Targeting Differentially Co-regulated Genes by Multiobjective and Multimodal Optimization

Oscar Harari[1], Cristina Rubio-Escudero[1], and Igor 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@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 in and between genetic networks. 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, however it is unknown which of these features increase or reduce expression of genes. In principle, this expression can be measured by microarray experiments, though they incorporate systematic errors, and moreover produce a limited classification (e.g. up/down regulated genes). In this work, we present an unsupervised machine learning method to tackle the complexities governing gene expression, which considers gene expression data as one feature among many. It analyzes features concurrently, recognizes dynamic relations and generates profiles, which are groups of promoters sharing common features. The method makes use of multiobjective techniques to evaluate the performance of profiles, and has a multimodal approach to produce alternative descriptions of same expression target. 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 profiles that were experimentally validated, suggesting correlations between promoter regulatory features and gene expression kinetics measured by green fluorescent protein (GFP) assays.

1 Introduction

Genetic and genomic approaches have been successfully used to assign genes to distinct regulatory networks. However, little is known about the differential expression of genes within a regulon. At its simplest, genes within a regulon are controlled by a common transcriptional regulator in response to the same inducing signal. Moreover it is suggested that subtle differences in the shared *cis*-acting regulatory elements are probably significant in the genes expression. However, it is not known which of these features, independently or collectively, can set expression patterns apart. Indeed, 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.

The material required for analyzing the promoter features governing bacterial gene expression is widely available. It consists of genome sequences, transcription data, and biological databases containing examples of preciously explored cases. In principle, genes could be differentiated by incorporating into the analysis quantitative and kinetic measurements of gene expression [1] and/or considering the participation of other transcription factors [2-4]. However, there are constraints in such analyses due to systematic errors in microarray experiments, the extra work required to obtain kinetic data and the missing information about additional signals impacting on gene expression. These constraints hitherto allow a relatively crude classification of gene expression patterns into a limited number of classes (e.g., up- and down-regulated genes [5, 6]), thus concealing distinctions among expression features, such as those that characterize the temporal order of genes or their levels of intensity

Here we describe an unsupervised machine learning method that discriminates among co-regulated promoters by simultaneously considering both cis-acting regulatory features and gene expression. By virtue of being an unsupervised method, it is neither constrained by a dependent variable [2, 7], such as expression data, which would restrict the classification to the dual expression classes reported by microarray experiments; nor it requires pre-existing kinetic data. Our method treats each of the promoter features with equal weight, because it is not known beforehand which features are important. Thus, it explores all of the possible aggregations of features; and applies multiobjective and multimodal techniques [8, 9] to identify alternative optimal solutions that describe target sets of genes from different perspectives.

We applied our methodology to the investigation of genes regulated by the PhoP protein of *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. We recovered several profiles that were experimentally validated [10] to establish that PhoP uses different configurations of promoter to regulate genes. We finally correlated these groups with more accurate independent experiments that measure gene expression over time by using GFP assays.

2 Methods

The purpose of this method is to identify all of the possible substructures, here termed profiles (i.e., groups of promoters sharing a common set of features), that characterize sets of genes. These common attributes can ultimately clarify the key *cis*-features that produce distinct kinetic patterns, shedding light in the transcriptional mechanisms that the cell employs to differentially regulate genes belonging to a regulon.

The identification of the promoter features that determine the distinct expression behavior of co-regulated genes is a challenging task because (i) the difficulty in ascertaining the role of the differences in the shared *cis*-acting regulatory elements of co-regulated promoters; (ii) detailed kinetic data that would help the classification of expression patterns is not always available, or it is available for a limited subset of genes; and (iii) the limited extent of genes regulated by a transcriptional factor. To circumvent these constrains, our method explores all of the possible *cis*-feature aggregations, looking for those that better characterize different subset of genes; uses an unsupervised approach, where pre-existing classes are not required; and allows a fuzzy incorporation of promoters to refined hypothesis which enables a same instance to support more than one hypothesis.

