Fusion of Domain Knowledge for Dynamic Learning in Transcriptional Networks

Oscar Harari[1], R. Romero-Zaliz[1], C. Rubio-Escudero[1], and I. Zwir[1,2]

[1] Dept. Computer Science and Artificial Intelligence, University of Granada, E-18071, Spain
[2] Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110-1093, USA
oharari@correo.ugr.es, rocio@decsai.ugr.es, crubio@decsai.ugr.es, zwir@borcim.wustl.edu

Abstract. A critical challenge of the postgenomic era is to understand how genes are differentially regulated even when they belong to a given network. Because the fundamental mechanism controlling gene expression operates at the level of transcription initiation, computational techniques have been developed that identify *cis*-regulatory features and map such features into differential expression patterns. The fact that such co-regulated genes may be differentially regulated suggests that subtle differences in the shared *cis*-acting regulatory elements are likely significant. Thus, we carry out an exhaustive description of *cis*-acting regulatory features including the orientation, location and number of binding sites for a regulatory protein, the presence of binding site submotifs, the class and number of RNA polymerase sites, as well as gene expression data, which is treated as one feature among many. These features, derived from different domain sources, are analyzed concurrently, and dynamic relations are recognized to generate profiles, which are groups of promoters sharing common features. We apply this method to probe the regulatory networks governed by the PhoP/PhoQ two-component system in the enteric bacteria *Escherichia coli* and *Salmonella enterica*. Our analysis uncovered novel members of the PhoP regulon as and the resulting profiles group genes that share underlying biological that characterize the system kinetics. The predictions were experimentally validated to establish that the PhoP protein uses multiple mechanisms to control gene transcription and is a central element in a highly connected network.

1 Introduction

One of the biggest challenges in genomics is the elucidation of the design principles controlling gene networks. However, knowing the connectivity of a given network is not sufficient to define the expression dynamics of a group of genes; one also needs to specify the strength of the connections in a network, which are determined by the *cis*-promoter features participating in the regulation (Fig. 1a-b).

This work describes a machine learning method [1, 2] that integrates heterogeneous domains of knowledge to identify, differentiate and group genes by their expression patterns within a regulatory network. We encapsulate each source of information into model-based features, including fix-length DNA sequence motifs from transcription factor binding sites encoded as position weight matrices; variable-length motifs

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 1067 – 1078, 2006.

from RNA polymerase encoded as neural network edges; locations of these regulatory elements in the chromosome as data distributions encoded into fuzzy sets; and gene expression patterns from multiple experiments encoded as temporal vectors. Furthermore, we account for the variability of the data by treating these features as fuzzy (i.e., not precisely defined) instead of categorical entities [3-5].

We use conceptual clustering techniques [1] to integrate the regulatory features by combining features and promoters[1] into dynamic profiles, which are sets of promoters sharing a common set of features. The features are treated with equal weight, because it is not known beforehand which features are important for a profile to explain a differential gene behavior. The formulation of this clustering problem would result in the generation of many profiles with small extent, as it is easier to explain or profile-match smaller data subsets than those that constitute a significant portion of the dataset. For this reason, our approach also considers additional criteria to extract broader profiles based on their size, the number of retrieved profiles, and their diversity and extent of overlap [3, 5]. These are conflicting criteria that are formulated as a multi-objective and multimodal optimization problem [6].

We applied our method to characterize a network controlled by the PhoP/PhoQ regulatory system of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. We could identify key features that enable the PhoP protein to produce distinct kinetic patterns in target genes and uncover novel members of the PhoP regulatory network [7]. Our approach provides resources for the annotation of genome regulatory regions and their compilation in predictive databases.

2 Methods

Regulatory networks constitute a typical case of structural data, where genes can be viewed as objects described by several features including expression patterns and particular *cis*-acting promoter elements. Promoters are inherently complex combinations of objects that, in turn, are described by a number of features. For example, binding sites for one or more transcriptional regulators are characterized by their match to the binding motif of the regulators, and their locations relative to each other and to that of the RNA polymerase binding site(s).

