Identifying the promoter features governing differential kinetics of co-regulated genes using fuzzy expressions

Rocio Romero Zaliz[1], Oscar Harari[1], Cristina Rubio Escudero[1] and Igor Zwir[1,2]

Abstract— **One of the biggest challenges in genomics is the elucidation of the design principles controlling gene expression. Current approaches examine promoter sequences for particular features, such as the presence of binding sites for a transcriptional regulator, and identify recurrent relationships among these features termed network motifs. To define the expression dynamics of a group of genes, the strength of the connections in a network must be specified, and these are determined by the *cis*-promoter features participating in the regulation. Approaches that homogenize features among promoters (e.g., relying on consensuses to describe the various promoter features) and even across species hamper the discovery of the key differences that distinguish promoters that are co-regulated by the same transcriptional regulator. Thus, we have developed a an approach based on fuzzy logic expressions to analyze proteobacterial genomes for promoter features that is specifically designed to account for the variability in sequence, location and topology intrinsic to differential gene expression. We applied our method to characterize network motifs controlled by the PhoP/PhoQ regulatory system of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. We identify key features that enable the PhoP protein to produce distinct kinetic patterns in target genes, which could not have been uncovered just by inspecting network motifs.**

I. Introduction

Whole 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 regulatory networks. These networks can be represented as directed graphs in which a node stands for a gene (or an operon in the case of bacteria) and an edge symbolizes a direct transcriptional interaction. Recurrent patterns of interactions, termed network motifs, occur far more often than in randomized networks, forming elementary building blocks that carry out key functions. This is a convenient representation of the topology of a set of regulatory Boolean (i.e. ON-OFF) networks, in which each gene is either fully expressed or not expressed at all, or that it has a binding site for a transcriptional regulator or lacks such a site. However, this approach has serious limitations because most genes are not expressed in a simple Boolean fashion. Indeed, genes that are co-regulated by the same transcription factor are often differently expressed with characteristic expression levels and kinetics. Therefore, a deeper understanding of regulatory networks demands the identification of the key features used by a transcriptional regulator to differentially control genes that display distinct behaviors despite belonging to networks with identical motifs.

The identification of the promoter features that determine the distinct expression behavior of co-regulated genes is a challenging task because: first, there are difficulties in discerning the sequence elements relevant to differential expression patterns (e.g., the binding sites for transcriptional regulators and RNA polymerase) from a background of variable DNA sequences that do not play a direct role in gene regulation. Second, the sequences recognized by a transcription factor may differ from promoter to promoter within and between genomes and may be located at various distances from other *cis*-acting features in different promoters [1]. Third, similar expression patterns can be generated from different or a mixture of multiple underlying features, thus, making it more difficult to discern the causes of analogous regulatory effects.

In this study, we present a method specifically aimed at handling the variability in sequence, location and topology that characterize gene transcription. Instead of using an overall consensus model for a feature, where important differences are often concealed because of intrinsic averaging operations between promoters and even across species (see Appendix), we decompose a feature into a family of models or building blocks. This approach maximizes the sensitivity of detecting those instances that weakly resemble a consensus (e.g., binding site sequences) without decreasing the specificity. In addition, features are considered using fuzzy assignments, which allow us to encode how well a particular sequence matches each of the multiple models for a given promoter feature. Individual features are then linked into more informative composite fuzzy expressions that can be used to explain the kinetic expression behavior of genes. We applied our method to analyze promoters controlled by the PhoP/PhoQ regulatory system of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. This system responds to the same inducing signal (i.e. low Mg^{2+}) in both species [1, 2]. Moreover, the *E. coli phoP* gene could complement a *Salmonella phoP* mutant [3]. The DNA-binding PhoP protein appears to recognize a tandem repeat sequence separated by 5 bp [1, 2], consistent with being a dimer. The PhoP/PhoQ system is an excellent test case because it controls the expression of a large number of genes, amounting to ca. 3% of the genes in the case of *Salmonella*. Furthermore, the PhoP/PhoQ regulon has been shown to employ a variety of network motifs including the single-input

[1]Department Computer Science and Artificial Intelligence, University of Granada, E-18071 Granada, Spain tel: (34) (958) 240469 fax: (34) (958) 243317. [2]Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, Campus Box 8230, 660 S. Euclid Ave., St. Louis, Missouri, 63110, USA Tel: (1) (314) 362-3691 fax: (1) (314) 7478228. E-mail: zwir@borcim.wustl.edu

1-4244-1210-2/07/$25.00 ©2007 IEEE.

module (Fig. 1A), the multi-input module (Fig. 1B), the bi-fan (Fig. 1C), the chained (Fig. 1D), and also the feedforward loop [1, 4]. Our analysis uncovered the salient features that distinguish genes co-regulated by PhoP belonging to similar networks. Gene transcription measurements provided experimental support for the investigated predictions.

