BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4073–4083
doi:10.1093/bioinformatics/bti672

Genome analysis

Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation

Igor Zwir, Henry Huang and Eduardo A. Groisman*

Department of Molecular Microbiology, Howard Hughes Medical Institute, Washington University
School of Medicine, Campus Box 8230, 660 S. Euclid Avenue, St Louis, MO 63110, USA

Received on June 30, 2005; revised on August 4, 2005; accepted on September 8, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: A critical challenge of the post-genomic 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 expression patterns to classify genes into distinct networks. However, these methods are not focused on distinguishing between differentially regulated genes within a given network. Here we describe an unsupervised machine learning method, termed GPS for gene promoter scan, that discriminates among co-regulated promoters by simultaneously considering both *cis*-acting regulatory features and gene expression. GPS is particularly useful for knowledge discovery in environments with reduced datasets and high levels of uncertainty.

Results: Application of this method to the enteric bacteria *Escherichia coli* and *Salmonella enterica* uncovered novel members, as well as regulatory interactions in the regulon controlled by the PhoP protein that were not discovered using previous approaches. The predictions made by GPS 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.

Availability: The scripts and programs used in this work are accessible from the gps-tools.wustl.edu website. Data and predictions are available by request.

Contact: groisman@borcim.wustl.edu

Supplementary information: http://gps-tools.wustl.edu/BIOINF-2005-1246R1-Supplemental.pdf

INTRODUCTION

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 suggests that subtle differences in the shared *cis*-acting regulatory elements are probably significant. However, it is not yet possible to predict the critical differences that govern the differential gene expression. Furthermore, while genes could in principle be differentiated by incorporating into the analysis quantitative and kinetic measurements of gene expression (Ronen *et al*., 2002) and/or the participation of other transcription factors (Bar-Joseph *et al*., 2003; Beer and Tavazoie, 2004; Conlon *et al*., 2003), 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 only allow a relatively crude classification of gene expression patterns into a limited number of classes [e.g. upregulated and downregulated genes (Oshima *et al*., 2002; Tucker *et al*., 2002)].

Here we describe a machine learning method (Cheeseman and Oldford, 1994; Cook *et al*., 2001; Cooper and Herskovits, 1992), termed GPS for Gene Promoter Scan, that identifies, differentiates and groups sets of co-regulated promoters by simultaneously considering multiple *cis*-acting regulatory features and gene expression. GPS carries 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, and the class and number of RNA polymerase sites. Moreover, it treats each of these promoter features with equal weight because it is not known beforehand which features are important. It further captures variability in the control of biological systems by treating the *cis*-acting features as fuzzy (i.e. not precisely defined) instead of categorical entities (Bezdek, 1998; Gasch and Eisen, 2002; Ruspini, 2001). To circumvent limitations imposed by back-correlating relatively few classes of gene expression measurements to *cis*-acting features, the GPS method treats gene expression data as one feature among many. The features are analyzed concurrently, and recurrent relations are recognized to generate profiles, which are groups of promoters sharing common features. GPS uses an unsupervised strategy, where pre-existing examples are not required, as well as multiobjective optimization techniques that recover all optimal feature associations rather than potentially biased subsets (Deb, 2001; Ruspini, 2001). The resulting profiles group promoters that may share underlying biological properties.

*To whom correspondence should be addressed.

Basis for GPS: conceptual clustering and machine learning

Cluster analysis—or simply clustering—is a data mining technique often used to identify various groupings or taxonomies in databases. Most existing methods for clustering are designed for linear feature-value data. However, sometimes we need to represent and learn structural data that not only contains descriptions of individual observations in databases, but also relationships among these observations. Therefore, mining into structural databases entails addressing both the uncertainty of which observations should be placed together, and also which distinct relationships among features best characterize different sets of observations. Typical clustering techniques (Everitt and Der, 1996) are not designed to do this, even when combined with global filter feature selection methods such as principal component analysis or stepwise descendent methods (Kohavi and John, 1997; Yeung and Ruzzo, 2001). In contrast, conceptual clustering techniques have been successfully applied to structural databases to uncover concepts that are embedded in subsets of structural data or substructures (Cheeseman and Oldford, 1994; Cook *et al*., 2001; Cooper and Herskovits, 1992). Consequently, conceptual learning can be formulated as the problem of searching through a predefined space of potential hypotheses (i.e. substructures or associations of features and observations) for those observations that best fit the training examples.

While most machine learning techniques applied directly or indirectly to structural databases exhibit methodological differences, they do share the same framework even though they employ distinct metrics, heuristics or probability interpretations (Cheeseman and Oldford, 1994; Cook *et al*., 2001; Cooper and Herskovits, 1992) as follows:

Structure representation. Structural data can be viewed as a graph containing nodes representing objects, which have features linked to other nodes by edges corresponding to their relations. A substructure consists of a subgraph of structural data (Cook *et al*., 2001). These data can be efficiently organized by taking advantage of a naturally occurring structure over feature space, which consists of a general to specific ordering of possible substructures (i.e. a lattice).

Structure learning. This process consist of searching through the lattice space for potential substructures, and returning either the best one found or an optimal sample of them. If the number of substructures is super-exponential in the number of nodes, different heuristic methods can be used [e.g. greedy (Cooper and Herskovits, 1992); hill climbing (Chickering, 2003); genetic algorithms (Larranaga, 1996)].

Subtructure evaluation. The formulation of the clustering problem in a lattice or graph-based structure would result in the generation of many substructures with small extent, as it is easier to explain or substructure-match smaller data subsets than those that constitute a significant portion of the dataset. For this reason, any successful methodology should also consider additional criteria to extract broader or more comprehensive substructures based on their size, the number of retrieved substructures, and their diversity and extent of overlap (Cook *et al*., 2001; Ruspini, 2001). These are conflicting criteria that can be formulated as a multiobjective optimization problem, analogous to minimum description-length methods (Rissanen, 1989), based on the combination of the individual criteria or objectives into a global measure of cluster quality. The basic challenge with this approach, however, is its potential bias and inflexibility caused by weighting of the objectives (Ruspini, 2001).

Inference. New observations can be predicted from previously learned substructures by using classifiers that optimize their matching based on distance (Bezdek, 1998) or probabilistic metrics (Cooper and Herskovits, 1992; Mitchell, 1997). When designed for labeled data, the approach is referred to as supervised learning (as opposed to unsupervised learning).

SYSTEMS AND 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 GPS is to identify interesting substructures or 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.

GPS represents, learns and infers from structural data by following three main phases (Fig. 1a):

Structure representation. Model the features of promoters (Zwir *et al*., 2005) by several steps, including constructing models of microarray expression data and *cis*-features (Bar-Joseph *et al*., 2003; Beer and Tavazoie, 2004) from available databases and describing promoters according to the detected features, allowing multiple occurrences of a feature and missing values (Bezdek, 1998; Gasch and Eisen, 2002).

Structure learning. (1) Initialize the profiles by grouping them into preliminary profiles by using clustering techniques. (2) Group the profiles by navigating into a hierarchical lattice structure (Fig. 1b) corresponding to the feature space and systematically use profile intersection to create compound profiles. The profiles are encoded as fuzzy models (Bezdek, 1998; Gasch and Eisen, 2002; Ruspini, 2001), which allows imprecise match of promoters with one or more profiles, and thus, promoter migrations are allowed even among sibling profiles in the lattice structure [i.e. optimization clustering (Falkenauer, 1998)]. GPS uses an exhaustive search in the *cis*-features and the expression feature space. In addition, we have successfully applied genetic algorithms (Cordon *et al*., 2002; Zwir *et al*., 2002) in more complex problems (data not shown). (3) Prototype the profiles by obtaining the centroid that best represents a group of promoters. (4) Search the profiles in a frontier of optimal solutions according to two opposing criteria or objectives, the probability of different sets of promoters to belong to a common profile, and the similarity between profile members. The GPS approach is less biased than weighting the objectives [e.g. minimum description length (Rissanen, 1989)] because it identifies all the profiles lying in the Pareto optimal frontier (Deb, 2001; Ruspini, 2001), 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.