The purpose of our proposed method is to identify interesting substructures, here termed profiles (i.e., groups of promoters sharing a common set of features), within a regulatory network, thus to suggest possible mechanisms by which the respective genes are controlled, which can further be used to classify additional (e.g., newly identified) promoters. Our method represents, learns and infers from structural data by following three main phases: (1) *Database representation by modeling the features of promoters*[8] ; (2) *Fusing distinct domain knowledge by dynamic learning profiles*; and (3) *Using the profiles to predict new members* [3].

[1] One gene can be regulated by the same transcription factor using more than one binding site. We consider each one of them and their corresponding relations with other regulatory elements as a promoter.

2.1 Dataset: Genes from *Escherichia Coli* and *Salmonella* Enterica Genomes

We built models based on microarray expression differed statistically between wild-type and *phoP E. coli* strains experiencing inducing conditions for the PhoP/PhoQ regulatory system and additional *S. enterica* promoters known to be regulated by the PhoP protein. This set of promoters constitutes the 70/30% training and test partitions (see [8] for a complete list of promoters as well as the codification for multiples promoters for a single gene). Expression values for *Salmonella* were inherited from its known orthologous genes in *E. coli*. Additional data for RNA polymerase and operons were obtained from RegulonDB database.

Representing Different Domain Knowledge: Modeling Promoter Features. We focused on four types of features [9] for describing our training set of promoters:

DNA Binding Site Motifs: (a) Fix-length Hierarchical Motifs: we modeled the PhoP box motifs by using position weight matrices[2] [10] (Fig. 1c) (see Consensus matrices in gps-tools.wustl.edu). Then, we used these preliminary models to describe promoters by using low thresholds corresponding to two standard deviations below the mean score obtained with the initial model [11]. We grouped the retrieved observations into subsets by using the possibilistic implementation of fuzzy C-means (PCM) [3] and rebuilt matrix models for each one (E-value < 10E-22), thus obtaining several more refined models, and increasing the sensitivity to departures from the consensus. These multiple matrices constitute the prototypes of the feature:

$$M_i(x_1,..,x_K) = \prod_{k=1}^{K} M(x_k) \tag{1}$$

where $M(x_k)$ is the marginal probability of each nucleotide x_k in the k'th position on motifs of length K, and i indexes a family of prototypes M_i [12]. The degree of matching between an observation and a feature is measured by its similarity with the prototype by using the informational content scores normalized as fuzzy values in the unit interval. The prototypes can be combined and arranged as a multiclassifier (see Bagging consensus in gps-tools.wustl.edu).

(b) Variable-length Motifs: we gathered sigma 70 promoters [13] from the RegulonDB database and built models of the RNA polymerase site using a neuro-fuzzy method (see Promoter search (CPR-MOSS) in gps-tools.wustl.edu), 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. The time delay neural network constitutes the feature prototype [5] and the scores were also normalized.

Transcription Factor Binding Site Orientation: categorical data. We classified PhoP boxes as either in direct or opposite orientation relative to the open reading frame (Fig. 1d), and the prototype is a simple Boolean function.

---

[2] A matrix of log-odd score $\log \frac{P(x_k)}{P_0(x_k)}$ where $P_0(x_k)$ is a background distribution.

RNA Polymerase Distances: data distributions modeled as fuzzy sets. We built histograms with the distance between RNA polymerase and transcription factor from information available in RegulonDB database [13]. We encoded these distributions by using fuzzy set representations [5] into close, medium, and remote sets (Fig. 1e). These fuzzy sets constitute the prototypes of the feature, and can be viewed as approximation of data distributions:

$$D_i(x) = \begin{cases} 0 & if\ x < a_0\ or\ x > x_2 \\ (x - a_0)/(a_1 - a_0) & if\ x < a_1 \\ (a_2 - x)/(a_2 - a_1) & if\ x > a_2 \\ 1 & otherwise \end{cases} \quad (2)$$

where x is any distance between the transcription start site of an RNA polymerase binding site and the center of a transcription factor binding site, and i indexes a family of distances D_{close}, D_{medium} and D_{remote}. Initial partitions are learned from the projection of the histograms onto the variable domains by simple regression and minimum squared methods [14]. The degree of matching between an observation and a prototype is calculated by specializing a value in a triangular fuzzy membership functions [15].

