1406–1415 *Nucleic Acids Research, 2013, Vol. 41, No. 3* *Published online 14 December 2012*
doi:10.1093/nar/gks1286

PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION

Mohamed Elati[1,*], Rémy Nicolle[1], Ivan Junier[1,2,3], David Fernández[4], Rim Fekih[1], Julio Font[4] and François Képès[1,*]

[1]Institute of Systems and Synthetic Biology, CNRS, University of Evry, Genopole, 91030 Evry, France, [2]Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain, [3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain and [4]Noray Bioinformatics, S.L.U.Parque Tecnológico 801A, 48160 Derio-Bizkaia, Spain

Received September 17, 2012; Revised November 5, 2012; Accepted November 10, 2012

ABSTRACT

Conventional approaches to predict transcriptional regulatory interactions usually rely on the definition of a shared motif sequence on the target genes of a transcription factor (TF). These efforts have been frustrated by the limited availability and accuracy of TF binding site motifs, usually represented as position-specific scoring matrices, which may match large numbers of sites and produce an unreliable list of target genes. To improve the prediction of binding sites, we propose to additionally use the unrelated knowledge of the genome layout. Indeed, it has been shown that co-regulated genes tend to be either neighbors or periodically spaced along the whole chromosome. This study demonstrates that respective gene positioning carries significant information. This novel type of information is combined with traditional sequence information by a machine learning algorithm called PreCisIon. To optimize this combination, PreCisIon builds a strong gene target classifier by adaptively combining weak classifiers based on either local binding sequence or global gene position. This strategy generically paves the way to the optimized incorporation of any future advances in gene target prediction based on local sequence, genome layout or on novel criteria. With the current state of the art, PreCisIon consistently improves methods based on sequence information only. This is shown by implementing a cross-validation analysis of the 20 major TFs from two phylogenetically remote model organisms. For *Bacillus subtilis* and *Escherichia coli*, respectively, PreCisIon achieves on average an area under the receiver operating characteristic curve of 70 and 60%, a sensitivity of 80 and 70% and a specificity of 60 and 56%. The newly predicted gene targets are demonstrated to be functionally consistent with previously known targets, as assessed by analysis of Gene Ontology enrichment or of the relevant literature and databases.

INTRODUCTION

Transcription factors (TF) regulate gene expression through their physical interaction with DNA at specific regulatory elements termed TF binding sites (TFBS). Genome-wide TFBS identification has drawn substantial interest in the recent years, as it represents a critical step in delineating transcription regulatory networks. Previous studies have used both experimental and computational techniques to identify or to predict TFBS. However, traditional experimental techniques, such as DNase I footprinting and gel-mobility shift assay, are time-consuming and are not suitable for genome-scale studies. Although current high-throughput approaches, such as ChIP-chip and ChIP-seq, are more efficient in determining the binding specificity at a large scale (1), they are too costly for daily applications. Efficient computational approaches using cheap and readily available genomic sequence data are therefore most welcome. Such methods can be used, in particular, to complement analysis of high-throughput data. Indeed, binding sites detected by high-throughput *in vitro* methods can be compared with predicted binding sites to prioritize studies aimed at confirming sites that are expected to regulate gene expression *in vivo*. A number of computational methods have been developed for predicting TFBS, given a set of known binding sites. Commonly used methods are based on the definition of a consensus sequence or the construction of a position-specific weight matrix (PWM), where DNA binding sites are represented as a sequence of letters

*To whom correspondence should be addressed. Tel: +33 169474443; Fax: +33 169474437; Email: mohamed.elati@issb.genopole.fr
Correspondence may also be addressed to Francois Kepes. Tel: +33 169474431; Fax: +33 169474437; Email: francois.kepes@epigenomique.genopole.fr

coming from the alphabet $\{A, T, C, G\}$. They then use the PWM of the binding site to scan new sequences for additional binding sites (2). As TFBSs are, in general, relatively short and degenerate, these approaches systematically lead to a high rate of false positives (FPs) (1,3). In spite of the wealth of research performed in the area of TFBS prediction, and the many insights gained, achieving a qualitative jump in this field would require information of a conceptually novel type, rather than improvements of methods, which all rely essentially on local sequence information (4). Here, we propose to additionally derive useful information from the respective positioning of co-regulated genes along the chromosome.

Proper genome-wide coordination of gene expression has been shown to be linked to the spatial organization of genes within the cell (5). In particular, transcriptional activity is often detected in discrete foci called transcription factories, rather than in a diffuse pattern (6,7). These transcription factories gather RNA polymerases, TFs and genes that can be far apart along the DNA (8). Recent experiments have further shown that genes within a given transcription factory share similar promoter sequences (9) and a homogeneous TF content (10). In this respect, the periodic organization observed for co-regulated genes (11) or evolutionarily correlated genes (12,13) in bacteria has been shown to be crucial for achieving chromosome conformations that favor the formation of these transcription factories. This was demonstrated using a thermodynamic model of chromosome folding where distal binding sites can be cross-linked by bivalent TFs (14). In particular, chromosomal proximity or periodicity were shown to respectively favor rosette-like or solenoid-like structures, consistent with previously published functional models [respectively, (6,15)]. In this manner, these two families of chromosomal conformations favor the spatial proximity of related TFBS, thus building local concentration effects, which in turn optimize transcriptional repression or activation (16).

The rationale proposed in this article is to combine TFBS nucleic sequence information with gene positional information to obtain an accurate and robust TFBS prediction model. This combination must itself be optimized to achieve a high classification performance. To our knowledge, no research exists addressing this question.

For this purpose, we model the TF–DNA binding problem as a multi-views classification problem (17), and we propose a variation of the AdaBoost algorithm (18) to fuse the classifiers. A key aspect of the boosting technique is that it forces some of the base classifiers to focus on the boundary between positive and negative examples, thus effectively reducing classification errors.

A preliminary outline of PreCisIon has been published in a workshop (19). In this article, we demonstrate the power of this approach by extensive empirical studies performed on a benchmark data set from two distinct bacteria: gram-negative *Escherichia coli* and gram-positive *Bacillus subtilis*. The PreCisIon software is made available at http://www.noraybio.com/en/gennetec.asp.

The article is organized as follows. First, the base classifiers used for TFBS prediction based on binding sequence and on chromosomal position are described. Next, the classifier fusion algorithm is introduced. Finally, results are presented and discussed.