Our method represents, learns and infers from structural data by following four main phases: (1) *Database conformation*; (2) *Profile learning*; (3) *Profile evaluation* (4) *Evaluation of external classes.*

2.1 Database Conformation

Biological Model. Multiple independent and interrelated attributes of promoters, naturally encoded into diverse data types, should be considered to perform an integrated analysis of promoter regulatory features. We focus on four types of features for describing our set of co-regulated promoters [2, 3, 10, 11]: *"submotifs"*, fix-length DNA motifs from transcriptional regulator binding sites, represented by position weight matrices *[12]* (Fig 1.a). 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; *"orientation"*, which characterizes the binding boxes as either in direct or opposite orientation relative to the open reading frame; *"RNA pol sites"*, represents the RNA polymerase: their location in the chromosome is studied as a distribution and encoded into fuzzy sets (*close*, *medium,* and *remote*). It also models the class of sigma 70 promoter [13]: *class I* promoters bind to upstream locations (Fig 1.b). By contrast *class II* promoters bind to sites that overlap the promoter region. [14](Fig 1.c); and *"expression"*, which considers gene expression from multiple experiments represented as vector patterns. See [15] for a detail description of the learning process of these features.

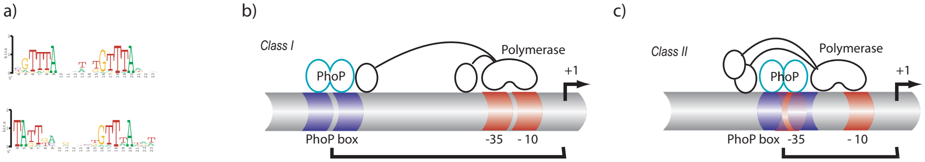

Fig. 1. Different *cis*-features participating in the regulation scheme. a) PhoP binding box modeled as position weight matrices shown as logos: The characters representing the sequence are stacked on top of each other for each position in the aligned sequences. The height of each letter is made proportional to its frequency.. **b-c)**Two transcription factors had binded to a DNA strain and recruited RNA polymerase (*Class I/II* respectively). A PhoP box might be located in the same strain as the polymerase (b) or in the opposite direction (c).

Representation Model. We use fuzzy sets as a common framework to represent the domain independent features. We cluster promoters considering each feature independently by using fuzzy C-means clustering (FCM) method and a validity index [16] to estimate the number of clusters, as an unsupervised discretization of the features [9, 17]. For example, we obtained three clusters for the "*expression*" feature (E_1^1: strong evidence of upregulation; E_2^1: mild evidence of upregulation; and E_3^1: evidence of downregulation). As a result of this process, we obtain initial prototypes of profiles, and are able to account for the variability of the data by treating these features as fuzzy (i.e., not precisely defined) instead of categorical entities. Thus, our database is conformed by the membership of each promoter to each of the cluster of every feature.

2.2 Profile Learning

Our method uses a conceptual clustering approach to incrementally find significant characterization of promoters (profiles) while exploring the features space [18-20]. Initial profiles are aggregated to create compound higher level profiles (i.e. offspring profiles) by using the fuzzy intersection[1]. In a hierarchical process, the number of features shared by a profile is increased, resulting in a lattice of profiles. Level *n* profiles are built by aggregating level *n-1* profiles (Fig. 2). This is because the method re-discretizes the original features:

$$V_{fj} = \sum\nolimits_{k=1}^{n} \mu_{jk} x_{fk} \, / \sum\nolimits_{k=1}^{n} \mu_{jk} \tag{1}$$

where μ_{jk} is the membership of the promoter *k* to cluster *j*; and x_{kf} is the original raw data for feature *f*. This allows to the prototypes of the profiles to be dynamically adapted to the promoters recovered by it. In account of these new prototypes, the membership of the entire database of promoters is re-evaluated:

$$\mu_{fij}(x) = \left[1 + \left(\left\| x_{if} - V_{fj} \right\|_f^2 \Big/ w_{fj} \right)^{1/m-1} \right]^{-1} \tag{2}$$

where w_{fj} is the “bandwith” of the fuzzy set V_{fj} [16]. This allows re-assignations of observations between sibling profiles [21], which is especially useful to gain support to hypothesis in problems, such as ours, that have a reduced number of samples.