Microarray Expression Data: collection of fuzzy sets encoded as a fuzzy centroid. We clustered PhoP-regulated gene expression levels (Fig. 1f) by using PCM and built models for each cluster by calculating its centroid. These models represent the prototypes, where the values of the expression feature for each promoter in E. coli is calculated by its similarity to the centroids $\overline{V_i}$ as a vector of fuzzy sets:

$$E_i(x) = \left[1 + \left(\left\| x - \overline{V_i} \right\|_A^2 \Big/ w_i \right)^{1/m-1} \right]^{-1} \quad (3)$$

where $x = \{x_1, ..., x_k\}$ corresponds to the expression of a gene in k microarray experiments; w_i is the "bandwidth" of the fuzzy set E_i; m is the degree of fuzzification which is initialized as 2; the type of norm, determinated by A, is Pearson correlation coefficient; and i indexes a family of prototypes E_i.

Composite Features. We combine several features with dependencies between each other into more informative models by using AND-connected fuzzy predicates:

$$C(F_i, F_j) = F_i^{\ i}\ AND\ F_j = F_i \cap F_j \quad (4)$$

where F_i and F_j are previously defined features. Fuzzy logic-based operations, such as *T-norm/T-conorm*, 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 [3, 15]. In this work we used the MINIMUN and MAXIMUM as *T-norm* and *T-conorm*, respectively. For example, the RNA polymerase motif, learned by using a neural network method, it *sigma class*, identified by

using an intelligent parser that differentiates *class* I from *class* II promoters, and the distance distributions ($D_{close}, D_{medium}, D_{remote}$) between RNA polymerase and transcription factor binding sites, learned by using fuzzy set representations [5], are normalized and combined into a single fuzzy vector (e.g., $P_i(x) = R_j \; AND \; D_k \; AND \; T_l$).

2.2 Fusing Distinct Domain Knowledge: Dynamic Learning Profiles

Initializing Profiles. Our method independently clusters each type of feature to build initial level-1 profiles (Fig. 1g) based on the PCM clustering method and a validity index [3] to estimate the number of clusters, as an unsupervised discretization of the features [5, 16]. For example, we obtained five level-1 profiles for the "submotifs" feature $(M_0^1,...,M_4^1)$ (The superscript denotes the level, 1 in this case. The subscript denotes the specific profile, with subscript 0 corresponding to profiles containing promoters that do not have the corresponding type of feature).

Grouping Profiles. We group profiles by navigating in a lattice corresponding to the feature searching space [1, 2] and systematically creating compound higher level profiles (i.e., offspring profiles) based on combining parental profiles, by taking the fuzzy intersection (Fig. 1h). For example: level-1: (E_1^1 , M_2^1 and P_3^1) $\mapsto$ level-2: ($E_1^2 M_2^2$, $M_2^2 p_3^2$ and $E_1^2 P_3^2$) $\mapsto$ level-3: ($E_1^3 M_2^3 P_3^3$), where level-3-profiles are obtained from intersection of the promoter members of level-2- profiles (e.g., $E_1^2 M_2^2$, $M_2^2 P_3^2$ and $E_1^2 P_3^2$) and not between those belonging to the initial profiles (E_1^1 , M_2^1 and P_3^1). This is because our approach dynamically re-discretizes the original features at each level and allows re-assignations of observations between sibling profiles. In this hierarchical process, each level of the lattice increases the number of features shared by a profile. After searching through the whole lattice space, the most specific profiles (i.e., the most specific hypothesis [17]) are found. As a result of this strategy, one promoter observation can contribute to more than one profile in the same or a different level of the lattice, with different degrees of membership. This differentiates our approach from a hierarchical clustering process where, once an observation is placed in a cluster, it can only be re-assigned into offspring clusters. In contrast, our approach is similar to optimization clustering methods [18] in that it allows transfers among sibling clusters in the same level.