MATERIALS AND METHODS

Outline of the method

PreCisIon is a general supervised method to infer new regulatory relationships between a known TF and all the genes of an organism. In its current form, it requires two types of data as inputs. First, each gene in the organism must be characterized by some properties (views), here two views: its promoter sequence and its chromosomal position. Second, for each TF, a list of its known targets genes and, if possible, of its known non-targets is needed. Such lists can be constructed from publicly available databases of experimentally characterized regulations, e.g. RegulonDB for *E. coli* genes (20). PreCisIon splits the problem of regulatory network inference into many binary classifications from disjoint views. For each view, PreCisIon trains a binary classifier to discriminate between genes known to be regulated and non-regulated by the TF. In this article, we introduce a new chromosomal position view to benefit from information pertaining to spatial chromosome conformation. The final step is to combine all individual classifiers that have been trained on disjoint views. Once trained, the model associated with a given TF is able to assign a class to each new gene, which has not been used during training.

Weight matrix-based TFBS

The Sequence classifier is structurally divided in two phases: PWM creation and TFBS Prediction. A PWM is generally learned from a collection of aligned DNA binding sites that are likely to bind a common TF. Given a learned PWM, the sum of the elements that correspond to a specific sequence s gives a total score for that sequence. This allows the model to provide a binding score BS to all possible binding sites for the protein:

$$BS(s) = \sum_{k=1}^{K} \sum_{b \in [A,C,G,T]} w_{b,k} I_{b,k}(s) \quad (1)$$

where $w_{b,k}$ is a weight assigned to each possible base $b \in [A,C,G,T]$ at each position k in the binding site and $I_{b,k}(s) = 1$ if base b occurs at position k of sequence s and 0 otherwise. The higher the score, the more likely a site will be bound by the TF.

For each phase, many algorithms have been developed (3). In our study, we use the classical packages called: 'MotifSampler' (21) for the first phase and 'Patser' (2) for the second phase.

Gene position along the chromosome

The positional regularities of a set of TF-target genes are assessed using the solenoidal coordinate method (22). In this method (see Figure 1), the score at a given period reflects the likelihood for the data set to present a periodic pattern with this period. A high score stems from (i) the remarkable alignment properties of periodic

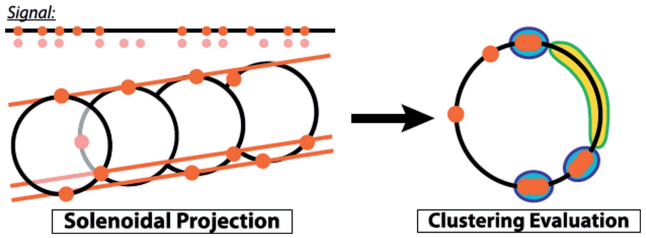

Figure 1. Principle of the Solenoidal Coordinate Method (SCM). A set of gene positions (red dots along horizontal line, upper left corner) derives from a perfectly *P*-periodic pattern (blurred red dots). Some of the initially periodic genes are missing (false negative) or have different positions (noise), and random genes have been added (false positive). Although the resulting pattern looks aperiodic, the position of the genes in a solenoidal coordinate of period *P* (lower left) reveals some alignment properties. The algorithm provides a score that is built using a distance-based information content for the organization of the genes on the solenoid face view (lower right), rewarding exceptionally dense or void regions (22). This information content is computed for all periods, which leads to a spectrum. The peaks that are abnormally high in this spectrum then reveal the periodic tendencies. Note that high scores at the period equal to full chromosome length reflect chromosomal proximity, as in this case, the solenoid is composed of only one loop, which is the whole chromosome itself. For a given period, SCM allocates a positional score to each gene, equal to the co-logarithm of the likelihood for this gene to be periodically positioned in dense regions with respect to the other genes of the data set.

positions when they are represented in a solenoidal coordinate system with the right period and (ii) the use of an information-theoretic measure *à la* Shannon that rewards both exceptionally dense and void regions of the solenoid [see (22) for details]. The period equal to full chromosome length plays a singular role in the analysis. Indeed, for this period, the 'solenoid' is composed of only one loop. Thus, the analysis does not bear on periodicity but on proximity along the chromosome. Accordingly, scores at this peculiar period are referred to as proximity scores. To build the positional classifier, both chromosomal proximity and periodicity of training genes are captured to generate a 'spectrum' of positional scores for all genes in the genome as a function of the period.

An algorithm on multi-view classifier fusion

Combination of multiple classifiers is an important research topic in the field of machine learning, and it is widely discussed in the literature (23). Methods for classifier fusion range from non-trainable combiners like the majority vote or simple functions, to sophisticated methods that require an additional training step. Other methods such as boosting and bagging (24) have been introduced to cope with the diversity in the classifier opinions. For instance, AdaBoost (18) has been shown to improve the prediction accuracy of weak classifiers using an iterative weight update process. The technique combines weak classifiers (classifiers having classification accuracy slightly greater than chance) in a weighted vote, resulting in an overall strong classifier.

A sequence classifier reported as a weight matrix assumes that different positions of the motif are independent. Under this assumption, a PWM is essentially a linear classifier when used with a cutoff value to predict binding sites in sequences. Same remarks for position classifier. Previous work, which only combined classifiers of a same type using boosting strategies to accommodate some non-linear factors in discriminating positive and negative examples, has produced TFBSs that achieve good accuracy in the context of building an integrated yeast regulatory network (25).

One of the ways boosting may be used for classifier fusion would be to run boosting separately on each view, obtain separate ensembles for each view and take a majority vote among the ensembles when presented with test data. In this case, separate training of classifiers is needed for each view, and the sampling distributions of the data points are also disjoint. Unlike this approach, we modify the boosting algorithm AdaBoost to train a TFBS classifier, as an ensemble model, on different views of the training data. We perform separate training for each view, but the cycle of error computation and example sampling is done using a shared distribution of example weights in a given iteration.

Suppose we have a set G of n training examples. Each example having two disjoint views (Sequence: S and Position: P) such that a given training gene g_i can be represented as $(g_i^{S,P},c_i)$, where $c_i = -1.1$ for correct and mis-classification, respectively. Weak Classifiers h^S and h^P will be trained on the training sets $G^P = \langle (g_1^S,c_1),(g_2^S,c_2),\cdots,(g_n^S,c_n)\rangle$ and $G^P = \langle (g_1^P,c_1),(g_2^P,c_2),\cdots,(g_n^P,c_n)\rangle$, respectively.