Inference. Predict new members of the profiles by searching genomes to discover new promoters that match the profiles, using an unsupervised classifier (Bezdek, 1998), which allows descriptions of new examples from multiple substructures or profiles by using fuzzy clustering techniques.

Dataset: *Escherichia coli* and *Salmonella enterica* promoters

We built models based on 33 genes whose microarray expression differed statistically between wild-type and *phoP E.coli* strains experiencing inducing conditions for the PhoP/PhoQ regulatory system (Zwir *et al*., 2005), and

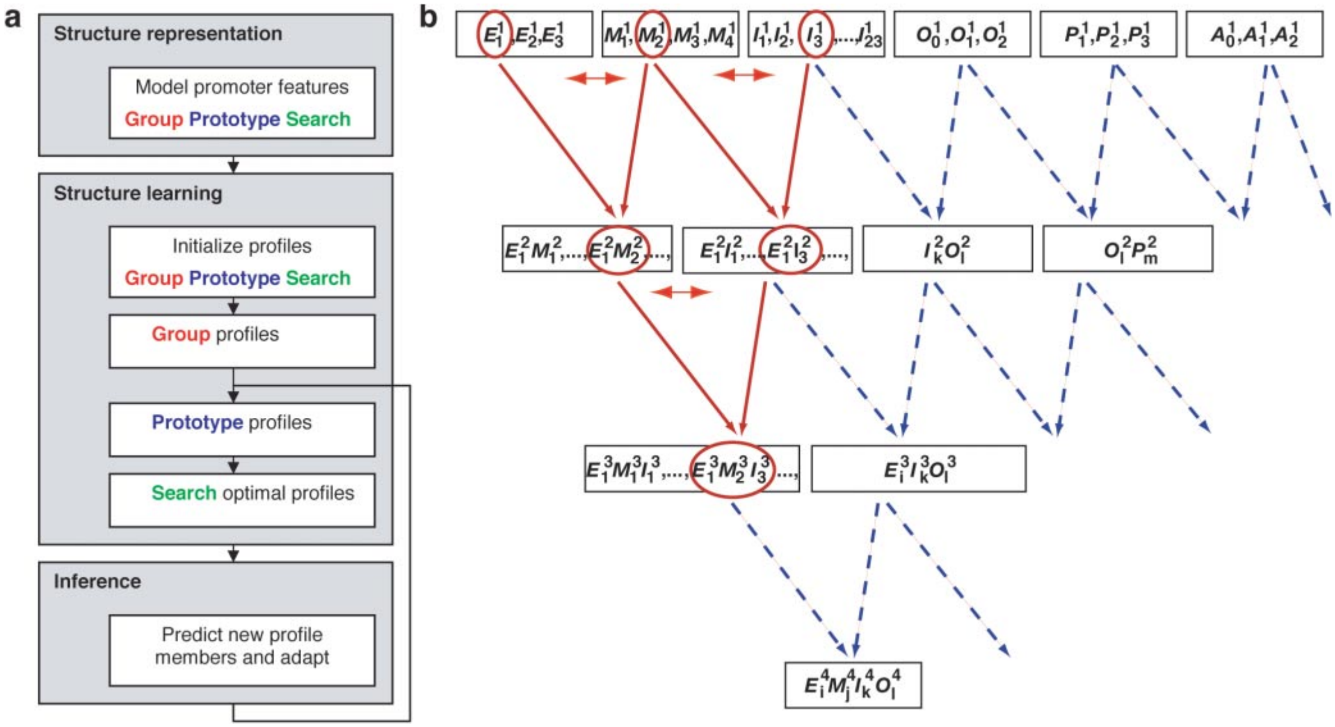


Fig. 1. The GPS method. (**a**) GPS is a machine learning technique that models promoter features as well as relations between them, uses them to describe promoters, combines such characterized promoters into groups termed profiles, evaluates the resulting profiles to select the most significant ones and performs genome-wide predictions based on such profiles. To accomplish this task, GPS carries out three basic operations: grouping observations from the dataset; prototyping such groups into their most representative elements (centroid); and searching in the set of optimal solutions (i.e. Pareto optimal frontier) to retrieve the most relevant profiles, which are used to describe and identify new objects by similarity with the prototypes. (**b**) GPS navigates through the feature-space lattice generating and evaluating profiles. For the analysis of promoters regulated by the PhoP protein, we identified up to five models for each type of feature, which are used to describe the promoters. Then, GPS generates profiles, which are groups of promoters sharing common sets of features. (The subscripts denote the different profiles for each feature, the superscripts denote the level in the lattice of the profile). For example, E_1^1 is a particular 'expression' profile that differs from E_2^1 and E_3^1. These 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. In addition, because of the fuzzy formulation of the clustering, any promoter that was initially assigned to a specific profile E_i^t, can participate in the profile of level-t where E_j^t is involved (i.e. indicated as a double-headed arrow). Thus, observations can migrate from parental to offspring clusters (i.e. hierarchical clustering), and among sibling clusters (i.e. optimization clustering). Here, we show a small part of the complete lattice, where the part that is highlighted in red is described in detail in Figure 2.

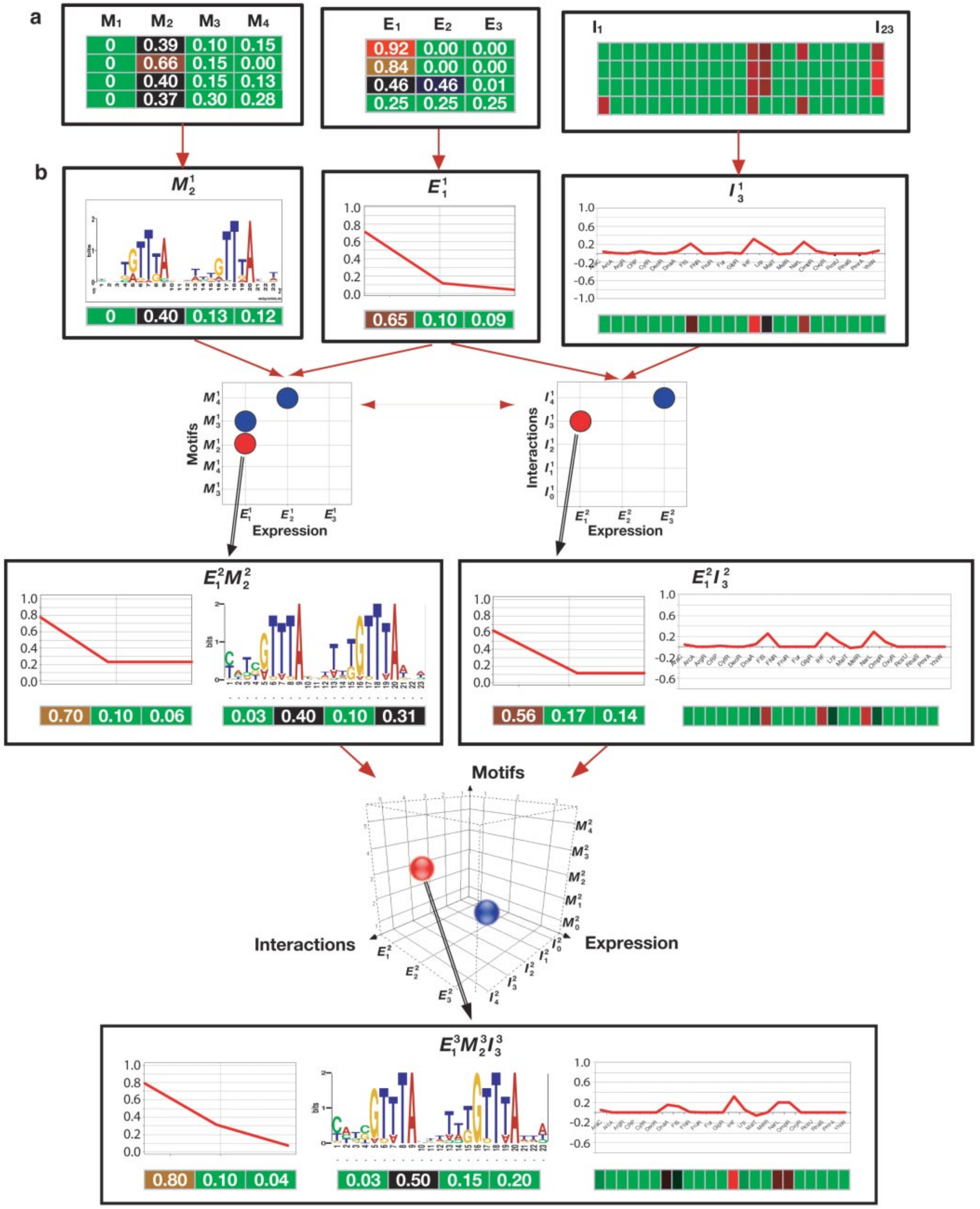
22 additional *S.enterica* promoters known to be regulated by the PhoP protein. This set of promoters constitutes 70% of the training partition. The remaining 30% constitutes the test partition dataset and contains genes known/assumed to be PhoP regulated (Zwir *et al*., 2005), compiled from the literature and our own lab information (Supplementary Table 1). Even if most of these genes were known to be PhoP-regulated, the mechanism by which the genes were regulated was not known in detail, and could differ from gene to gene. Missing expression values in the *Salmonella* genome were inherited from the *E.coli* orthologs.

