BIOINFORMATICS

Vol. 14 no. 5 1998
Pages 391–400

Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli *K-12*

Denis Thieffry, Heladia Salgado, Araceli M. Huerta and Julio Collado-Vides[1]

Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, AP 565-A, Cuernavaca, Morelos 62100, México

Received on December 16, 1997; revised on January 30, 1998; accepted on February 2, 1998

Abstract

Motivation: *As one of the best-characterized free-living organisms, Escherichia coli and its recently completed genomic sequence offer a special opportunity to exploit systematically the variety of regulatory data available in the literature in order to make a comprehensive set of regulatory predictions in the whole genome.*

Results: *The complete genome sequence of E.coli was analyzed for the binding of transcriptional regulators upstream of coding sequences. The biological information contained in RegulonDB (Huerta,A.M. et al., Nucleic Acids Res., **26**, 55–60, 1998) for 56 different transcriptional proteins was the support to implement a stringent strategy combining string search and weight matrices. We estimate that our search included representatives of 15–25% of the total number of regulatory binding proteins in E.coli. This search was performed on the set of 4288 putative regulatory regions, each 450 bp long. Within the regions with predicted sites, 89% are regulated by one protein and 81% involve only one site. These numbers are reasonably consistent with the distribution of experimental regulatory sites. Regulatory sites are found in 603 regions corresponding to 16% of operon regions and 10% of intra-operonic regions. Additional evidence gives stronger support to some of these predictions, including the position of the site, biological consistency with the function of the downstream gene, as well as genetic evidence for the regulatory interaction. The predictions described here were incorporated into the map presented in the paper describing the complete E.coli genome (Blattner,F.R. et al., Science, **277**, 1453–1461, 1997).*

Availability*: The complete set of predictions in GenBank format is available at the url: http://www.cifn.unam.mx/Computational_Biology/E.coli-predictions*

Contact: *ecoli-reg@cifn.unam.mx, collado@cifn.unam.mx*

[1]*To whom correspondence should be addressed*

Introduction

Several complete genomic sequences have already been published. A large number of other genome sequencing projects are already well advanced. Together, these projects have already generated interesting and even sometimes surprising results [see, for example, Dujon (1996)]. They provide a strong impulse to computational biology, aiming to interpret the vast amounts of molecular information produced. In particular, computational biology has already helped to define putative genes and their corresponding functions. However, much work remains to be done in order to decipher the genomic information already available, especially regarding the prediction of gene expression and its regulation.

The recent completion of the *Escherichia coli* genomic sequence constitutes a special opportunity in this respect. Indeed, as a long-standing model system for the study of gene regulation, *E.coli* is certainly the free-living organism about which we know the most, regarding the mechanisms of gene regulation, metabolism, etc. [see the classical books edited by Neidhart *et al.* (1987, 1996)]. In fact, this wide knowledge has already motivated the development of several models or theories related to gene regulation (e.g. Savageau, 1977; Thomas and D'Ari, 1990; Collado-Vides, 1992), as well as several dedicated databases (Karp *et al.*, 1996; Huerta *et al.*, 1998). In addition, several computational methods have been developed to predict the occurrence of promoters or regulatory sites in *E.coli* DNA sequences (e.g. Staden, 1984; Schneider *et al.*, 1986; O'Neill, 1989; Goodrich *et al.*, 1990; Hertz *et al.*, 1990).

In this paper, we predict transcriptional *cis*-regulatory sites in the genome of *E.coli* using all the molecular and genetic information collected in RegulonDB. These predictions are then discussed in the context of promoter annotations, as well as operon organization as predicted in the complete genome (see Blattner *et al.*, 1997). This paper complements another paper dealing with the prediction of promoters in the *E.coli* genome (Huerta *et al.*, in preparation).

© Oxford University Press

Materials and methods

For several years, our group has been systematically collecting genetic and sequence information dealing with the regulation of transcription in *E.coli*. This information has been compiled and organized in a relational database, RegulonDB (Huerta *et al.*, 1998). This database contains, among others, the sequences of 388 *cis*-regulatory sites of σ^{70} promoters, corresponding to 56 regulatory proteins.

It is important to note that the information associated with the regulatory sites in RegulonDB (Version 1.0) is diverse: for 248 sites, there is information on their position relative to transcription initiation; for 140 sites, the sequence is known, but not the position relative to transcription initiation because there is no corresponding promoter yet characterized; finally, 258 sites are supported only by genetic evidence suggesting a direct regulatory interaction, but lack information about their sequence and their position. All the information gathered was exploited in our prediction strategy and compared with the putative sites distributed in the genome.

