Research

Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA

Esperanza Benítez-Bellón, Gabriel Moreno-Hagelsieb and
Julio Collado-Vides

Address: Program of Computational Genomics, CIFN, UNAM, A.P. 565-A, Cuernavaca, Morelos 62100, Mexico.

Correspondence: Gabriel Moreno-Hagelsieb. E-mail: moreno@cifn.unam.mx. Julio Collado-Vides. E-mail: collado@cifn.unam.mx

Published: 21 February 2002

Genome ***Biology*** 2002, **3(3)**:research0013.1–0013.16

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/3/research/0013

Received: 16 August 2001
Revised: 20 November 2001
Accepted: 28 January 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Sites in DNA that bind regulatory proteins can be detected computationally in various ways. Pattern discovery methods analyze collections of genes suspected to be co-regulated on the evidence, for example, of clustering of transcriptome data. Pattern searching methods use sequences with known binding sites to find other genes regulated by a given protein. Such computational methods are important strategies in the discovery and elaboration of regulatory networks and can provide the experimental biologist with a precise prediction of a binding site or identify a gene as a member of a set of co-regulated genes (a regulon). As more variations on such methods are published, however, thorough evaluation is necessary, as performance may differ depending on the conditions of use. Detailed evaluation also helps to improve and understand the behavior of the different methods and computational strategies.

Results: We used a collection of 86 regulons from *Escherichia coli* as datasets to evaluate two methods for pattern discovery and pattern searching: dyad analysis/dyad sweeping using the program Dyad-analysis, and multiple alignment using the programs Consensus/Patser. Clearly defined statistical parameters are used to evaluate the two methods in different situations. We placed particular emphasis on minimizing the rate of false positives.

Conclusions: As a general rule, sensors obtained from experimentally reported binding sites in DNA frequently locate true sites as the highest-scoring sequences within a given upstream region, especially using Consensus/Patser. Pattern discovery is still an unsolved problem, although in the cases where Dyad-analysis finds significant dyads (around 50%), these frequently correspond to true binding sites. With more robust methods, regulatory predictions could help identify the function of unknown genes.

Background

As a consequence of the availability of whole-genome expression methodologies, regulation of gene expression is at the core of current post-genomic studies [1]. Once a set of genes is clustered on the basis of similar expression profiles, a logical next step is that of searching their upstream regions for potential binding sites for transcriptional regulators. The predicted binding sites in DNA can then be mutated or used to fish out the DNA-binding regulatory protein. Different methods exist for finding binding sites [2-6], with a recent rapid increase in different methods with small variations and improvements [7-9]. However, as the computational biology community has long been aware, a common limitation of such methods is the high rate of false-positives that they generate as a result of the low degree of conservation of the DNA sequences of binding sites.

This work is a contribution towards a more detailed evaluation of the performance of these methods, with the aim of finding the best selection of thresholds to provide reliable predictions. On the basis of our evaluations, we suggest improved methods to search for novel binding sites that give a much lower rate of false positives. We use information gathered in RegulonDB, a database on regulation of transcription in *Escherichia coli* compiled from the literature [10,11]. The database contains data on regulons - sets of genes in transcription units whose expression is regulated by the same regulatory proteins - with different types of evidence and different levels of description. For instance, at the time of writing, the database contains information on 112 regulatory proteins, but binding sites in DNA are only described for 60 of these. The data for 26 of the regulatory proteins includes information on at least three regulated genes, with at least one binding site per gene (Table 1). The total number of regulatory binding sites listed is 505.

As explained below, we distinguish between pattern discovery and pattern search and evaluate each separately. We evaluate two methodologies. One is Dyad-analysis [12], a program developed to find over-represented small words separated by a given distance. We also describe and evaluate an elaboration of this method that aims to search for probable binding sites using the dyads generated (dyad sweeping). The other method uses Consensus [13], a program that generates optimized ungapped multiple alignments for sets of known or suspected regulatory sequences and builds matrices representing the frequency of each base at each position of the aligned sequences. Its companion program 'Patser' uses the matrices generated to scan for similar new sequences. The evaluations take into account the interest in minimizing the false-positive rate, as even a very small false-positive rate can overshadow true positives because of the small number of genes expected to be part of each regulon (see below).

Description of datasets

As most regulatory sites for DNA-binding proteins are found 200 to 400 base-pairs (bp) upstream of the regulated genes [14], we built two sets of upstream regions. One contained 200 bp of the region upstream of the genes' start sites plus 50 bp downstream (200+50 set); the other contained 400 bp upstream plus 50 bp downstream of the start sites (400+50 set). Repressor sites are located near the promoter site, whereas activators tend to occupy a larger region upstream of the promoter. It is therefore potentially useful to evaluate the performance of the methods with these two different ranges of sequence. Additional information can also influence the decision of the experimentalist to select the length of upstream region to analyze. For instance, some proteins tend to have a single binding site per promoter, which has to be proximal to the promoter (for example LexA), whereas other proteins tend to have several binding sites per upstream region, with some of them farther upstream of the promoter (for example AraC, Lrp and MetJ). Another factor that influences the size of region to analyze is whether the precise site of transcription initiation (the +1 position) is known. When the promoter is known, the search can be limited to 200 bp upstream from the +1 position. If it is not known, then the reference point has to be the start codon and the 400 bp upstream of this are used - which assumes an average of 50 to 100 bp between the promoter and the beginning of the gene.

We used the total set of upstream regions containing at least one reported binding site in RegulonDB as the basic data for evaluation. In each case, upstream regions of genes regulated by the same protein (regulons) were separated from the collection and constituted the 'training sets'. For each set, the remaining upstream regions, known to be regulated by other proteins, are assumed to be the collection of 'known negatives'. Though there is still a risk that the known negatives contain genes that also pertain to the regulon we are contrasting them with, the fact that they have been the subject of experimental work allows us to think that this risk is minute.

Because of the small amount of data for each protein, we could not leave out a set of known positives to evaluate the rate of true positives, except in the case of the regulatory protein CRP. For those families having at least five upstream regions we were able to apply a 'leave one out' procedure as described below. We also have information, in some cases, on genes regulated by a given protein in the regulons analyzed, but with no reported binding site. The upstream regions of these genes were used to search for binding sites and provide further evaluation. A more detailed analysis was performed for LexA, comparing our predictions with a recent report in the literature [15].

Levels of analysis

Depending on the information available, there are basically two computational approaches to predicting binding sites for transcription initiation factors in DNA. In the best cases, there is information on experimentally determined examples of binding sites for a given regulatory protein. In such cases, the search programs can be trained using the sequences corresponding to the binding sites, and the information obtained (dyads, weight matrices) can then be used to find similar sequences, and thus other genes that might be under the control of the same regulatory protein. This is pattern searching.

On the other hand, a common scenario at present is that a set of apparently co-regulated genes is identified from transcriptome experiments. In this case, a program would be

Table 1

Summary of the datasets in RegulonDB

Regulatory protein	Number of binding sites	Site size (bp)	Regions with sites	Regions without sites
Ada	2	28	2	2
AlpA	-	-	-	1
AppY	-	-	-	2
AraC	15	17	5	1
ArcA	20	61	11	9
ArgR	12	16	6	1
AsnC	-	-	-	2
AtoC	-	-	-	1
BetI	2	21	2	-
BirA	2	40	2	-
CRP	109	19	65	15
CadC	-	-	-	1
Cbl	1	45	1	-
CsgD	-	-	-	2
CspA	3	5	1	1
CynR	2	60	2	-
CysB	7	42	5	1
CytR	12	40	6	2
DeoR	7	16	2	1
DnaA	8	9	2	-
DsdC	-	-	-	3
EbgR	-	-	-	1
EnvY	-	-	-	2
ExuR	-	-	-	1
FIS	29	16	25	-
FNR	30	22	20	4
FadR	6	17	4	-
FarR	2	21	1	-
FecI	1	7	1	1
FhlA	-	-	-	3
FruR	8	14	7	4
FucR	-	-	-	2
Fur	9	19	4	6
GalR	4	17	1	1
GalS	2	16	1	1
GatR	-	-	-	1
GcvA	4	29	2	-
GlpR	17	20	4	1
GntR	-	-	-	5
GutM	-	-	-	1
GutR	-	-	-	1
Hns	-	-	-	5
IHF	21	13	14	12
IclR	1	34	1	-
IlvY	4	26	2	-
KdpE	1	12	1	-
LacI	3	20	1	-
LeuO	-	-	-	1
LexA	9	20	8	1
Lrp	22	12	11	3
LysR	1	13	1	2
MalI	4	12	2	-
MalT	9	10	4	-
MarA	-	-	-	5
MarR	-	-	-	1
MelR	6	18	2	1
MetJ	5	8	2	1
MetR	3	24	2	1
Mlc	2	26	1	-
MtlR	-	-	-	1
NR_I	10	15	3	-
NadR	-	-	-	2
NagC	8	26	4	-
NarL	20	19	9	3
NhaR	-	-	-	1
OmpR	14	10	4	3
OxyR	4	45	4	-
PdhR	1	21	1	-
PhoB	7	17	4	1
PurR	16	16	14	3
RbsR	-	-	-	1
RcsB	2	25	2	-
RhaR	3	20	1	-
RhaS	3	17	2	-
Rob	-	-	-	4
SdiA	-	-	-	1
SoxR	2	19	2	-
SoxS	4	18	3	2
TdcA	1	15	1	-
TdcR	1	12	1	-
TorR	4	10	1	-
TrpR	5	27	5	-
TyrR	15	22	8	-
UhpA	1	39	1	-
XapR	2	13	1	-
XylR	4	16	2	-