Prototyping Profiles. We learn profiles by using the PCM clustering method [3, 4], where promoters can belong to more than one cluster with different degrees of membership, and are not forced to belong to any particular cluster. This consists of individually evaluating the membership of the promoters to each feature, at each level in the lattice, and combining the results (equation (4)).

Selecting Profiles. Profile search and evaluation is carried out as a multi-objective optimization problem [5, 6], between the extent of the profile and the quality of matching among its members and the corresponding features. The extent of a profile is calculated by using the hypergeometric distribution that gives the chance probability (i.e.,

probability of intersection (PI)) of observing at least *p* candidates from a set V_i of size *h* within another set V_j of size *n*, in a universe of *g* candidates:

$$PI(V_{i,j}) = 1 - \sum_{q=0}^{p} \binom{h}{q}\binom{q-h}{n-q} \Big/ \binom{g}{h} \tag{5}$$

where V_i is an alpha-cut of the offspring profile and V_j is an alpha-cut of the union of its parents. The PI [19] is a more informative measure than the number of promoters belonging to the profile, such as the Jaccard coefficient, in being an adaptive measure that is sensitive to small sets of examples, while retaining specificity with large datasets.

The quality of matching between promoters and features of a profile (i.e., similarity of intersection (SI)) is calculated using the following equation:

$$SI(V_i) = \frac{1}{f}\left(1 - \frac{\sum_{k \in U_\alpha} \mu_{ik}}{n_\alpha}\right) \quad U_\alpha = \{\mu_{ik} : \mu_{ik} > \alpha\} \tag{6}$$

where n_α is the number of elements in an arbitrary *alpha*-cut U_α.

The tradeoff between the opposing objectives (i.e., PI and SI) is estimated by selecting a set of solutions that are non-dominated, in the sense that there is no other solution that is superior to them in all objectives (i.e., Pareto optimal frontier) [5, 6]. The dominance relationship in a minimization problem is defined by:

$$a \prec b \; iif \; \forall i \; O_i(a) \le O_i(b) \; \exists j O_j(a) < O_j(b) \tag{7}$$

where the O_i and O_j are either PI or SI. The method applies the non-dominance relationship only to profiles in the local neighborhood or niche [6] by using the hypergeometric metric (equation **(5)**) between profiles and selecting an arbitrary threshold; in this way combining both multi-objective and multimodal optimization concepts [6].

2.3 Using the Profiles to Predict New Members

The method uses a fuzzy k-nearest prototype classifier (FKN) to predict new profile members using an unsupervised classification method [3] applied to regulatory regions of genomes described by regulatory features. First, we determine the lower-boundary similarity threshold for each non-dominated profile. This threshold is calculated based on the ability of each profile to retrieve its own promoters and to discard promoters from other profiles [20]. Second, we calculate the membership of a query observation x_q to a set of *k* profiles previously identified and apply a fuzzy OR logic operation:

$$FKN(x_q, V_1, ..., V_k) = i, \; i \in \{1, .., k\} \tag{8}$$

where $\mu_{i,q} = OP_{OR}\{\mu_{1,q},...,\mu_{k,q}\}$, μ is calculated based on (equation **(4)**) in which w_i (equation **(3)**) is initialized as:

$$w_i = \frac{r_1\ PI(V_i\ ) + r_2\ (f\ /\ t')SI(V_i\ )}{r_1\ + r_2} \tag{9}$$

with t' being the number of distinct features observed in x_q and V_i, and f is the number of features in common between x_q and V_i, which are combined to obtain a measure of belief or rule weight [2]; r_1 and r_2 are user-dependent parameters, initialized as 1 if no preference exists between both objectives; and OP_{OR} is the Maximum fuzzy operator [3, 4].