II. Results

We investigated five types of *cis*-acting promoter features by extracting the maximal amount of useful information from datasets and then creating models that describe promoter regulatory regions. This entailed applying three key strategies (Appendix Fig. S1): first, we conducted an initial survey of the data provided from different available sources, capturing and distinguishing between broad and easily discernable patterns. We then used these patterns as models to re-visit the data with greater sensitivity and specificity, which allowed the detection of those instances where a binding site sequence resembles the consensus only weakly or where the distances between the transcription factor and the RNA polymerase are unusual. Second, we utilized fuzzy clustering methods [5, 6] to encode how a promoter matches each of the multiple models for a given promoter feature, which avoided having to make premature categorical assignments, thus producing an initial classification of the promoters into multiple subsets. Finally, we applied fuzzy logic to link basic features into more informative composite models that explain the distinct expression behavior of genes belonging to similar networks (Appendix Fig. S1 and S2). Additional features are described in the Appendix.

A. Transcription factor binding site submotifs

Many genes are controlled by a single-input network motif where the affinity of a transcription factor for its promoter sequences is a major determinant of gene expression (Fig. 2A). Thus, co-regulated genes displaying distinct expression patterns are likely to differ in the binding site for such a transcription factor. Methods that look for matching to a consensus sequence have been successfully used to identify promoters controlled by particular transcription factors. However, the strict cutoffs used by such methods increase specificity but decrease sensitivity, which makes it difficult to detect binding sites with weak resemblance to a consensus sequence [7].

To circumvent the limitation of consensus methods [8], we decomposed the binding site motif of a transcription factor into several submotifs and then combined the submotifs into a multi-classifier (see Methods), which increased the sensitivity to weak sites without losing specificity. In the case of PhoP, we identified four submotifs (Appendix Fig. S3), and used them to search both strands of the intergenic regions of the *E. coli* and *Salmonella* genomes (Appendix Fig. S2). 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 consensus position weight matrix model [7] despite being footprinted by the PhoP protein [1, 4].

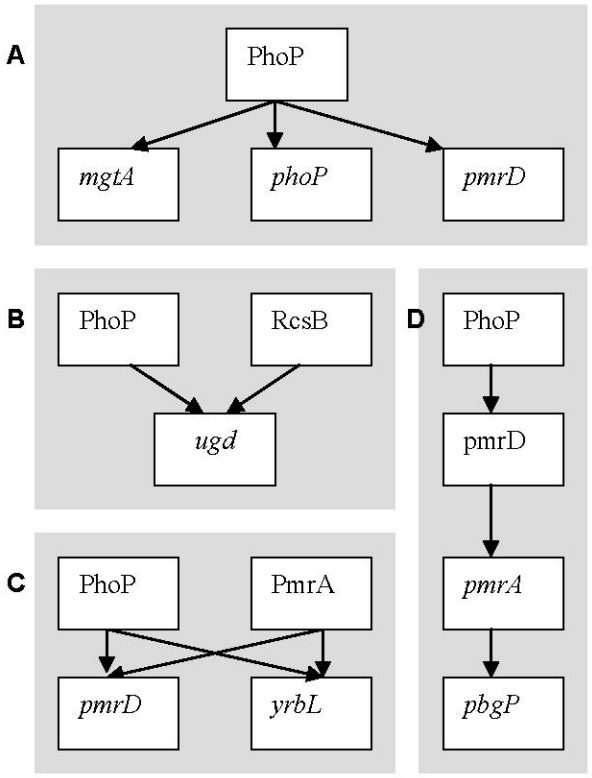

Fig. 1. The PhoP/PhoQ system employs a variety of network motifs to regulate gene transcription. (**a**) In the single-input module, PhoP as a single transcription factor regulates a set of genes (i.e. *mgtA*, *phoP* and *pmrD*). (**b**) In the multi-input module, two or more transcription factors (e.g., PhoP and RcsB) regulate a target gene (i.e. *ugd*). (**c**) In the bi-fan module, a set of genes (i.e. *pmrD* and *yrbL*) 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.

To test the notion that PhoP binding to promoters with different PhoP box submotifs is a determinant promoter activity, we compared the gene expression patterns of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene to different PhoP-activated promoters. Faster GFP expression kinetics were observed when transcription was driven by the *phoP* promoter, which has the M_2 submotif, than when it was driven by the *pmrD* promoter, which has the M_1 submotif, (Fig. 2B-C). Thus, the binding site for a transcriptional regulator is a key determinant in gene expression.

Performance. To evaluate the ability of the resulting models to describe PhoP-regulated promoters, we extended the dataset by including 772 promoters (RegulonDB V3.1 database [11]) that are regulated by transcription factors other than PhoP (see "Search known transcription factor motifs" in gps-tools.wustl.edu), by selecting the promoter region corresponding to the respective transcription factor binding site ± 10 bp. We considered the compiled list of PhoP regulated genes as true positive examples (Appendix Table S1) and the binding sites of other transcriptional regulators as true negative examples to evaluate the performance of the submotif feature. We used a leave-one-out crossvalidation process (Crossvalind, Matlab r2006a), which is appropriate for reduced datasets, as a procedure to estimate the variance error on the training set (correct test estimation of 94% vs.