In the initialization step of Algorithm 1, all the views for a given training gene are initialized with the same weight. We modify the boosting algorithm by adding more initial weights to the minority class examples such that the initial total weights of two classes are equal.

As the sampling distribution for all views of a given example is shared, the sampling weight of the S and P views of example g_i in iteration t are given by $w_t^{S,P}(i) = w_t^S(i) = w_t^P(i)$. After a classifier h_t^* with lowest error rate ϵ_t^* is selected in step 4 of Algorithm 1 and combination weight α_t^* is obtained, the sampling weights for the S and P views will be updated (step 5 of Algorithm 1).

Weights of S and P views of a training example g_i are updated based on whether the winning weak classifier h_t^* classifies g_i correctly. As a result, the sampling distribution of the weights remains the same for all views. This allows the most consistent data type to dominate over time, thereby significantly reducing sensitivity to noise. The selected weak classifiers are then combined, using the approach of weighted majority vote, into a unique classifier. This unique non-linear classifier is strong because (i) its components each fit well to a particular region of the landscape and (ii) it contains classifiers that are trained to focus on different views of the data.

Independently, AdaBoost has recently been investigated in multi-view learning with application to image analysis, e.g. face detection (26,27). In particular, there is a close relationship between all these variants of multi-view Adaboost. If we have a single view then all variants reduce to AdaBoost. The slight difference of the abovementioned variants is how to combine the best classifiers of each view in the same iteration. In principle, there are two approaches to combining classifiers, namely

classifier fusion and classifier selection. In classifier fusion, the two best classifiers for each view are added to the output hypothesis, and the weight update process is adjusted using further parameters to express their concordance. However, if the selected classifiers are completely redundant or too strong, it may converge too rapidly, inhibiting the positive effects of boosting. In classifier selection, only the best weak classifier is added to the output hypothesis in each iteration such that the least redundancy is maintained. Importantly, (27) show that a lower training and generalization error bound can be achieved if a shared sampling distribution is used and a weak classifier from the lowest error view is selected. We support this theoretical conclusion with our empirical studies (see the 'Results' section).

Algorithm 1: Pseudo-code description of the classifier fusion algorithm
Input:

- N training examples (genes), with 2 views (promoter sequence s and chromosome position p) available for each gene i and hence 2 training sets $\langle(s_1,y_1),(s_2,y_2),\cdots,(s_N,y_N)\rangle$ and $\langle(p_1,y_1),(p_2,y_2),\cdots,(p_N,y_N)\rangle$; of which a genes have $y_i = +1$ (TF target) and b genes have $y_i = -1$ (not TF target);
- The maximum number k_{max} of individual classifiers to be combined;

Initialize example weights:

$\omega_1(i) = \frac{1}{2a}, \frac{1}{2b}$ for $y_i = +1, -1$, respectively.
For $k = 1,...,k_{max}$ **do:**

(1) Train Sequence and Position classifiers, using the distribution ω_k;
(2) Obtain the error rates $\in_k$ of each classifier C: $\epsilon_k^C = P_{i\sim\omega_k}[C_k(g_i) \neq y_i]$;
(3) Select the individual classifier C_k^* with the lowest error rate;
(4) Compute the value $\alpha_k^* = \frac{1}{2} ln \frac{1-\epsilon_k^*}{\epsilon_k^*}$;
(5) Update examples' weights: $\omega_{k+1}(i) = \frac{\omega_k(i)e^{x\alpha_k^*}}{Z_k^{(*)}}$, where $x = -1,1$ for correct and mis-classification, respectively, and $Z_k^{(*)}$ is the normalizing factor so that $\omega_{k+1} = 1$.

end For
Final hypothesis:

$H(i) = sign(\sum_{k=1}^{kmax} \alpha_k^* C_k^*(i))$

Implementation

A public version of the PreCisIon tool is available at http://www.noraybio.com/en/gennetec.asp. PreCisIon is implemented in JAVA, and the program is available through a user friendly interface connected to a MySQL database.

Experimental protocol

Data preparation

As a proof of concept, the 10 TFs having the highest number of gene targets were selected for both *E. coli* and *B. subtilis*. The use of these particularly well-studied model organisms ensures that the available annotations are among the most complete. It also allows to cover a broad range of the bacterial phylogeny, as *E. coli* is gram-negative and *B. subtilis* is gram-positive. For each organism, the classifiers were built on the positions and the sequences of their transcription units (TU). One TU expresses one mRNA. It may encode one protein or several proteins. In the latter case, it is called an operon and contains several cistrons, each encoding one protein. To focus on transcriptional regulation unhindered by the operonic organization of bacterial chromosomes, the data set was reduced to TUs, i.e operons plus non-operonic genes. Otherwise, the data set would contain replicates (cistrons of the same operon) and would thus artificially inflate the classifiers. In other words, if cistrons within an operon are split between a training and a test set (see the 'Performance evaluation' section), then the prediction is likely to be correct simply because the classifier will predict that a test cistron with a promoter similar to a training cistron should be in the same class.

The list of operons in *E. coli* and *B. subtilis* was downloaded from RegulonDB (20) and DBTBS (28), respectively. It contains 899 (resp. 633) operons. Non-operonic genes were added, resulting in a total of 3360 (resp. 3426) TUs from the initial 4345 (resp. 4100) genes. For each TU, two features were associated, i.e. the promoter sequence and the start position. The tool 'retrieve-seq' of the 'Regulatory Sequence Analysis Tools' [(29), http://rsat.ulb.ac.be/rsat/] was used to retrieve upstream regulatory sequences ('promoters') defined here by the DNA sequence between position −400 and −1. Experimentally validated TF–gene regulations were downloaded from RegulonDB and DBTBS as well.

