Fusing genetic knowledge by dynamic learning regulatory profiles: visualizing the strategy

Oscar Harari[a], Coral del Val[a], Igor Zwir[a,b1]
[a] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[b] Howard Hughes Medical Institute, Washington University School of Medicine, St. Louis, MO, USA

A critical challenge of the postgenomic era is to understand how genes are regulated in and between genetic networks. The fact that such co-regulated genes may be differentially expressed 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 from microarray experiments, which is treated as one feature among many. These features, derived from different domain sources, are analyzed concurrently by an unsupervised machine learning method. This method uncovers dynamic relations and generates profiles, which are groups of promoters sharing common features. We apply this method to explore the regulatory networks governed by the PhoP/PhoQ two-component system in the enteric bacteria *Escherichia coli* and *Salmonella enterica*.. Our analysis recognizes novel members of the PhoP regulon and the resulting profiles share underlying biological mechanisms that were experimentally validated. We provide visualizations for a better understanding of each phase of the learning process and results.

Keywords: gene regulation, *cis*-features, fuzzy sets, conceptual clustering, multi-objective optimization

1 Introduction

One of the biggest challenges in genomics is the elucidation of the design principles controlling gene networks [1]. However, knowing the connectivity of a given network is not sufficient to define the expression dynamics of a group of genes; it is also required to specify the strength of the connections in a network, which are determined by the *cis*-promoter features participating in the regulation [2]. 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 behaviours despite belonging to networks with identical wiring design.

We initially report a model-based approach to analyze genomes for promoter[2] features, which is specially designed to account for sequence variability, location and topology intrinsic to differential gene expression. We use these features to generate genome-wide descriptions developing a predictive transcriptional database. This information constitutes the input for a machine learning method [3] that integrates regulatory features knowledge from different sources, comprehensively exploring the space for all possible combinations. Moreover, the method uses an unsupervised strategy and conceptual clustering techniques [3, 4], where pre-existing examples are not required. The features are analyzed concurrently, and recurrent dynamic relations are recognized to generate profiles (i.e., groups of promoters sharing common features).

The formulation of the conceptual 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, diversity and their overlap [3, 4]. These are conflicting criteria that are formulated as a multi-objective and multimodal optimization problem [5, 6], where several solutions can be optimal and only few of them are biologically meaningful. Therefore, we have developed visualization techniques using Spotfire visual analysis tools [7] to illuminate relationships otherwise not noticed and to identify gene profiles that are biologically significant and difficult to detect [8]. This visualization technique sets a framework for decision making in genomics.

Application of our method to the enteric bacteria *Escherichia coli* and *Salmonella enterica* uncovered novel members of, as well as regulatory interactions in the regulon controlled by the PhoP protein that were not discovered using previous approaches. Our 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 [9].

[1] To whom correspondence should be addressed: zwir@borcim.wustl.edu

[2] 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 promote

patterns and particular *cis*-acting promoter elements. The purpose of our method is to identify interesting substructures; here termed profiles, within a regulatory network, which can provide possible mechanisms by which the respective genes are controlled. These profiles can further be used to classify additional promoters (e.g., newly identified).

Our method represents and learns from structural data by following four main phases:

Database creation by modelling promoter features. We focus on four types of features for describing our set of co-regulated promoters that are naturally encoded into diverse data types: fix-length DNA motifs from transcriptional regulator binding sites, represented by position weight matrices ("*Submotif*"); variable-length motifs from RNA polymerase encoded into a neural net, their location in the chromosome is studied as a distribution and encoded into fuzzy sets ("*RNA Pol site*"); categorical data from the motif orientation ("*Orientation*") (Fig. 1); and gene expression from multiple experiments represented as vector patterns ("*Expression*") [9]. We account for the variability of the data by treating these features as fuzzy (i.e., not precisely defined) instead of categorical entities [10-12].