RegulonDB contains information for the 86 regulons shown in this table. Of these, only 60 have at least three known binding sites for their corresponding regulatory protein. The second column indicates the total number of known sites, which are distributed in upstream regions (fourth column). The last column indicates the number of upstream regions for which there is experimental evidence suggesting regulation, but no direct proof of binding of the regulator to the upstream site is yet available. For instance, there are12 known sites for ArgR located in only six regions (with two sites per region), plus one region for a different gene for which there is evidence of regulation by ArgR.

trained with a collection of upstream regions from these genes with the goal of identifying probable shared regulatory sites. This is the problem of pattern discovery. If the data come from transcriptome experiments, the collection of co-regulated genes might not be complete. Because of the noise inherent to such experiments, and/or to the limitations of clustering algorithms, a researcher might wish to try to find other genes likely to be under the control of the same protein. However, other genes regulated by the same protein might display a different pattern of expression as a result of complications such as regulation by more than one regulatory protein.

On the basis of these considerations, the analyses we present contemplate the use of experimentally determined binding sites as training sets to study pattern search, and the use of upstream regions of co-regulated genes to study pattern discovery. More precisely, we use the set of binding sites in DNA for each regulatory protein reported in RegulonDB to try to find additional genes in the genome with similar sites. We also use the data on known co-regulated genes to try to find the binding site within the genes' upstream regions.

As training sets, we ran the dyad or matrix search programs on the sequences of known regulatory binding sites and on upstream regions of 200+50 and 400+50 bp from genes regulated by a given regulatory protein. Families corresponding to a given regulatory protein were evaluated only if there were at least three sequences in the corresponding training set (40 in the collection of binding sites; 26 in the 200+50 and the 400+50 datasets). Subsequently, the dyads and matrices were evaluated against the complete collections of 200+50 and 400+50 upstream regions. This gives a total of 3 x 2 = 6 evaluations for each regulon analyzed. The evaluations included regions 200+50 or 400+50 only if there was at least one reported binding site within that range; thus, the total set of 200+50 regions contained 172 sequences, and the 400+50 set contained 189.

Dyad analysis

We used the Dyad-analysis program [12] to find dyads within each training set. The options used were to find dyads of 3 bp long separated by distances of 0 to 16 bp, with any kind of dyad (direct repeat, inverted repeat, asymmetric), searching in both DNA strands [12,16]. Further analyses were limited to the training sets where the program found at least one dyad with a significance equal to or above 1.0 (see [12] for a detailed description of significance). This left 19 families from the binding-sites training sets, 11 from the 200+50 regions, and 14 from the 400+50 regions (the program Dyad-analysis did not find any dyad in about 75% of the rejected families, and found just one in most of the rest of them).

The program Consensus was run to obtain alignments and matrices 20 bp long - the most frequent size among binding sites for regulatory proteins. To assign match scores, we used an 'alphabet' based on the frequency of each base at upstream regions of 200+50 and 400+50 of all genes in *E. coli*. The search was done in a single strand. Although we also ran the program to find symmetric patterns, no clear improvement was observed. In the Results section, we first present results of pattern discovery, then concentrate on the selection of the best thresholds, analyzing their performance on the basis of the evaluation criteria described above. Finally, we present some specific predictions.

Results

Pattern discovery

Pattern discovery starts with a collection of co-regulated genes for which no binding sites are yet known. To evaluate the methodology, we counted the number of times a sensor can locate a known binding site in a collection of 200+50 or 400+50 regions.

The Dyad-analysis program is designed to find over-represented small words. Over-represented words would be expected to occur at the binding sites, and thus the first step was to determine if the resulting dyads match the binding sites. We found that there are significant dyads all along the sequences analyzed, with most of them matching at or near the known binding sites. Figure 1 shows, using the PurR family, that most dyads were found at distances very close to or overlapping the true binding sites. We observed the same tendency for all families. We thus decided to search for stretches of contiguous matches, which we call 'regions of overlapping matches' (ROMs), in the upstream sequences being analyzed by counting (sweeping), base by base, the number of matching dyads. As seen in Figure 2, the ROMs with the highest number of matching dyads overlap the true known binding sites in the DNA. This result motivated us to use the highest number of matches within a ROM as the score. We call this method dyad sweeping.

As the highest-scoring ROMs frequently overlap reported binding sites (Figure 2, Table 2), we decided to keep, for subsequent analyses, the dyads found within the highest-scoring ROMs of each upstream region, as long as the ROM contained at least two dyads. In Table 3 it can be seen that, except in a few of the regulons, the fraction of regions with known binding sites found is quite high. In other words, the set of dyads that result after keeping only those that contribute to the highest ROM in each family is able to recover a large fraction of all the known binding sites in the family. It is important to keep in mind that a given dyad can match several positions - and therefore sites - in a single region or family. Thus, selecting only those dyads appearing in the

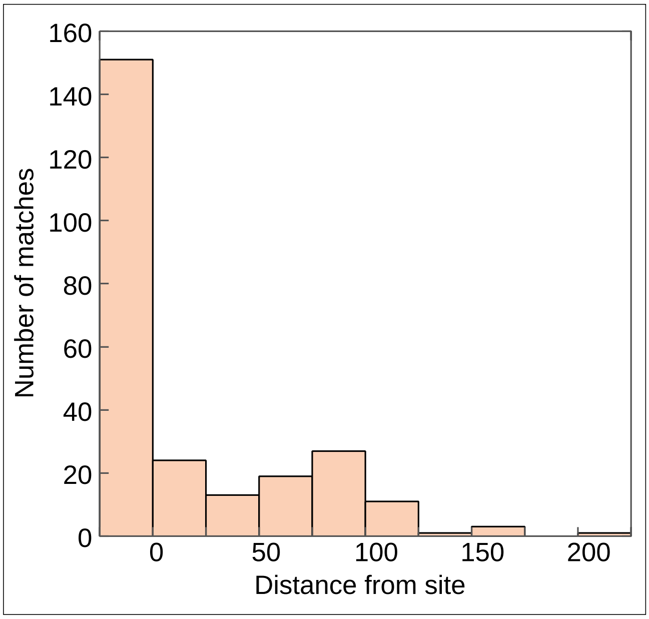

Figure 1
Position of dyads found by the Dyad-analysis program in relation to the binding sites in DNA for the whole PurR family. The graph shows the distances between all the dyads found in relation to the known binding sites of the PurR regulon. Distances below zero mean that the dyad is overlapping the binding site.

Table 2

Pattern discovery using ROMs (regions of overlapping matches) with maximal score to find binding sites in DNA

Regulatory protein	Number of genes in the regulon*	Touched by max†	Percent touched by max‡	Touched by other§	Total percent touched¶	Not touched**	Without dyads††
ArgR	6	2	33.33	4	100.00	0	0
CRP	54	42	77.78	8	92.59	4	0
FNR	17	10	58.82	0	58.82	2	5
GlpR	4	3	75.00	1	100.00	0	0
IlvY	2	2	100.00	0	100.00	0	0
LexA	8	8	100.00	0	100.00	0	0
MalI	2	2	100.00	0	100.00	0	0
MalT	4	4	100.00	0	100.00	0	0
MelR	2	2	100.00	0	100.00	0	0
NR_I	3	3	100.00	0	100.00	0	0
NarL	6	3	50.00	0	50.00	2	1
PhoB	4	4	100.00	0	100.00	0	0
PurR	12	11	91.67	0	91.67	1	0
TorR	1	1	100.00	0	100.00	0	0
TrpR	5	4	80.00	1	100.00	0	0
TyrR	8	7	87.50	0	87.50	1	0

*The total number of genes in the regulon with a known binding site (in the 400+50 upstream regions). †The number of regions where a ROM (region of overlapping matches) with the highest number of matches (max ROM) touches a known binding site. ‡This value expressed as a percentage. §Number of regions where either a ROM or dyad touches a known binding site, but the max ROM does not. ¶The percentage of all upstream regions in which any ROM touches a binding site. **Number of regions with dyads, but no match between known binding sites and ROMs. ††Number of regions with no dyads at all.

highest peak does not restrict their ability to find more than one site per region.

The number of dyads that describe the set of known binding sites in a given regulatory family is quite variable. For instance, if we use the known binding sites as training sets, the TyrR family involves 14 different dyads whereas ArcA has 65. There is no clear correlation between the number of dyads per site and the total number of sites in the training set for any given family, or any other property of the regulatory site, such as its size.