Choice of negative examples

Although regulatory interactions reported in databases such as RegulonDB can safely be taken as positive training examples, the choice of negative examples is more difficult for two reasons. First, few information is published and archived regarding the fact that a given TF is found not to regulate a given target gene. Hence, there is no systematic source of negative examples for our problem. A natural choice is then to consider genes not reported to have regulatory relations in databases as negative examples, mixing both true and false negatives. In that case, we are confronted with the second problem, which is that, once a classifier is trained on positive and negative examples, it always predicts significantly negative scores on negative examples used during training. To overcome this difficulty, we propose the following scheme. Let us suppose we want to predict whether genes in a set are regulated by a given TF. All genes known to be regulated by this TF form a class of

positive examples, and no prediction is needed for them. The remaining genes are split in three subsets of roughly equal size. In turn, each subset is taken apart, and PreCisIon is trained on all the positive examples plus all genes in the two other subsets, considered as negative examples. PreCisIon is then tested on the third subset, which has not been used during training. Rotating three times over the three subsets allows PreCisIon to attribute a prediction to each unlabelled gene by using an independent model. Second, the resulting number of negative sequences (usually in thousands) is often much larger than the positive ones (usually <100). Without proper adjustments, negative examples would overwhelm a classifier and reduce its capability to recognize positive examples. As a remedy, we constrain the total weight of the positive examples to be equal to that of the negative examples. The sum of weights within each class by default equals 0.5 so that the overall sum is 1 (see Algorithm 1). This in effect imposes a higher penalty for misclassifying a positive sequence than misclassifying a negative one. In initialization step, within each of those two classes, all examples have equal weight.

Performance evaluation and comparison

To assess the performance of PreCisIon, and compare it with other existing methods, it was tested on a benchmark of *E. coli* and *B. subtilis* TFs. We adopt a 3-fold cross-validation strategy, coherent with the PreCisIon protocol used to make predictions as explained in the previous section. Given a positive set of known targets and a negative set containing all other TUs of the organism, we split randomly these two sets of TUs in three parts, train PreCisIon on two of these subsets and evaluate their prediction quality on the third subset, i.e. on both positive and negative TUs that were not used during training. This process is repeated three times, testing successively each subset, and the prediction qualities of all folds are averaged and used to compute performance measures:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, FPs and false negatives, respectively. The receiver operating characteristic (ROC) curve, which is one of the most robust approach for classifier evaluation and comparison (30), can then be drawn by plotting the true positive rate (i.e. sensitivity) against the FP rate (i.e. 1 – specificity). The ROC curve was generated by varying the output threshold of a classifier and plotting true positive rate against FP rate for each threshold value, using the ROCR package (31). The area under the ROC curve (AUC) can be used as a reliable measure of classifier performance (32). As the ROC plot is a unit square, the maximum value of AUC is 1, which is achieved by a perfect classifier. Weak classifiers have AUC values slightly >0.5.

Individual classifiers parameters

As Sequence and Position classifiers each provide an output score for any TU, a threshold is used to discretize the score values to obtain binary decisions (class -1 for non-target genes and class 1 for target ones). ROC curves can be used to select the optimal decision threshold by maximizing any pre-selected measure of efficiency (e.g. accuracy). In this study, we used the true rate as proposed by (33), which is equal to the sum of the true positive rate and the true negative rate as follows:

$$TrueRate = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \quad (4)$$

The true rate is more relevant than the accuracy whenever the ratio of positive instances versus negative ones is large, or whenever it is small as in our case (33).

RESULTS

Chromosomal position carries information and PreCisIon exploits it

The 10 TFs with the highest number of gene targets were analysed for both *E. coli* and *B. subtilis*. To assess the relevance of chromosomal relative position of the target genes for TFBS prediction, the TF showing the highest Position score, Lrp, was subjected to ROC curve analysis (Figure 2a). The more accurate a classifier is, the closer is its ROC curve from the left and top borders of the plot, and consequently the larger the area under ROC curve (AUC). Although the best Sequence classifier is rather weak for Lrp (AUC 0.56), the best Position classifier is very strong (AUC 0.74), and boosted combination of the two views does not improve the AUC (0.73). Hence, this is a case where chromosomal position carries highly significant, irreplaceable information for TFBS prediction. However, there may be cases where the Position view alone brings little information; yet, it synergistically combines with the Sequence view. This appears to be the case for SigD (Figure 2b). Although the best Position classifier is ineffective (AUC 0.51), the best Sequence classifier is effective (AUC 0.76); yet, PreCisIon clearly surpasses them (AUC 0.90). Even when both views are little informative, their optimized combination may be effective. The case of CRP illustrates this possibility, with AUCs of 0.54 and 0.53 for Sequence and Position, respectively, and a PreCisIon AUC of 0.70 (Figure 2c).

The above three diverse cases illustrate the power of using chromosomal gene position for TFBS prediction. The two last cases further emphasize the crucial importance of building a view on several weak classifiers rather than on only the best one of them, and the value of fusing Sequence and Position views. In this way, the Boosting algorithm is able to draw great benefit from minute Sequence and/or Position informations.

To evaluate how general this prediction improvement is, all 20 TFs with highest number of gene targets were separately subjected to ROC analysis. The averaged results are shown in Figure 2d. Overall, the Position classifiers are weaker than the Sequence ones, but their combination with PreCisIon significantly exceeds both. The most