Structure representation: modeling promoter features

To describe promoters controlled by the PhoP/PhoQ regulatory system of *E.coli* and *S.enterica* serovar Typhimurium, we built model-based features that encode many promoter properties. The data include observations from RegulonDB, our microarray experiments and a survey of the promoter regions of the *E.coli* and *S.enterica* genomes. For each feature, we assigned/calculated the degree of matching (in a scale to the unit-interval) between an observation and a model, or a family of models. This approach allows an observation to be described by multiple features, and the models are allowed to combine different individual features, reflecting the fact that promoters may harbor multiple features.

We focused on six types of features (Bar-Joseph *et al*., 2003; Beer and Tavazoie, 2004; Li *et al*., 2002; Zwir *et al*., 2005) for describing our training set of promoters (Supplementary Table 2), which are briefly described here, and more in detail in the study by I. Zwir, R. Romero-Zaliz, H. Huang and E. A. Groisman (in preparation):

'Submotifs'. We modeled the PhoP box motifs by using position weight matrices (Stormo, 2000). Then, we used these preliminary models to describe promoters by using low thresholds. We grouped the retrieved observations into subsets and rebuilt matrix models for each of them, thus obtaining several more refined models and increasing the sensitivity to departures from the consensus and the specificity of submotif recognition.
'Orientation'. We classified PhoP boxes as either in direct or opposite orientation relative to the open reading frame.
'RNA pol sites'. We studied (1) the RNA polymerase motif by using a neural network method (Cotik *et al*., 2005), (2) the class of sigma 70 promoter by using an intelligent parser (Romero Zaliz *et al*., 2004) that differentiates *class I* from *class II* promoters and (3) the distance distributions (*close*, *medium* and *remote*) between RNA polymerase and transcription factor binding sites in activated and repressed promoters by using fuzzy set representations (Ruspini, 2001) from information available in the RegulonDB database (Salgado *et al*., 2004). We derived models to describe


a
M1 M2 M3 M4
0 0.39 0.10 0.15
0 0.66 0.15 0.00
0 0.40 0.15 0.13
0 0.37 0.30 0.28
E1 E2 E3
0.92 0.00 0.00
0.84 0.00 0.00
0.46 0.46 0.01
0.25 0.25 0.25
I1
I23
b
M_2^1
0 0.40 0.13 0.12
E_1^1
0.65 0.10 0.09
I_3^1
Motifs
M_4^1 M_3^1 M_2^1 M_4^1 M_3^1
E_1^1 E_2^1 E_3^1
Expression
Interactions
I_4^1 I_3^1 I_2^1 I_1^1 I_0^1
E_1^2 E_2^2 E_3^2
Expression
$E_1^2M_2^2$
0.70 0.10 0.06
0.03 0.40 0.10 0.31
$E_1^2I_3^2$
0.56 0.17 0.14
Motifs
M_4^2 M_3^2 M_2^2 M_1^2 M_0^2
Interactions
E_1^2 E_2^2 E_3^2
Expression
I_0^2 I_1^2 I_2^2 I_3^2 I_4^2
$E_1^3M_2^3I_3^3$
0.80 0.10 0.04
0.03 0.50 0.15 0.20

putative relationships between PhoP and RNA polymerase binding sites by using AND-connected fuzzy predicates. The equivalent probabilistic interpretation of this is the posterior probability obtained by following Bayes' rule [(e.g. *p(class II/close)* that, given a *class II* RNA polymerase site, it interacts at *close* distance with the PhoP protein)].

'Activated/repressed'. We learned activation and repression distributions by compiling distances between RNA polymerase and transcription factor binding sites from the RegulonDB database. We modeled and used them to estimate whether the position of the PhoP box suggests that a promoter is activated or repressed by PhoP, when gene expression data may be unavailable (Collado-Vides *et al*., 1991).

'Interactions'. We built motifs for all transcription factor-binding sites in the RegulonDB database. We also modeled the distance distributions between motifs co-located in the same promoter regions. Then, we connected them with models of the PhoP binding sites and used them to describe putative relationships among the PhoP protein and other 23 regulators.

'Expression'. We clustered PhoP-regulated gene expression levels and built models for each cluster by calculating its centroid (Li and Wong, 2001). Then, we described gene expression of each promoter in *E.coli* by its similarity to each model.

Structure learning: *g*rouping, *p*rototyping and *s*earching profiles in the lattice space

Initializing profiles GPS independently clusters each type of feature to build initial level-1 profiles based on the fuzzy C-means clustering method (Fig. 2) and a validity index (see below) (Bezdek, 1998) to estimate the number of clusters, as an unsupervized discretization of the features (Kohavi and John, 1997; Mitchell, 1997; Ruspini, 2001). A dummy variable representation of the nucleotides was used to cluster binding-site motifs (Rosner, 1986). For example, we obtained five level-1 profiles for the 'submotifs' feature $\left(M_0^1, \ldots, M_4^1\right)$ (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); three level-1 profiles for the 'orientation' feature $\left(O_0^1, \ldots, O_2^1\right)$; four level-1 profiles for the 'RNA pol sites' feature $\left(P_0^1, \ldots, P_3^1\right)$; three level-1 profiles for the 'activated/repressed' feature $\left(A_1^1, \ldots, A_3^1\right)$; five level-1 profiles for the 'interactions' feature $\left(I_0^1, \ldots, I_4^1\right)$; and three level-1 profiles for the 'expression' feature $\left(E_1^1, \ldots, E_3^1\right)$.

Grouping profiles GPS groups profiles by navigating in a lattice corresponding to the feature searching space (Cheeseman and Oldford, 1994; Cook *et al*., 2001; Cooper and Herskovits, 1992) (Fig. 1b and 2b) and systematically creating compound higher level profiles (i.e. offspring profiles) $V_{(i,j)}$ based on combining parental profiles V_i and V_j, by taking the intersection $V_{(i,j)} = V_i \cap V_j$. For example, level-1: (E_1^1, M_2^1 and I_3^1) → level-2: ($E_1^2M_2^2$, $M_2^2I_3^2$ and $E_1^2I_3^2$) → level-3: ($E_1^3M_2^3I_3^3$), where level-3-profiles are obtained from intersection of the promoter members of level-2- profiles (e.g. $E_1^2M_2^2$, $M_2^2I_3^2$ and $E_1^2I_3^2$) and not between those belonging to the initial profiles (E_1^1, M_2^1 and I_3^1) (Fig. 2). This is because GPS locally re-discretizes the original features at each level and allows re-assignations of observations between sibling profiles (Fig. 2b, see below). 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 (Mitchell, 1997)] are found. Profiles combined by GPS constitute local partitions of the datasets (Ruspini, 2001) that are not constrained by an arbitrary fixed number of clusters and features (see below), thus permitting a better characterization of the relationship between promoters and clusters.