Sequence alignment

Consensus is a program designed to find and align shared stretches of sequence among a given set of sequences. The searching method based on the results of Consensus is already available [13]; the weight matrix generated can be used to search, with the companion program Patser, for sites in other upstream regions. The search using Patser was made using the first matrix (highest informational content) obtained in the final cycle of Consensus. This cycle requires all regions to contribute at least one sequence to the matrix. Using Patser, we searched for the highest-scoring sequence in each region in the training set. The lowest value among these results was set as the minimal score and a second search was performed with this threshold in order to find new sites above this limit within each upstream region in *E. coli* for further searches and analyses.

The capacity for pattern discovery of the two methods can be estimated by calculating the fraction of binding sites found when the training sets were the 200+50 or 400+50 bp regions, as shown in Table 4. A site was considered found when the predicted pattern overlaps 20% of the binding site.

We also show the results of using the sequences of the binding sites with 10 bp extensions on each side as training sets, so we could distinguish between pattern discovery and pattern abstraction or identification. In the case of Dyad-analysis/sweeping we evaluated whether the filtered dyads overlap the set of true sites. In the case of Consensus/Patser we evaluated whether the set of sites selected by Consensus/Patser overlaps the set of known sites. Consensus/Patser is able to abstract a pattern for each of the 25 families, whereas Dyad-analysis/sweeping can only do it for 19 of the families. In 11 of these 19 families Consensus/Patser finds more sites, in two families Dyad-analysis/sweeping finds more sites, and in the remaining six both methods perform equally well.

The real pattern discovery situation is that of the 450/sites cases (see legend to Table 5 for definition), where Consensus generates matrices for 24 of the families and Dyad-analysis finds significant dyads for 11 of them. Dyad sweeping finds on average more than 70% of the binding sites (when Dyad-analysis obtains significant dyads) as compared to around 60% with Patser. Note that using shorter regions to search for DNA binding sites (200+50), improves the performance of both methods by about 5-7%.

Once Table 4 was generated, we estimated the fraction of upstream regions recovered (Table 3). A region is considered found when at least one site in that region is found. Therefore,

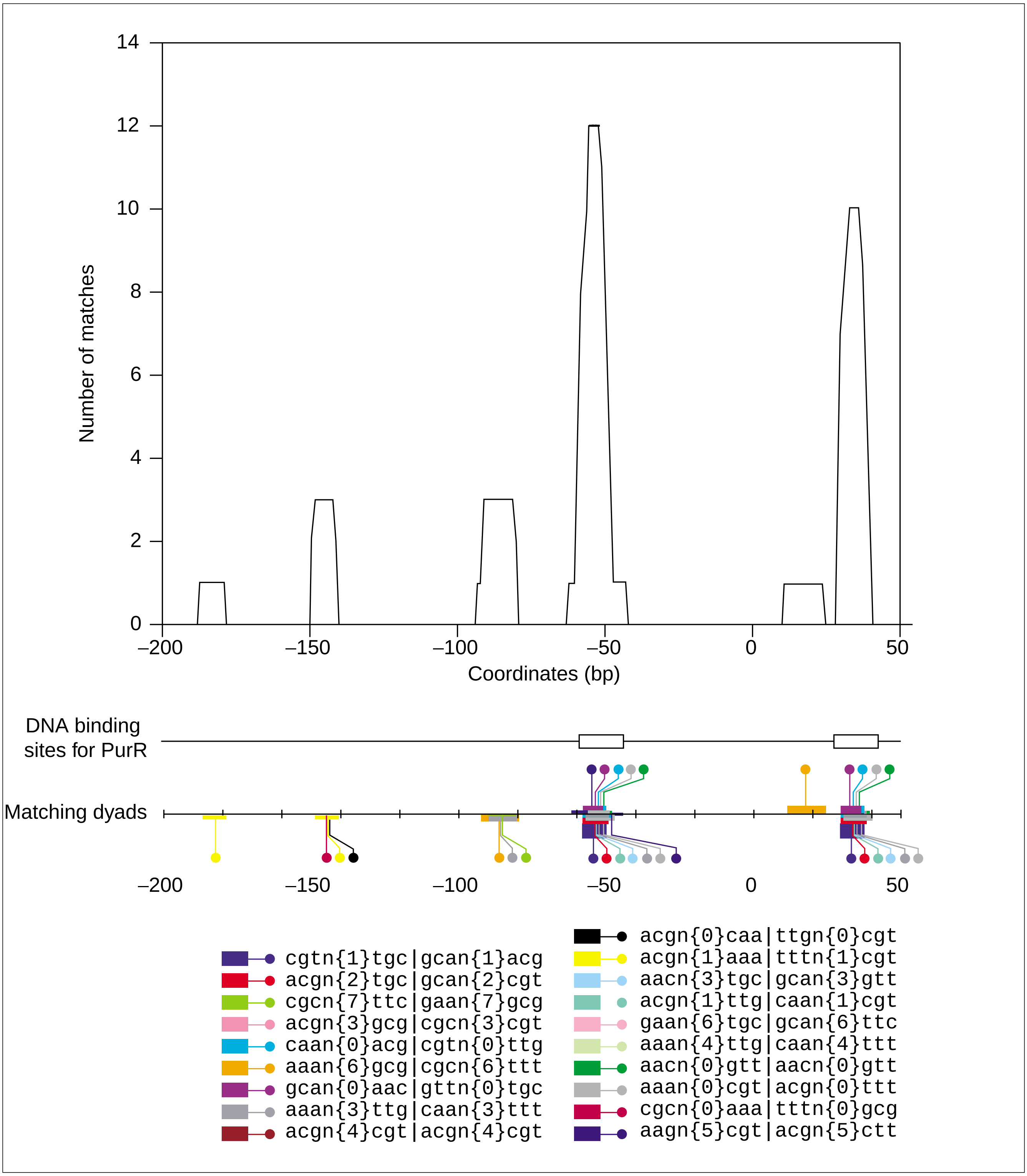

Figure 2

Dyad sweeping along the upstream region of the *purR* gene. Contiguous regions of overlapping matching dyads (ROMs) frequently overlap with the known binding sites. This example shows results after finding significant dyads in the 200+50 regions of the PurR regulon, and finding the ROMs within the same regions. The two ROMs with the highest peaks completely overlap with the two reported regulatory binding sites in this region (sites lie at positions -59 to -43 and at 29 to 45). The coordinates here are relative to the annotated first coding nucleotide of the gene. The known binding sites are illustrated as boxes below the figure. The lower line shows the different dyads coded in different colors. It can be seen, for instance, that blue dyads occur only in the two true binding sites.

Table 3

Pattern discovery at the level of upstream regions

	Consensus/Patser			Dyad-analysis/sweeping		
Regulon	Sites/sites	250/sites	450/sites	Sites/sites	250/sites	450/sites
AraC	100.00	20.00	20.00	100.00	-	80.00
ArcA	80.00	60.00	60.00	90.00	-	80.00
ArgR	100.00	100.00	100.00	100.00	33.33	100.00
CRP	95.24	93.65	95.24	90.48	66.67	65.08
CysB	100.00	60.00	40.00			
CytR	100.00	16.67	16.67			
FIS	60.00	60.00	-			
FNR	90.00	75.00	70.00	80.00	60.00	60.00
FadR	100.00	75.00	-			
FruR	100.00	14.29	71.43	71.43	-	-
Fur	100.00	100.00	25.00	75.00	-	-
GlpR	100.00	100.00	100.00	100.00	75.00	100.00
IHF	100.00	75.00	33.33	58.33	-	-
LexA	100.00	87.50	87.50	100.00	100.00	100.00
Lrp	80.00	60.00	50.00	50.00	-	-
MalT	100.00	50.00	50.00	100.00	100.00	100.00
NR_I	100.00	100.00	100.00	100.00	100.00	100.00
NagC	75.00	25.00	25.00	-	-	50.00
NarL	100.00	55.56	22.22	77.78	-	-
OmpR	100.00	-	25.00	100.00	-	-
OxyR	75.00	25.00	-			
PhoB	100.00	100.00	100.00	100.00	100.00	75.00
PurR	92.31	84.62	84.62	100.00	84.62	84.62
TrpR	100.00	100.00	100.00	100.00	100.00	100.00
TyrR	100.00	100.00	87.50	87.50	87.50	100.00
Average	94.14	68.22	61.98	88.45	82.47	85.34

For each family, we show the results with Dyad-analysis/sweeping and with Consensus/Patser. The data shown are obtained using different training sets - the 200+50 and 400+50 regions (250 and 450) and a comparison with training sets of known binding sites (sites) as a reference standard. Results are given as the number of regions where at least one binding site was found divided by the total number of regions, and expressed as percentages. Note that only the dyads extracted from the max ROMs within each region are used here. In each column heading, the first word refers to the training set and the second refers to the regions where the patterns were searched. For instance, columns headed 450/sites show the results of pattern discovery when Consensus or Dyad-analysis has as input the 450+50 bp regions, and the sensor is evaluated with the files of known sites. We counted only those regions containing known binding sites within the range covered (that is, if a known binding site is present more than 200 bp upstream of the gene start site, the corresponding 200+50 region is not counted). Averages count only the lines where the programs provided a result. Dashes mean that either there was no binding site within the region, or the programs failed to provide a matrix (Consensus) or significant dyads (Dyad-analysis). A region is considered found if at least one of its binding sites is matched.