The annotations of open reading frames (ORFs) generated by Fred Blattner's group (Blattner *et al.*, 1997) were used in order to generate a set of 4288 putative regulatory regions, covering 400 bp upstream from each ORF and 50 bases downstream. This length was selected on the basis of the known distribution of a large collection of regulatory sites in σ^{70} promoters (Gralla and Collado-Vides, 1996).

Taking advantage of existing agrep and gais unix programs, Perl scripts were written to localize and characterize perfect and imperfect matches in the set of putative *cis*-regulatory regions. The upper limit imposed by gais and agrep of eight mismatches is sufficient to search for significant regulatory sites since they average 20 bp in size. Whenever the available regulatory sites were too small (e.g. TorR), we extended the original sites slightly in order to reach at least 13 bases, to keep the numbers of matches low. Such extended length is reasonable, recalling that most transcriptional regulators in *E.coli* are dimers with a helix–turn–helix domain defining operator sites of around 20 nucleotides long.

The overall strategy to predict regulatory sites can be described as being formed by three phases. In the first phase, we identified all potential sites allowing a given number of mismatches when compared to any of the known sites. This number was defined based on two criteria. First, we identified the allowed mismatches to increase by about an order of magnitude the number of predicted sites, compared to the number of known sites. Second, we defined an upper limit

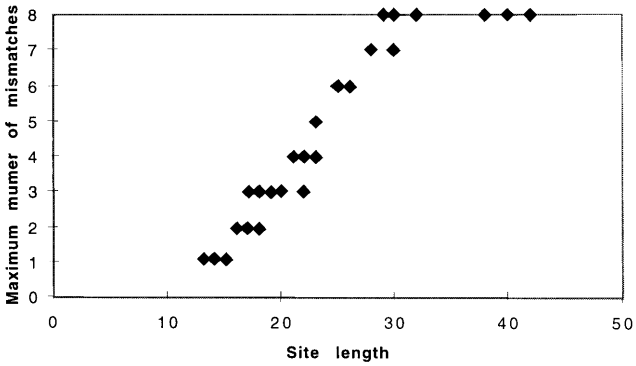

Fig. 1. Maximum numbers of mismatches allowed in the first phase of the string search as a function of the length of the sites. The names of the corresponding proteins are found in Table 1.

that increases linearly with the size of the sites (see Figure 1). This provides uniformity to the search for a large variety of sites for different regulatory proteins. For each protein (first column), Table 1 gives the number of known functional sites supporting this search (second column), the length of the sites (third column), as well as the maximum number of mismatches allowed (fourth column).

The second phase was to filter the set of sites previously found by string match with a weight matrix. Whenever the number of sites for a regulatory protein is sufficient (≥4), these sites have been aligned and cut using Wconsensus in order to define a weight matrix. Wconsensus is a program generating optimized ungapped multiple alignments of unknown prior width (Hertz and Stormo, 1995). Subsequently, the program Patser was used to scan the regulatory regions and identify sites using weight matrices (Hertz *et al.*, 1990; Hertz and Stormo, 1995). Patser was run with thresholds defined as the lowest score within each set of known functional sites. The result of this second phase, affecting only a fraction of the proteins, is a reduced set of predicted sites satisfying both a limited number of mismatches and a score higher than the threshold of the weight matrix ('Filtered' column in Table 1).

The third phase consists of annotating and evaluating these predictions of sites in terms of additional independent information. This includes their position in relation to the known or predicted site for transcription initiation, the functional description of the downstream gene, and genetic evidence supporting the regulatory interactions.

Table 1. Regulatory sites predicted after filtering with the weight matrix. The names of the proteins are listed in alphabetical order in the first column, followed in subsequent columns by the number of characterized sites (Nseq), their length in nucleotides (Length), the maximum number of allowed mismatches (Mismatches), the number of (perfect and imperfect) matches found by string match (S-Matches), the threshold for those weight matrices used in the second phase (Threshold), the number of sites found by Patser using the weight matrix (M-Matches), the number of sites after filtering (Filtered) and the number of regions where the final predictions are found (Nregion). Note that for TetR, whose native sites belong to a plasmid, no sites are found in the chromosome