Prototyping profiles GPS learns profiles by using an extension of the fuzzy C-means clustering method (Bezdek, 1998; Gasch and Eisen, 2002), where promoters can belong to more than one cluster with different degrees of membership, as it is done in the possibilistic clustering implementation (Bezdek, 1998), and are not forced to belong to any particular cluster. This consists of individually applying fuzzy C-means clustering to each type of feature at each level in the lattice, and combining the results: the membership of a promoter k with a feature value x_k of a specific type of feature f from a particular profile V_i, is calculated as

$$u_{i,k,f} = \left[1 + \left(\frac{\|x_{k,f} - \overline{V_{i,f}}\|_A^2}{w_i}\right)^{1/m-1}\right]^{-1}$$
$$\forall i \in \{1, \ldots, c\}, \quad \forall k \in \{1, \ldots, n\}, \quad \forall f \in \{1, \ldots, t\}, \quad 1 < m \leq \infty, \quad (1)$$

where $u_{i,k,f}$ is taken as the degree of membership of the value $x_{k,f}$ in the i-th partitioning fuzzy subset of type of feature f; $\overline{V_{i,f}}$ is the profile prototype or centroid of partition $V_{i,f}$ (see below); m is the degree of fuzzification; A determines the type of norm commonly used in pattern recognition [e.g. $A = 2$ is the Euclidean norm (Bezdek, 1998)]; and w_i is a weight for penalty initialized as 1 in the absence of prior information. Thus, each level of the lattice is arrayed as ($c \times n \times t$) matrix $U = \{u_{1,1,1}, \ldots, u_{c,1,t}; \ldots u_{1,n,1}, \ldots u_{c,n,t}\}$ containing the vector representation of matching between c-profiles, n promoters and t features.

If the interpretation of the partition U is probabilistic, $u_{i,k,f}$ is usually the posterior probability $p(V_{i,f}/x_{k,f})$ that, given $x_{k,f}$, it comes from class $V_{i,f}$, by following the Bayes' rule (Bezdek, 1998; Bezdek *et al*., 1992; Everitt and Der, 1996). If U is fuzzy, $u_{i,k,f}$ is taken as the membership of $x_{k,f}$ to one or more fuzzy subsets $V_{i,f}$ of X, which is calculated based on distances measurements. Previous interpretations constrain the sum of memberships (or probabilities, as the case may be) to be one. However, if the interpretation is possibilistic (Bezdek, 1998), this constraint is relaxed, and thus, a more realistic situation can be represented where we do not force each observation to belong to a profile (Ruspini, 2001).

Fig. 2. Using GPS to build promoter profiles. GPS generation of the $E_i^3M_2^3I_3^3$ profile is shown here. It corresponds to the highlighted substructure of the lattice shown in Figure 1b. (**a**) GPS starts by using information from databases and microarray data to construct a family of models for each feature (e.g. expression levels E_1 to E_3, PhoP box submotif M_1 to M_4 and presence of binding sites from other transcription factors I_1 to I_{23}) (Zwir *et al*., 2005). The promoters are described using the modeled features, the degree of matching between features and promoters being encoded as a vector of independent values (Alon *et al*., 1999), where 1 (red color) corresponds to maximum matching and 0 (green color) corresponds to the absence of the feature. For each feature, the promoters are then grouped into subsets that share similar patterns, using fuzzy clustering. (**b**) Each subset shown in (a) is prototyped by locating the centroid that best represent the group, to generate the initial, level-1 profiles (e.g. E_1^1, M_2^1 and I_3^1). The centroids are encoded as a vector, and also visualized by graphical plots for the 'expression' and the 'interactions' features, and by a sequence logo (Crooks *et al*., 2004) for the 'submotifs' feature. These level-1 profiles are combined to generate level-2 profiles [e.g. $E_1^2M_2^2$ and $M_2^2I_2^2$ (red circles)], by the intersection of the ancestor profiles, and then prototyped. (Blue circles represent profiles containing other subsets of promoters. The absence of a circle signifies that no promoters are classified into these profiles.) Further navigation through the feature-space lattice generates the level-3 profiles, e.g. $E_1^3M_2^3I_3^3$. The $E_1^3M_2^3I_3^3$ profile thus encompasses promoters that share the same expression pattern, PhoP submotif and RNA polymerase sites. Note that the vectors of the daughter profiles are built anew from the constituent promoters, and are slightly different from those of their ancestors, which is due to the refinement that takes place during the profile learning process. The double-headed arrow indicates that observations can migrate among sibling clusters (i.e. optimization clustering).

The profile prototype or centroid of partition V_i of type of feature f is calculated as

$$\overline{V_{i,f}} = \frac{\sum_{k=1}^{n} \left(u_{i,k,f}\right)^m x_{k,f}}{\sum_{k=1}^{n} \left(u_{i,k,f}\right)^m} \quad \forall i, \quad i \in \{1,\ldots,c\}; \quad \forall f, f \in \{1,\ldots,t\}. \tag{2}$$

(Supplementary Fig. 1). We note that each type of feature is locally re-discretized in each profile of the lattice based on the original x_k values of the promoters included in that profile, thereby producing a dynamic discretization of the dataset (Kohavi and John, 1997; Quinlan, 1993). Then, a profile is represented as a set of prototypes, each corresponding to a different type of feature included in the profile:

$$\overline{V_i} = \{\overline{V_{i,1}}, \ldots, \overline{V_{i,t}}\}. \tag{3}$$

Therefore, the membership of x_k to a profile V_i is calculated by linking the individual memberships for each type of feature, by using fuzzy AND logic operations (Bezdek, 1998):

$$u_{i,k} = \mathrm{OP}_{\mathrm{AND}}\{u_{i,k,1}, \ldots, u_{i,k,t}\}, \tag{4}$$

where $\mathrm{OP}_{\mathrm{AND}}$ represents fuzzy logic-based operations, such as Minimum or Product (Bezdek, 1998) that link the degrees of matching between a promoter (the observation) and the prototypes of different type of features composing a profile. One promoter observation x_k can contribute to more than one profile V_i in the same or a different level of the lattice, with different degrees of membership $u_{i,k}$. 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 (Falkenauer, 1998) in that it allows tranfer among sibling clusters in the same level (Fig. 2b). Position weight matrices are used for prototyping groups of binding site motifs instead of Equation (2). The resulting scores are normalized by the maximum information content of the matrix (Stormo, 2000) and incorporated in Equations (3) and (4). Finally, only those promoters that exhibit a profile membership >60% (i.e. $u_{i,k} > 0.6$) are used in the generation of the next level profiles.

Searching profiles in the Pareto optimal frontier Profile search and evaluation is carried out as a multiobjective optimization problem (Deb, 2001; Rissanen, 1989; Ruspini, 2001), between the extent of the profile and the quality of matching among its members and the corresponding features. For computational convenience, both objectives are minimized. 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 profile V_i within another profile V_j of size n:

$$\mathrm{PI}\left(V_{(i,j)}\right) = P\left(V_i \cap V_j\right) = 1 - \sum_{q=0}^{p-1} \frac{\binom{h}{q}\binom{g-h}{n-q}}{\binom{g}{h}}, \tag{5}$$

where h is the total number of elements within profile V_i and g is the total number of candidates, such that the lower the p-value the better the size of the profile association (Supplementary Fig. 1a and b) (Tavazoie *et al.*, 1999). The multivariate hypergeometric distribution is used to combine more than two profiles (Requena and Ciudad, 2000). The quality of matching between promoters and features of a profile (i.e. similarity of intersection SI) is normalized by the number of features f considered in the profile, and calculated using the following equation:

$$\mathrm{SI}\left(V_{(i,j)}\right) = \frac{1}{f}\left(1 - \frac{\sum_{k=1}^{n} u_{(i,j),k}}{n}\right), \tag{6}$$

where n is the number of promoters in profile $V_{(i,j)}$. Good profiles in current implementation are those that minimize both PI and SI (Supplementary Fig. 1a and 1c).