2.3 Profile Evaluation

We applied multiobjective and multimodal techniques to evaluate the performance of the profiles [8, 9, 22], considering the conflicting criteria of the extent of the profile, and the quality of matching among its members and the corresponding features.

The extent of the profile is calculated by using the hypergeometric distribution that gives the probability of intersection (PI) of an offspring profile and its parents:

$$PI(V_{i,j}) = 1 - \sum\nolimits_{q=0}^{p} \binom{h}{q} \binom{q-h}{n-q} \Big/ \binom{g}{h} \tag{3}$$

where V_i is an alpha-cut of the offspring profile, of size h; V_j is an alpha-cut of the union of its parents, of size *n*; *p* is the number of promoters of the intersection; and *g* is the number of candidates. The PI is an adaptive measure that is sensitive to small sets of examples, while it retains specificity with large datasets [23].

[1] Fuzzy logic-based operations, such as T-norm/T-conorm, include operators which are used as basic logic operators, such as AND or OR, [16]. In this work we used the MINIMUN and MAXIMUM as T-norm and T-conorm, respectively.

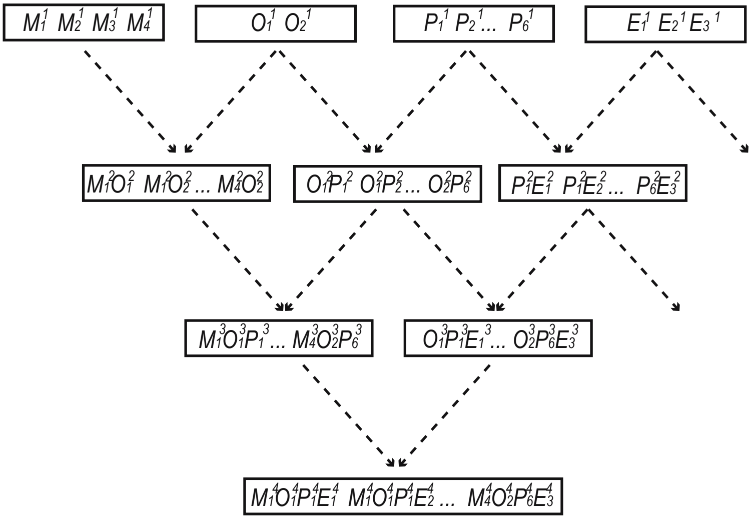
$M_1^1\ M_2^1\ M_3^1\ M_4^1$
$O_1^1\ O_2^1$
$P_1^1\ P_2^1 \ldots\ P_6^1$
$E_1^1\ E_2^1\ E_3^1$
$M_1^2O_1^2\ M_1^2O_2^2 \ldots\ M_4^2O_2^2$
$O_1^2P_1^2\ O_1^2P_2^2 \ldots\ O_2^2P_6^2$
$P_1^2E_1^2\ P_1^2E_2^2 \ldots\ P_6^2E_3^2$
$M_1^3O_1^3P_1^3 \ldots\ M_4^3O_2^3P_6^3$
$O_1^3P_1^3E_1^3 \ldots\ O_2^3P_6^3E_3^3$
$M_1^4O_1^4P_1^4E_1^4\ M_1^4O_1^4P_1^4E_2^4 \ldots\ M_4^4O_2^4P_6^4E_3^4$

Fig. 2. Schematic view of the method. The method navigates through the feature-space lattice generating and evaluating profiles. Hierarchically, profiles of one level are combined to generate the profiles of the following one. Observations can migrate from parental to offspring clusters (i.e., hierarchical clustering), and among sibling clusters (i.e., optimization clustering).