Protein	Nseq	Length	Mismatches	S-Matches	Threshold	M-Matches	Filtered	Nregion
Ada	2	13	1	14				11
AraC	9	13	1	47	4.75	3577	32	26
ArcA	5	13	1	28	5.16	1607	15	12
ArgR	19	23	4	49	4.38	2741	41	26
ArsR	1	26	6	3				3
BioB	2	40	8	4	38.07	2	4	2
CRP	60	20	3	90	0.07	44679	86	70
CynR	2	25	6	18				18
CysB	3	15	1	3				3
CytR	4	22	4	14	10.48	22	4	4
DeoR	3	28	7	17				17
DnaA	3	21	4	9				8
FIS	6	28	7	26	4.41	13049	6	6
FNR	9	19	3	26	8.59	101	12	10
FadR	6	13	1	16				14
FarR	2	20	3	5				4
FhlA	1	16	2	3				3
FruR	1	14	1	2				2
Fur	7	23	4	13				13
GalR	1	29	8	24				24
GalS	2	13	1	6				6
GcvA	4	23	5	40	10.53	24	7	5
GlpR	12	23	5	114	6.34	891	31	22
HipB	4	30	8	34	14.52	7	7	2
IHF	31	42	8	12	5.76	2012	12	11
IclR	1	14	1	3				3
IlvY	4	21	4	21				19
KdpE	1	23	5	3				3
LacI	3	32	8	4				3
LexA	13	21	4	96	6.61	160	50	31
Lrp	21	15	1	27	4.15	6962	27	21
MalT	11	13	1	31	3.45	8893	19	12
MelR	1	18	2	3				3
MetJ	13	13	1	43	1.37	32780	36	26
MetR	2	13	1	10				10
NAC	2	16	2	4				3
NR_I	5	17	2	6	11.20	12	6	5
NagC	7	23	5	72	11.62	15	11	8
NarL	12	17	2	10	6.39	279	9	6

Table 1. *Continued.*

Protein	Nseq	Length	Mismatches	S-Matches	Threshold	M-Matches	Filtered	Nregion
OmpR	8	15	1	13	5.25	3056	12	7
OxyR	3	14	1	6				6
PapB	2	13	1	13				13
PdhR	1	17	3	12				12
PhoB	12	18	3	111	2.16	10274	70	54
PurR	11	17	2	32	9.93	32	25	15
PutA	8	13	1	49	5.41	2498	25	25
RafR	2	13	1	3				2
RhaR	2	32	8	3				2
RhaS	2	17	3	37				37
Rob	3	23	5	21				21
SoxS	9	30	7	15	4.44	4444	7	5
TetR	4	17	3	33	18.09	0	0	0
TorR	4	14	1	7	9.52	20	5	2
TrpR	5	22	3	6	18.26	6	6	6
TyrR	15	19	3	61	2.68	4067	36	27
UhpA	1	38	8	1				1

Results

Experimentally characterized sites

We localized most of the sites described in RegulonDB, that is to say, sites that have been characterized at the molecular level, leading to an up-to-date annotation of the *E.coli* sequence. Among the 248 well-characterized original sites (see Gralla and Collado-Vides, 1996), 202 were unambiguously identified without any mismatch in the putative regulatory regions. In addition, among 140 additional sites characterized at the molecular level but lacking precise localization, 91 were found in the putative regulatory regions and thus precisely positioned.

In other words, only ~75% of these well-characterized sites could be precisely localized in the 4288 putative regulatory regions 450 bases long. The other sites are not found for various reasons, such as errors in the sequences themselves, sites located at more remote locations, as well as operons and sites in our collection that belong to plasmids of *E.coli* (e.g. TetR sites) or to other related (bacterial or viral) genomes [e.g. some integration host factor (IHF) sites in our collection].

Predicted regulatory sites

Set of predictions: first and second phases. A total of 807 known and predicted regulatory sites were found within the 4288 regulatory regions upstream of all *E.coli* annotated ORFs. Of these, 293 are known and 514 are predictions. For each protein, Table 1 contains the total number of (perfect and imperfect) matches found by the string search strategy (fifth column). These correspond to the sites found in the first phase.

As already described, the second phase dealt only with the proteins with a sufficient number of known sites to generate a consensus matrix and filter the results of the first phase. For each of these proteins, the threshold used, and the number of matches found with Patser, are shown in the sixth and seventh columns of Table 1. The final number of filtered sites, i.e. after the combined string and weight matrix selection, are shown in the column of 'Filtered' sites. The last column indicates the number of putative regions where the final set of predicted sites are found. This number is smaller than the total number of sites in cases where multiple sites occur in a single regulatory region.