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 (Deb, 2001; Ruspini, 2001); see Supplementary Fig. 1d]. The dominance relationship in a minimization problem is defined by

$$a \prec b \ \mathit{iff} \ \forall i O_i(a) \leq O_i(b) \exists j O_j(a) < O_j(b), \tag{7}$$

where the O_i and O_j are either PI or SI.

Another objective indirectly considered by GPS 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 different angles. 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 (Supplementary Fig. 1d) (Deb, 2001; Ruspini, 2001). This strategy, which combines multiobjective and multimodal optimization concepts (Deb, 2001), relies on competition of solutions for determining their search space 'niches' (i.e. to keep all important solutions without the need to be exhaustive).

GPS calculates niches (i.e. classes of equivalence) by using the hypergeometric metric between profiles:

$$\mathrm{PI}\left(V_i, V_j\right) < \delta, \tag{8}$$

where PI is calculated by using Equation (5), profiles V_i and V_j can be any profile in the lattice and δ is a small initialized value. PI is distinguished from other metrics, such as the Jaccard coefficient (Saporta, 1996), in being an adaptive measure that is sensitive to small sets of examples, while retaining specificity with large datasets (Supplementary Fig. 2). Thus, GPS is designed to identify profiles that might contain very few member promoters. It also maintains diversity by constraining niches to members sharing the same type of features, by not applying Equation (8) to profiles located in disjointed branches of the hierarchical lattice searching space. For example, the profile $P_2^2 O_1^2$ (PI = 0.37, SI = 0.20) is dominated by $E_1^3 I_3^3 O_1^3$ (PI = 0.03, SI = 0.18) only if the niching strategy is solely based on the PI metric, but it is just dominated by profile $P_2^4 M_1^4 I_3^4 O_1^4$ (PI = 0.36, SI = 0.10) if the niches are constrained to contain the same type of features (i.e. $P_2^2 O_1^2 \subseteq P_2^4 M_1^4 I_3^4 O_1^4$), where the profiles are located in the same branches of the hierarchical lattice.

Inference: unsupervised fuzzy k-nearest-prototype classifier

GPS uses a fuzzy k-nearest prototype classifier (FKN) to predict new profile members using an unsupervised classification method (Bezdek, 1998; Ruspini, 2001) applied to annotated regulatory regions of genomes (I. Zwir, R. Romero-Zaliz, H. Huang and E. A. Groisman, manuscript in preparation). First, we determine the lower-boundary similarity threshold for each profile finally selected by GPS. This threshold is calculated based on the ability of each profile to retrieve its own promoters and to discard promoters from other profiles (Benitez-Bellon *et al.*, 2002) (see below). 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:

$$\mathrm{FKN}\left(x_q, V_1, \ldots, V_k\right) = i, \quad i \in \{1, \ldots, k\}, \tag{9}$$

where $u_{i,q} = \mathrm{OP}_{\mathrm{OR}}\{u_{1,q}, \ldots, u_{k,q}\}$, u is calculated based on Equation (4) in which w_i [Equation (1)] is initialized as

$$w_i = \frac{r_1 \mathrm{PI}(V_i) + r_2 (f/t') \mathrm{SI}(V_i)}{r_1 + r_2}, \tag{10}$$

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 (Cooper and Herskovits, 1992; Mitchell, 1997) or rule weight (Cordon *et al.*, 2002); r_1 and r_2 are user-dependent parameters, simply initialized as 1 if no preference exist between both objectives; and $\mathrm{OP}_{\mathrm{OR}}$ is the Maximum fuzzy operator (Bezdek, 1998; Gasch and Eisen, 2002).

Fuzzy C-means clustering method (Bezdek, 1998; Gasch and Eisen, 2002) (0) Initialize $L_0 = \{\bar{V}_1, \ldots, \bar{V}_c\}$, (1) while (s < S and $\|L_s - L_{s-1}\| > \varepsilon$), where S is the maximum number of iterations, (2) calculate the membership of U_s in L_{s-1} as in Equation (1), (3) update L_{s-1} to L_s with U_s as in Equation (2) and (4) iterate.

Xie-Beni validity index (Bezdek, 1998) The minimization of this index through different number of clusters (i.e. $c = 2$ to $c = \sqrt{n}$) detects compact representations of fuzzy C-means partitions:

$$\mathrm{XB}(U,L) = \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} u_{i,k}^2 \|x_k - \bar{V}_i\|^2}{n\left(\min\limits_{i \neq j}\{\|\bar{V}_i - \bar{V}_j\|^2\}\right)}. \quad (11)$$

Metrics for evaluating the Pareto optimal frontier The metric C in Equation (12) (Zitzler and Thiele, 1999) measures the dominance relation between a set of solutions X from one method over another set provided by another method X′ in the unit interval, where $C(X, X') = 1$ means that all points in X′ are dominated by solutions in X, and $C(X, X') = 0$ represents the situation where none of the X′ solutions are dominated by X.

$$C(X, X') = \frac{\sum_{a \in X} |\{a' \in X'; a \prec a'\}|}{|X'||\{a \in X; \exists a' : a \prec a'\}|}. \quad (12)$$

This metric is not symmetric, thus, both *C(X, X′)* and *C(X, X′)* have to be measured.

Profile similarity thresholds We evaluated GPS performance in retrieving desired and discarding undesired observations (see Results) by determining a lower-boundary similarity threshold for each profile. These values are calculated based on optimizing the overall performance measurement (Benitez-Bellon *et al*., 2002) for each profile finally selected by GPS: OP = (AC + PPV)/2, where (positive predictive value) PPV = TP/(TP + FP) and (accuracy) AC = (TP + TN)/(TP + TN + FP + FN) were defined on the basis of specificity = TN/(TN + FP) and sensitivity = TP/(TP + FN), where P = positive examples for a specific profile, N = negative examples from another different profile, T = true and F = false. A final constraint requires the sensitivity to be >60%, otherwise, the closest to this constraint is used.

Programming resources The scripts and programs used in this work are accessible at the website gps-tools.wustl.edu, were based on Perl, Matlab 6.1 and C++ interpreters/languages, and the visualization routines were performed on the Spotfire DecisionSite 8 software.

RESULTS

We investigated the utility of GPS by exploring the regulatory targets of the PhoP protein in *E.coli* K-12 and *S.enterica* serovar Typhimurium, 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 (Groisman, 2001). For PhoP analysis, we identified six types of features: gene expression levels ('expression'), PhoP box submotifs ('submotifs'), the presence of potential binding sites for 23 transcription factors ('interactions'), the orientation of the PhoP box ('orientation'), the distance of the PhoP box relative to the RNA polymerase site and the class of sigma 70 promoter ('RNA pol sites'), and whether the position of the PhoP box suggests that a promoter is activated or repressed ('activated/repressed') (I. Zwir, R. Romero-Zaliz, H. Huang and E. A. Groisman, manuscript in preparation) (Zwir *et al*., 2005). A detailed description of the profile generation process performed by GPS is presented in Supplementary Figure 3, and a comprehensive list of profiles predicted for PhoP-regulated genes is presented in Supplementary Table 3.

We demonstrated that GPS makes predictions at three levels (Zwir *et al*., 2005): (1) it recovers the canonical PhoP-regulated promoters; (2) it detects new candidate promoters for a regulatory protein; and (3) it indicates possible mechanisms by which genes previously known to be controlled by a regulator are expressed.