75% between submotifs and single position weight matrices, respectively). Then, each matrix threshold has been optimized for classification purposes by using the correlation coefficient measurement (see below) based on the extended dataset (Appendix Table S2). (See the complete evaluation of genomes in gps-tools.wustl.edu). We found that the PhoP-binding site model increases its sensitivity from 66% to 91% when submotifs are used instead of a single consensus, while its specificity went from 98% to 97% (correlation coefficient 73% vs. 87%).

We also obtained substantial improvements for other transcription factors from RegulonDB. For example, by considering the CRP regulator, we used 130 promoters regulated by this protein in RegulonDB as the true positive values and 642 regulated by other proteins than CRP as negative examples. We found that the sensitivity of the CRP model for binding sites increases from 29% to 50%, by using submotifs instead of a single consensus, while the specificity remains the same at 98% (correlation coefficient 39% vs. 62%). Overall, by considering transcription factors with more than ten reported binding sequences in the RegulonDB data base (including CRP, Lrp, FIS, IHF, FNR, ArcA, NarL, GlpR, PurR, OmpR, TyrR, AraC, Fur, CytR, FruR,Hns, ArgR, DnaA, PhoB, and LexA), we could increase the sensitivity in an average of 35%, while retain almost the same sensitivity than a single position weight matrix (average correlation coefficient 87%).

B. Transcription factor binding site orientation

Functional binding sites for a transcription factor may be present in either orientation relative to the RNA polymerase binding site. This is due to the possibility of DNA looping and to the flexibility of the alpha subunit of the bacterial RNA polymerase in its interactions with transcriptional regulators [9]. Analysis of PhoP-regulated promoters revealed that the PhoP box could be found with the same probability in either orientation in the intergenic regions of the *E. coli* and *Salmonella* genomes (Appendix Fig. S7). For example, the *E. coli ompT* and *yhiW* promoters and the *Salmonella mig-14, pipD*, *pagC* and *pagK* promoters harbor putative PhoP binding sites in the opposite relative orientation to that described for the prototypical PhoP-activated *mgtA* promoter [1] (Appendix Fig. S2). Yet other promoters (i.e. those of the *ybjX*, *slyB*, *yeaF* genes in *E. coli* and the *virK*, *ybjX,* and *mgtC* genes in *Salmonella*) contain sequences resembling the PhoP box in both orientations. The demonstration that PhoP does bind to the *mgtC, mig-14* and *pagC* promoters [1], which harbor the PhoP binding site in the opposite orientation as in the *mgtA* promoter, validates our predictions and argues against alternative network designs where these promoters would be regulated by PhoP only indirectly [10].

To assess the contribution of PhoP box orientation to gene expression, we determined the fluorescence of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene to PhoP-regulated promoters that differed in the orientation of the PhoP box. Promoters with the PhoP box in the direct orientation, such as those corresponding to the *yobG* and *slyB* genes, were transcribed earlier and faster than the *pagK* and *pagC* promoters in which the PhoP box is in the opposite relative orientation (Fig. 3A-C). This is in spite of the fact that *yobG* and *pagK* promoters are equally divergent from the PhoP binding site consensus (60% and 66% of the consensus information content (Appendix Fig. S3), respectively). Furthermore, promoters sharing the same PhoP binding site submotif but arranged in different orientations (e.g. the *ugd* and *mig-14* promoters) produced distinct rise times and expression levels (data not shown).