Table 4

Pattern discovery at the level of binding sites

	Consensus/Patser			Dyad-analysis/sweeping		
Regulon	Sites/sites	250/sites	450/sites	Sites/sites	250/sites	450/sites
AraC	100.00	8.33	8.33	58.33	-	41.67
ArcA	76.47	76.47	47.06	76.47	-	76.47
ArgR	75.00	58.33	58.33	100.00	16.67	50.00
CRP	90.53	87.37	88.42	66.32	46.32	47.37
CysB	100.00	42.86	28.57	-	-	-
CytR	100.00	12.50	12.50	-	-	-
FIS	55.56	66.67	-	-	-	-
FNR	93.10	65.52	51.72	58.62	41.38	41.38
FadR	83.33	50.00	-	-	-	-
FruR	100.00	14.29	71.43	71.43	-	-
Fur	100.00	66.67	11.11	77.78	-	-
GlpR	60.00	40.00	40.00	80.00	33.33	46.67
IHF	72.22	66.67	22.22	50.00	-	-
LexA	100.00	88.89	88.89	100.00	100.00	100.00
Lrp	78.95	47.37	26.32	36.84	-	-
MalT	66.67	22.22	22.22	55.56	44.44	77.78
NR_I	77.78	55.56	55.56	77.78	77.78	77.78
NagC	57.14	14.29	14.29	-	-	28.57
NarL	75.00	35.00	15.00	55.00	-	-
OmpR	71.43	-	7.14	42.86	-	-
OxyR	75.00	25.00	-	-	-	-
PhoB	71.43	57.14	57.14	71.43	71.43	42.86
PurR	92.86	85.71	78.57	92.86	78.57	85.71
TrpR	100.00	100.00	100.00	100.00	100.00	100.00
TyrR	86.67	73.33	66.67	53.33	46.67	60.00
Average	88.59	60.36	53.07	75.19	70.76	65.64

For each family, we show the results of applying Dyad-analysis/sweeping and Consensus/Patser to the problem of discovering binding sites. The results contain pattern discovery data similar to those in Table 3, but this time counting the number of binding sites found per total number of sites. Again, only dyads extracted from max ROMs are used. The names of columns are as described in Table 3.

the results differ from those in Table 4 because of the occurrence of multiple sites in some upstream regions. A clear case of this is the ArgR regulon, where each of the six regions has two binding sites. The methods detect from 17% to 58% of the sites, but find from 33 to 100% of the regions.

Detection of new members of regulons by pattern matching

Detection of new members of regulons requires the selection of an optimal threshold to accept a sequence as a predicted binding site, and the genes downstream of such sequences as new members of the regulon family. The selection of the best threshold requires the evaluation of the following parameters: sensitivity (rate of true positives), specificity (rate of true negatives), accuracy (overall rate of true results), and, very important in this case, the positive predictive value (rate of true positives among the total number of positives, true and false). Definitions of these terms are given in the legend to Table 5.

We used a leave one out (LOO) procedure to evaluate the true-positive and false-negative rates with families containing at least five reported genes with binding sites. The LOO method consists of leaving one gene at a time out of the training set; then, with the matrix or dyads built with the remaining sites, a search is made for a probable binding site within the upstream region of the gene that was left out. We combined the results of the left-out regions to build the total set of known positives for evaluation of true positives and false negatives. The evaluation of true negatives and false positives was carried out using the whole set of known positives as training sets and all the

Table 5

Definitions of parameters used in evaluating the predictions

Evaluation	Formula
Sensitivity	TP/(TP + FN)
Specificity	TN/(TN + FP)
Accuracy	(TP + TN)/(TP + TN + FP + FN)
Positive predictive value (PPV)	TP/(TP +FP)
Overall performance (OP)	(Accuracy + PPV)/2

FN, false negative; FP, false positive; TN, true negatives; TP, true positives.

remaining regions, known to be regulated by any other protein, as known negatives. Instead of calculating an average of the scores, and defining the threshold on the basis of standard deviations, we scanned the scores scale from the minimum score obtained in the collection of positives, to the maximum one, calculating the evaluation parameters noted above at each point of the scale. There is no point in searching at lower scores as there is no effect on sensitivity at such values.

In Figure 3 we show the results of the analyses of the PurR regulon using dyad sweeping. Here, the minimum number of matches evaluated was one. Note that, as the dataset of known negatives exceeds that of known positives, high accuracy coexists with a large number of true negatives. Nevertheless, at the threshold of 10 matches, despite a very low false-positive value (less than 10%), and a very high accuracy (approximately 95%) and sensitivity (90%), the positive predictive value (PPV) shows that the total true positives in the whole 'predicted' set is about 60%. This is a very important issue. As most regulatory proteins regulate just a few genes in comparison with the whole set of genes in a given organism, such a difference means that false positives might dilute reliable predictions even at very low false-positive rates. The PPV alone would leave results with very little recovery of true binding sites. Therefore, calculating an optimal point for prediction requires the use of a balanced evaluation criterion. After examining several graphs, we noticed that the average between accuracy and PPV (which we call the overall performance or OP) would be a good criterion. This makes

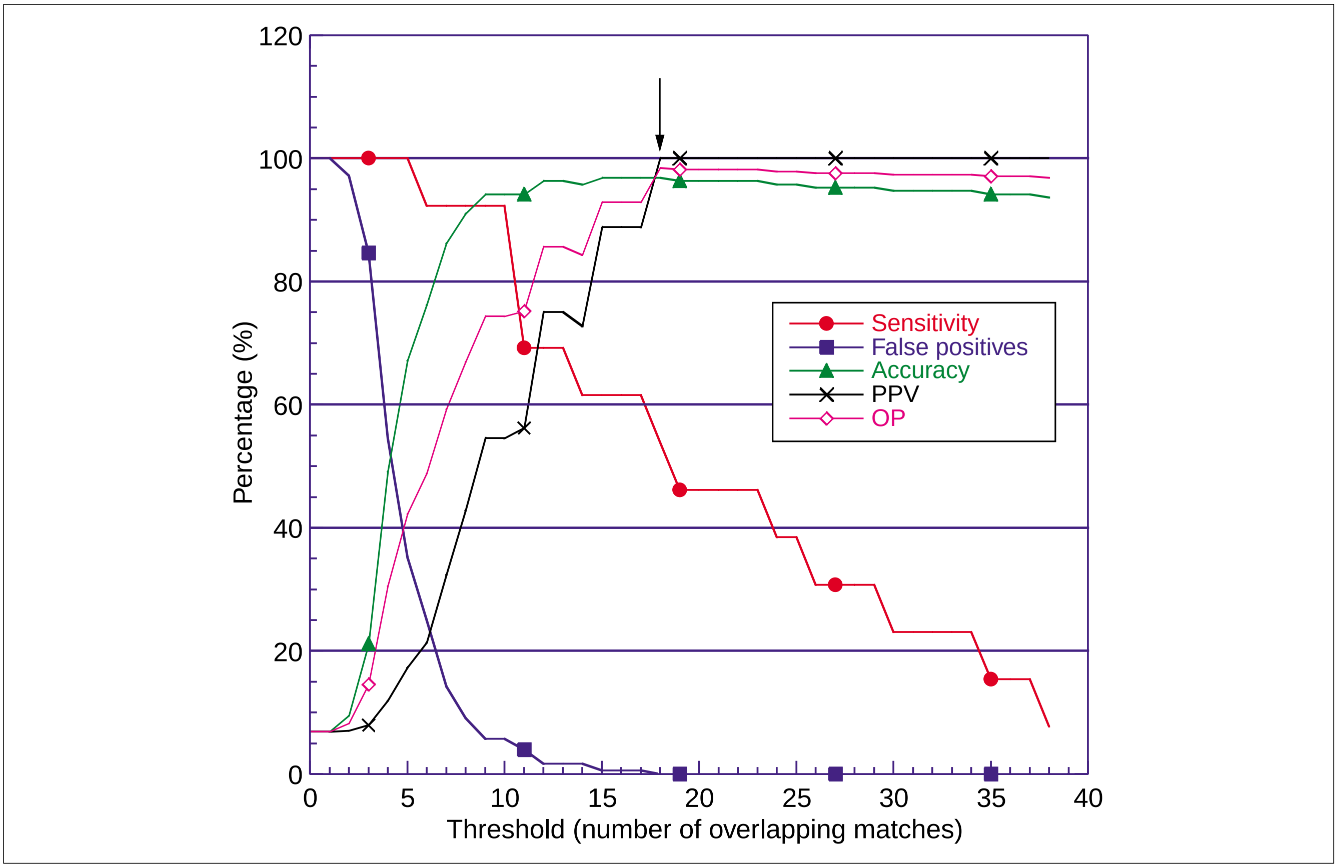

Figure 3
Evaluation of predictive capabilities as a function of the threshold using dyad sweeping. Different thresholds, defined as number of overlapping matches, were evaluated for all regulons. This graph shows the case of the PurR regulon when the dyads are obtained from the known binding sites and the evaluation is carried out on the 400+50 regions. The only dyads used in the search were those found at the ROMs with the highest value per region in the PurR regulon. The statistical parameters (see Table 5) are plotted as percentages instead of fractions. The arrow indicates the point of maximum overall performance (OP) (see text).

sense, as OP represents a trade-off between those two statistical measures. Other criteria, such as the product of accuracy and PPV, might be used instead, but OP worked well for our purposes. In a few cases, the point of highest OP leaves a very small sensitivity value (around 50% in PurR, for instance). If the sensitivity value was less than 60%, we used the last point where the sensitivity was above 60%. In Figure 4 we show the results of sensitivity and false-positive rate for all regulons at their best OP value using dyad sweeping.