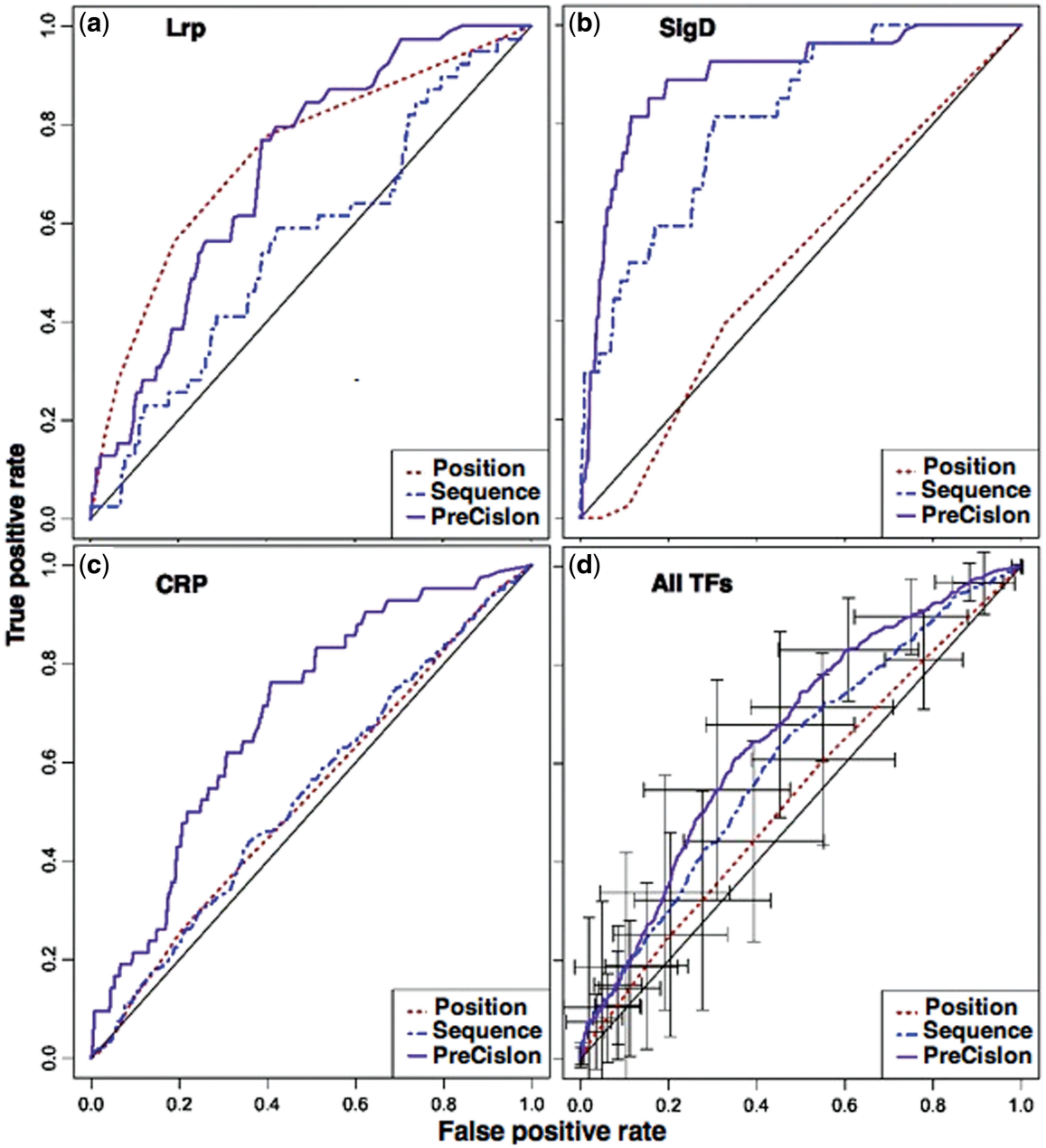

Figure 2. ROC curves on a test set for Position classifier, Sequence classifier and PreCisIon. The curves shown here are the individual curves of three TFs: Lrp, SigD and CRP (panels **a–c**); the average of all the individual ROC curves obtained for each single TF, using the 'threshold averaging' method (32). The standard deviation bars indicate the variation around the average curve (panel **d**). The gray diagonal denotes the ROC curve of a random classifier.

effective position classifiers were observed to capture gene periodicity rather than proximity (not shown). These 20 TFs were also studied one by one (Table 1). For four TFs only (SigA, SigE, SigG, CcpA), the combined AUC is slightly decreased over Sequence view alone. In such cases, because PreCisIon allows the most consistent data type to dominate over time ('Methods' section), it yields results based solely on sequence information. An important limitation of classical sequence-based TFBS prediction is the high proportion of false predicted targets (FPs). The most significant contribution of PreCisIon to this problem is to consistently reduce the number of FPs while loosing very few or no true positives. Moreover, true positives may even be sometimes retrieved from false negatives, as demonstrated by the Fur case (Table 1). Beyond the universal increase in specificity, Fur also displays an increase in sensitivity from 0.71 for Sequence alone to 0.86 for PreCisIon. Accordingly, an analysis of the 45 Fur gene targets indicates that PreCisIon identifies seven true targets that were classified as negative by Sequence alone.

Overall, the Sequence classifier achieved on average an AUC of 56 and 68%, a sensitivity of 87 and 92% and a specificity of 15 and 30%, for *E. coli* and *B. subtilis*, respectively. Importantly, a weak classifier can be built using

Table 1. Prediction performance

TF	NG	Sequence classifier			Position classifier			PreCisIon		
		AUC	Sn	Sp	AUC	Sn	Sp	AUC	Sn	Sp
Escherichia coli										
CRP	293	0.54	0.83	0.17	0.53	0.47	0.59	0.70	0.74	0.64
FNR	132	0.56	0.86	0.12	0.50	0.49	0.50	0.68	0.53	0.51
IHF	107	0.55	0.83	0.21	0.48	0.38	0.61	0.56	0.75	0.43
Fis	86	0.60	0.85	0.08	0.53	0.21	0.76	0.59	0.73	0.52
ArcA	80	0.49	1.00	0.13	0.54	0.41	0.62	0.53	0.60	0.48
H-NS	75	0.59	0.84	0.11	0.48	0.16	0.75	0.58	0.68	0.43
Fur	45	0.66	0.71	0.15	0.54	0.35	0.8	0.64	0.86	0.43
Lrp	41	0.56	0.89	0.13	0.74	0.56	0.82	0.73	0.71	0.65
CpxR	36	0.56	0.69	0.23	0.46	0.05	0.85	0.56	0.55	0.59
NarL	36	0.55	0.58	0.33	0.57	0.14	0.87	0.65	0.58	0.62
Bacillus subtilis										
SigA	277	0.55	1.00	0.12	0.47	0.57	0.40	0.51	0.52	0.60
SigE	65	0.60	0.98	0.07	0.54	0.57	0.65	0.50	0.78	0.38
SigB	63	0.65	0.98	0.14	0.51	0.27	0.72	0.68	0.79	0.30
SigG	52	0.65	0.98	0.07	0.51	0.27	0.72	0.60	0.62	0.49
SigK	46	0.69	0.81	0.19	0.45	0.26	0.77	0.71	0.71	0.51
Ccpa	38	0.72	0.92	0.30	0.50	0.16	0.81	0.68	0.92	0.52
LexA	33	0.79	0.97	0.44	0.54	0.18	0.88	0.82	0.81	0.76
AbrB	33	0.64	0.78	0.49	0.52	0.27	0.80	0.68	0.51	0.70
SigW	32	0.64	0.80	0.32	0.51	0.20	0.79	0.75	0.76	0.66
SigD	29	0.76	0.74	0.38	0.51	0.12	0.84	0.90	0.73	0.70

Area under ROC curve (AUC), sensitivity (Sn) and Specificity (Sp) of all the tested TFs from *E. coli* and *B. subtilis* on test set (3-fold cross-validation). NG: Number of target Genes.

only Position information. This classifier achieved an AUC of 54 and 51%, a sensitivity of 38 and 25% and a specificity of 78 and 80%, for *E. coli* and *B. subtilis*, respectively. By combining Position and Sequence views, PreCisIon gave rise to the highest performance by achieving on average an AUC of 60 and 70%, a sensitivity of 70 and 80% and a specificity of 56 and 60%, for *E. coli* and *B. subtilis*, respectively (Table 1).