Third phase: additional independent support for predictions. In several cases, we have additional sources of independent information that help to give stronger support to some of the putative regulatory sites predicted. For instance, we determined the relative distance of the center of the site in relation to the predicted initiation of transcription, and compared that distance with the distances contained in RegulonDB for the same protein. The predicted sites were also evaluated based on the function of the first downstream regulated gene. Finally, we analyzed the set of regulatory regions for which there is genetic evidence supporting the regulatory interaction.

It was found that 41% of the predicted sites fall within the 100 bp upstream from the predicted initiation of transcription, a fraction comparable to the 56% found for sites within the well-characterized collection of sites for σ^{70} promoters (Gralla and Collado-Vides, 1996). More precisely, 129 predicted sites are found at a similar position (± 8) to some homologous known functional site. This corresponds to 24% of the total set of predicted sites. The window of ± 8 was used given that the distance of the +1 in relation to the –10 box can vary ± 4 and to allow for variation in the position of the site itself. These positionally restricted predictions are shown in Table 2.

As mentioned above, we systematically looked for the functional description of the downstream putatively regulated gene. Table 3 contains sites making biological sense with the function of the first downstream gene. For instance, it makes biological sense to find that FNR may regulate *narX*, the gene for the nitrate/nitrite sensor protein, or that GalR may regulate *galP*, the gene for the galactose transporter.

Table 2. Predicted sites at known functional positions in relation to transcription initiation. The sites are listed in alphabetical order of the regulatory protein followed by the name of the first downstream ORF, the centered coordinate of the site in relation to (known or predicted) transcription initiation, the name of the operon with the site known that has the closest relative position (within ± 8 bp) and the centered relative position of this known site. Sites with an even length have central positions with '.5' values

Regulator	Predicted gene	Predicted position	Known gene	Known position
AraC	csgG	–101	araC	–101
AraC	leuO	–42	araE	–43
AraC	rpoH	–50	araE	–43
AraC	rpsT	102	araC	107
AraC	w0024	–40	araE	–43
AraC	w0664	–266	araBAD	–274
AraC	w2042	–36	araE	–43
AraC	yfiL	–28	araC	–29
AraC	yfiL	–42	araE	–43
AraC	ygiL	–52	araC	–50
ArgR	argG	–6.5	carABp2	–9
ArgR	w1756	–10.5	carABp2	–9
ArgR	w1756	10.5	carABp2	14
ArgR	w3376	24.5	argCBH	19
ArgR	ybfH	26.5	argCBH	19
CRP	frwC	–147	malE	–139
CRP	glnS	–163	malK	–167
CRP	mrr	–46	tsxp2	–41.5
CRP	mtlA	4	cya	–2
CRP	w1657	–219	papB	–215
CRP	w2456	–61	tnaA	–61.5
CRP	ycbG	1	cya	–2
CRP	ygjH	–33	pBRp4	–40.5
CRP	ygjU	–138	malK	–133
CRP	yhfC	–79	tsxp2	–78
CRP	yiiU	–141	malK	–133
CRP	yjbA	–4	cya	–2
FIS	rna	–158	nrdA	–156
FIS	rna	–48	fis	–42
FNR	menG	–58.5	pflp7	–58.5
FadR	yaaH	–19	fadL	–17
FadR	yaaH	–44	fabA	–40
GcvA	w2332	6.5	gcvA	1
GlpR	w1483	38.5	glpACB	42.5
GlpR	w1954	–46.5	glpACB	–51.5
GlpR	yjbP	–137.5	glpTQ	–133.5

