doi:10.1016/j.jmb.2005.09.037 J. Mol. Biol. (2005) 354, 184–199

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Comparative Studies of Transcriptional Regulation Mechanisms in a Group of Eight Gamma-proteobacterial Genomes

Vladimir Espinosa[1], Abel D. González[1], Ana T. Vasconcelos[2]

Araceli M. Huerta[3] and Julio Collado-Vides[3]*

[1]*National Bioinformatics Center, Industria y San José Capitolio Nacional, CP. 10200 Habana Vieja, Ciudad de la Habana, Cuba*

[2]*National Laboratory for Scientific Computing, Av. Getulio Vargas 333 Quitandinha, CEP 25651-075 Petropolis, Rio de Janeiro Brazil*

[3]*Center for Genomic Sciences UNAM, AP 565-A Cuernavaca CP 62100, Morelos, Mexico*

**Corresponding author*

Experimental data on the *Escherichia coli* transcriptional regulation has enabled the construction of statistical models to predict new regulatory elements within its genome. Far less is known about the transcriptional regulatory elements in other gamma-proteobacteria with sequenced genomes, so it is of great interest to conduct comparative genomic studies oriented to extracting biologically relevant information about transcriptional regulation in these less studied organisms using the knowledge from *E. coli*.

In this work, we use the information stored in the TRACTOR_DB database to conduct a comparative study on the mechanisms of transcriptional regulation in eight gamma-proteobacteria and 38 regulons. We assess the conservation of transcription factors binding specificity across all the eight genomes and show a correlation between the conservation of a regulatory site and the structure of the transcription unit it regulates. We also find a marked conservation of site-promoter distances across the eight organisms and a correspondence of the statistical significance of co-occurrence of pairs of transcription factor binding sites in the regulatory regions, which is probably related to a conserved architecture of higher-order regulatory complexes in the organisms studied. The results obtained in this study using the information on transcriptional regulation in *E. coli* enable us to conclude that not only transcription factor-binding sites are conserved across related species but also several of the transcriptional regulatory mechanisms previously identified in *E. coli*.

Keywords: *Escherichia coli*; transcriptional regulation; gamma-proteobacteria; regulon; TF binding site

Introduction

Since more bacterial genomes are being sequenced, it has become important to extend the transcriptional regulation studies to new organisms, in the search for deciphering their transcriptional regulatory networks and establishing comparative regulatory studies.[1–8] These studies may be used as the starting point for a multi-genomic prokaryotic regulation database, structured as RegulonDB.[9] The experimental studies carried out with *Escherichia coli* in the past years have produced an important group of databases containing detailed information on its genome organization and physiology,[9–12] which can be exploited to understand the mechanisms of transcriptional regulation in this enterobacterium. However, a lot of work remains to be done to describe the structure and characteristics of gene regulation in organisms related to *E. coli* whose genomes are already sequenced, most of which are

Abbreviations used: TF, transcription factor; TU, transcription unit; PWMs, positional weight matrices; *SOS*, site-orthology score; LGT, lateral gene transfer; TSS, transcription start site.

E-mail address of the corresponding author: collado@ccg.unam.mx

important pathogens. So it is of great interest to generate methodologies to extract regulatory information from those sequenced genomes starting from the knowledge of *E. coli*.

The evolution of DNA-binding sites is correlated to that of the proteins that bind to them,[6] which means that in order to conserve the recognition event, a change in the domain of interaction of the protein will impose a corresponding change of the DNA operator site.[13] A study made in *E. coli* and *Vibrio cholerae* revealed that, at least in those two organisms, there is a tendency for large regulons to evolve more slowly than small regulons.[6] Trying to generalize the behavior observed in this study is complicated because of the diversity and the incomplete understanding of the mechanisms of regulation of transcription in *E. coli*. For example, the functioning of the best studied system of regulation, σ70, is quite different from that of σ54,[14–17] about which far less is known in organisms other than *E. coli*. In addition to this, the interactions that might occur between transcription factors (TFs) and the cooperative relationships established in complex regulatory regions to exert different regulatory outputs[18] increase the complexity of the regulatory process significantly.