Initialization of promoter profiles for each type of feature. Our method clusters independently promoters considering individually each type of feature to build level-1 initial profiles based on the fuzzy C-means clustering method and a validity index [10] to estimate the number of clusters, as an unsupervised discretization of the features [12, 13]. For example, we obtained three level-1 profiles for the "*expression*" feature $(E_1^1,...,E_3^1)$. The superscript denotes the level-1 in this case; and the subscript denotes the specific profile. Each profile is represented by its prototype (e.g., binding motifs as a weighted matrix). Then, different types of original features can be unified by converting the promoter values for a feature into degrees of matching with the prototype of that feature (i.e., membership value to a cluster) (Fig. 2).

Dynamically profile learning by domain knowledge fusion. We group profiles by navigating in a lattice which is the feature search space [3, 4] and create systematically higher level profiles (i.e., offspring profiles) based on the combination by fuzzy intersection of parental profiles. For example, level-1 profiles: (E_1^1 , M_2^1 and P_3^1) produce level-2 profiles ($E_1^2M_2^2$, $M_2^2P_3^2$ and $E_1^2P_3^2$). As the exploration process continues level-3-profiles are obtained from intersection of the promoter members of level-2- profiles and not between those belonging to the initial profiles (e.g., $E_1^3M_2^3P_3^3$ is product of aggregating $E_1^2M_2^2$, $M_2^2P_3^2$ and $E_1^2P_3^2$; where initial profiles E_1^1 , M_2^1 and P_3^1 are not involved). This is because our approach dynamically re-discretizes the original features at each level, adapting each feature for the set of promoters recovered by the profile, and allows re-assignations of observations between sibling profiles.

Profiles evaluation using a multimodal and multi-objective context. Profile evaluation is carried out as a multi-objective optimization problem between the extent of the profile and the quality of matching among its members and the corresponding features [5, 12]. 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, within a universe of g candidates:

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

$$SI(V_i) = \left(1 - \sum_{k \in U_\alpha} \mu_{ik} / n_\alpha\right) \Big/ f \quad U_\alpha = \{\mu_{ik} : \mu_{ik} > \alpha\} \qquad (2)$$

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 [14] 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 equation (2), where μ_{ik} is the degree of membership of promoter k to an arbitrary alpha-cut U_α of the profile i and n_α is the number of elements of U_α.

The trade-off 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, 12]. The dominance relationship in a minimization problem is defined as $a \prec b\, iif\ \forall i\ O_i\,(a) \leq O_i\,(b)\ \exists j O_j\,(a) < O_j\,(b)$ where O_i and O_j are either PI or SI (Fig. 3).

The other property that characterizes good clusters is diversity, which is accounted in situations where we need to describe a system from different points of view [5]. We addressed this problem by identifying all non-dominated optimal profiles that have no better solution in the local neighbourhood of the decision variable space. This strategy, which combines multi-objective and multimodal optimization concepts, relies on competition of solutions for determining their search space 'niches' (i.e. to keep all important solutions without the need to be exhaustive). Thus the non-dominance relationship is only applied to profiles in the local neighbourhood [5] by using the hypergeometric metric (equation (1)) between profiles. For example, profile $D_1^3 E_2^3 P_3^3$ (PI=6.4E-6; SI=0.029) retrieves different promoters than profile $D_1^3 E_3^3 P_5^3$ (PI=2.84E-4; SI=0.035) that would be dominated by the first one if no niching strategy were applied (Fig. 4ab).

3 Results and Discussion

We investigated the utility of our approach by exploring the regulatory targets of the PhoP protein in *E. coli* and *Salmonella*, 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 [15] (see [9] for a complete list of promoters). Genetic and genomic approaches have been successfully used to assign genes to distinct regulatory networks both in prokaryotes and eukaryotes. 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. The fact that such co-regulated genes may be differentially regulated is often concealed by microarray gene expression experiments, which sometimes hither to only allow a relatively crude classification of gene expression patterns into a limited number of classes (e.g., up- and down-regulated genes [16]). Therefore, and because the fundamental mechanism controlling gene expression operates at the level of transcription initiation, subtle differences in co-regulated genes could be caused by the *cis*-acting regulatory elements.