Fusion using boosting approach outperforms other types of combined classifiers

To compare the boosting approach adopted here to other classifier fusion methods, PreCisIon was compared with other approaches ranging from simple non-trainable methods, e.g. linear combination, to more sophisticated ones based on an additional training step, e.g. Stacked generalization (34). Stacked generalization applies a learning algorithm, e.g. a 'Naive Bayes' classifier, to learn how to combine the predictions of the base-level classifiers. The resulting meta-level classifier is then used to obtain the final prediction. An ROC analysis shows that PreCisIon is more efficient than the two other combination methods (Figure 3a). Importantly, the enhanced predictive quality of PreCisIon is not only due to the boosting procedure as proposed by (25). It definitely results from combining both Sequence and Position views, as each view separately subjected to the boosting procedure is less accurate as shown in Figure 3b.

PreCisIon predicted gene targets are functionally relevant

As new TFBSs were predicted purely by computational methods, additional evidences were sought to support the functionality of the new regulatory links obtained for the *E. coli* TFs. These lines of evidence were based on available functional genomics information.

Functional analysis of predicted gene targets of global TFs

A Gene Ontology (GO) enrichment analysis was used to characterize the biological functions of newly predicted targets of global regulators. We next compared these results against those coming from the set of known curated targets. For each of the five TFs with the highest number of targets in RegulonDB (CRP, FNR, IHF, FIS and ArcA), the analysis was carried out using UniProt GO annotations from the European Bioinformatics Institute at http://www.ebi.ac.uk/GOA/proteomes.html and the GO enrichment analysis software DAVID (35). In Table 2 is listed for each TF the top-ranked GO category among its predicted targets along with the enrichment *P*-value, as well as the *P*-value for the same category among the known targets. It appears that for CRP, ArcA and IHF, the top ranked GO category based on the predicted targets is also significant for the known targets. These results support the assignments made by PreCisIon and indicate that the newly predicted targets for most TFs can be used to correctly extend our understanding of the function of these TFs. The FNR case is complicated by its numerous functional categories, but 'metal ion transport' is indeed a major one among them (36). By contrast, the newly predicted FIS targets are functionally biased with respect to its known targets. Interestingly, this bias is towards the main functional category of CRP: carbohydrate catabolic process. This suggests a partial overlap of the functions of these two TFs, which turns out to be corroborated by the

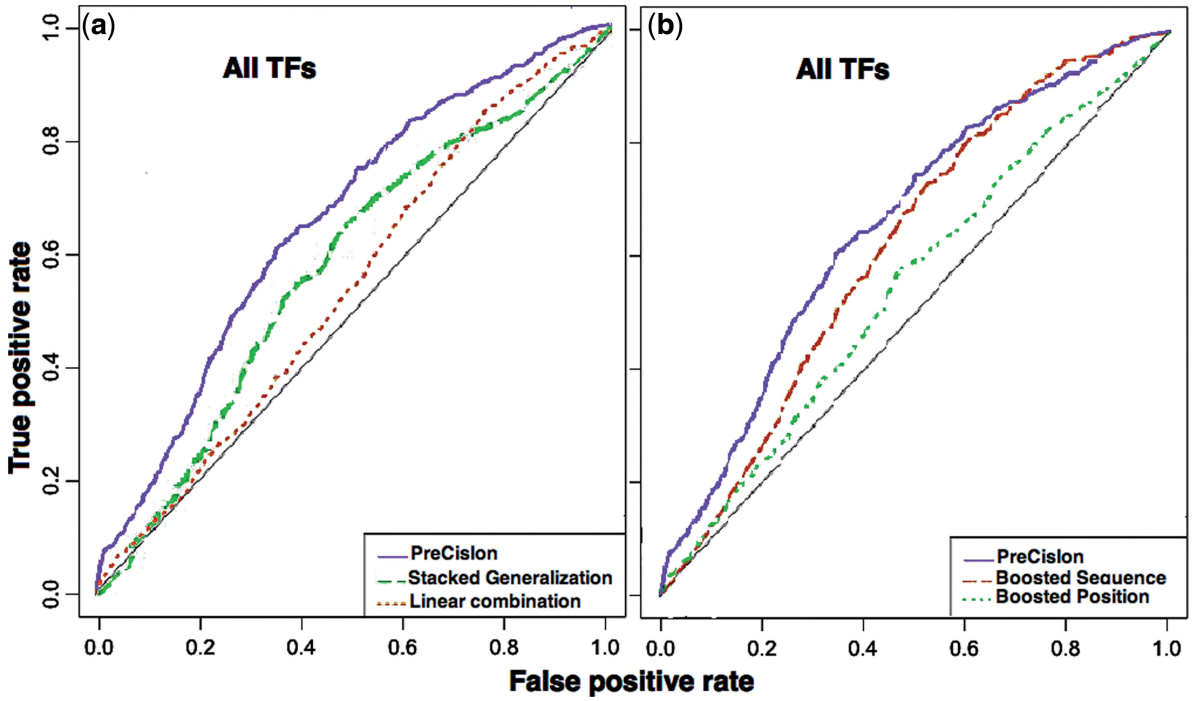

Figure 3. Methods combining Sequence and Position classifiers. (**a**) ROC analysis comparing three classifier fusion algorithms: PreCisIon (classifier fusion using boosting), linear combination based on average and Stacked generalization based on Naive Bayes learning; (**b**) ROC analysis of the boosted contributions of the individual Sequence or Position view, and of their combination into PreCisIon.

