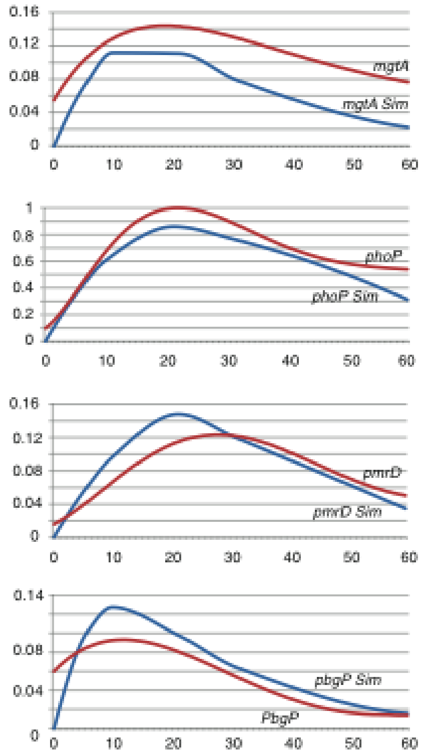

Figure 5: Simulated and experimentally validated gene kinetics. This charts reflects the high correlation between the predicted behavior and the experimentally obtained results (*i.e.,* mRNA expression quantified by real-time PCR).

3. CONCLUDING REMARKS

In this work, we proposed a framework to cope to the difficulties that arise at every inference stage. We provided tools to improve the sensitivity of detecting the *cis*-acting elements that determine the interaction among genes and their products, which in turn conform the whole set of network motifs instances. Employing an incremental approach, we integrated into a unique architecture the autoregulation, simple, bi-fan and chained modules, without compromising neither the *flexibility* nor the *robustness* of the final architecture. The predictions produced by the entire network for each of the network motifs correlated to the experimentally observed ones. We proved that the entire network preserves and recovers the network motifs kinetics, resulting in an adequate approach that solves the difficulties that arise when connecting previously identified network motifs. Moreover, simulating the dynamics of genes belonging to any of these modules independently of the reaming ones would allow the evaluation of the *realism* and *flexibility* of the proposed architectures, but would produce a cumbersome study of the *robustness* qualitative measures (*e.g.,* initial concentration of governing genes is a common parameter).

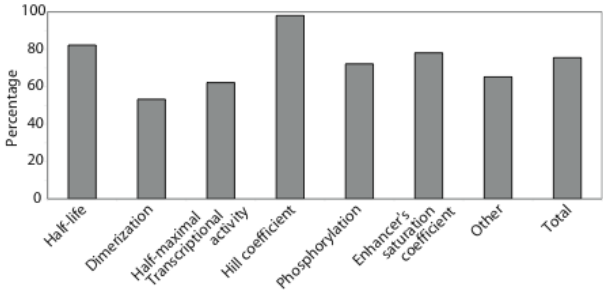

Figure 6: Robustness analysis. This chart shows the percentage of fulfillment for the biological meaningful range for parameters. The parameters are grouped by their type; the values represent the obtained average for 10 random solutions that satisfied the constraints).

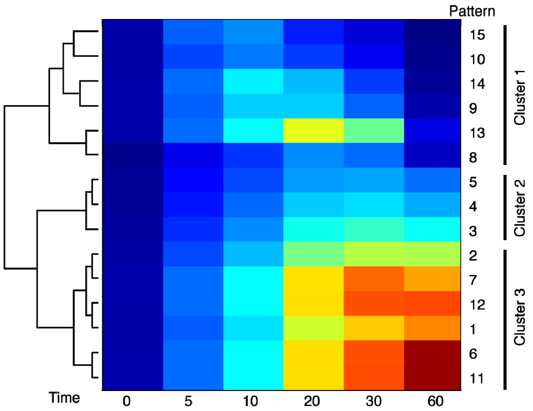

Figure 7. Identifying different kinetics obtained by scanning parameters of feasible solutions for simple network motif. The rows correspond to patterns obtained by different configuration; and the columns to the values (blue: low; red: high) obtained during the simulated time. The dendrogram (left panel) indicates the similarity (based on the correlation) of the patterns, revealing three characteristic clusters: low and short levels (cluster 1); low but longer peaks (cluster 2); and high level (cluster 3).

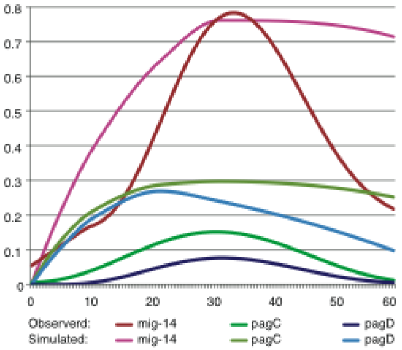

Figure 8. Simple network motif simulated dynamics vs. real-time PCR measured mRNA levels. Levels are normalized to the maximum one observed for PhoP gene (Fig. 5).

4. Materials and Methods

Tables and supplemental figures are available online at http://gps-tools2.wustl.edu/Sim/Appendix.pdf

Acknowledgment

This research was supported, in part, by Howard Hughes Medical Institute, and by The Spanish Ministry of Science and Technology grant TIN2009-13950 and GREIB grant to I.Z.

References

1. Kærn, M., *Regulatory dynamics in engineered gene networks.*, in *4th International Systems Biology Conference*. 2003: Washington University, St. Louis.
2. Alon, U., *An introduction to System Biology*. Mathematical and Computational Biology Series, ed. C. Hall/CRC. 2007, London: CRC Press, Taylor & Francis Group.
3. Alon, U., *Biological networks: the tinkerer as an engineer.* Science, 2003. **301**(5641): p. 1866-7.
4. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.
5. Meir, E., et al., *Ingeneue: a versatile tool for reconstituting genetic networks, with examples from the segment polarity network.* J Exp Zool, 2002. **294**(3): p. 216-51.
6. von Dassow, G., et al., *The segment polarity network is a robust developmental module.* Nature, 2000. **406**(6792): p. 188-92.
7. Groisman, E.A., *The pleiotropic two-component regulatory system PhoP-PhoQ.* J Bacteriol, 2001. **183**(6): p. 1835-42.
8. Harari, O., et al., *Identifying promoter features of co-regulated genes with similar network motifs.* BMC Bioinformatics, 2009. **100 Suppl IEEE**.
9. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
10. Kato, A., T. Latifi, and E.A. Groisman, *Closing the loop: the PmrA/PmrB two-component system negatively controls expression of its posttranscriptional activator PmrD.* Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4706-11.
11. Barnard, A., A. Wolfe, and S. Busby, *Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes.* Curr Opin Microbiol, 2004. **7**(2): p. 102-8.
12. Salgado, H., et al., *RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.* Nucleic Acids Res, 2004. **32**(Database issue): p. D303-6.
13. Zwir, I., et al., *Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica.* Proc Natl Acad Sci U S A, 2005. **102**(8): p. 2862-7.
14. Shin, D., et al., *A positive feedback loop promotes transcription surge that jump-starts Salmonella virulence circuit.* Science, 2006. **314**(5805): p. 1607-9.