Profiles with canonical PhoP-regulated promoters

One of the profiles, $P_1^4E_1^4M_2^4I_3^4$ (PI = 0.39, SI = 0.07), encompasses promoters (e.g. those of the *phoP*, *mgtA*, *ybcU* and *yhiW* genes of *E.coli* and the *slyB* gene of *Salmonella*) that share the same type of RNA polymerase sites, expression patterns, PhoP box submotif and the same pattern for other transcription factor binding sites. This profile includes not only the prototypical *phoP* and *mgtA* promoters (Minagawa *et al*., 2003), but also the promoters of the *yhiW* gene, which was not known to be under PhoP control. Another profile, $P_1^3M_r^3O_2^3$ (PI = 0.23, SI = 0.13), includes promoters (e.g. those of the *hdeD*, *ompX*, *rstA*, *slyB* and *yiaG* genes of *E.coli*, and the *nmpC*, *ompX* and *pagP* genes of *Salmonella*) with a similar PhoP box orientation and type of RNA polymerase sites as the profile described above. These two profiles differ in the number of features because GPS uses a multivariate environment, where feature selection is locally performed for each profile, as not every feature is relevant for all profiles. The two profiles are also distinguished by the values of two of the features: they have distinct PhoP box submotifs and different distances of the PhoP box to the RNA polymerase sites. Thus, the canonical PhoP-regulated promoters consist of at least two distinct subsets.

Profiles with PhoP boxes in the opposite orientation of the canonical PhoP-regulated promoters

One profile, $P_2^4E_1^4I_3^4O_1^4$ (PI = 0.40, SI = 0.12), includes promoters (e.g. those of the *ompT* gene of *E.coli* and the *pipD*, *ugtL* and *ybjX* genes of *Salmonella*) that share the type of RNA polymerase sites, expression patterns and other transcription factor binding site patterns. Strikingly, the PhoP box in these promoters is in the opposite orientation relative to that found in the prototypical *phoP* and *mgtA* promoters. A second profile, $P_3^2O_1^2$ (PI = 0.07, SI = 0.17), includes promoters also with the PhoP box in the opposite orientation (e.g. those of the *slyB* and *yhiW* genes of *E.coli* and the *ybjX*, *mig-14*, *virK*, *mgtC* and *pagC* genes of *Salmonella*) but differs from the former profile in that the PhoP box is located further upstream from the RNA polymerase site than the typical PhoP-regulated gene. Both of these profiles were preserved and distinguished from each other because GPS uses a multiobjective optimization method that considers non-dominance relationships between PI (e.g. PI = 0.40 versus PI = 0.07, respectively) and SI (e.g. SI = 0.12 versus SI = 0.17, respectively).

Notably, the promoters of the latter profile could be assigned to a profile even in the absence of expression data. By virtue of being an unsupervised method, GPS is not constrained by a dependent variable (Beer and Tavazoie, 2004; Mitchell, 1997), such as expression data, which would condition the classification to the available number of expression classes. Despite the unusual orientation of the PhoP box in the promoters of the genes belonging to profiles $P_2^4E_1^4I_3^4O_1^4$ and $P_3^2O_1^2$, the identified PhoP boxes are bona fide PhoP-binding sites (Shi *et al*., 2004; Shin and Groisman, 2005; Zwir *et al*., 2005).

Profiles revealing interactions with other regulatory proteins

One profile, $P_1^3I_4^3O_2^3$ (PI = 0.21, SI = 0.17), includes promoters (e.g. those of the *yeaF* and *yrbL* genes of *E.coli* and the *pmrD*, *udg* and *yrbL* genes of *Salmonella*) that share the type of RNA polymerase sites, PhoP-box orientation and the presence of potential binding sites for the regulatory protein PmrA (Kato and Groisman, 2004). Interestingly, promoters of the *Salmonella pmrD* and *yrbL* genes and the *E.coli yrbL* gene have similarly arranged binding sites for the PhoP and PmrA proteins, suggesting that the *pmrD* and *yrbL* genes may be regulated in a similar fashion, which has been verified experimentally (Kato *et al*., 2003; Zwir *et al*., 2005). This profile was recovered by GPS, despite its potential domination by another profile [i.e. $P_1^3I_4^3O_2^3$ (PI = 0.21, SI = 0.17) versus $P_3^2O_1^2$ (PI = 0.07, SI = 0.17)] because GPS uses a multimodal optimization strategy (i.e. niching) that retrieves local optimal profiles that describe the system from different points of view.

By using gene expression as one feature among many, GPS could distinguish between promoters of the acid resistance genes (Masuda and Church, 2003; Tucker *et al*., 2002) that, otherwise, would have stayed undifferentiated within the same expression group. These promoters were found to belong to one of the three distinct profiles: $E_2^3M_0^3I_1^3$ (PI = 0.11, SI = 0.03), includes promoters for acid resistance structural genes lacking a recognizable PhoP box (e.g. those of the *dps* and *gadA* genes of *E.coli*); $E_2^2M_4^2$ (PI = 0.25, SI = 0.10), comprises promoters of a different set of structural genes that include *hdeD* and *hdeAB*; and $E_2^2P_3^2$ (PI = 0.419, SI = 0.185), harbors promoters of the acid resistance regulatory genes *yhiE* and *yhiW* (also termed *gadE* and *gadW*, respectively (Tucker *et al*., 2002; Zwir *et al*., 2005). The promoters in the latter two profiles harbor PhoP boxes but these profiles differ in the type of RNA polymerase sites and their distance to the PhoP box.

Three attributes of GPS enabled the classification of the acid resistance promoters into the three profiles. First, GPS encodes features in a flexible format, thus, it can capture promoters that would be discarded otherwise. For example, GPS found that the promoter of the *E.coli hdeA* gene has an atypical PhoP box submotif that does get footprinted by the PhoP protein (Zwir *et al*., 2005), but would have been discarded by consensus approaches (Martinez-Antonio and Collado-Vides, 2003; McCue *et al*., 2001). Second, GPS allows individual promoters (e.g. those of the *yhiW* and *hdeD* genes that were also assigned to other profiles) to belong to more than one profile, by using the fuzzy method instead of crisp clustering. Third, GPS detects cohesion even in small groups of promoters (e.g. the group that includes *yhiE* and *yhiW* genes) by evaluating profiles based on both the PI and SI instead of by the number of promoters or features. Moreover, initial assignments of features and profiles are continuously revisited by GPS due to the refinement performed during the profile learning process (e.g. the level-2 profile $E_2^2M_4^2$ is built anew from its constituent promoters, and differs slightly than those of its ancestors E_2^1 and M_4^1. See Fig. 2). For example, GPS was able to capture promoters that were initially left out: the *E.coli hemL* promoter was not considered initially because its expression level did not surpass a statistical threshold typically used in microarray experiments (Li and Wong, 2001). However, it was retrieved by its similarity with the profile comprising promoters with 'expression' E_1 and PhoP box 'submotif' M_3 (Supplementary Fig. 3), and shown to bind the PhoP protein *in vitro* (Eguchi *et al*., 2004). Moreover, GPS could also recover promoters that had been identified as PhoP-regulated using different inducing conditions than those considered in the experiments that provided the original dataset (Zwir *et al*., 2005). This allows GPS to improve upon initial decisions, based on the subsequent analysis.

GPS performance

To evaluate the ability of GPS to retrieve PhoP-regulated promoters, we analyzed the statistical significance of GPS predictions in comparison with random classifications, and then, we evaluated the ability of GPS to discriminate between promoters regulated by PhoP and by other transcription factors. First, we compared GPS prediction of the test set with a typical statistical approach consisting of randomly assigning two classes to 100 000 sets of observations with the same size of the test partition (Beer and Tavazoie, 2004). This experiment retrieved an expected ca. 50% of 'correct' classifications, following a distribution close to normal and providing a standard deviation of 9.76%. Therefore, GPS prediction of 92% for the test set is 4.3 standard deviations away from the mean obtained by random assignment, which corresponds to a *P*-value $<10^{-5}$, determined by using paired *t*-test with Bonferroni correction (Matlab statistical toolbox). These results are in agreement with a sample size >23, a power of 92% and significance level given by the stated P-value (Rosner, 1986). Second, we extended the test set by including 487 promoters from the RegulonDB database (Salgado *et al*., 2004) 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. GPS had a false positive rate of 5.3% and a 93.92% of overall performance measurement (Benitez-Bellon *et al*., 2002) as a particular correlation coefficient implementation, with a 94 and 92% specificity and sensitivity on the extended set, respectively (Supplementary Table 4).