Table 2. Functional validation of predicted gene targets against GO

TF	Top GO category	p_v PT	p_v KT
CRP	carbohydrate catabolic process	3×10^{-9}	6×10^{-7}
FNR	metal ion transport	4×10^{-3}	9×10^{-1}
IHF	nitrogen biosynthetic process	2×10^{-11}	6×10^{-9}
FIS	carbohydrate catabolic process	6×10^{-11}	5×10^{-2}
ArcA	cellular biosynthetic process	2×10^{-9}	1×10^{-8}

This table shows the most significant GO categories for newly predicted gene targets for *E. coli* TFs obtained by applying PreCisIon to the most curated known targets (KTs) from RegulonDB. The table compares the enrichment *P*-value (p_v) of this category for the newly predicted targets (PTs) and known targets. The reported uncorrected *P*-values are based on the 'EASE Score' (35), a modified Fisher test, for gene-enrichment analysis.

literature. Indeed, FIS indirectly regulates expression of several operons involved in catabolism of sugars and nucleic acids, which are also under the direct control of CRP (37). FIS also affects CRP expression (38).

Validation against functional genomics data

The significativity of PreCisIon improvements can be tested by comparing its predicted TFBSs with those determined by genome-wide condition-specific assays such as ChIP-chip and ChIP-seq. The three *E. coli* TFs Lrp, FNR and CRP were chosen because (i) they are in the top five with the most significant periodical distributions of their gene targets (not shown) and (ii) they have a large number of known targets that can be used for training. The list of targets identified by ChIP-chip was taken from (39) for Lrp and from (40) for FNR and CRP. These targets were assumed to form, to a first approximation, the complete set of TFBS. For both Sequence classifier and PreCisIon, the approximate number of FPs was determined by subtracting the number of predictions that matched experimentally defined TF targets from the total number of predicted targets (Table 3). Compared with the Sequence classifier alone, PreCisIon generated a much reduced proportion of FPs, under 15%, whereas the overall number of hits was much less reduced to 21%. Altogether, it appears that PreCisIon strongly boosts the specificity of TF gene target prediction, and that newly predicted targets are functionally and experimentally consistent with prior knowledge about their TFs.

DISCUSSION

Inference of TFBSs is a difficult bioinformatics problem because it relies on short and degenerate DNA binding sequences and because for most TFs, the small number of their characterized gene targets impedes the learning procedure. An ability to better predict TFBSs from

Table 3. Validation against ChIP-chip data

TF	TG	P	Sequence classifier			PreCisIon		
			AUC	PT	FP	AUC	PT	FP
CRP	293	19 015	0.54	1862	1698	0.70	428	264
FNR	132	107 899	0.56	1068	937	0.68	162	82
Lrp	41	149 666	0.56	986	838	0.73	259	171

This table shows the number of documented target genes (TG) and the identified periods in the data (P) for the TFs CRP, FNR and Lrp. It also shows for both Sequence classifier and PreCisIon the area under the ROC curve (AUC), the number of predicted targets (PT) genes and the number of FPs using ChIP-chip data as references.

small training data sets would therefore advance our understanding of TF–DNA binding specificity, regulation and coordination of gene expression, and ultimately gene function. It would also contribute to fill in part the remaining *terra incognita* of transcriptional interaction maps, thus improving our ability to understand and ultimately control regulatory disorders and disease.

All computational approaches so far have relied on local sequence information only, in a way or another. In this article, we show that for bacteria, respective gene positioning along the chromosome carries significant information for TFBS prediction. This global positional information fundamentally differs from local sequence information. As a result, they can be combined to significantly improve genetic network inference. On this basis, we set up a tool named PreCisIon to optimize this combination using a powerful machine-learning algorithm. Validation on data sets from two phylogenetically remote bacteria shows that indeed, PreCisIon improves the specificity of TFBS prediction. In most cases, this improvement goes well beyond that achieved by the best classifier of positional information alone. It is also more effective than the boosted classifiers of either sequence or position view alone. This demonstrates the importance of classifier fusion with the Boosting algorithm to greatly benefit from even small sequence and/or position informations. Importantly, PreCisIon appears to be an appropriate first step in bacterial TFBS detection study. Indeed, its specificity consistently surpasses the sequence view alone. Furthermore, it is most often superior to either view according to ROC analysis, and when it is not, its output directly points to the view, which should preferentially be used. The learning set size ranged from 293 (CRP) to 29 (SigD) known targets genes. As the performance evaluation was done using a 3-fold cross-validation, these results were computed from training sets as small as two thirds of the initial data. Even so, the performance on SigD with a training set of ~20 targets was good. Indeed, PreCisIon and classifier fusion techniques in general do not require a larger training set than the individual classifier does.

Any advance in genome description or sequence-based TFBS prediction algorithmics can be readily incorporated into PreCisIon. Future work should focus at (i) extending the multi-view learning algorithm to other existing features of transcriptional control, for instance conservation information (41), the positioning of sites with respect of the Transcription Start Site of the TU (42), co-expression information (43), TF co-operativity (4,44); (ii) adapting PreCisIon to eukaryotic organisms. Indeed, we showed in 2003 that target genes tended to be periodically positioned also in eukaryotic yeast (45). Furthermore, recent studies have linked the three-dimensional structuring of eukaryotic chromosomes to their gene expression (46,47). Finally, our preliminary results indicate potential success in applying the principle of global gene position to TF binding site prediction in yeast. However, they also show that a reformulation of the algorithm will be required in the future for eukaryotes compared with the present analysis of prokaryotes, as gene targets are now spread over several chromosomes, TF co-operativity is more extensive and TF translation is uncoupled from its transcription by the nuclear envelope.

ACKNOWLEDGEMENTS

The authors thank Joan Hérisson and Marc Schoenauer for their early input into this study. They also thank the anonymous reviewers for their stimulating comments.

FUNDING

Sixth European Research Framework (GENNETEC project number 034952) and Seventh European Framework Programme (SYSCILIA project number 241955); PRES UniverSud Paris, CNRS, Genopole and French National Institute of Cancer (in part) [project PL-2010-196 to M.E.]. R.N. is supported by a fellowship from the French Ministry of Higher Education and Research. I.J. is supported by a Novartis grant (CRG). Funding for open access charge: SYSCILIA.

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
2. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
3. MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, **2**, e36.