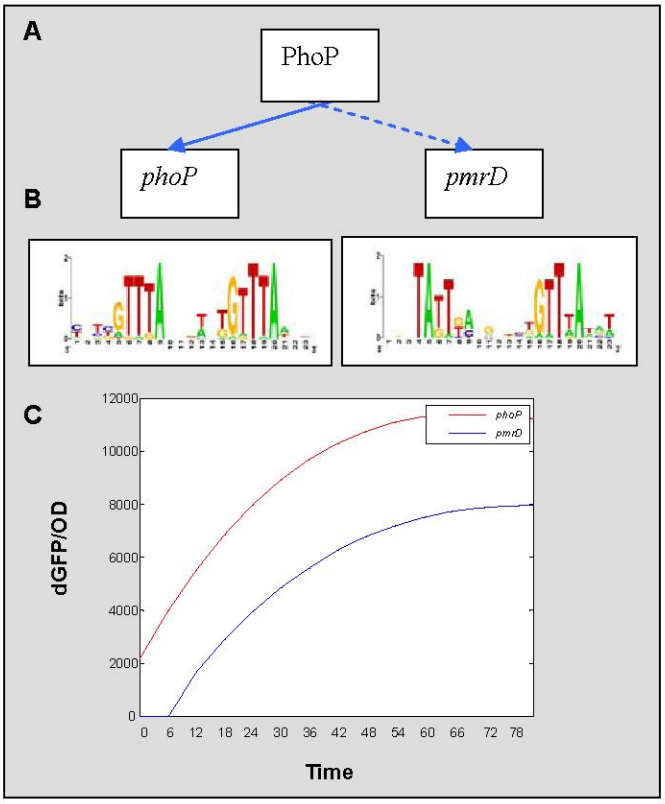

Fig. 2. The PhoP protein achieves differential expression using the single-input network motif by controlling genes that differ in their binding site submotifs. (**a**) PhoP regulates several promoters (i.e. *phoP* and *pmrD*) using a single-input network motif. (**b**) The PhoP protein recognizes a binding site motif consisting of a hexameric direct repeat separated by 5 bp, but distinguishes between different submotifs with different specificities. We identified four of these classes (M_1- M_4; Appendix Fig. S3), and tested the influence of this *cis*-feature in the *phoP* and *pmrD Salmonella* promoters corresponding to class M_2 and M_1, respectively. (**c**) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella phoP* (red color) or *pmrD* (blue color) promoters. The activity of each promoter is proportional to the number of GFP molecules produced per unit time per cell $[dG_i(t)/dt]/OD_i(t)]$, where $G_i(t)$ is GFP fluorescence from wild-type *Salmonella* strain 14028s culture and conditions described in Methods, and $OD_i(t)$ is the optical density. The activity signal was smoothed by a polynomial fit (sixth order). Faster and earlier GFP expression was observed when transcription was driven by the *phoP* promoter, which has the M_2 submotif, than by the *pmrD* promoter, which has the M_1 submotif.

C. RNA polymerase site

The distance of a transcription factor binding site to the RNA polymerase binding site(s) and the class of sigma 70 promoter are critical determinants of gene expression [9]. These classes

correspond to the different types of contacts that can be established between a transcription factor and RNA polymerase. We identified seven patterns among PhoP-regulated promoters of *E. coli* and *Salmonella* (Appendix Fig. S2) that combine promoter class and distance between the PhoP box and the RNA polymerase site (Appendix Fig. S5). These patterns may correspond to different kinetic behaviors within a network motif [9]. For example, the *ugtL* and *pagC* promoters share the orientation of the PhoP box but differ in the distance of the PhoP box to the RNA polymerase binding site (Fig. S4A-B). This may account for the different dynamic behavior of these promoters when tested in a wild-type strain harboring plasmids with promoter fusions to the promoterless *gfp* gene (Fig. S4C). In addition, some PhoP-regulated promoters (e.g. the *hemL* and *phoP* promoters of *E. coli*) contain several putative RNA polymerase binding sites located at different positions and belonging to different classes, suggesting that such promoters may be regulated by additional signals and/or transcription factors [2].

Performance. The RNA polymerase site feature was evaluated using 721 RNA polymerase sites from RegulonDB as positive examples and 7210 random sequences as negative examples. We obtained an 82% sensitivity and 95% specificity for detecting RNA polymerase sites. These values provide an overall performance measurement (see below) of 92% corresponding to a false discovery rate <0.001 and a correlation coefficient of 82%. In addition, we selected 34 examples of RNA polymerase sites reported to be of class II, which all differ from the typical class I promoter by exhibiting a degenerate -35 sequence motif [2, 9], and obtained 74% sensitivity and 95% specificity.

D. *Binding sites for other transcription factors.*

Certain promoters harbor binding sites for more than one transcription factor. This could be because transcription requires the concerted action of such proteins, or because the promoter is independently activated by individual transcription factors, each responding to a distinct signal. We analyzed the intergenic regions of the *E. coli* and *Salmonella* genomes for the presence of binding sites for 54 transcription factors [11]. 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 network motifs involving PhoP-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 that can be described by the bi-fan network motif [1, 12] (Fig. 4A). In addition, the relative position of transcription factor binding sites (Appendix Fig. S6D) can play a critical role because the PmrA-box in the *Salmonella pmrD* and *yrbL* promoters is located closer to the PhoP-box (~38 bp and ~24 bp, respectively) than in the *udg* promoter (~65 bp), which could account for the different expression patterns exhibited by their respective genes (Fig. 4B-C). By analyzing both the binding site quality and the location of transcription factor binding sites, we increase the chances of identifying co-regulated promoters. By considering the presence of binding sites for multiple transcription factors, it is possible to generate hypotheses about potential network motifs. This notion was experimentally verified [1], validating our prediction.