Table 2. *Continued*

Regulator	Predicted gene	Predicted position	Known gene	Known position
IlvY	w1416	–19.5	ilvY	–18
IlvY	w1416	–37.5	ilvC	–31
IlvY	ydjE	–13.5	ilvY	–18
KdpE	w3000	–6.5	kdp	1
KdpE	w3000	8.5	kdp	1
LexA	dinD	–12.5	lexA	–10
LexA	dinI	–8.5	sulA	–2
LexA	dinI	–9.5	sulA	–2
LexA	ftsK	–31.5	uvrA	–33
LexA	ftsK	14.5	uvrD	11
LexA	recN	–19.5	uvrBp2	–21
LexA	recN	2.5	sulA	–2
LexA	yebG	–0.5	sulA	–2
LexA	yjgN	–20.5	uvrBp2	–21
LexA	yjiW	23.5	colE1p122	22
LexA	yjiW	24.5	colE1p122	22
Lrp	fes	–179.5	leuABCD	–183
Lrp	leuO	–134.5	livJ	–127
Lrp	leuO	–69.5	tdh	–77
Lrp	w0212	–60.5	lysU	–58
Lrp	yjeJ	–72.5	tdh	–77
MalT	cbpA	–49.5	malE	–44
MetJ	metE	–15	metF	–8.5
MetJ	w1416	–79	metJp1	–75.5
MetJ	w2573	–7	metF	–8.5
MetJ	w2946	–25	metB	–31.5
MetJ	ybiC	–46	metJp1	–47.5
OmpR	mopB	–43.5	ompF	–46
OmpR	w1515	–365.5	ompF	–368
PhoB	aslA	–56	ugpp1	–52.5
PhoB	phoB	–29	ugpp1	–30.5
PhoB	phoH	–28	ugpp1	–30.5
PhoB	putA	–45	ugpp1	–52.5
PhoB	w0562	–55	ugpp1	–52.5
PhoB	yaeQ	–51	ugpp1	–52.5
PhoB	yaeQ	–73	ugpp1	–74.5
PhoB	yaeQ	–95	phoE	–88.5
PurR	purE	–37.5	purL	–38
PurR	purT	–18.5	pyrC	–23.5
PurR	yicE	19.5	glnBp2	20.5
PutA	aroP	12	putPp1	14
RhaS	w1845	6.5	rhaBAD	2
RhaS	yiiQ	7.5	rhaBAD	2
Rob	w1538	–2.5	fumC	1
TyrR	aroL	–46.5	tyrR	–50
TyrR	w4177	–84.5	mtr	–77
TyrR	ydbA_1	–29.5	aroF	–30
TyrR	yfiL	–26.5	aroF	–30
TyrR	yfiL	20.5	tyrB	20
TyrR	yheE	–77.5	mtr	–77

Table 3. Selected regulatory predictions: predictions where the regulatory interaction and the function of the regulated gene make physiological sense. These were obtained by visual inspection of the function of the regulated gene

Protein	Gene	Function of the regulated gene
AraC	yabJ	hypothetical ABC transporter in araC–tbpA region
	yfiL	hypothetical protein in aroF–rplS intergenic
CRP	fucA	fuculose-1-phosphate aldolase
	pgk	phosphoglycerate kinase
	malX	pts system, maltose and glucose-specific II ABC component
	malY	degrades the inducer of the maltose sstem or prevents its synthesis
	gdhA	NADP-specific glutamate dehydrogenase
	pgk	phosphoglycerate kinase
FNR	ndh	NADH dehydrogenase
	narX	nitrate/nitrite sensor protein NarX
	ordL	putative oxidoreductase
	dcuA	anaerobic c4-dicarboxylate transporter—membrane transport of aspartase
FadR	narY	respiratory nitrate reductase 2 beta chain
	cfa	cyclopropane-fatty-acyl-phospholipid synthase
	fadL	long-chain fatty acid transport protein precursor
Fur	fepB	ferrienterobactin-binding periplasmic protein precursor
GalR	galP	galactose-proton symport (galactose transporter)
LexA	recN	recombination and DNA repair
GcvA	proV	glycine betaine/l-proline transport ATP-bindingG protein ProV
GlpR	glpR	glycerol-3-phosphate regulon repressor
	glpE	glpe protein, gene of glp regulon
Lrp	leuO	probable activator protein in leuABCD operon
MalT	rhaA	l-rhamnose isomerase
	rhaT	rhamnose permease
MetJ	metE	5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase
	metC	beta-cystathionase
NagC	nagD	NagD protein
PhoB	phoA	alkaline phosphatase precursor
	pstS	periplasmic phosphate-binding protein
PurR	purT	phosphoribosylglycinamide formyltransferase 2
	yebG	hypothetical 10.7 kDa protein in purT 5 region
RhaS	yhhI	h repeat-associated protein in rhsB–pit intragenic region
	rhaT	rhamnose permease
TorR	torR	torCAD operon transcriptional regulatory protein TorR
TyrR	glnP	glutamine transport system permease protein GlnP
	yecH	hypothetical 7.3 kDa protein in tyrP–rsgA intergenic region
	yfiL	hypothetical protein in aroF–rplS intergenic region

Finally, recall that we had also collected genetic evidence supporting the existence of a regulatory interaction for 258 genes, i.e. interactions still lacking an associated binding sequence. Out of these, 20 have been corroborated by the site search in the genome. These predictions, together with a representative reference in Medline, are shown in Table 4.