We fused the *cis*-features from different domains into a common framework by representing features as fuzzy sets (Fig. 2). The detailed analysis of the gene behaviour would not be possible to be obtained neither by just inspecting each feature nor by using all features in a typical clustering technique. This happens because beforehand it is not known which aggregations of features are biologically meaningful for the different set of promoters. Therefore, we used a conceptual clustering approach to search through the space of all potential hypotheses. The fuzzy representation allowed us to homogenize features and was specifically designed to account for the variability in sequence, location and topology intrinsic to differential gene expression. The final data was transformed into an effective visual form (Fig. 4b), which improved our interaction with the large volume of profiles produced, and helped the overview of the behaviour of promoters according to features and profiles.

We recovered several optimally evaluated profiles (Fig. 4a), thus, revealing distinct putative profiles that can describe the PhoP regulation process from different angles. The predictions made by our method were experimentally validated [9] to establish that the PhoP protein uses multiple mechanisms to control gene transcription, and moreover, these profiles can be used to effectively explain the different kinetic behaviour of co-regulated genes measured by GFP reporter strains with high-temporal resolution (Fig. 5). For example, the profile $E_2^3 M_4^3 P_2^3$ (PI=1.95E^{-6}; SI=0.006) corresponds to the canonical PhoP-regulated promoters, 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 [17], but also other promoters, which was not known to be under PhoP control. We found that the promoters sharing this profile produced the earlier rise times and the higher levels of transcription (Fig. 5).

Another uncovered profile $O_2^3 E_3^3 P_1^3$ (PI=1.95E-5, SI=0.05), product of the aggregation of different features than the previous example because not every feature is relevant for all profiles, includes promoters (e.g., those of the *mgtC*, *mig-14*, *pagC*, *pagK*, and *virK* genes of *Salmonella*) that share a PhoP box in the opposite orientation of the canonical PhoP-regulated promoters as well as a class I RNA polymerase sites situated at medium distances from the PhoP boxes. We tested this profile by GFP and found that effectively differs from the previous canonical profile, exhibiting the latest genes with the lowest levels of expression (Fig. 5). Notably, this profile was recovered despite its potential domination by another profile (i.e., PI=1.95E^{-6}; SI=0.006 vs. PI=1.95E-5, SI=0.05) because we use a multimodal optimization strategy (i.e., niching) that retrieves local optimal profiles that describe the system from different points of view (Fig. 4a).

The optimization strategy was visualized as a heat map, where the different features spaces (i.e., groups of genes covered by a profile) can be identified by an optimal profile (i.e., non-dominated), and several of these solutions can be possible and even biologically significant (Fig. 4b).

We also uncovered another slightly different profile $O_2^2 P_{41}^2$ (PI=0.033, SI=0.044), which includes promoters (e.g., those of the *ompT* gene of *E. coli* and the *pipD, ugtL* and *ybjX* genes of *Salmonella*) that exhibit a PhoP binding site in the opposite orientation, but preserves the RNA polymerase of the canonical PhoP regulated promoters. We tested the kinetic behaviour of genes in this profile and found that present an intermediate value between previously described regulatory profiles (Fig. 5).

We showed that our method can make precise mechanistic predictions even with incomplete input dataset and high levels of uncertainty by fusing heterogeneous domains of knowledge into regulatory profiles. In addition, we exemplified how diversity is required to obtain biologically significant regulatory profiles. Therefore, we provided an optimal selection strategy based on multi-objective and multimodal optimization techniques that describe the system from different points of view. We also showed that the uncertainty of exploration process can be narrowed down by applying regular visualization tools.