III. Materials and Methods

Our method consists of three phases (Appendix Fig. S1): first, encoding the available information into preliminary model-based features, which includes identifying *cis*-features from DNA sequences and information from available databases; performing initial modeling of each individual feature, allowing the process of multiple occurrences of a feature and using relaxed thresholds and permitting missing values. A *model-based* feature is generated by the identification of a feature in a subset of observations (F) in the dataset, based on measuring the degree of match (Q) between an observation and a model, or a family of models ($M=\{M_\alpha\}$), at some degree (α) defined in a unit-interval scale (i.e., fuzzy values, $Q(F, M_\alpha)$) [13-15]. Second, grouping the results into subsets, thus, decomposing the preliminary models into a family of models or building blocks by using fuzzy clustering. Third, combining the same or different types of features by using fuzzy logic expressions and describing new promoters using the resulting models.

A. *Dataset*

We initially used the intergenic regions of *E. coli* and *Salmonella* operons from -800 to +50 because >5% are larger than 800 bp in bacterial genomes (as described in the RegulonDB database or generously provided by H. Salgado); however, predictions have been performed in whole coding and non coding regions (see gps-tools.wustl.edu). The promoter and transcription factor information was taken from RegulonDB database. We compiled from the literature and our own lab information (Appendix Table S1) genes whose expression (using microarrays) differed statistically between wild-type and *phoP E. coli* strains experiencing inducing conditions for the PhoP/PhoQ regulatory system [1], as well as a list of genes known/assumed to be PhoP regulated (Appendix Table S2). However, this information did not explicitly indicate whether these genes were regulated directly or indirectly by the PhoP protein. The learned features were used to make genome-wide predictions in the *E. coli* and *Salmonella* genomes.

B. *Binding site submotifs and orientation*

(1) We built an initial model for the PhoP binding site by learning a position weight matrix (*E-value* < 10E-12) based on the upstream sequences of genes corresponding to the training set of the *E. coli* and *Salmonella* genomes (Appendix Table S1). **(2)** We searched the intergenic regions of the genes in both orientations, using low thresholds corresponding to two standard deviations below the mean score obtained with the initial model [16]. Multiple PhoP

binding site candidates were allowed in a given promoter operator region. **(3)** After transforming nucleotides into dummy variables, we grouped sequences matching the PhoP position weight matrix using the fuzzy C-means clustering method with the Xie-Beni validity index (see below) to estimate the number of clusters [5].

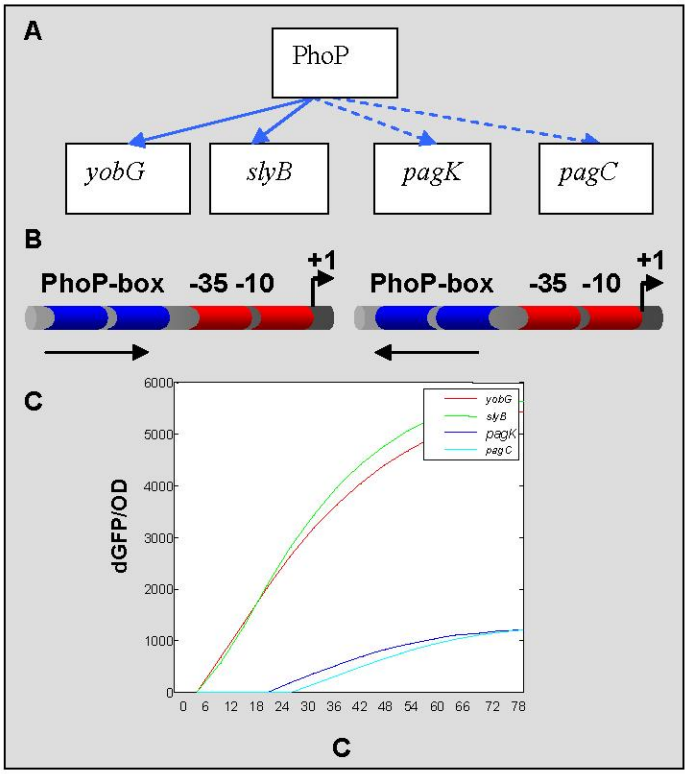

Fig. 3. Expression of PhoP-regulated promoters that differ in the orientation of the PhoP-binding site. (**a**) PhoP regulates a set of promoters including those of the *Salmonella yobG*, *slyB*, *pagK* and *pagC* genes using a single-input network motif. (**b**) We established that when *Salmonella* experiences low Mg^{2+}, the PhoP protein binds to both the archetypal directly oriented *yobG* and *slyB* promoters as well as the oppositely oriented *pagK* and *pagC* promoters using chromatin immunoprecipitation (ChIP) *in vivo*. (**c**) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella yobG* (red color) or *slyB* (green color) promoters reveals a much earlier an higher levels of activity than the isogenic strains with fusions to the *pagK* (blue color) and *pagC* (cyan color) promoters. Promoter activity was determined as described in the legend to Fig. 2. Thus, the orientation of the binding site for a transcriptional regulator contributes to the kinetic behavior as well as the maximum expression levels achieved by the promoters.