The quality of matching between promoters and features of a profile (i.e., similarity of intersection (SI)) is calculated using the equation (4), where U_α is an alpha-cut of the profile *i* and n_α is its number of elements.

$$SI(V_i) = \left(1 - \sum\nolimits_{k \in U_\alpha} \mu_{ik} \big/ n_\alpha\right) \big/ f \quad U_\alpha = \{\mu_{ik} : \mu_{ik} > \alpha\} \tag{4}$$

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) [8, 9]. The dominance relationship in a minimization problem is defined by:

$$a \prec b\, iif\ \forall i\ O_i\,(a) \le O_i\,(b)\ \exists j O_j\,(a) < O_j\,(b) \tag{5}$$

where the O_i and O_j are either PI or SI. This approach is less biased than weighting the objectives because it identifies the profiles lying in the Pareto optimal frontier [8, 9], which is the collection of local multiobjective optima in the sense that its members are not worse than (i.e. dominated by) the other profiles in any of the objectives being considered.

Another objective indirectly considered is the profile diversity, which consists of maintaining a distributed set of solutions in the Pareto frontier, and thus, identifying clusters that describe objects from alternative regulatory scenarios. Therefore, our approach applies the non-dominance relationship locally, that is, it identifies all non-dominated optimal profiles that have no better solution in the local neighborhood

[8, 9]. We evaluate niches by applying equation (3) to every pair of solution and establish a small threshold value as boundaries of neighborhoods.

2.4 Evaluation of External Classes

This proposed unsupervised method, in contrast to supervised approaches, does not need the specification of output classes. Consequently, the discovered profiles can be used for independently explain external classes as a process often termed labeling [7]

Instead of choosing a single profile to characterize an external target set, the method selects all of the profiles that are correlated enough to the query set. To find its classes of equivalence it applies equation (3) to the target set and the entire collection of profiles previously produced. In this way, the method can recover all of the alternative profiles that match the external class, including the most specific and general solutions.

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 [24]. As little is known about the mechanism by which *cis*-regulatory features govern gene expression, we searched through the space of all potential hypotheses; evaluated them, by considering both their extent and similarity of the recovered promoters; and obtained alternative descriptions for target set of genes. Moreover, to tackle constrains of the crude classification obtained by microarray experiments -which would not have allowed finding detail topologies of promoters- in an unsupervised approach we modeled gene expression as one feature among many.

We demonstrated that our method makes predictions at two levels: it detects new candidate promoter for a regulatory protein; and it indicates alternative possible configurations by which genes previously identified as controlled by a regulator are differentially expressed. We recovered several optimally evaluated profiles, thus, revealing distinct putative profiles that can describe the PhoP regulation process:

One profile ($O_1^4E_2^4M_3^4P_2^4$: PI=1.57E-4, SI=0.002) corresponds to canonical PhoP-regulted promoters (e.g., those of the *phoP, mgtA, rstA, slyB, yobG, ybjX, ompX, PagP, pdgL, pipD, and pmrD* genes) characterized by a class II RNA polymerase sites situated close to the PhoP boxes, high expression patterns and a typical PhoP box submotif in a direct orientation. Notably, this profile recovers promoters previously not known to be directly regulated by PhoP. The method was also able to describe this target by using other profiles, being the most general ones composed of only two features (Fig 3.a)

Another profile ($E_3^4M_1^4O_2^4P_1^4$: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 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. As expected, the method was able to identify this target set by more general hypothesis that aggregates again only two features (Fig 3.b).

Finally, another profile ($E_2^3 O_2^3 P_4^3$: PI=6.48E-06, SI=0.070), 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*) is defined by a PhoP binding site in the opposite orientation, the RNA polymerase of the canonical PhoP regulated promoters and a mild evidence of upregulation. The method was also able to characterize this target by a specific Phop box submotif and the same type of RNA polymerase.