Comparison of GPS with other methods

We evaluated the performance of GPS by comparing the set of solutions retrieved by three other well known machine learning techniques that have been used for data mining of structural databases: Bayesian Network (BN) (Cooper and Herskovits, 1992), Association Rules (AR) (Agrawal and Shafer, 1996) and Decision Trees (DT) (Quinlan, 1993) (Supplementary Text). The comparison criteria is the Pareto optimal frontier, which essentially measures the quality of the profiles in terms of their extent (PI), quality of descriptions (SI) and diversity (niches). We have illustrated this comparison between GPS and BN in Supplementary Figure 4 (similar results were obtained for comparisons with the other methods).

We summarized the comparison results between profiles retrieved by GPS, BN, AR, and DT, given in Supplementary Table 5 using a quantitative metric [see Equation (12)] [given in System and Methods] that measures the quality of the solutions selected by each method in the Pareto (Deb, 2001). According to this criterion GPS produces a better distribution of the identified profiles along the Pareto optimal frontier than the implementation used for the BN, AR and DT methods (e.g. it dominates 71% of the ones obtained by BN, while BN dominates 6% of the solutions provided by GPS; see Supplementary Information). Thus, this avoids the convergence to solutions corresponding to a single or limited regions of the search space. The diversity of these profiles provides descriptions of the PhoP regulatory network from different points of view, being mostly influenced by the presence of cohesive profiles, even those containing small sets of promoters.

Although quantitative metrics provide one estimation of the performance of the methods, the qualitative evaluation of these methods provides a more realistic assessment of their usefulness. Consequently, we use some of those profiles predicted by GPS and experimentally validated (Zwir *et al*., 2005) as a seed to evaluate the specificity and sensitivity of each method to group together promoters and features with demonstrated biological significance (Supplementary Fig. 5; Supplementary Table 6).

First, we considered the set of promoters included in the profile $P_3^2O_1^2$. The AR and BN methods grouped these promoters together, however, BN included additional promoters in its retrieved group, exhibiting downregulated values for the 'expression' feature. (The group slightly differs from the profile detected by GPS because the latter locally re-discretizes the features.) The DT method could not group together all promoters included in the profile, and also included additional promoters with PhoP boxes in the opposite orientation without discriminating between the type of RNA polymerase sites. This happens because it re-discretizes the 'RNA pol sites' into an excessively general feature (Supplementary Fig. 5a).

Second, we considered the set of promoters included in the profile $P_1^3I_4^3O_2^3$. BN retrieved the promoters of this profile, however, it was unable to describe them by the 'interactions' feature, even though this feature was crucial for identifying promoters regulated by both PhoP and PmrA proteins. As a consequence, BN retrieved a large list of other promoters that do not specifically address the biological mechanism described by the $P_1^3I_4^3O_2^3$ profile. AR retrieved the promoters in the $P_1^3I_4^3O_2^3$ profile, but only via the 'interaction' feature. Thus, by missing the 'RNA pol sites' and 'orientation' features in the group, AR produced an unspecific group that is less informative about the regulatory mechanism. DT did not group these promoters when its default parameters were used; however, it did so after customizing them (see below and Supplementary Fig. 5b).

Third, we considered the set of promoters included in the profile $E_2^3M_0^3I_1^3$. BN specifically identified together the set of promoters included in this profile. AR and DT combined the promoters in the profile with those in $E_2^2M_0^2$ profile, because neither methods were able to distinguish between the two sets of promoters by the 'interactions' feature, which identifies the acid resistance regulatory genes that regulate the promoters in the $E_2^3M_0^3I_1^3$ but not those in the $E_2^2M_0^2$ profile (Supplementary Fig. 5c and Supplementary Table 6).

Fourth, we considered the set of promoters included in the profile $E_2^2M_4^2$. Only GPS characterized them by both their 'expression' and 'submotifs' features, which are the most relevant features distinguishing these promoters from other acid resistance genes, as their expression only slightly differs from that of the canonical PhoP regulated genes, and they are directly regulated by PhoP, by the presence of PhoP binding site motifs (Supplementary Fig. 5d).

Fifth, we considered the set of promoters included in the profile $E_2^2P_3^2$. Neither BN nor AR identified the promoters included in the profile, which are crucial for inferring the architecture of the regulatory network that control acid resistance genes. DT did not group these promoter using its default parameters (see below, Supplementary Fig. 5e).

Examination of the results obtained by BN, AR and DT suggests that their deficiencies can be attributed to (1) the constrained architecture of BN, which only retrieved profiles based on their parents (i.e. profiles that only include features immediately linked in its learned structure) instead of a wider set of features, as well as their need for a sufficiently large amount of data to apply conditional probabilities successfully; (2) the thresholds used in AR, which discard profiles with few members, without considering the features that they do share; and (3) the strict dependence on output classes (e.g. 'expression') of DT, which evaluate feature partitions solely based on their ability to create distinguishable output classes, thus, producing additional partitions when a more general description is in fact more informative. DT can be customized to identify more general groups of promoters by pruning the trees at lower levels than the default implementation. However, this improvement for some groups would degrade the performance of other groups. Moreover, even if some promoters can be finally grouped together, they cannot avoid the inconsistencies caused by the forced inclusion of the 'expression' feature predicted by the method, which often differ from their original values (Quinlan, 1993).

DISCUSSION

We showed that GPS can make precise mechanistic predictions even with incomplete input dataset and high levels of uncertainty. For example, it had been suggested that PhoP regulates the *mig-14*, *virK*, *mgtC* and *pagC* genes of *Salmonella* indirectly, because a PhoP binding site could not be identified at a location typical of other PhoP-activated genes (Lejona *et al*., 2003). However, GPS grouped the promoters of these genes into a profile that shared an atypical orientation and distance of the PhoP box to the RNA polymerase site (Supplementary Figs 1 and 3). Likewise, GPS separated the PhoP-regulated acid resistance genes into three distinct profiles, allowing us to infer that the PhoP protein controls transcription of acid resistance genes, using both a feedforward loop and a classical transcriptional cascade (Zwir *et al*., 2005). The experimental verification of these predictions (Zwir *et al*., 2005) illustrates the utility of the GPS method, and demonstrates that PhoP uses multiple mechanisms for the differential regulation of genes within a regulon.

Several characteristics of GPS contribute to its power. First, it considers gene expression as one feature among many, thereby allowing classification of promoters even in its absence (Beer and Tavazoie, 2004; Conlon *et al*., 2003). Particularly, GPS differs from supervised learning methods (Mitchell, 1997) that group features and observations based on explicitly defined dependent variables (Beer and Tavazoie, 2004; Conlon *et al*., 2003; Quinlan, 1993). Second, GPS performs a local feature selection for each profile because not every feature is relevant for all profiles (Kohavi and John, 1997), and, a priori, we do not know which feature is biologically meaningful for a given promoter. This is in contrast to approaches that filter or reduce features for all possible clusters (Yeung and Ruzzo, 2001). Third, GPS finds all optimal solutions among multiple criteria (Pareto optimality) (Deb, 2001), which avoids the biases that might result from using any specific weighing scheme (Rissanen, 1989). This can detect cohesion within a small number of promoters that would remain undetected by methods that emphasize the number of promoters in a profile (Agrawal and Shafer, 1996). Fourth, GPS has a multimodal nature that allows alternative descriptions of a system by providing several adequate solutions (Deb, 2001; Ruspini, 2001), thus recovering locally optimal solutions, which have been shown to be biologically meaningful (Azevedo *et al*., 2005; Zwir *et al*., 2005). This differentiates GPS from methods that focus on a single optimum (Gutierrez-Rios *et al*., 2003; Martinez-Antonio and Collado-Vides, 2003). And fifth, GPS allows promoters to be members of more than one profile by using fuzzy clustering (Bezdek, 1998; Cordon *et al*., 2002; Gasch and Eisen, 2002), thus explicitly treating the profiles as hypotheses, which are tested and refined during the analysis (Mitchell, 1997). This distinguishes GPS from clustering approaches that prematurely force promoters into disjointed groups (Qin *et al*., 2003). In addition, GPS recognizes that not every profile is meaningful (Bezdek, 1998), which avoids the constraints of methods that force membership even to uninteresting groups because the sum of membership is required to be one (Cooper and Herskovits, 1992).