The use of weight matrices derived from Consensus (with Patser) is not illustrated, as the selection of the best threshold is the same as in dyad sweeping. In Figure 5 we show the results of sensitivity and false-positive rates of each regulon at the best overall performance point of each regulon analyzed using Patser.

In Table 6 we give the fraction of sites found per family in regions of 400+50 bp when starting from different training sets using the threshold chosen as described above. Dyad-detection/sweeping still performs better at finding the sites within an upstream region, while Consensus/Patser trained with binding sites finds the sites at an average of almost 77%. An interesting finding here was that, when trained with all the upstream 400+50 sequences, Consensus finds an alignment and matrix that clearly discriminates between the sequences used in the training set, or regulon, from any other upstream sequence in *E. coli*. However, in some families, the matrix matches at sites different from the experimentally determined DNA binding site of the regulon under analysis (Figure 6), and such sites do not correspond to any known site, motif or region annotated in RegulonDB in the upstream sequence. We also verified that they do not match conserved regions in between pairs of sites. It will be indeed interesting to find out if these sequences have any biological meaning.

Predictions

Once the optimal threshold was obtained, we proceeded to predict other members of each regulon using the complete collection of upstream regions (200+50 and 400+50) of the *E. coli* genome [17]. In order to further evaluate the predictions obtained, we used the recent annotations of cellular functions assigned by Monica Riley and her group to known *E. coli* genes [18]. About 30% of the genes in *E. coli* have no function assigned, and each gene or gene product can be

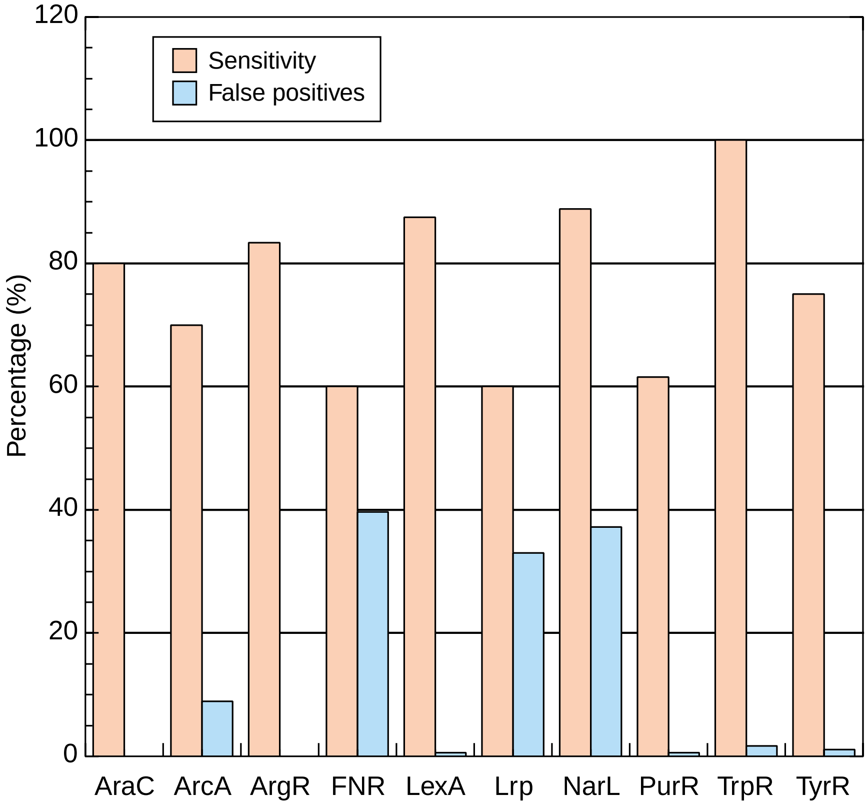

Figure 4
Performance of Dyad-analysis/dyad sweeping at the best threshold defined for each family. Sensitivity and false-positive rate (expressed as percentages) at the highest overall performance for each regulon are shown, using the binding sites as training sets, and the 400+50 regions as evaluation sets. We do not show the regulons where the methods did not provide significant results.

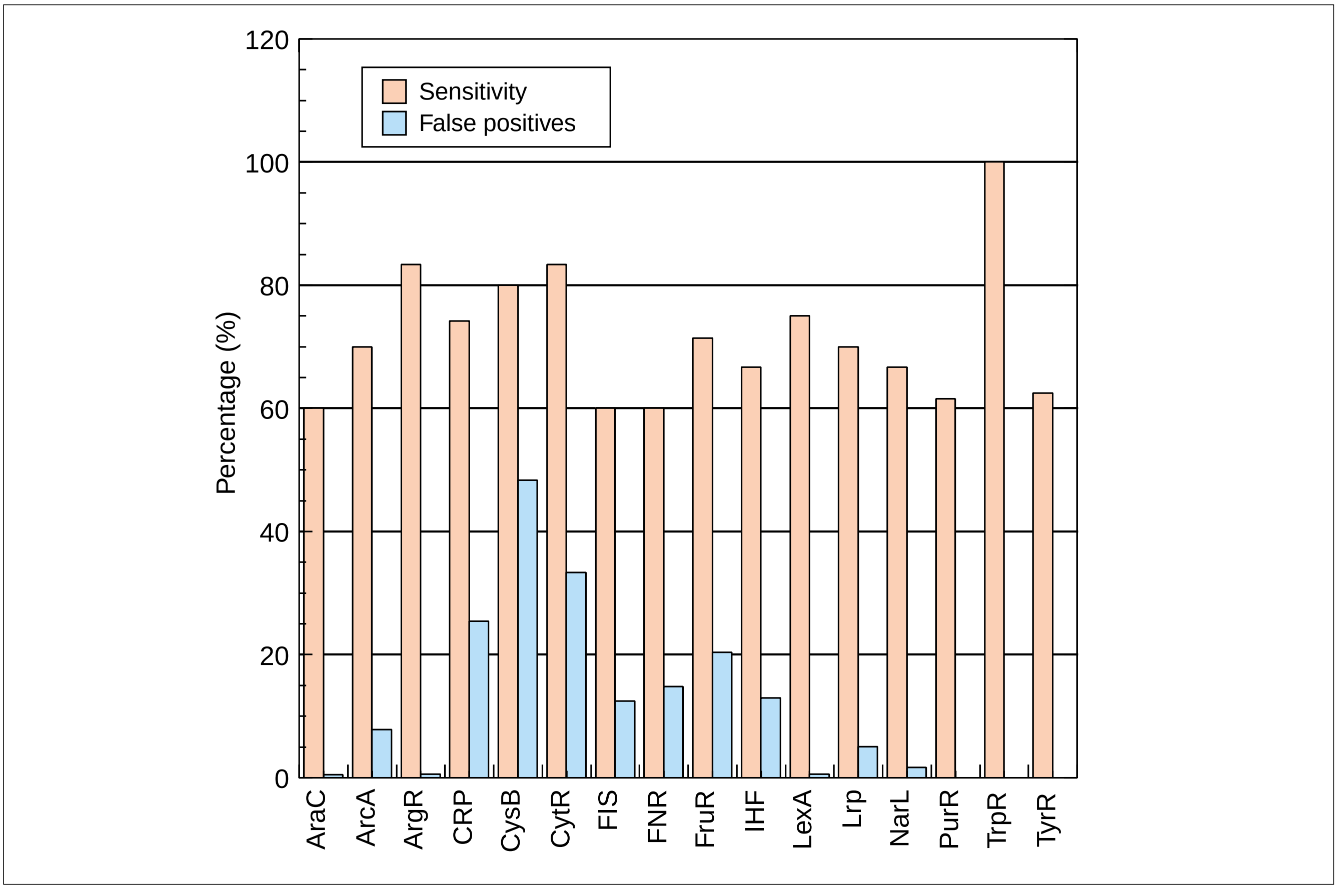

Figure 5
Performance of Consensus/Patser at the best threshold defined for each family. Sensitivity, and false-positive rate (expressed as percentages) at the highest overall performance for each regulon are shown, using the binding sites as training sets, and the 400+50 regions as evaluation sets. We do not show the families where the methods did not provide significant results.

assigned to more than a single cellular role. In Table 7 we show the consistency between the functional annotations of genes experimentally demonstrated to belong to each regulon as compared with the functional annotations of the set of predicted genes. In the cases of predictions of high confidence (for example, ArgR, CRP and PurR - all with correspondences above 90%), a putative function can be reliably assigned to genes of unknown function. For instance, in the case of the PurR family, the genes without functional annotations might be assigned to macromolecule (DNA/RNA) biosynthesis. This is an example of functional gene prediction based on analysis of its regulatory elements. Annotations like 'active transporter' would require other kinds of evidence (see Additional data files). Functional annotations might be quite helpful in cleaning up wrong predictions, or adjusting the proposed thresholds, although limited by the genomic coverage of the functional assignments.