Possibilistic Fuzzy C-means Clustering Method [3, 4]: (i) Initialize $L_0 = \{\overline{V_1},..,\overline{V_c}\}$; (ii) while ($s<S$ and $\|L_s - L_{s-1}\| > \varepsilon$), where S is the maximum number of iterations; (iii) calculate the membership of U_s in L_{s-1} as in (equation (3)); (iv) update L_{s-1} to L_s with U_s and $\overline{V_i} = \sum_{k=1}^{n} \mu_{ik} x_k \,/\, \sum_{k=1}^{n} \mu_{ik}$; (v) iterate.

3 Results

We investigated the utility of our approach by exploring the regulatory targets of the PhoP protein in *E. coli* and *S. enterica*, which is at the top of a highly connected network that controls transcription of dozens of genes mediating virulence and the adaptation to low Mg^{2+} environments [7]. We demonstrated that our method makes predictions at three levels [8]: (i) it makes an appropriate use of the regulatory features to perform genome-wide predictions; (ii) it detects new candidate promoters for a regulatory protein; and (iii) it indicates possible mechanisms by which genes previously known to be controlled by a regulator are expressed.

Performance of the Features. We illustrated the performance of the encoded features by analyzing three of them.. We evaluated the ability of the resulting models to describe PhoP-regulated promoters, we extended the dataset by including 772 promoters (RegulonDB V3.1 database [13]) 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. We considered the compiled list of PhoP regulated genes as true positive examples and the binding sites of other transcriptional regulators as true negative examples to evaluate the performance of the submotif feature. Each matrix threshold has been optimized for classification purposes by using the overall performance measurement [20] based on the extended dataset. We found that the PhoP-binding site model increases its sensitivity from 66% to 90% when submotifs are used instead of a single consensus, while its specificity went from 98% to 97%. This allowed the recovery of promoters, such as that corresponding to the E. coli hdeD gene or the

Salmonella pmrD, that had not been detected by the single consensus position weight matrix model [10] despite being footprinted by the PhoP protein [7, 8].

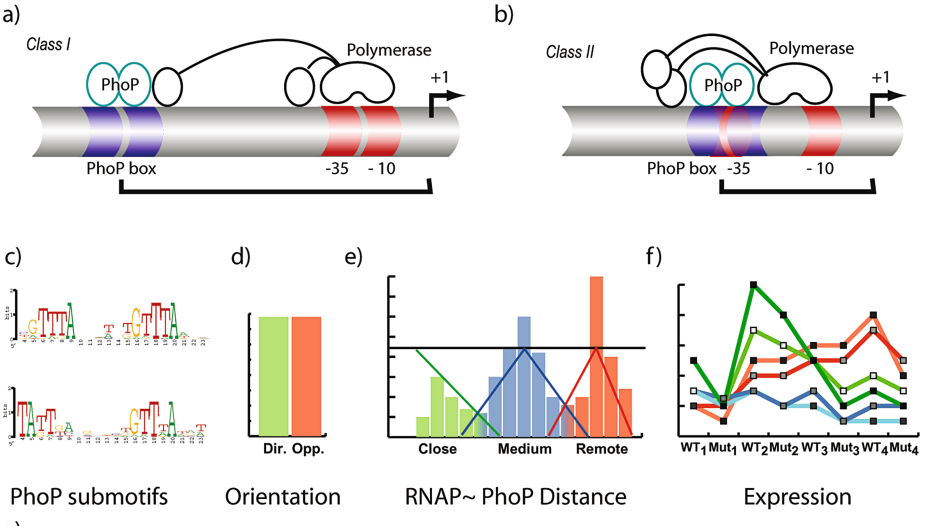
The RNA polymerase site feature was evaluated using 721 RNA polymerase sites from RegulonDB as positive examples and 7210 random sequences as negative examples. We obtained an 82% sensitivity and 95% specificity for detecting RNA polymerase sites. These values provides an overall performance measurement [20] of 92% corresponding to a false discovery rate <0.001. 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 [21], and obtained 74% sensitivity and 95% specificity.