Another interesting problem is the structure of the genome and its relationship with transcriptional regulation. Structural changes in complete genomes have been examined in several eubacteria,[19–22] and the gene order has been shown to be generally unstable. It is not clear, however, what happens to the regulation of the fragments generated by the excision of a transcription unit (TU) during evolution or the resulting regulation of a TU formed by the fusion of two or more smaller TUs. The issue of structural changes of TUs across genomes has been assessed by Itoh *et al.*,[23] finding little restriction to the conservation of gene order within TUs, except for a group of operons like those encoding ribosomal proteins.[23] However, the unavailability of regulatory information about the organisms studied in the aforementioned work, limited their ability to correlate these comparative genome organization findings with their regulatory counterparts.

Many researches in the last decade have addressed the molecular organization of the regulatory system in *E. coli*. It is known that there is a correlation between the functionality of a TF binding site and the characteristics of its surroundings and its relative distance to the corresponding promoter.[24] In fact, in regulatory regions with binding sites for a single TF, it is possible to differentiate the distributions of site positions of activator and repressor sites. The first are concentrated in positions around and upstream from −40 and the second concentrate between −60 and +20. Those intervals correspond to positions in which the TF binds to sites that are close enough to interact directly with the transcription machinery, in the case of activator sites, or to binding sites that overlap the recognition area of the RNA polymerase near or between the −10 and −35 boxes in the case of repressor sites.

In complex regulatory regions, on the other hand, multiple sites for the same or different TFs form higher-order structures that exert a combined and non-additive regulatory output. In those cases it is more complicated to infer the functionality of a site given its distance to the promoter, because of the interactions that may exist among the TFs that bind different sites in the regulatory region. Actually, a TF-binding site with a given function may exist in different positions, depending on the function of the other sites that coexist in the same region.[25]

Functional interactions occurring among TFs closely located in the regulatory regions of genes, although more common in eukaryotes, occur also in prokaryotes.[26–32] Most of the methods used to discover putative TF-binding sites in genomic sequences use statistical models built from a set of known sequences that are used to search for occurrences of single sites resembling the starting model. However, a group of algorithms have been developed to search for occurrences of more complex patterns of TF-binding sites, ranging from pairs of sites[26,33] to grammatical representations of regulatory regions using computational linguistics.[18,34] Bulyk *et al*. have recently described an interesting method to estimate the statistical significance of the co-occurrence of TF-binding sites in *E. coli* regulatory regions.[35] The functionality of the most statistically significant predictions obtained using this approach were tested *in vivo* using RT-PCR, which proves the possible regulatory importance of the predictions made using such methodology.

In a previous work we developed the TRACTOR_DB database†,[36] which stores information about transcriptional regulation in 17 gamma-proteobacteria using two different predictive methodologies. Using the publicly accessible information about transcriptional regulation in the enterobacterium *E. coli*, we constructed general positional weight matrices (PWMs) starting from training sets enriched in known binding sites of each *E. coli* TF and containing the regulatory regions of orthologous TUs in the other 16 organisms. The resulting matrices were then used to scan the genomes, and the putative binding sites rescued in each organism were used to construct organism-specific matrices employed to re-scan the genomes for new TF-binding sites. We also constructed TF-binding site models using a Gibbs sampling method, which were used to scan the genomes independently. All the predictions rescued using both kind of models were merged and filtered using orthology information, which allowed us to score the biological significance of each putative site with respect to its conservation throughout the

† http://www.bioinfo.cu/Tractor_DB; http://www.tractor.lncc.br; http://www.ccg.unam.mx/Computational_Genomics/tractorDB

organisms studied. As a result of this predictive methodology, we were able to identify new members of 74 regulons in the 17 organisms studied. These results were then extended to 86 regulons using another methodology based on a pattern matching approach.[37]