Finally, the GPS method, termed gene promoter scan here, can be generalized to a method for Grouping, Prototyping and Searching in the lattice space of hypotheses, which can be used in different structural domains. For example, it is being applied to mine the Gene Ontology database (Ashburner *et al*., 2000) to discover and annotate profiles across biological processes, cellular components and molecular functions, to identify molecular pathways that provide insight into the host response over time to systemic inflammatory insults.

ACKNOWLEDGEMENTS

The authors thank G. Stormo, E. Ruspini and C. Gu for methodological suggestions and H. Salgado for the *Salmonella* operon information. This work was supported, in part, by the grant AI49561 from the National Institutes of Health to E.A.G., who is an Investigator of the Howard Hughes Medical Institute. I.Z. is also a member of the Computer Science Department at the University of Granada, Spain, and supported, in part, by the Spanish Ministry of Science and Technology under project TIC 2003-00 and BIO2004-0270E. Funding to pay the Open Access publication charges for this article was provided by the Howard Hughes Medical Institute.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Shafer,J.C. (1996) Parallel mining of association rules. *IEEE Trans. Knowl. Data Eng*., **8**, 962–969.

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Azevedo,R.B. *et al*. (2005) The simplicity of metazoan cell lineages. *Nature*, **433**, 152–156.

Bar-Joseph,Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol*., **21**, 1337–1342.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Benitez-Bellon,E. *et al*. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA. *Genome Biol*., **3**, RESEARCH0013.

Bezdek,J.C. (1998) Pattern analysis. In Pedrycz,W., Bonissone,P.P. and Ruspini,E.H. (eds), *Handbook of Fuzzy Computation*. Institute of Physics, Bristol, pp. F6.1.1–F6.6.20.

Bezdek,J.C. and Pal,S.K. and IEEE Neural Networks Council (1992) *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*. IEEE Press, New York.

Cheeseman,P. and Oldford,R.W. (1994) *Selecting Models from Data: Artificial Intelligence and Statistics IV*. Springer-Verlag, New York.

Chickering,D.M. (2003) Optimal structure identification with greedy search. *J. Mach. Learn. Res*., **3**, 507–554.

Conlon,E.M. *et al*. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.

Consortium,E. (2002) Elvira: an environment for probabilistic graphical models. In Gamez,J.A.a.S.A. (ed.), *1st European Workshop on Probabilistic Graphical Models*, pp. 222–230.

Cook,D.J. *et al*. (2001) Structural mining of molecular biology data. *IEEE Eng. Med. Biol. Mag*., **20**, 67–74.

Collado-Vides,J. *et al*. (1991) Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev*., **55**, 371–394.

Cooper,G.F. and Herskovits,E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Cordon,O. *et al*. (2002) Linguistic modeling by hierarchical systems of linguistic rules. *IEEE Trans. Fuzzy Syst*., **10**, 2–20.

Cotik,V. *et al*. (2005) A hybrid promoter analysis methodology for prokaryotic genomes. *Fuzzy Set. Syst*., **152**, 83–102.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.

Deb,K. (2001) *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester, New York.

Eguchi,Y. *et al*. (2004) Signal transduction cascade between EvgA/EvgS and PhoP/PhoQ two-component systems of *Escherichia coli*. *J. Bacteriol*., **186**, 3006–3014.

Everitt,B. and Der,G. (1996) *A Handbook of Statistical Analysis using SAS*. Chapman & Hall, London.

Falkenauer,E. (1998) *Genetic Algorithms and Grouping Problems*. John Wiley & Sons, New York.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*., **3**, RESEARCH0059.

Groisman,E.A. (2001) The pleiotropic two-component regulatory system PhoP-PhoQ. *J. Bacteriol*., **183**, 1835–1842.

Gutierrez-Rios,R.M. *et al*. (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res*., **13**, 2435–2443.

Kato,A. and Groisman,E.A. (2004) Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor. *Genes Dev*., **18**, 2302–2313.

Kato,A. *et al*. (2003) Closing the loop: the PmrA/PmrB two-component system negatively controls expression of its posttranscriptional activator PmrD. *Proc. Natl Acad. Sci. USA*, **100**, 4706–4711.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artif. Intell*., **97**, 273–324.

Larranaga,P. and Poza,M. (1996) Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters. *IEEE J. Pattern Anal. Mach. Intell*., **18**, 912–926.

Lejona,S. *et al*. (2003) Molecular characterization of the Mg2+-responsive PhoP-PhoQ regulon in *Salmonella enterica*. *J. Bacteriol*., **185**, 6287–6294.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Li,H. *et al*. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.

Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol*., **6**, 482–489.

Masuda,N. and Church,G.M. (2003) Regulatory network of acid resistance genes in *Escherichia coli*. *Mol. Microbiol*., **48**, 699–712.

McCue,L. *et al*. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*., **29**, 774–782.

Minagawa,S. *et al*. (2003) Identification and molecular characterization of the Mg2+ stimulon of *Escherichia coli*. *J. Bacteriol*., **185**, 3696–3702.

Mitchell,T.M. (1997) *Machine Learning*. McGraw-Hill, New York.

Oshima,T. *et al*. (2002) Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12. *Mol. Microbiol*., **46**, 281–291.

Qin,Z.S. *et al*. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol*., **21**, 435–439.

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Requena,F. and Ciudad,N.M. (2000) Characterization of maximum probability points in the Multivariate Hypergeometric distribution. *Stat. Probab. Lett*., **50**, 39–47.

Rissanen,J. (1989) *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

Romero Zaliz,R., Zwir,I. and Ruspini,E.H. (2004) Generalized analysis of promoters: a method for DNA sequence description. In Coello Coello,C.a.L.G. (ed.), *Applications of Multi-Objective Evolutionary Algorithms*. World Scientific, pp. 427–450.

Ronen,M. *et al*. (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl Acad. Sci. USA*, **99**, 10555–10560.

Rosner,B. (1986) *Fundamentals of Biostatistics*. Duxbury Press, Boston, MA.

Ruspini,E.H.a.Z.I. (2001) Automated generation of qualitative representations of complex objects by hybrid soft-computing methods. In Pal,S.K. and Pal,A. (eds), *Pattern Recognition: from Classical to Modern Approaches*. World Scientific, NJ, pp. 612.

Salgado,H. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.

Saporta,G. (1996) Probabilits, analyse des donnes et statistiques, Technip, France.

Shi,Y. *et al*. (2004) Transcriptional control of the antimicrobial peptide resistance ugtL gene by the *Salmonella* PhoP and SlyA regulatory proteins. *J. Biol. Chem*., **279**, 38618–38625.

Shin,D. and Groisman,E.A. (2005) Signal-dependent binding of the response regulators PhoP and PmrA to their target promoters *in vivo*. *J. Biol. Chem*., **280**, 4089–4094.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., **22**, 281–285.

Tucker,D.L. *et al*. (2002) Gene expression profiling of the pH response in *Escherichia coli*. *J. Bacteriol*., **184**, 6551–6558.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

Zitzler,E. and Thiele,L. (1999) Multiobjective evolutionary algorithms: a comparative case study and the Strength Pareto approach. *IEEE Trans. Evol. Comput*., **3**, 257–271.

Zwir,I. *et al*. (2002) Automated biological sequence description by genetic multiobjective generalized clustering. *Ann. N. Y. Acad. Sci*., **980**, 65–82.

Zwir,I. *et al*. (2005) Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl Acad. Sci. USA*, **102**, 2862–2867.