(4) We built models for these clusters using position weight matrices (*E-value* < 10E-22) and searched the *E. coli* and *Salmonella* genomes to characterize each gene according to its similarity to each model as a fuzzy partition (Appendix Fig. S2 and S2).

C. RNA polymerase sites

(1) We gathered sigma 70 class I and class II promoters [11] from the RegulonDB database. Then, we built models of the RNA polymerase site using a neuro-fuzzy method (see HPAM in gps-tools.wustl.edu [17]), and used the resulting models to perform genome-wide descriptions of the intergenic regions of the *E. coli* and *Salmonella* genomes with a false discovery rate <0.001 (see Promoter search in gps-tools.wustl.edu). **(2)** We used an intelligent parser to differentiate class I and class II promoters that evaluate the quality of the -35 motif [9], based on fuzzy logic and genetic algorithms techniques (see MOSS in gps-tools.wustl.edu [18]). **(3)** To characterize the distance relationship between transcription factors binding sites and RNA polymerase binding sites, we built models of such distances from the examples reported in the RegulonDB database. **(3.1)** We modeled activated and repressed promoters (see below *Activated or repressed* feature). **(3.2)** We re-built histograms for each group of distances (i.e. activated and repressed), distinguishing three overlapping distributions for each of them (Appendix Fig. S5). **(3.3)** We built models for distances by fitting their distributions into models based on fuzzy membership functions, which were termed close, medium and remote distances for each set of activated and repressed genes. Finally, to characterize the distance relationship between the PhoP box and putative RNA polymerase binding site, we connected (2) and (3) by using fuzzy logic-based operations (see below).

This process allowed us to retrieve the most representative RNA polymerase binding site candidates for each promoter region relative to the PhoP binding site (e.g., best class II RNA polymerase site, which is located close to the PhoP box in an activated promoter), which were arrayed and constituted the value of the RNA polymerase site feature in Appendix Fig. S2.

D. Binding sites for other transcription factors

We developed models for different transcription factor binding sites from the RegulonDB database as follows: **(1)** We built position weight matrices for each transcription factor using the Consensus/Patser program, choosing the best final matrix for motif lengths between 14-30 bps if the corresponding length had not been previously specified (see "Consensus matrices" in gps-tools.wustl.edu). We accounted for the motif symmetry (e.g., asymmetric, direct, inverted [11]) if available (see "Search known transcription factor motifs" in gps-tools.wustl.edu). **(2)** We searched the intergenic regions of the *E. coli* and *Salmonella* genomes with these models, using the *overall performance* measure (see below) and additional 772 promoters from the RegulonDB database [11] to establish a threshold (average *E-value* < 10E-10) for each matrix [19] (see "Threshold consensus" in gps-tools.wustl.edu). **(3)** We accounted for the distances between distinct transcription factor binding sites occurring in the same promoter region (e.g., the distance between the CRP and FIS sites in the *proP* promoter) in promoters reported in RegulonDB database and built a histogram with the obtained results (Appendix Fig. S6D). **(4)** We fitted the histogram using a fuzzy membership function (see below) and used this model as a fuzzy cluster to characterize the distances between a putative PhoP box and another putative transcription factor binding site detected in the same region. **(5)** Finally, we connected **(2)** and **(4)** by using fuzzy logic-based operations as described above to characterize PhoP regulated candidates promoters.