4. van Hijum,S.A., Medema,M.H. and Kuipers,O.P. (2009) Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol. Mol. Biol. Rev.*, **73**, 481–509.
5. Fraser,P. and Bickmore,W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413–417.
6. Cook,P.R. (2002) Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.*, **32**, 347–352.
7. Cabrera,J.E. and Jin,D.J. (2003) The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues. *Mol. Microbiol.*, **50**, 1493–1505.
8. Osborne,C.S., Chakalova,L., Brown,K.E., Carter,D., Horton,A., Debrand,E., Goyenechea,B., Mitchell,J.A., Lopes,S., Reik,W. *et al.* (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.*, **36**, 1065–1071.
9. Xu,M. and Cook,P.R. (2008) Similar active genes cluster in specialized transcription factories. *J. Cell. Biol.*, **181**, 615–623.
10. Schoenfelder,S., Sexton,T., Chakalova,L., Cope,N.F., Horton,A., Andrews,S., Kurukuti,S., Mitchell,J.A., Umlauf,D., Dimitrova,D.S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.*, **42**, 53–61.
11. Képès,F. (2004) Periodic transcriptional organization of the *Escherichia coli* genome. *J. Mol. Biol.*, **340**, 957–964.
12. Wright,M., Kharchenko,P., Church,G. and Segrè,D. (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *PNAS*, **104**, 10559–10564.
13. Junier,I., Herisson,J. and Képès,F. (2012) Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. *J. Mol. Biol.*, **419**, 369–386.
14. Junier,I., Martin,O. and Képès,F. (2010b) Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Comput. Biol.*, **6**, e1000678.
15. Képès,F. and Vaillant,C. (2003) Transcription-based solenoidal model of chromosomes. *Complexus*, **1**, 171–180.
16. Vilar,J.M. and Leibler,S. (2003) DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.*, **331**, 981–989.
17. Blum,A. and Mitchell,T. (1998) Combining labeled and unlabeled data with co-training. In *Proceeding of 11th conference on Computational Learning Theory*. ACM, Madison, Wisconsin, United States, pp. 92–100.
18. Schapire,R.E. (1999) A Brief Introduction to Boosting. *IJCAI '99: Proceeding of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., Stockholm, Sweden, pp. 1401–1406.
19. Elati,M., Fekih,R., Nicolle,R., Junier,I., Hérisson,J. and Képès,F. (2011) Boosting binding sites prediction using gene's positions. *Algorithms in Bioinformatics (WABI'11). LNCS*, **6833**, 92–103.
20. Gama-Castro,S. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli k-12* beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, 120–124.
21. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., Moor,B.D., Rouzé,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
22. Junier,I., Herisson,J. and Képès,F. (2010a) Periodic pattern detection in sparse boolean sequences. *Algorithms for Molecular Biology*, **5**, 31.
23. Lam,L. and Suen,C.Y. (1995) Optimal combinations of pattern classifiers. *Pattern Recogn. Lett.*, **16**, 945–954.
24. Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.
25. Hong,P., Liu,X.S., Zhou,Q., Lu,X., Liu,J.S. and Wong,W.H. (2005) A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636–2643.
26. Xu,Z. and Sun,S. (2010) An algorithm on multi-view adaboost. *ICONIP*, 355–362.
27. Peng,J., Barbu,C., Seetharaman,G., Fan,W., Wu,X. and Palaniappan,K. (2011) Shareboost: boosting for multi-view learning with performance guarantees. In *ECML/PKDD (2)*, pp. 597–612.
28. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, 93–96.
29. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, 119–127.
30. Swets,J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
31. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
32. Fawcett,T. (2004) ROC graphs: notes and practical considerations for researchers. *Technical report HPL-2003-4*.
33. Hong,C.S. (2009) Optimal threshold from ROC and CAP curves. *Commun. Stat.*, **38**, 2060–2072.
34. Wolpert,D.H. (1992) Stacked generalization. *Neural Netw.*, **5**, 214–259.
35. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
36. Spiro,S., Roberts,R.E. and Guest,J.R. (1989) Fnr-dependent repression of the ndh gene of escherichia coli and metal ion requirement for fnr-regulated gene expression. *Mol. Microbiol.*, **3**, 601–608.
37. González-Gil,G., Bringmann,P. and Kahmann,R. (1996) FIS is a regulator of metabolism in *Escherichia coli*. *Mol. Microbiol.*, **22**, 21–29.
38. González-Gil,G., Kahmann,R. and Muskhelishvili,G. (1998) Regulation of CRP transcription by oscillation between distinct nucleoprotein complexes. *EMBO J.*, **17**, 2877–2885.
39. Cho,B.K., Barrett,C.L., Knight,E.M., Park,Y.S. and Palsson,B. (2008) Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. *PNAS*, **105**, 19462–19467.
40. Ernst,J., Beg,Q.K., Kay,K.A., Balzsi,G., Oltvai,Z.N. and Bar-Joseph,Z. (2008) A Semi-supervised method for predicting transcription factor gene interactions in *Escherichia coli*. *PLoS Comput. Biol.*, **4**, e1000044.
41. Bauer,A.L., Hlavacek,W.S., Unkefer,P.J. and Mu,F. (2010) Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Comput. Biol.*, **6**, e1001007.
42. Yus,E., Guell,M., Vivancos,A.P., Chen,W.H., Lluch-Senar,M., Delgado,J., Gavin,A.C., Bork,P. and Serrano,L. (2012) Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.*, **8**, 8.
43. Coppe,A., Ferrari,F., Bisognin,A., Danieli,G.A., Ferrari,S., Bicciato,S. and Bortoluzzi,S. (2009) Motif discovery in promoters of genes co-localized and co-expressed during myeloid cells differentiation. *Nucleic Acids Res.*, **37**, 533–549.
44. Elati,M., Neuvial,P., Bolotin-Fukuhara,M., Barillot,E., Radvanyi,F. and Rouveirol,C. (2007) LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics*, **23**, 2407–2414.
45. Képès,F. (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *Journal of Molecular Biology*, **329**, 859–865.
46. Janga,S.C., Collado-Vides,J. and Babu,M.M. (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *PNAS*, **105**, 15761–15766.
47. Xiao,G., Wang,X. and Khodursky,A.B. (2011) Modeling three-dimensional chromosome structures using gene expression data. *J. Am. Stat. Assoc.*, **106**, 61–72.