RegulonDB contains information on a few genes belonging to some of the regulons studied but with no mapped binding site for the relevant regulatory protein. As further evaluation, we show the results of dyad sweeping and Patser, trained with the known binding sites of each regulon, for all of these genes (Tables 8,9). In the tables we indicate whether the gene would be included in the corresponding predictions, the highest scoring ROM (dyad sweeping, Table 8) or pattern match (Patser, Table 9) found in the 400+50 region of the gene, and the actual sequence suggested as part of the possible binding site. Some genes would be rejected as predictions, but the small amount of data makes it impossible to appropriately evaluate this problem. A researcher might choose to use a different, perhaps lower, threshold if the intention is to find every gene for a given regulon experimentally, and such a decision would depend on how many confirmatory experiments it is possible to perform (an example is shown in the next section). Lower thresholds can also be used if the intention is to confirm new members suggested by other data, like clustering of a gene or genes with known members of a regulon. The latter case is exemplified by the results with those regulon members lacking a mapped binding site. Most contain ROMs or patterns scoring above the minimal score obtained for a known member of the

Table 6

Binding sites remaining at best threshold

Regulon	Consensus/Patser		Dyad-analysis/dyad sweeping	
	Sites/sites	450/sites	Sites/sites	450/sites
AraC	60.00	20.00	80.00	100.00
ArcA	80.00	90.00	90.00	90.00
ArgR	83.33	100.00	100.00	66.67
CRP	80.95	63.49	90.48	95.24
CysB	100.00	80.00	-	-
CytR	100.00	16.67	-	-
FIS	40.00	-	-	-
FNR	65.00	70.00	50.00	65.00
FruR	100.00	85.71	-	-
Fur	-	-	75.00	-
GlpR	-	-	100.00	100.00
IHF	83.33	33.33	-	-
LexA	87.50	87.50	87.50	100.00
Lrp	40.00	60.00	-	-
MalT	-	-	75.00	100.00
NR_I	-	-	100.00	100.00
NagC	-	-	-	50.00
NarL	77.78	22.22	-	-
PhoB	-	-	75.00	75.00
PurR	69.23	84.62	92.31	84.62
TrpR	100.00	100.00	100.00	-
TyrR	62.5	87.50	75.00	37.50
Average	76.85	66.74	62.65	76.00

For each family, we show the results with Dyad-analysis/sweeping and with Consensus/Patser accepting a match only if its score exceeds the defined best threshold. This threshold corresponds to the highest overall performance, see Figures 3 and 4. Otherwise the results are treated as explained in Table 4. The column names are as described in Tables 3 and 4.

regulon (no search is performed below this lower limit), often just below our suggested threshold. Thus, if there is additional evidence that a gene belongs to a given regulon, the ROMs found can be proposed as the putative binding sites.

Comparison of results with recently examined members of the LexA regulon

A recent attempt has been made by Fernandez De Henestrosa *et al.* to locate all the members of the LexA regulon by a combined strategy that included prediction of probable binding sites and experimental confirmation [15]. Their predictions were based on similarity to known sites. Experimental confirmation showed that only 10 of the 49 predicted new members responded to LexA. The authors also give a table of previously found members of the LexA regulon, which includes a few genes not annotated in RegulonDB. We could analyze only five of their experimentally confirmed genes and 31 of their wrong predictions (predictions they later found experimentally not to be regulated by LexA) because of the lack of updating of the *E. coli* K12 genome annotations. In Table 10 we present our results for the genes noted as previously determined members of the LexA regulon in [15], plus the five new members found by this study. In Table 11 we show our results with their wrong predictions. Using dyad sweeping, we find 20 out of the 23 confirmed members of the LexA regulon, whereas we would reject 20 out of their 31 wrong predictions. With Consensus, we detect 18 of the 23 confirmed members of the regulon, while rejecting 19 of their wrong predictions.

Table 7

Correspondence between the functional annotations of predicted genes and of known genes.

Regulon	Consensus/Patser		Dyad-analysis/sweeping	
	Percent with related function	Percent without functional annotation	Percent with related function	Percent without functional annotation
AraC	73.33 (66.67)	37.50 (42.86)	53.75 (51.32)	36.51 (37.70)
ArcA	72.61 (70.75)	45.86 (47.50)	74.07 (70.83)	40.88 (43.75)
ArgR	100.00 (100.00)	18.18 (33.33)	56.25 (36.36)	51.52 (60.71)
CRP	93.90 (93.38)	40.87 (43.63)	-	-
CysB	56.78 (56.56)	43.03 (43.16)	-	-
CytR	53.90 (53.58)	38.56 (38.72)	-	-
FIS	45.52 (45.11)	41.23 (41.28)	-	-
FNR	82.85 (81.99)	45.85 (47.06)	83.77 (83.43)	41.43 (41.92)
FruR	51.96 (51.29)	36.57 (36.89)	-	-
IHF	80.95 (80.33)	46.34 (47.14)	-	-
LexA	69.70 (61.54)	42.11 (48.00)	65.79 (58.06)	38.71 (43.64)
Lrp	82.61 (81.95)	43.67 (44.58)	-	-
NarL	73.38 (72.11)	45.20 (46.35)	-	-
PurR	95.24 (92.31)	8.70 (7.14)	77.14 (70.37)	22.22 (25.00)
TrpR	85.71 (50.00)	30.00 (60.00)	46.94 (40.91)	42.35 (45.00)
TyrR	69.23 (50.00)	18.75 (27.27)	73.68 (61.54)	32.14 (40.91)

A comparison between the functional annotations of genes known to be regulated by a given protein and the functional annotations of the predicted set of genes. The percentage with related function is calculated against all the genes with functional annotations, while the percentage without functional annotations is calculated against the whole set of predicted genes. The number in parentheses excludes genes known to be part of the corresponding regulon. In cases with high correlation of functional annotations we can propose a related function for genes without functional annotations, as in the Consensus/Patser predictions of ArgR, CRP and PurR (all with correspondences above 90%). Detailed tables are provided as Additional data files.

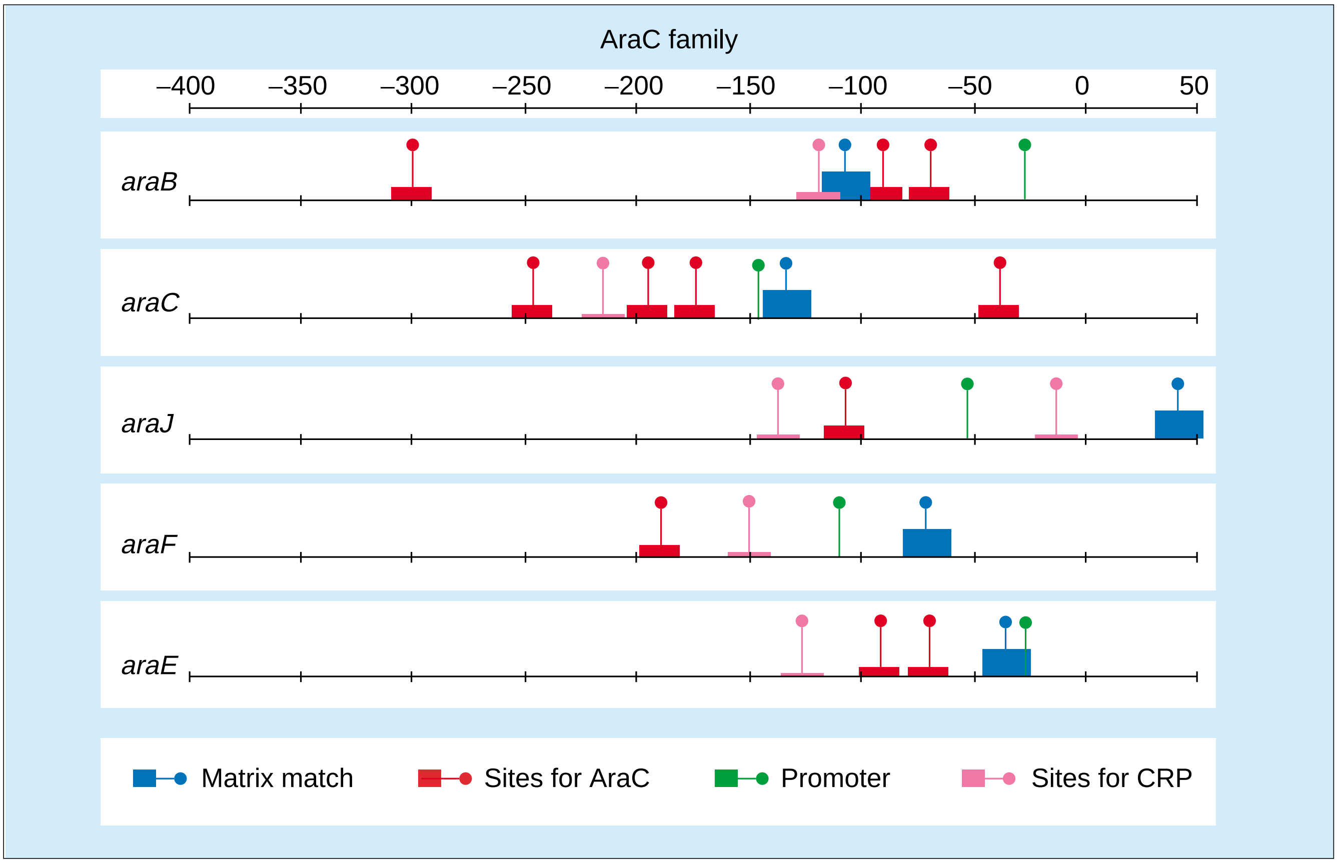

Figure 6
The positions found by Consensus/Patser. If Consensus is run to find an alignment within the 400+50 regions, the resulting matrix finds sites within each region (indicated here by the sites labeled 'matrix') that do not always match the binding sites for the relevant regulatory protein (AraC in the case illustrated here), but are very specific to the gene family. The sequence found does not correspond to known binding sites for other regulatory proteins (for example CRP) within the regions nor to the promoter.