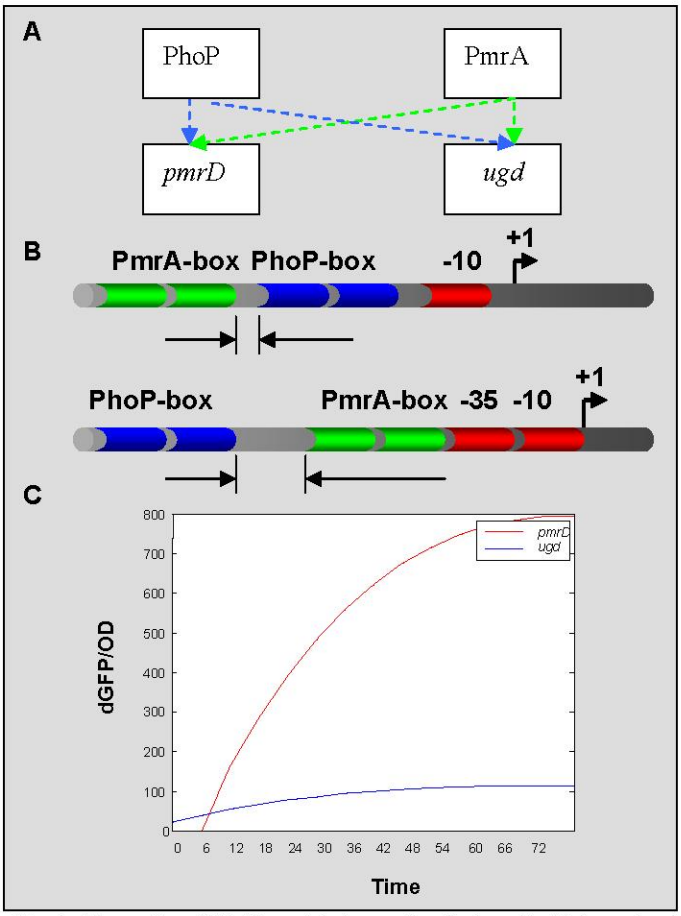

Fig. 4. Expression of PhoP-regulated promoters that use the bi-fan network motif. (**a**) The *Salmonella pmrD*, and *ugd* promoters harbor experimentally verified PhoP- and PmrA-binding sites that can be described by the bi-fan network motif. (**b**) The distance between the PhoP and PmrA boxes in the *Salmonella pmrD* and *ugd* promoters are different (~38 bp and ~65 bp, respectively). (**c**) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella pmrD* and *ugd* promoters. Promoter activity was determined as described in the legend to Fig. 2. The two promoters confer different expression and kinetic patterns.

E. Fuzzy logic expressions

Propositional calculus logic expressions can be extended by incorporating predicates having fuzzy variables, which are manipulated using various theorems/axioms and methods. This approach, which has been widely used in several fields including decision-making, artificial intelligence and electrical engineering for many years, was applied to model related features that describe different regulatory objects. Thus, given a dataset $X=\{x_1,...,x_n\}$, the feature that characterizes it can be best described as a set $F_1(X)=\{d_{11}/x_1,...,d_{1n}/x_n\}$, where $\{d_{11},...,d_{1n}\}\in\{0,1\}$ in classical set theory and $[0,1]$ in fuzzy set theory. These fuzzy values represent the degree of matching between an observation of the dataset and a fuzzy set. The degree of matching is defined in the unit interval and can be obtained from evaluating the membership function of the corresponding fuzzy set (see below). Then, given $F_2(X)=\{d_{21}/x_1,...,d_{2n}/x_n\}$ and the Minimum as an intersection operator, we define the expression:

$$F_1(X)\ AND\ F_2(X)=F_1\cap F_2=MIN(F_1,F_2)=$$
$$\{MIN(d_{11},d_{21})/x_1,...,MIN(d_{1n},d_{2n})/x_n\}$$

Fuzzy logic-based operations, such as T-norms/conorms, include operators like *MINIMUM, PRODUCT*, or *MAXIMUM*, which are used as basic logic operators, such as AND or OR, or their set equivalents *INTERSECTION* or *UNION* [5]. We used in this work the Minimum and Maximum as T- and Tconorms, respectively.

F. Fuzzy membership functions

They can be viewed as approximation of data distributions, where the degree of matching in the [0,1] scale is calculated using triangular functions. These functions were learned from the projection of the histograms onto the variable domains (Fig. S4) by simple regression and minimum squared methods [20].

G. Performance Measurement

We use a correlation coefficient implementation to establish best local thresholds for transcription factor binding site motifs. That is, from a range of possible thresholds applied over a particular motif, we choose the one that maximizes this coefficient defined as:

$$CC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}},$$

where *specificity* = *TN/(TN + FP)* and *sensitivity* = *TP/(TP + FN)*; *P= positive, N=negative, T = true and F=false* [19]. We constrained the sensitivity of the selected threshold to be above the 60%. The false positive rate for binding site analysis was calculated by detecting binding sites from other transcription factors different from the one being evaluated (RegulonDB database).

IV. Conclusions

We demonstrated that a transcription factor can mediate differential expression of genes that are described even by the same network motif. This is because of the functional significance of variability in sequence, location and topology that exists among promoters that are co-regulated by a given transcription factor. We developed a flexible computational framework to encode and to combine these promoter features, which allows matching of *cis*-observations to multiple models for a given promoter feature. This enables the description of regulatory elements from different angles and the generation of composite models that can be used to explain the different kinetic behavior of co-regulated genes.

Finally, unlike regulators such as the LacI and MelR [1] proteins of *E. coli* that govern expression of single promoters, many transcriptional regulators control multiple promoters that express products required in different amounts or for different extents of time. This is clearly the case for the

regulatory protein PhoP, which controls transcription of a large numbers of genes, that can be described by a variety of network motifs (Fig. 1). Our finding argues that understanding a cell's behavior in terms of differential expression of genes controlled by a transcription factor requires a detailed analysis of a promoter's regulatory features. As a single nucleotide difference in the binding site for a transcription factor can dictate the requirement for co-activator proteins [21], we feel that by considering multiple models (as opposed to the relying on consensuses) it will be possible to uncover subtle differences between regulatory targets and to capture the salient properties of co-regulated promoters.