The above profiles differ in the number of features because our method uses a multivariate environment, where feature selection is locally performed for each profile, as not every feature is relevant for all profiles. The predictions made by our method were experimentally validated [10] to establish that the PhoP protein uses multiple mechanisms to control gene transcription.

Furthermore, as these profiles can be used to effectively explain the different kinetic behavior of co-regulated genes, we measured the promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution (Fig. 4); and obtained independent target sets by clustering them by using FCM. We found that the cluster that recovers those promoters that expressed earlier rise times and higher levels of transcription (e.g. mgtA, ompX, pagP, phoP, pmrD, rstA, slyB, ybjX, yobG) is correlated to profile $O_1^4 E_2^4 M_3^4 P_2^4$ (*p*-value < 0.03) (Fig. 3.a). Another target set includes those promoters that expressed the latest rise time and lowest levels of transcription (e.g. mgtC, mig-14, pagC, pagK, pipD, ugtL, virK, pagD); and it is correlated to profile $E_3^4 M_1^4 O_2^4 P_1^4$ (*p*-value < 0.013) (Fig. 3.b). The cluster which contains the promoters that showed intermediate values (e.g., those of the *ompT* gene of *E. coli* and the *pipD, ugtL* and *ybjX* genes of *Salmonella*) is correlated to profile $E_2^3 O_2^3 P_4^3$ (*p*-value < 0.025)

This detailed analysis of the gene expression behavior would not be possible to be obtained by applying a supervised machine learning approach because of the lack of kinetic data for some promoters.

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 crude gene expression as one feature among many (unsupervised approach), thereby allowing classification of promoters even in its absence; (ii) it has a multimodal nature that allows alternative descriptions of a system by providing several adequate solutions [9] that characterize a target set of genes; (iii) 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 (iv) it is particularly useful for knowledge discovery in environments with reduced datasets and high levels of uncertainty.

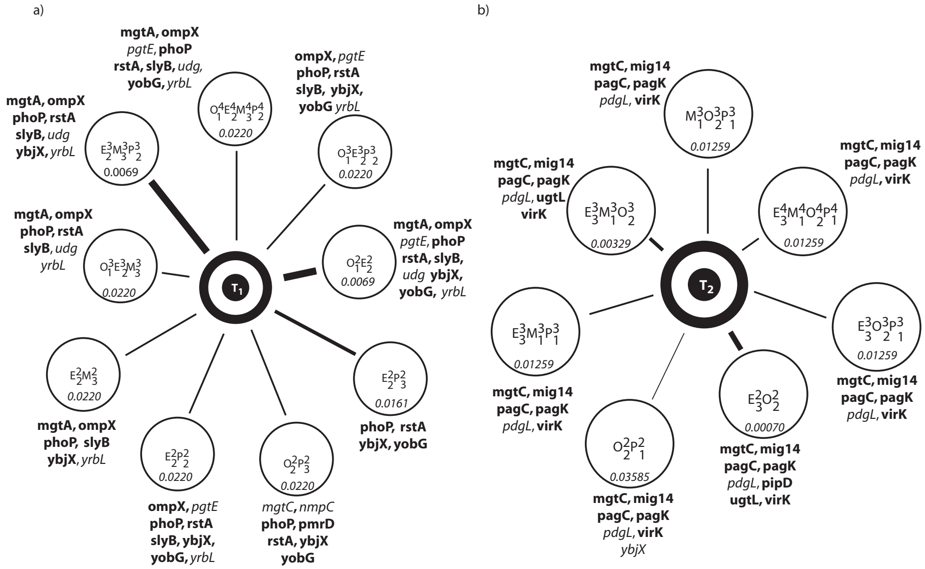

Fig. 3. Chart of Correlated Profiles. Targets are display at the center of each chart, surrounded by the profiles that hit them. Optimal profiles are situated closer to the targets. For each profile it is displayed the features that characterizes it, the promoters that recovers (boldface belonging to the target, and italic not belonging to it) and the correlation to the target set. *E* stands for *"Expression"*, *P* for *"RNA Pol. Sites"*, *O* for *"Orientation"* and *"M"* for *"Submotif"*; subscripts denote the cluster and superscripts the re-discretized level.