Conclusions

Stringent evaluations of pattern discovery and pattern searching methods should be carried out to establish the confidence of a given prediction. Here we take advantage of the availability of reasonable negative samples - all other known regulons described in RegulonDB, except the one under study - in order to use standard statistical measurements of performance such as specificity and PPV. The PPV allowed us to stress how important even low rates of false positives might become in a large population. The small proportion of genes expected to be regulated by a given regulatory protein makes it important to emphasize the need for a stringent threshold to admit new members of regulons, as the true positives might be diluted in a high number of false positives. Nevertheless, if additional independent evidence is available, thresholds can be relaxed to include as many predictions as the confirmation procedure (genetic evidence of the regulatory effect, for instance) would allow. For instance, if the two computational methods were combined, only one of the genes known to be regulated by LexA (see previous section) would be rejected by both methods (*ybfE* in Table 10), while 16 of the wrong predictions are rejected by both methods (Table 11).

A very striking observation that deserves experimental analysis is the behavior of Consensus when identifying binding sites versus upstream regions. The program discovers patterns that discriminate, very specifically, the upstream regions used as training sets from the other regions. However, the patterns found do not always match the DNA binding sites. What are these specific motifs? These results imply the existence of new sequence elements specific to each family, different from those reported in the literature. We have not yet found (data not shown) any additional property that could suggest their function; their distance from the start site of transcription to known binding sites is not conserved; in some cases the predicted motif occurs upstream of the known sites in some promoters and downstream in other promoters. We have, of course, verified these observations twice, and find no additional property to associate with such families.

Table 8

Dyad-analysis/sweeping predictions in regions without binding sites reported in RegulonDB

Regulatory protein and genes shown to be regulated by it	Above threshold	Site coordinates and number of matches	Site sequence
ArcA (12)			
aceB	-	i:-144;f:-85;m:8	TTATCAAGTATTTTTAATTAAAATGGAAATTGTTTTTGATTTTGCATTTTAAATGAGTAG
fadB	-	i:-96;f:-51;m:6	ATTTCTTTAATCTTTTGTTTGCATATTTTTAACACAAAATACACAC
fumA	+	i:-77;f:-56;m:14	TATTGTTACTCGCTTTTAACAG
fumC	-	i:-92;f:-53;m:10	ATTTGTTATCAAATGGTAAATAATAAGTGAGCTAAAAGTT
glpA	-	i:-248;f:-232;m:8	TTATTTATGATTAACAG
hyaA	-	i:-168;f:-150;m:12	TACGCTTTATTAACAATAC
lpdA	+	i:-232;f:-204;m:15	TGTTTAAAAATTGTTAACAATTTTGTAAA
sucA	-	i:-323;f:-300;m:5	TGTTGTTGCAACGTAATGCGTAAA
ArgR (11)			
argD	-	i:-63;f:-51;m:5	TTTTTATGCATAT
FNR (4)			
aspA	-	i:-156;f:-144;m:3	TGATCTATTTCAC
cyoA	+	i:-25;f:2;m:5	GATCCCGTGGAATTGAGGTCGTTAAATG
icdA	+	i:-306;f:-292;m:6	ATTGAACAGGATCAC
sdhC	-	i:-340;f:-326;m:2	GATGATTAAAAATTA
LexA (9)			
umuD	+	i:-57;f:11;m:24	CTGCTGGCAAGAACAGACTACTGTATATAAAAACAGTATAACTTCAGGCAGATTATTAT GTTGTTTATC
Lrp (1)			
livK	+	i:-277;f:-269;m:2	CAGCATAAT
sdaA	-	-	-
serA	-	i:-146;f:-139;m:1	CAGCATAT
NarL (1)			
adhE	-	i:-215;f:-201;m:1	TACCCAGAAGTGAGT
caiF	-	-	-
torC	+	i:-209;f:-195;m:2	TACCCCTCCTGAGTG
PurR (15)			
codB	+	i:-82;f:-64;m:16	ACGAAAACGATTGCTTTTT
prsA	+	i:-356;f:-344;m:21	GAAAACGTTTTCG
speA	-	i:-132;f:-119;m:6	GAAACCGGTTGCGC

Sequences and positions of binding sites predicted by dyad sweeping in genes with experimental evidence for co-regulation in RegulonDB, but with no binding site experimentally identified. Genes follow the alphabetic order of the regulatory proteins, with the name of the protein separating each group. The number in parentheses after the regulator is the value of the threshold - derived from requesting best overall performance. The site coordinates are 'i' for initial base, 'f' for final position relative to the start codon. The score is given as the maximum number of matching ('m') dyads within a ROM. The number of families used was the same for any method, but we only show families where the methods provided significant results.

In the comparison of the two methods we have not found that one of them performs better in all the evaluations and scenarios considered (pattern search, pattern abstraction and pattern discovery). This implies that one could consider combining the different methods to make the best use of their respective strengths. For instance, if there is evidence of co-regulation only, we would suggest using Dyad-analysis/sweeping first to find the binding sites. If Dyad-analysis finds significant dyads, the dyad sweeping methodology can be used to extract possible binding sites. After that, the predicted sites can be used to train Consensus and search for further co-regulated genes. In cases where the DNA binding sites are known, Consensus/Patser, which are both very fast and simple to use, can give very reliable results in a short time.

The combination of computationally more confident predictions, together with additional independent evidence - for example, functional classes or operon organization - is an intelligent strategy for making more robust predictions. These more robust upstream regulatory analyses can be used to assign function to unknown genes, as illustrated here with the ArgR, CRP and PurR regulons. One can envisage highly relevant genomic applications of these predictions, such as distinguishing orthologs within families of paralogous genes, based on their differential regulation, or identifying non-orthologous gene displacement on the basis of regulatory comparisons.

The goal in computational biology is twofold: to provide, on the one hand, methods that generate useful and evaluated

Table 9

Consensus/Patser prediction in regions without binding sites reported in RegulonDB