Regarding the expression feature, results suggest that the sensitivity of the "expression" feature can be increased from 45% to the 76% by using the model-based approach in a complementary fashion to the original statistical approach, by just admitting a limited decrease in specificity. This approach allowed us to recover additional genes (e.g., the *hemL* and the *proP* promoters of *E. coli*) that have expression levels too weak to be initially detected using strict statistical filters (35). (see gps-tools.wustl.edu for predicted features in *E. coli* and *Salmonella*).

Performance of the Profiles. We recovered several profiles, some of which were experimentally validated [8]. In addition, here we measured the promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution to evaluate the behavior of the profiles. For example, one of the profiles corresponds to the canonical PhoP-regulated promoters (PI=$1.57E^{-4}$, SI=0.002), and encompasses promoters (e.g., those of the *phoP, mgtA, rstA, slyB, yobG and yrbL* genes) that share the class II RNA polymerase sites situated close to the PhoP boxes, high expression patterns, and typically PhoP box submotif. This profile includes not only the prototypical *phoP* and *mgtA* promoters [22], but also other promoters, which was not known to be under PhoP control. The promoters sharing this profile produced the earlier rise times and the higher levels of transcription (Fig. 1i). Particularly, *phoP* itself, perhaps affected by its autoregulation, generates the top levels of expression during time. Another profile (PI=$3.53E^{-4}$, SI=0.032) includes promoters (e.g., those of the *mgtC*, *mig-14*, *pagC*, *pagK*, and *virK* genes of *Salmonella*) that share a PhoP boxes in the opposite orientation of the canonical PhoP-regulated promoters, as well as class I RNA polymerase sites situated at medium distances from the PhoP boxes, all of the features dynamically adapted for the current set of genes. This profile, exhibit the latest genes with the lowest levels of expression (Fig. 1i). Finally, another profile (PI=0.033, SI=0.044), which is slightly different from the former includes promoters (e.g., those of the *ompT* gene of *E. coli* and the *pipD, ugtL* and *ybjX* genes of *Salmonella*) that although exhibit a PhoP binding site in the opposite orientation, preserves the RNA polymerase of the canonical PhoP regulated promoters and present an intermediate kinetic behavior. The detailed analysis of the gene behavior would not be possible to be obtained by just inspecting each features independently, or by considering simple consensus of these features.

Predictions. To evaluate the ability of the method to retrieve PhoP-regulated promoters, we extended the test set by including 487 promoters from the RegulonDB database [13] that are regulated by transcription factors other than PhoP, by selecting the promoter region corresponding to the respective transcription factor binding site ±10

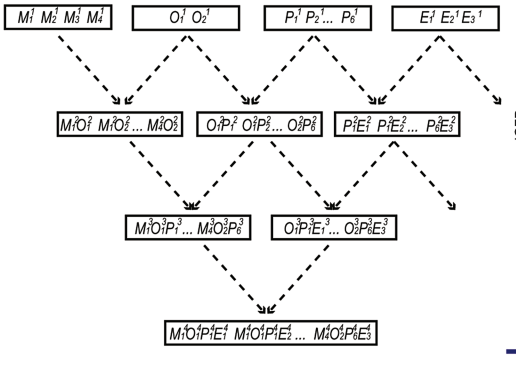
bp, its corresponding RNA polymerase site ±10 bp and expression levels from our own experiments. The method had a false positive rate of 5.3% and a 93.92% of overall performance measurement [20] as a particular correlation coefficient implementation, with a 94 and 92% specificity = TN/(TN + FP) and sensitivity = TP/(TP + FN) respectively, where P is positive examples, N is negative examples, T is true and F is false.