Table 4. Predicted sites supported by independent genetic evidence. Regulatory proteins in alphabetical order are followed by the name of the regulated gene and the Medline number supporting the genetic evidence for this regulatory interaction. Altogether, 24 regulatory sites are found within the 16 regions indicated

Regulator	Regulated gene	Medline
CRP	fucA	96306444
CRP	mtlA	93023871
CynR	cynT	95050221
Fur	fepB	89053871
GlpR	glpE	96306444
LexA	umuD	96306444
Lrp	gcvT	96063908
Lrp	ompC	96063908
MetJ	metC	96306444
MetJ	metR	96306444
NagC	manX	91171292
PhoB	pstS	94069315
PurR	cvpA	96306444
RhaS	rhaT	95055724
SoxS	fumC	95198541
SoxS	nfo	95198541

Anatomy of gene regulation: known versus predicted comparisons. The database that we used to predict operons, promoters and sites is heterogeneous in the sense that basically all combinations of incomplete knowledge exist. For instance, there are cases of known promoters where the clustering of downstream genes into an operon is not known. Also, some operons are described without specification of an upstream promoter. As a consequence, the final annotations in the *E.coli* genome are a mixture of all sorts of predictions and characterized operons, promoters and regulatory sites. For the purpose of comparing the known organization with a set of clearly predicted cases, we decided to evaluate the set of regulatory sites that were found within the regulatory regions where a promoter has been predicted. Given the general strategy of the annotation work, we know that these regions contain no experimentally characterized regulatory sites. Additional predicted sites are found in other regions, but are excluded in the following comparisons. These predicted sites are compared against a well-characterized and analyzed collection of 140 promoters (see Gralla and Collado-Vides, 1996).

Figure 2 shows the distributions of regulatory regions as a function of the number of regulatory proteins, and as a function of the number of *cis*-regulatory sites. Figure 2A shows exclusively predictions, whereas Figure 2B refers to experimental data dealing with the 140 well-characterized promoters and their associated regulatory sites.

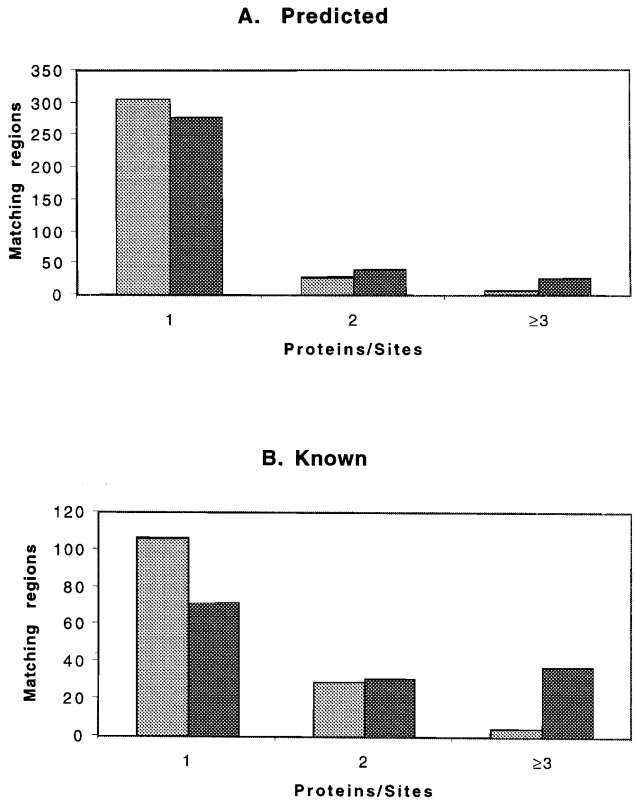

Fig. 2. Number of matching regions as a function of the number of matches. Histograms showing (**A**) the distribution of putative regulatory regions with respect to the number of predicted regulatory proteins and predicted *cis*-regulating sites and (**B**) the distribution of known regulatory regions with respect to the number of known regulatory proteins and known sites. Bars in the left describe the number of regions with one, two and three or more proteins. Bars in the right describe the number of regions with one, two and three or more sites.