Regulatory protein and genes thought to be regulated by it	Above threshold	Site coordinates and score	Site sequence
ArcA (8)			
aceB	+	i:-164;f:-144;sc:9.48	TTCATATTGTTATCAACAAG
fadB	-	-	-
fumA	-	-	-
fumC	+	i:-74;f:-54;sc:10.82	AATAATAAGTGAGCTAAAAG
glpA	-	i:-184;f:-164;sc:6.11	AAGAAAACATTCATAAATTA
hyaA	-	-	-
lpdA	+	i:-228;f:-208;sc:8.43	TAAAAATTGTTAACAATTTT
sucA	-	-	-
ArgR (13)			
argD	-	i:-70;f:-50;sc:11.12	TAGTGATTTTTTATGCATAT
CRP (6)			
cirA	+	i:-51;f:-31;sc:6.39	ATGTGAGCGATAACCCATTT
dsdA	-	-	-
ebgA	+	i:-91;f:-71;sc:7.53	TCGTGATCCAGTTAAAGTAA
flhD	+	i:-269;f:-249;sc:10.86	GTGTGATCTGCATCACGCAT
fucA	+	i:-399;f:-379;sc:10.18	ATATGACGGCGGTCACACTT
fucP	+	i:-205;f:-185;sc:9.74	AAGTGATGGTAGTCACATAA
glgC	-	i:-166;f:-146;sc:3.60	TCGCAATTAACGCCACGCTT
gntK	+	i:-169;f:-149;sc:11.49	ATTTGAAGTAGCTCACACTT
lpdA	-	i:-335;f:-315;sc:3.73	TGGTGATGTAAGTAAAAGAG
melA	-	i:-228;f:-208;sc:3.79	CTGCGAGTGGGAGCACGGTT
speC	-	i:-16;f:4;sc:5.55	GTTTGACCCATATCTCATGG
srlA	+	i:-91;f:-71;sc:8.89	TTGCGATCAAAATAACACTT
ubiG	-	i:-234;f:-214;sc:5.93	CAATGACCGACATCGCATAA
CysB (4)			
cysP	+	i:-237;f:-217;sc:7.04	TTTATTTGTCATTTTGGCCC
CytR (1)			
nupC	-	-	-
FNR (7)			
aspA	-	i:-367;f:-347;sc:3.88	CATGGGCAACCTGAATAAAG
cyoA	-	i:-182;f:-162;sc:4.31	TTTGTTATAACGCCCTTTTG
icdA	-	i:-105;f:-85;sc:6.30	AATCATTAACAAAAAATTGC
sdhC	-	i:-4;f:16;sc:3.89	ATTCATGATAAGAAATGTGA
FruR (5)			
aceB	+	i:-253;f:-233;sc:11.44	GATCGTTAAGCGATTCAGCA
fruB	+	i:-38;f:-18;sc:13.85	GAGGCTGAATCGTTTCAATT
ppsA	+	i:-105;f:-85;sc:6.29	TTTGCTTGAACGATTCACCG
IHF (7)			
caiT	+	i:-83;f:-63;sc:8.41	AATAATAATTATATTAAATG
ecpD	+	i:-39;f:-19;sc:8.96	ATTATTCCCTGTTTTAATTA
himA	-	-	-
himD	-	i:-136;f:-116;sc:6.43	ATTCCGAAGTTTGTTGAGTT
hycA	-	i:-73;f:-53;sc:6.57	TAATAACAATAAATTAAAAG
hypA	-	i:-155;f:-135;sc:6.75	TTAATTTATTGTTATTAAAG
narK	-	i:-106;f:-86;sc:6.66	AAATATCAATGATAGATAAA
ompR	-	i:-135;f:-115;sc:6.39	TATACTTAAGCTGCTGTTTA
sucA	-	-	-
LexA (9)			
umuD	+	i:-40;f:-20;sc:10.80	CTACTGTATATAAAAACAGT
Lrp (8)			
livK	+	i:-235;f:-215;sc:8.35	TGCCGTTATTTTATGCTGAC
sdaA	-	i:-317;f:-297;sc:4.95	ATCACCCTTTAGATATCTAC
serA	-	i:-79;f:-59;sc:6.62	TGCCGCAATATTATTTTTTG

Table 9 *(continued)*

Regulatory protein and genes thought to be regulated by it	Above threshold	Site coordinates and score	Site sequence
NarL (7)			
adhE	-	i:-160;f:-140;sc:6.11	ATAACTCTAATGTTTAAACT
caiF	-	i:-163;f:-143;sc:5.62	CAAATAATAGCGTGTCATGG
torC	-	i:-20;f:0;sc:4.98	ATAATTCTACAGGGGTTATT
PurR (11)			
codB	+	i:-84;f:-64;sc:13.10	CCACGAAAACGATTGCTTTT
prsA	+	i:-360;f:-340;sc:12.33	GCAAGAAAACGTTTTCGCGA
speA	-	i:-136;f:-116;sc:7.05	AAAAGAAACCGGTTGCGCAG

Data and analysis as described in Table 8. sc, Score as obtained from Patser. The number of families used was the same for any method, but we only show families where the methods provided significant results.

Table 10

Predictions in experimentally characterized binding sites for LexA

Gene	Consensus/Patser		Dyad sweeping	
b1728	+	sc:10.89	+	m:13
b1741	-	-	+	m:9
dinD	+	sc:14.01	+	m:20
dinG	+	sc:9.11	+	m:16
dinI	+	sc:12.22	+	m:12
dinP	-	-	+	m:14
ftsK	-	sc:7.82	+	m:11
lexA	+	sc:17.45	+	m:21
molR_1	+	sc:10.46	-	m:8
polB	-	sc:8.30	+	m:14
recA	+	sc:14.71	+	m:32
recN	+	sc:13.56	+	m:23
ruvA	+	sc:11.12	-	m:8
sbmC	+	sc:14.08	+	m:27
ssb	+	sc:12.99	+	m:29
sulA	+	sc:15.64	+	m:22
umuD	+	sc:10.80	+	m:24
uvrA	+	sc:16.30	+	m:29
uvrB	+	sc:14.94	+	m:12
uvrD	+	sc:16.07	+	m:15
ybfE	-	sc:8.18	-	m:3
yebG	+	sc:14.37	+	m:23
yjiW	+	sc:12.21	+	m:14

Fernandez De Henestrosa *et al.* [15] experimentally characterized new LexA-binding sites, which are not included in RegulonDB. The table shows our binding-site predictions with dyad sweeping and with Patser, using their corresponding best overall performance thresholds. sc, Score as obtained by Patser; m, maximum number of matching dyads. Note that most genes clearly have ROMs with 10 or more matches and with scores of Patser above 10.

Table 11

Contrasting predictions: regions known to lack LexA sites

Gene	Consensus/Patser		Dyad sweeping	
b3020	+	sc:13.66	-	m:6
brnQ	-	-	-	m:2
creA	+	sc:11.67	-	m:6
dinJ	+	sc:11.52	+	m:13
ecpD	-	-	+	m:9
hofQ	-	-	-	m:4
ilvD	-	-	-	m:6
ivbL	+	sc:13.23	+	m:21
metE	-	-	+	m:13
metR	+	sc:9.10	+	m:13
minC	+	sc:12.22	+	m:11
pshM	-	-	-	m:4
rfaJ	-	-	-	m:8
rob	-	-	-	m:6
xylE	+	sc:9.15	+	m:11
yafL	+	sc:15.21	+	m:13
ybiA	+	sc:12.07	+	m:16
ybiT	-	-	-	m:5
ycgJ	+	sc:13.42	+	m:11
ycgL	-	-	-	m:3
yciG	-	-	-	m:3
ydbH	-	-	-	m:3
ydeJ	+	sc:9.05	-	m:6
yecS	+	sc:9.89	-	m:6
yfiE	-	-	-	m:6
yfiK	-	-	-	m:6
ygjF	-	-	-	m:4
yhiX	-	-	-	m:4
yiaO	-	-	-	m:2
yigN	-	-	-	m:5
yjgN	-	sc:7.66	+	m:10

After experiment, Fernandez De Henestrosa *et al.* [15] rejected this set of genes in which they had predicted LexA sites using other computational methods. We tested the capacity of dyad sweeping and Patser to also reject these false positives. sc, Score as obtained by Patser; m, maximum number of matching dyads**.** Note that for both methods, most of the genes here show much smaller scores than genes belonging to the LexA regulon (see Table 10).

predictions, and, on the other hand, to use such methods as models of the biology under study. This latter virtue could generate new ways of understanding fundamental processes in gene regulation, along with, as suggested here, new properties of gene regulation at the genomic level. We cannot rely on a single methodology to solve the problems. Each algorithm should be tested on well-defined problems in order to find their strengths. Thus it should be possible to choose which method, or combination of methods, is best suited for the problem at hand.

comment
reviews
reports
deposited research
refereed research
interactions
information

Additional data files

Additional data files containing the functional annotations associated to the genes within each regulons, and of those genes downstream of predicted binding sites are available with the online version of this article.

Acknowledgements

E.B.B. has been supported by a fellowship from Conacyt, grant 0028. This research was supported by grants from CONACYT No. 0028 and from DGAPA to J.C.V. We acknowledge fruitful discussions with Jacques van Helden and suggestions by an anonymous referee. We also appreciate computational support from Victor del Moral.

References

1. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88:**243-251.
2. Hertz GZ, Hartzell GW 3rd, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6:**81-92.
3. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7:**41-51.
4. Waterman MS, Arratia R, Galas DJ: **Pattern recognition in several sequences: consensus and alignment.** *Bull Math Biol* 1984, **46:**515-527.
5. Wolfertstetter F, Frech K, Herrmann G, Werner T: **Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm.** *Comput Appl Biosci* 1996, **12:**71-80.
6. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
7. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214
8. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8:**1202-1215.
9. Crowley EM: **A Bayesian method for finding regulatory segments in DNA.** *Biopolymers* 2001, **58:**165-174.
10. Huerta AM, Salgado H, Thieffry D, Collado-Vides J: **RegulonDB: a database on transcriptional regulation in *Escherichia coli*.** *Nucleic Acids Res* 1998, **26:**55-59.
11. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12.** *Nucleic Acids Res* 2001, **29:**72-74.
12. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28:**1808-1818.
13. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15:**563-577.
14. Gralla JD, Collado-Vides J: **Organization and function of transcription regulatory elements.** In *Cellular and Molecular Biology:* Escherichia coli *and* Salmonella. Edited by Neidhardt FC, Curtiss III R, Ingraham J, Lin ECC, Low KB, Magasanik B, Reznikoff W, Schaechter M, Umbarger HE, Riley M. Washington, DC: American Society for Microbiology, 1996, 1232-1245.
15. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R: **Identification of additional genes belonging to the LexA regulon in *Escherichia coli*.** *Mol Microbiol* 2000, **35:**1560-1572.
16. **Regulatory sequence analysis tools** [http://embnet.cifn.unam.mx/~jvanheld/rsa-tools/]
17. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.
18. Serres MH, Riley M: **MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products.** *Microb Comp Genomics* 2000, **5:**205-222.