g)

	M_1^1	M_2^1	M_3^1	M_4^1	O_1^1	O_2^1	P_1^1	P_2^1	P_3^1	P_4^1	P_5^1	P_6^1	E_1^1	E_2^1	E_3^1
ompT	0.25	0.39	0.10	0.15	1	0	0.42	0.39	0.10	0.15	0.10	0.15	0.92	0.00	0.00
rstA	0	0.66	0.15	0.00	0	1	0	0.66	0.15	0.70	0.15	0.00	0.84	0.00	0.00
hilA	0	0.40	0.15	0.62	0	1	0	0.12	0.78	0.13	0.15	0.90	0.46	0.46	0.08
ybjX	0	0.37	0.76	0.28	1	0	0	0.18	0.30	0.41	0.30	0.28	0.25	0.25	0.25

h)



i)

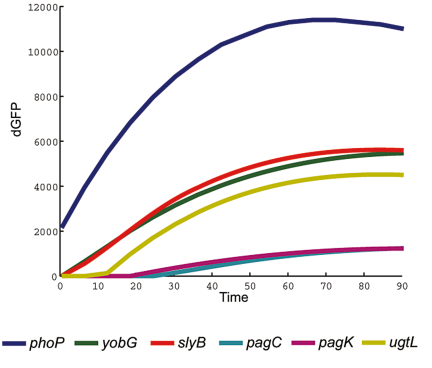


Fig. 1. Different *cis*-features participating in the regulation scheme. **a-b)**Two PhoP proteins had binded to a DNA strain and recruited RNA polymerase. Class I activators bind to upstream locations. By contrast Class II activators bind to sites that overlap the target promoter -35 region. A PhoP box might be located in the same strain as the polymerase (a) or in the opposite direction (b). **c)** PhoP binding box modeled as position weight matrices shown as logos: The characters representing the sequence are stacked on top of each other for each position in the aligned sequences. The height of each letter is made proportional to its frequency, and the letters are sorted so the most common one is on top. We used these matrices to prototype DNA sequences, where its elements are the weights used to score a test sequence to measure how close that sequence word matches the pattern described by the matrix. **d)** Orientation: The PhoP box can be located either in the direct or opposite direction, thus it is modeled as a categorical set. **e)** Distance between PhoP box and transcription start site (+1): The distance is usually between 20 and 100 bases. This graph represents the distance histogram and the distribution approximated by triangular functions. **f)** Microarray expression data: The gene expression difference between wild-type and *phoP E. coli* strains experiencing PhoP/PhoQ inducing condition were modeled as a vector of fuzzy sets. **g)** Database representation: The regulatory features model heterogeneous domains corresponding to different *cis*- and expression descriptions of the PhoP regulated promoters by using fuzzy membership values. Here we exemplify data from DNA sequences where the cells represent the degree of matching between a promoter value and the model of a feature (red: high; green: low). This framework facilitates the application of machine learning methods to extract profiles, which are sets of promoters sharing a common set of features. **h)** Part of the complete lattice: The method navigates through the feature-space lattice generating and evaluating profiles. Level-1 profiles of each feature are combined to identify level-2 profiles, and similarly, level-2 profiles are combined to create level-3 profiles; the observations can migrate from parental to offspring clusters (i.e., hierarchical clustering), and among sibling clusters (i.e., optimization clustering). **i)** Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella* promoters including *phoP* (blue), *yobG* (green), *slyB* (red), *pagC* (cyan), *pagK* (magenta) and *ugtL* (yellow). 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). The genes are evaluated by their rise time and levels of transcriptions.