From a total of 2502 upstream regions where a promoter has been predicted, regulatory sites were found in 343 regions. Taking into consideration the stringency of the method and the incompleteness of the information available on regulatory binding sites, the small number of regions where a match was found is reasonable. These numbers are fairly consistent with the distribution of well-characterized regulatory sites where transcriptional regulation has been well studied. In this collection of 132 promoters, 73% are regulated by one protein and 43% contain only one site for the binding of a regulator (Gralla and Collado-Vides, 1996).

Finally, when comparing the predictions of regulatory sites with the predictions of operon organization in the whole genome, it is interesting to observe that binding sites are found in 16% of the total known and predicted regions upstream of operons and in 10% of the intra-operon regions (Blattner *et*

al., 1997). If regulatory sites were equiprobably distributed upstream of genes, they would occur in 14% of operon and intra-operon regions. These numbers provide independent evidence supporting the set of regulatory site predictions as a whole.

All the predicted sites and, of course, all experimentally determined sites found in the genome were included in the analysis of the *E.coli* genome presented in the paper published by Blattner *et al.* (1997). We plan in the future to include detailed annotations within a new version of RegulonDB. These include the absolute position in the genome, its sequence, its relative position to the proximal (real or putative) promoter, the name of the promoter and/or the name of the proximal ORF, the name of the matching site, the number of mismatches, and the name of the corresponding protein.

Discussion

A systematic and detailed search for transcriptional regulatory sites in the whole *E.coli* sequence is presented. Several methods are available to perform such searches, including consensus and matrix-based methods. However, these methods have known limitations. For instance, we have satisfactory collections of sites supporting a good consensus or a reasonable frequency matrix for only a limited subset of *E.coli* regulatory proteins (~20). Although it is known that matrix methods often generate high numbers of false-positive matches (e.g. see Thieffry *et al.*, 1996), combining results using weight matrices (when available) with a string search allowing for a limited number of mismatches produces a smaller set of final predictions which should be of better quality.

The additional advantage of the string search strategy is that it could also be applied in the case of small and poorly characterized collections of sites, generating a reduced number of new matches. This conservative methodology generates a reasonable number of predictions that were further characterized using complementary biological information, including relative position, consistency with the function of the downstream gene, as well as independent genetic evidence.

At first glance, it is puzzling to obtain such a low number of matched regulatory regions (i.e. with at least one predicted site). However, this is consistent with the fact that the 56 sets of sites, each for one transcriptional factor, correspond only to a small fraction of the estimated number of regulatory proteins. Indeed, considering that the total number of genes in *E.coli* is roughly estimated at 4000, and assuming a 1:10 ratio of regulatory to regulated genes (see Riley, 1993), one would expect ~400 regulatory genes in *E.coli*. Our collection of sites would thus represent only about an eighth of the regulatory proteins of *E.coli*.

There is no doubt that important improvements may be achieved in the future on the computational methods for the recognition of regulatory signals. Nonetheless, the work presented here on the prediction of regulatory sites, together with the associated prediction of promoter sites and operon organization, constitutes a new approach where predictions of one type affect the evaluation of predictions of a different type. For example, the predicted operon organization is consistent with the independent search for regulatory sites. As discussed elsewhere, a similar consistency is found in operons and promoter predictions. Similarly, predicted sites were used as a guide to select putative promoters inside operons (Huerta *et al.*, in preparation).

Methods that use integrated information on several sites have been developed, such as a syntactic recognizer for sites in *E.coli* (Rosenblueth *et al.*, 1996; Thieffry *et al.*, 1996). The applicability of this method depends on prior knowledge of the transcription initiation since it makes use of the relative position of sites in relation to the beginning of transcription. We have shown that in some cases positional information can lower the number of false positives by around one order of magnitude. This advantage is, however, limited by the adequate determination of transcription initiation on the one hand and, on the other, it may also limit the prediction of sites only to previously identified positions. An extensive application of these approaches is currently limited by the performance of prediction of transcription initiation. In fact, as suggested by Hertz and Stormo (1995), promoter prediction itself can benefit from the identification of putative upstream activator sites, as well as from the precise combinations of homologous or heterologous sites in a strategy similar to that developed by the group of T.Werner (see, for example, Quandt *et al.*, 1995). We did not follow this strategy to analyze the complete genome given the reduced number of activator proteins currently known compared to the complete set of expected regulators in *E.coli*, as already mentioned.