ACKNOWLEDGMENT

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. Brenner, S., Genomics. The end of the beginning. Science, 2000. 287(5461): p. 2173-4.
2. Barash, Y., et al., CIS: Compound importance sampling method for protein-DNA binding site p-value estimation. Bioinformatics, 2004.
3. Cook, D.J., et al., Structural mining of molecular biology data. IEEE Eng Med Biol Mag, 2001. 20(4): p. 67-74.
4. 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.
5. 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.
6. Zwir, I., R.R. Zaliz, and E.H. Ruspini, Automated biological sequence description by genetic multiobjective generalized clustering. Ann N Y Acad Sci, 2002. 980: p. 65-82.
7. Asher, B., Decision analytics software solutions for proteomics analysis. J Mol Graph Model, 2000. 18(1): p. 79-82.
8. Borner, K., C. Chen, and K.W. Boyack, Visualizing Knowledge Domains. ARIST, 2003. 37: p. 179-258.
9. 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.
10. 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.
11. 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.
12. 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.
13. Kohavi, R. and G.H. John, Wrappers for feature subset selection. Artificial Intelligence, 1997. 97(1-2): p. 273-324.
14. Tavazoie, S., et al., Systematic determination of genetic network architecture. Nat Genet, 1999. 22(3): p. 281-5.
15. Groisman, E.A., The pleiotropic two-component regulatory system PhoP-PhoQ. J Bacteriol, 2001. 183(6): p. 1835-42.
16. 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.
17. Minagawa, S., et al., Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli. J Bacteriol, 2003. 185(13): p. 3696-702.

FIGURE LEGENDS

Figure 1. Heterogamous domain features. The PhoP proteins binds to a DNA strain and recruits the RNA polymerase using a class I activator approach (class II activators bind to sites that overlap the target promoter -35 region) that can bind at different upstream locations (e.g., close, medium and remote distance to RNA polymerase) and at different orientations from the open reading frame. One of the putative PhoP submotifs is detailed as a logos chart, where the characters representing the sequence are stacked on top of each other for each position in the aligned sequences and the height of each letter is made proportional to its frequency.

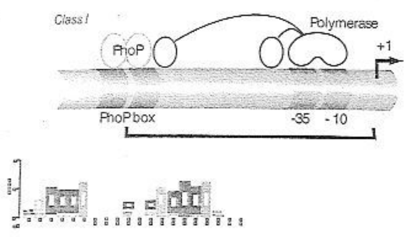
Figure 2. Database representation. The regulatory features model heterogeneous domains corresponding to different *cis*- and expression descriptions of the PhoP regulated promoters by using fuzzy membership values. Here we exemplify data from DNA sequences representing PhoP binding sites, the orientation of this site, the class and distance of the RNA polymerase that interacts with the PhoP protein, and gene expression patterns. The heat map 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.

Figure 3. Knowledge fusion. A node representation of the lattice that fuses profiles containing different types of features (e.g., "RNA *Pol sites*" and "*orientation*"). The profiles are evaluated by using probability (PI) which measures the extent of a profile (low size of the circles: small; high: big) and the similarity (SI) which measures the explanatory quality of a profile (low p-value: green; high: red).

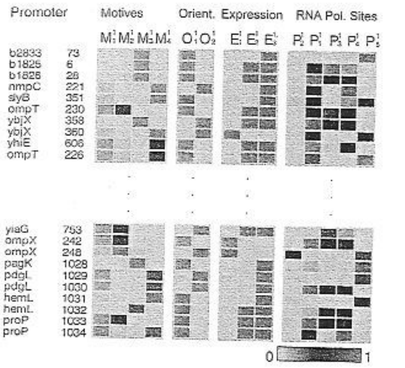
Figure 4. Selection of the most representative profiles. a) Non-dominance optimization approach (Non-dominated solutions red; Dominated ones in green) between two conflicting objectives PI and SI. This guideline is applied in local neighbourhood to support diversity. b) Heat map corresponding to promoters (columns) recognized at different degrees of matching (green: low; red: high) by the profiles (rows) divided in neighbourhood (clusters). These localities are dominated by a representative profile (left columns). This guideline prevents the population of solutions to converge to a single region and obtains optimal and diverse solutions.

Figure 5. Independent validation of profiles using kinetic classes. Transcriptional activity of wild-type *Salmonella* harbouring plasmids with a transcriptional fusion between a promoterless gfp gene and the promoters.. The activity of each promoter is proportional to the number of GFP molecules produced per unit time per cell [dGi(t)/dt]/ODi(t)], where Gi(t) is GFP fluorescence from wild-type Salmonella strain 14028s, and ODi(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.