Here, we report a comparative genomics study made by our group using the information stored in TRACTOR_DB[36] for 38 regulons in a subset of eight closely related gamma-proteobacteria, including *E. coli*. While other groups have reported that TF binding site sequences are conserved in the genomes of closely related organisms,[3,8] our results show for the first time that other important properties that are essential to transcriptional regulation are conserved across phylogenetically close organisms. Firstly, we made some variations and extensions to the work reported by Rajewsky *et al*.[6] to make an estimation of the degree of conservation of TF binding specificity across the organisms studied. We also found a correlation between the conservation of a given prediction across all the organisms and the conservation of the structure of the TU it regulates, complementing the findings reported by Itoh *et al*.[23] Secondly, we discuss the conservation of site-promoter distance profiles in the closest related organisms with marked over-representation of sites occurring at specific positions, as described by Collado-Vides *et al*. in *E. coli*.[24] Our results show that as the phylogenetic distance between the organisms compared increases, the conservation of the distance profiles tends to disappear. Lastly, we assess the statistical significance of the co-occurrence of sites predicted by us in the TRACTOR_DB work, and we find a conservation in the number of occurrences and statistical significance of co-occurring pairs of TFs in subsets of organisms. In other words, our results show that at this evolutionary scale both TF binding site sequences and higher-order transcriptional regulation complexes are probably conserved.

Results

Comparison of TF binding specificity in different genomes

In order to compare the characteristics of the real regulon size across the eight genomes, we use the TF binding specificities calculated using an information-based approach,[6] instead of making a comparison of the total number of the sites found for each TF (see Methods). We have used this estimator instead of the mere counting of sites because the latter is more variable among organisms. For example, the CRP regulon in *E. coli* contains 487 putative transcription units, while the same regulon in *Yersinia pestis*, with a genome similar in size to that of *E. coli*, contains 267 transcription units and that of *Shewanella oneidensis*, with a genome almost 1.1 times larger than that of *E. coli*, contains 235 TUs. On the other hand, the