It is important to note that predictions of regulatory sites could only be definitively confirmed based on experimental grounds. In this sense, it would be interesting to compare them with results obtained with global experimental studies such as those developed by Chuang *et al.* (1993), Appel *et al.* (1996) and Van Boggelen *et al.* (1996). Regulatory predictions at the level of a complete genome, such as those shown here, should eventually be compared with global regulatory experimental analyses of gene expression in cells with completed genomes.

Acknowledgements

We are grateful to Fred Blattner for sharing *E.coli* sequences and annotations during the last phase of the sequencing process. We also acknowledge the computer support provided

by Víctor Del Moral. This work was supported by grants from DGAPA-UNAM and Conacyt to J.C.-V.

References

Appel,R.D., Sanchez,J.C., Bairoch,A., Golaz,O., Ravier,F., Pasquali,C., Hughes,G.J. and Hochstrasser,D.F. (1996) The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis, its status in 1995. *Nucleic Acids Res.*, **24**, 180–181.

Blattner,F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1461.

Chuang,S.E., Daniels,D.L. and Blattner,F.R. (1993) Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.*, **175**, 2026–2036.

Collado-Vides,J. (1992) A grammatical model of the regulation of gene expression. *Proc. Natl Acad. Sci. USA*, **89**, 9405–9409.

Dujon,B. (1996) The yeast genome project: what did we learn? *Trends Genet.*, **12**, 263–270.

Goodrich,J.A., Schwartz,M.L. and McClure,W.R. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res.*, **18**, 4993–5000.

Gralla,J.D. and Collado-Vides,J. (1996) Organization and function of transcription regulatory elements. In Neidhardt,F.C. *et al*. (eds), *Cellular and Molecular Biology: Escherichia coli and Salmonella*, 2nd edn. American Society for Microbiology, Washington, DC, pp. 1232–1245.

Hertz,G.Z. and Stormo,G.D. (1995) Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical byasis for penalizing gaps. In Lim,H.A. and Cantor,C.R. (eds), *Bioinformatics and Genome Research.* World Scientific, Singapore, pp. 201–216.

Hertz,G.Z., Hartzell III,G.W. and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Applic. Biosci.*, **6**, 81–92.

Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: A database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res.*, **26**, 55–60.

Karp,P.D., Riley,M., Paley,S.M. and Pelligrini-Toole,A. (1996) EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **24**, 32–39.

Neidhardt,F.C., Ingraham,J., Low,K.B., Magasanik,B., Schaechter,M. and Umbarger,H.E. (eds) (1987) *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC.

Neidhardt,F.C. *et al*. (eds) (1996) *Escherichia coli and Salmonella Cellular and Molecular Biology. Vols 1 and 2*, 2nd edn. American Society for Microbiology, Washington, DC.

O'Neill,M.C. (1989) Consensus methods for finding and ranking DNA binding sites. Application to *Escherichia coli* promoters *J. Mol. Biol.*, **207**, 301–310.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.

Rosenblueth,D.A., Thieffry,D., Huerta,A.M., Salgado,H. and Collado-Vides,J. (1996) Syntactic recognition of regulatory regions in *Escherichia coli*. *Comput. Applic. Biosci.*, **12**, 415–422.

Savageau,M.A. (1977) Design of molecular mechanisms and the demand for gene expression. *Proc. Natl Acad. Sci. USA*, **74**, 5647–5651.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences *J. Mol. Biol.*, **188**, 415–431.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences *Nucleic Acids Res.*, **12**, 505–519.

Thomas,R. and D'Ari,R. (1990) *Biological Feedback.* CRC Press, Boca Raton, FL.

Thieffry,D., Rosenblueth,D.A., Huerta,A.M., Salgado,H. and Collado-Vides,J. (1996) Definite-clause grammars for the analysis of *cis*-regulatory regions in *E. coli.* In Altman,R., Dunker,K., Hunter,L. and Klein,T. (eds), *Proceedings of the Pacific Symposium on Biocomputing '97.* World Scientific, Singapore, pp. 441–452.

Van Bogelen,R.A., Abshire,K.Z., Pertsemlidis,A., Clark,R.L. and Neidhardt,F.C. (1996) Gene-protein database of *Escherichia coli* K-12, Edition 6. In Neidhardt,F.C. *et al*. (eds), *Cellular and Molecular Biology: Escherichia coli and Salmonella*, 2nd edn. American Society for Microbiology, Washington, DC.