4 Discussion

We showed that our method can make precise mechanistic predictions even with incomplete input dataset and high levels of uncertainty; making use of several characteristics that contribute to its power: (i) it considers gene expression as one feature among many (unsupervised approach), thereby allowing classification of promoters even in its absence; (ii) it performs a local feature selection for each profile because not every feature is relevant for all profiles [16], and, a priori, it is not know which feature is biologically meaningful for a given promoter; (iii) it finds all optimal solutions among multiple criteria (Pareto optimality) [6], which avoids the biases that might result from using any specific weighing scheme; (iv) it has a multimodal nature that allows alternative descriptions of a system by providing several adequate solutions [5]; (v) it allows promoters to be members of more than one profile by using

fuzzy clustering thus explicitly treating the profiles as hypotheses, which are tested and refined during the analysis; and (vi) it is particularly useful for knowledge discovery in environments with reduced datasets and high levels of uncertainty. The predictions made by our method were experimentally validated [8] to establish that the PhoP protein uses multiple mechanisms to control gene transcription, and is a central element in a highly connected network. These profiles can be used to effectively explain the different kinetic behavior of co-regulated genes.

Acknowledgments

This work was partly supported by the Spanish Ministry of Science and Technology under Project BIO2004-0270-E, and I.Z. is also supported by and by Howard Hughes Medical Institute.

References

1. Cook, D.J., et al., Structural mining of molecular biology data. IEEE Eng Med Biol Mag, 2001. 20(4): p. 67-74.
2. Cooper, G.F. and E. Herskovits, A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 1992. 9(4): p. 309-347.
3. Bezdek, J.C., Pattern Analysis, in Handbook of Fuzzy Computation, W. Pedrycz, P.P. Bonissone, and E.H. Ruspini, Editors. 1998, Institute of Physics: Bristol. p. F6.1.1-F6.6.20.
4. Gasch, A.P. and M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol, 2002. 3(11): p. RESEARCH0059.
5. Ruspini, E.H. 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, Editors. 2002, World Scientific: New Jersey. p. 454-474.
6. Deb, K., Multi-objective optimization using evolutionary algorithms. 1st ed. Wiley-Interscience series in systems and optimization. 2001, Chichester ; New York: John Wiley & Sons. xix, 497.
7. Groisman, E.A., The pleiotropic two-component regulatory system PhoP-PhoQ. J Bacteriol, 2001. 183(6): p. 1835-42.
8. 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.
9. Beer, M.A. and S. Tavazoie, Predicting gene expression from sequence. Cell, 2004. 117(2): p. 185-98.
10. Stormo, G.D., DNA binding sites: representation and discovery. Bioinformatics, 2000. 16(1): p. 16-23.
11. Robison, K., 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, 1998. 284(2): p. 241-54.
12. Barash, Y., Elidan, G., Friedman, N., Kaplan, T. Modeling Dependencies in Protein-DNA Binding Sites. in RECOMB'03. 2003.
13. 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.

14. Sugeno, M. and T. Yasukama, A Fuzzy-logic-based Approach to Qualitative Modeling. IEEE Transactions on Fuzzy Systems, 1993. 1(1): p. 7-31.
15. Klir, G.J. and T.A. Folger, Fuzzy sets, uncertainty, and information. 1988, London: Prentice Hall International. xi,355.
16. Kohavi, R. and G.H. John, Wrappers for feature subset selection. Artificial Intelligence, 1997. 97(1-2): p. 273-324.
17. Mitchell, T.M., Machine learning. 1997, New York: McGraw-Hill. xvii, 414.
18. Falkenauer, E., Genetic Algorithms and Grouping Problems. 1998, New York: John Wiley & Sons.
19. Tavazoie, S., et al., Systematic determination of genetic network architecture. Nat Genet, 1999. 22(3): p. 281-5.
20. Benitez-Bellon, E., 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, 2002. 3(3): p. RESEARCH0013.
21. 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.
22. Minagawa, S., et al., Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli. J Bacteriol, 2003. 185(13): p. 3696-702.