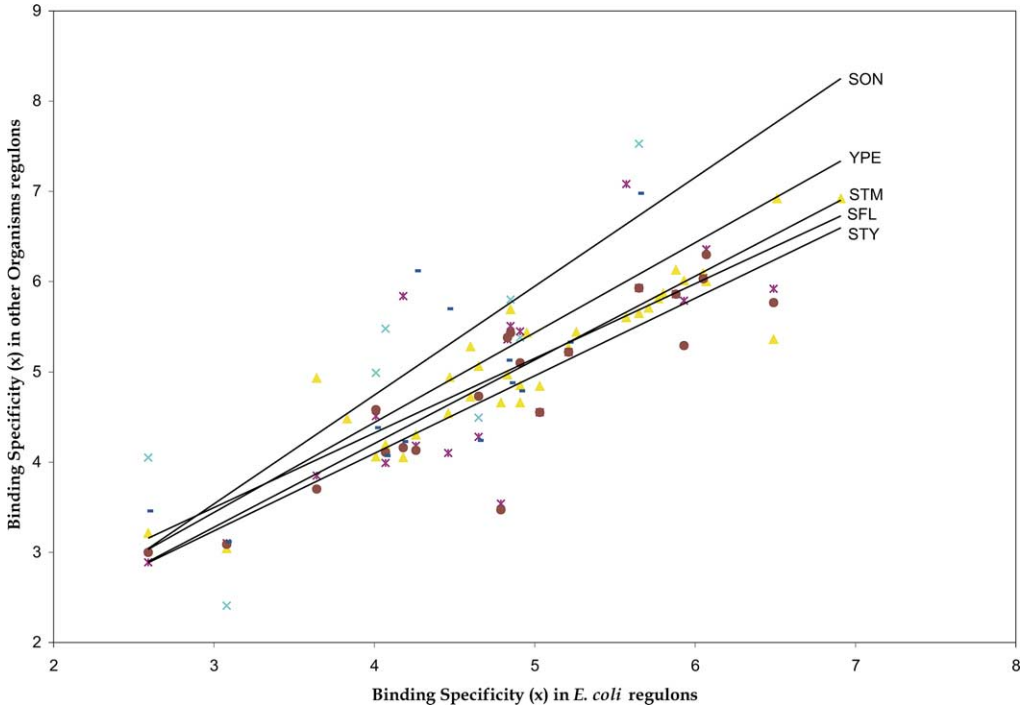

Figure 1. The binding specificities of TFs in different gamma-proteobacteria relative to *E. coli* regulons. In the graph: STY, *Salmonella typhi* ($y=0.8592x+0.6598$; $R^2=0.7983$); STM, *Salmonella typhimurium* ($y=0.9271x+0.4952$; $R^2=0.683$); SON, *Shewanella oneidensis* ($y=1.2062x-0.0816$; $R^2=0.6783$); SFL, *Shigella flexneri* ($y=0.8273x+1.0126$; $R^2=0.834$); YPE, *Yersinia pestis* ($y=0.9959x+0.4547$; $R^2=0.5998$).

CRP binding specificities in these three organisms are rather similar: 3.08, 3.12 and 2.41, respectively. The TF binding specificity is a term related to the signal-to-noise ratio that a TF must discriminate to exert its function of recognizing the set of sequences that belongs to its regulon.

A comparison of TF binding specificity in the different organisms studied in this work is shown in Figure 1, with *E. coli* regulons as reference. *V. cholerae* and *Haemophilus influenzae* were excluded from this comparison, since binding sites are found in these organisms for only a few regulons. In each organism, the lower values of the binding specificity distribution correspond to TFs that regulate a wide group of TUs, while small and compact regulons usually have higher values of binding