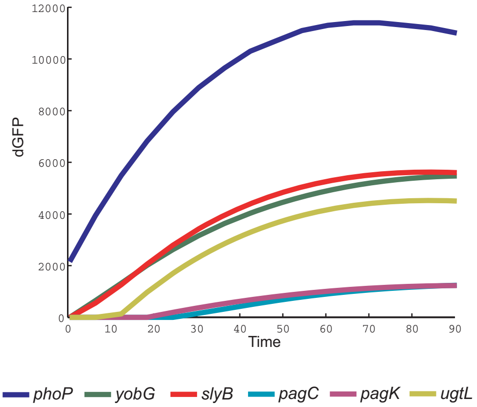

Fig. 4. Rise time and levels of transcription. Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella* promoters. The activity of each promoter is proportional to the number of GFP molecules produced per unit time per cell [$dG_i(t)/dt]/OD_i(t)$], where $G_i(t)$ is GFP fluorescence from wild-type *Salmonella* strain 14028s, and $OD_i(t)$ is the optical density. The activity signal was smoothed by a polynomial fit (sixth order). Details about genetic experiments can be found in http://www.pnas.org/ and about GFP assays available under requirements to the authors.

The predictions made by our method were experimentally validated [10] 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. O.H. acknowledges the doctoral MAEC- AECI fellowship.

References

1. Ronen, M., et al., *Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics.* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10555-60.
2. Beer, M.A. and S. Tavazoie, *Predicting gene expression from sequence.* Cell, 2004. **117**(2): p. 185-98.
3. Bar-Joseph, Z., et al., *Computational discovery of gene modules and regulatory networks.* Nat Biotechnol, 2003. **21**(11): p. 1337-42.
4. Conlon, E.M., et al., *Integrating regulatory motif discovery and genome-wide expression analysis.* Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3339-44.
5. Oshima, T., et al., *Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12.* Mol Microbiol, 2002. **46**(1): p. 281-91.
6. Tucker, D.L., N. Tucker, and T. Conway, *Gene expression profiling of the pH response in Escherichia coli.* J Bacteriol, 2002. **184**(23): p. 6551-8.
7. Mitchell, T.M., *Machine learning*. 1997, New York: McGraw-Hill. xvii, 414.
8. 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.
9. 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.
10. 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.
11. Li, H., et al., *Identification of the binding sites of regulatory proteins in bacterial genomes.* Proc Natl Acad Sci U S A, 2002. **99**(18): p. 11772-7.
12. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
13. Romero Zaliz, R., 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., Editor. 2004, World Scientific: Singapore. p. 427-450.
14. 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.

15. Zwir, I., H. Huang, and E.A. Groisman, *Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation 10.1093/bioinformatics/bti672.* Bioinformatics, 2005. **21**(22): p. 4073-4083.
16. 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.
17. Kohavi, R. and G.H. John, *Wrappers for feature subset selection.* Artificial Intelligence, 1997. **97**(1-2): p. 273-324.
18. Cheeseman, P. and R.W. Oldford, *Selecting models from data : artificial intelligence and statistics IV.* 1994, New York: Springer-Verlag. x, 487.
19. Cook, D.J., et al., *Structural mining of molecular biology data.* IEEE Eng Med Biol Mag, 2001. **20**(4): p. 67-74.
20. 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.
21. Falkenauer, E., *Genetic Algorithms and Grouping Problems*. 1998, New York: John Wiley & Sons.
22. Rissanen, J., *Stochastic complexity in statistical inquiry*. World scientific series in computer science, 15. 1989, Singapore: World Scientific. 177.
23. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.
24. Groisman, E.A., *The pleiotropic two-component regulatory system PhoP-PhoQ.* J Bacteriol, 2001. **183**(6): p. 1835-42.