APPENDIX

Additional text, methods, tables and figures are available online at gps-tools2.wustl.edu/IEEE-FUZZY07/Appendix_IEEE. pdf.

ACKNOWLEDGMENT

We thank U. Alon (Weizmann Institute of Science, Israel) for plasmid pMS20 and E.A. Groisman for his suggestions and lab support. This research was supported, in part, by Howard Hughes Medical Institute, and by The Spanish Ministry of Science and Technology grants TIN2006-12879 to I.Z.

REFERENCES

[1] I. Zwir, D. Shin, A. Kato, K. Nishino, T. Latifi, F. Solomon, J. M. Hare, H. Huang, and E. A. Groisman, "Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica," *Proc Natl Acad Sci U S A,* vol. 102, pp. 2862-7, Feb 22 2005.

[2] S. Minagawa, H. Ogasawara, A. Kato, K. Yamamoto, Y. Eguchi, T. Oshima, H. Mori, A. Ishihama, and R. Utsumi, "Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli," *J Bacteriol,* vol. 185, pp. 3696-702, Jul 2003.

[3] E. A. Groisman, "The pleiotropic two-component regulatory system PhoP-PhoQ," *J Bacteriol,* vol. 183, pp. 1835-42, Mar 2001.

[4] I. Zwir, H. Huang, and E. A. Groisman, "Analysis of Differentially-Regulated Genes within a Regulatory Network by GPS Genome Navigation," *Bioinformatics,* vol. 21, pp. 4073-83, Nov 15 2005.

[5] J. C. Bezdek, S. K. Pal, and IEEE Neural Networks Council, *Fuzzy models for pattern recognition : methods that search for structures in data.* New York: IEEE Press, 1992.

[6] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol,* vol. 3, p. RESEARCH0059, Oct 10 2002.

[7] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics,* vol. 16, pp. 16-23, Jan 2000.

[8] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites," *Nat Biotechnol,* vol. 23, pp. 137-44, Jan 2005.

[9] A. Barnard, A. Wolfe, and S. Busby, "Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes," *Curr Opin Microbiol,* vol. 7, pp. 102-8, Apr 2004.

[10] S. Lejona, A. Aguirre, M. L. Cabeza, E. Garcia Vescovi, and F. C. Soncini, "Molecular characterization of the Mg2+-responsive PhoP-PhoQ regulon in Salmonella enterica," *J Bacteriol,* vol. 185, pp. 6287-94, Nov 2003.

[11] H. Salgado, S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jimenez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martinez, and J. Collado-Vides, "RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12," *Nucleic Acids Res,* vol. 32, pp. D303-6, Jan 1 2004.

[12] A. Kato and E. A. Groisman, "Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor," *Genes Dev,* vol. 18, pp. 2302-13, Sep 15 2004.

[13] E. H. Ruspini and I. Zwir, "Automated Qualitative Description of Measurements," in *Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conf.*, Venice, Italy, 1999.

[14] E. H. Ruspini and I. Zwir, "Automated generation of qualitative representations of complex objects by hybrid soft-computing methods," in *Pattern recognition : from classical to modern approaches*, S. K. Pal and A. Pal, Eds. New Jersey.: World Scientific, 2002, pp. 454-474.

[15] I. Zwir, R. R. Zaliz, and E. H. Ruspini, "Automated biological sequence description by genetic multiobjective generalized clustering," *Ann N Y Acad Sci,* vol. 980, pp. 65-82, Dec 2002.

[16] K. Robison, A. M. McGuire, and G. M. Church, "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome," *J Mol Biol,* vol. 284, pp. 241-54, Nov 27 1998.

[17] V. Cotik, R. R. Zaliz, and I. Zwir, "A hybrid promoter analysis methodology for prokaryotic genomes," *Fuzzy Sets and Systems,* vol. 152, pp. 83-102, MAY 16 2005.

[18] R. Romero Zaliz, I. Zwir, and E. H. Ruspini, "Generalized analysis of promoters: a method for DNA sequence description," in *Applications of Multi-Objective Evolutionary Algorithms*, C. a. L. Coello Coello, G., Ed. Singapore: World Scientific, 2004, pp. 427-450.

[19] E. Benitez-Bellon, G. Moreno-Hagelsieb, and J. Collado-Vides, "Evaluation of thresholds for the detection of binding sites for regulatory proteins in Escherichia coli K12 DNA," *Genome Biol,* vol. 3, p. RESEARCH0013, 2002.

[20] M. Sugeno and T. Yasukama, "A Fuzzy-logic-based Approach to Qualitative Modeling," *IEEE Transactions on Fuzzy Systems,* vol. 1, pp. 7-31, 1993.

[21] T. H. Leung, A. Hoffmann, and D. Baltimore, "One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers," *Cell,* vol. 118, pp. 453-64, Aug 20 2004.