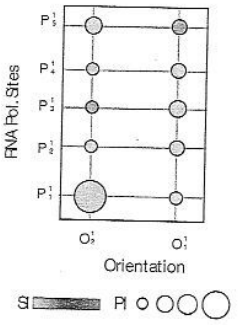
1)



2)

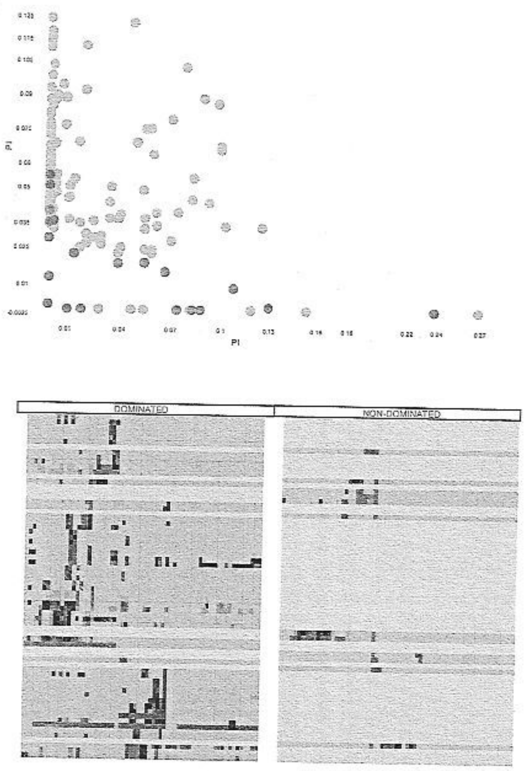


3)



4a)

4b)



5)

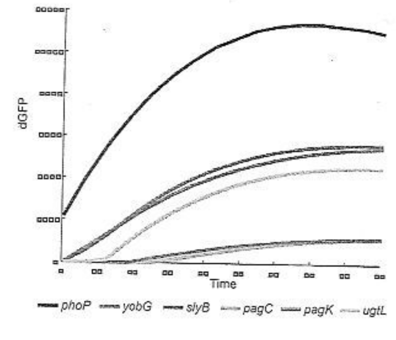


TABLE OF CONTENTS

DATA MINING

HUMAN COMPUTER INTERACTION

Current Research in Information Sciences and Technologies

Multidisciplinary approaches to global information systems

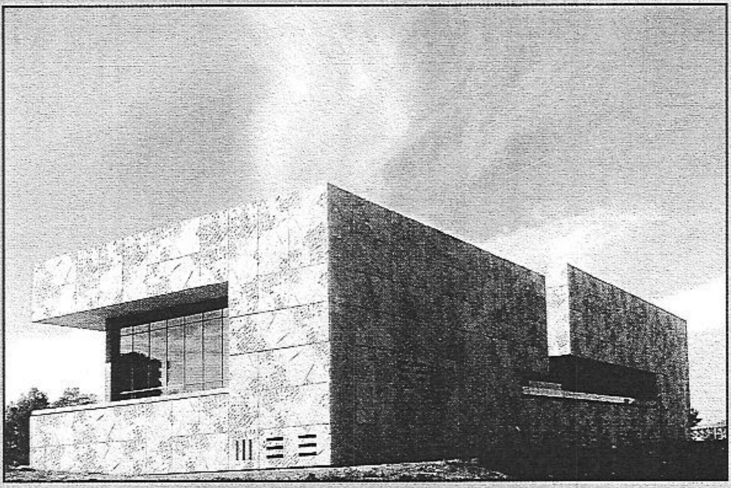
Volume II

Vicente P. Guerrero-Bote (Editor)

InSciT2006

I International Conference on Multidisciplinary Information Sciences & Technologies

Mérida's Conferrnce Hall
October 25-28th, 2006



Instituto
Abierto del Conocimiento
Open Institute of Knowledge