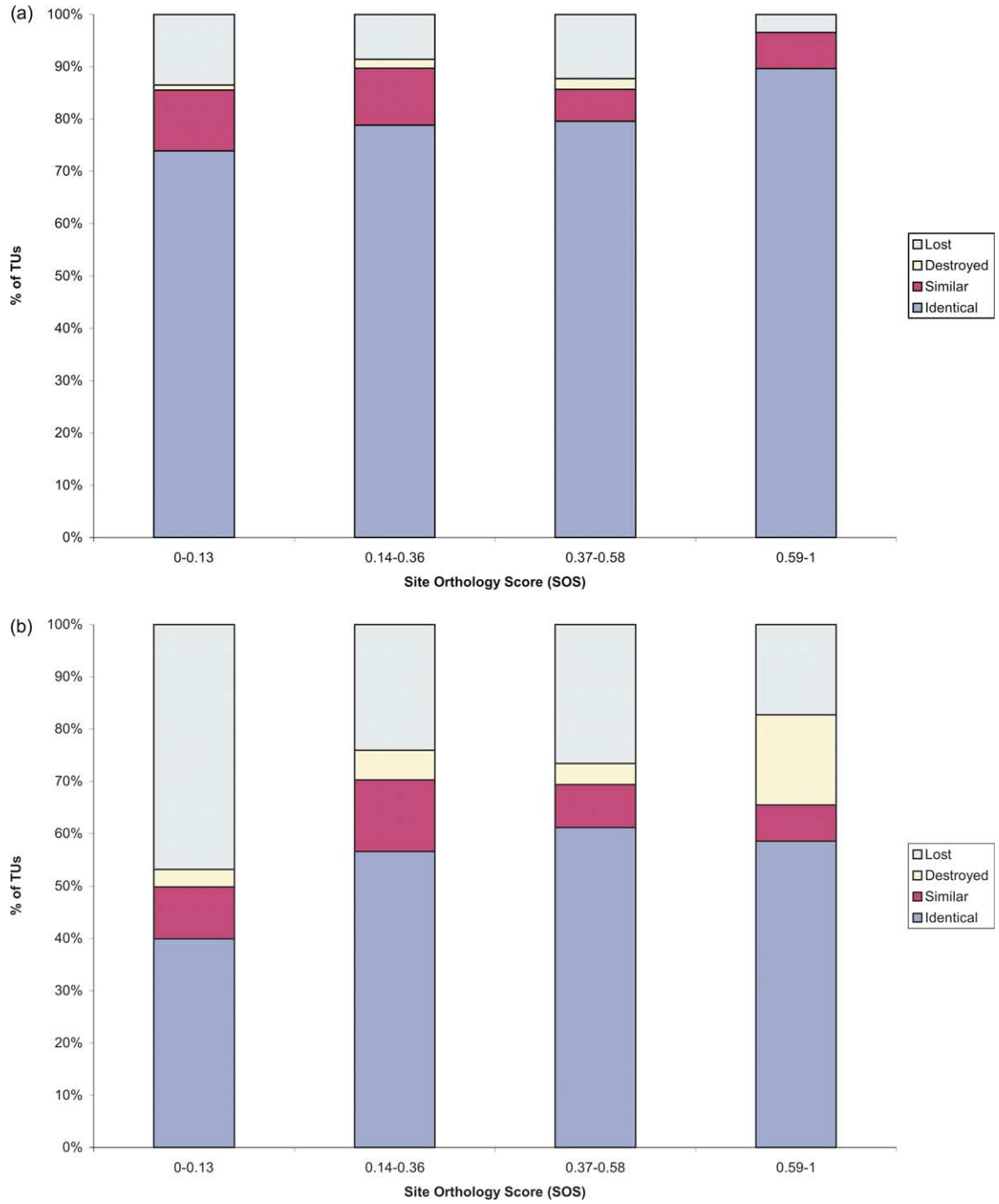

Figure 2. Conservation of TU structure and TU regulation in *S. flexneri* and *Y. pestis*. Fraction of operons in (a) *S. flexneri* and (b) *Y. pestis* resulting in each one of the four TU comparison categories when compared to *E. coli* TUs and the conservation of the sites existing in the regulatory regions of those TUs in the organisms studied estimated through the site orthology score (*SOS*). Each bar represents the percentage of TUs in the given organism belonging to each category containing sites in their regulatory regions within the given site orthology score range.

specificity, as in the case of *E. coli*, where FNR (2.59) and TrpR (6.91) are the bottom and top extremes of the binding specificity distribution, respectively. While the total number of putative sites in a given regulon is more variable across genomes, depending on the size and the base distribution of the genome and the characteristics of the organism-specific model used in the search, the binding specificity of a regulator tends to be conserved, as proved by the linear behavior observed in the plots depicted in Figure 1.

The conservation of the binding specificity calculated for all the TFs included in this study in each one of the eight genomes means that the distance between the two distributions composed of TF-specific binding sequences and background sequences (represented by the entire genome) is rather constant among all organisms. As can be seen in Figure 1, the slopes corresponding to each line, all of the five organisms against *E. coli*, are rather similar and very close to 1, which may be explained as the result of the conservation of the signal-to-noise ratio of each *E. coli* regulon and all orthologous regulons in the other organisms.

Conservation of regulatory signals and the structure of the TUs they regulate

An important point arising from comparative TU structure analyses,[22,23] is that TUs tend to have a lower degree of conservation as the evolutionary distance between organisms increases. An interesting question deriving from these studies that may be explored using our predictions stored in TRACTOR_DB is whether regulation is conserved despite the relative low degree of conservation of operon structure. So, we decided to find out if putative sites that are more conserved across gamma-preoteobacterial genomes occur upstream of TUs whose structure is more conserved within this group. We tracked down, for each *E. coli* TU, its fate in all other genomes (using the classification of TU structure conservation introduced by Itoh *et al.*,[23] Identical, Similar, Destroyed and Lost, see definitions in Methods). On the other hand, we designed a site-orthology score (*SOS*) to assess the conservation of the regulatory sites across all genomes (see Methods).

The results of this analysis in *Y. pestis* and *Shigela flexneri* are shown in Figure 2; the results obtained for the other five organisms are available as Supplementary Data. Each bar represents the percentage of TUs classified as Identical, Similar, Lost or Destroyed when compared to the *E. coli* orthologous TUs that have binding sites for all the TFs under study organized by ranges of *SOS*. As can be seen in Figure 2(a) and (b), and the rest of the Figures of the Supplementary Data, there is a tendency for predictions more conserved across organisms, those with higher *SOS*, to occur in the regulatory regions of TUs with more conserved structure. This corresponds to an increment of the fraction of predictions in the Identical-Similar

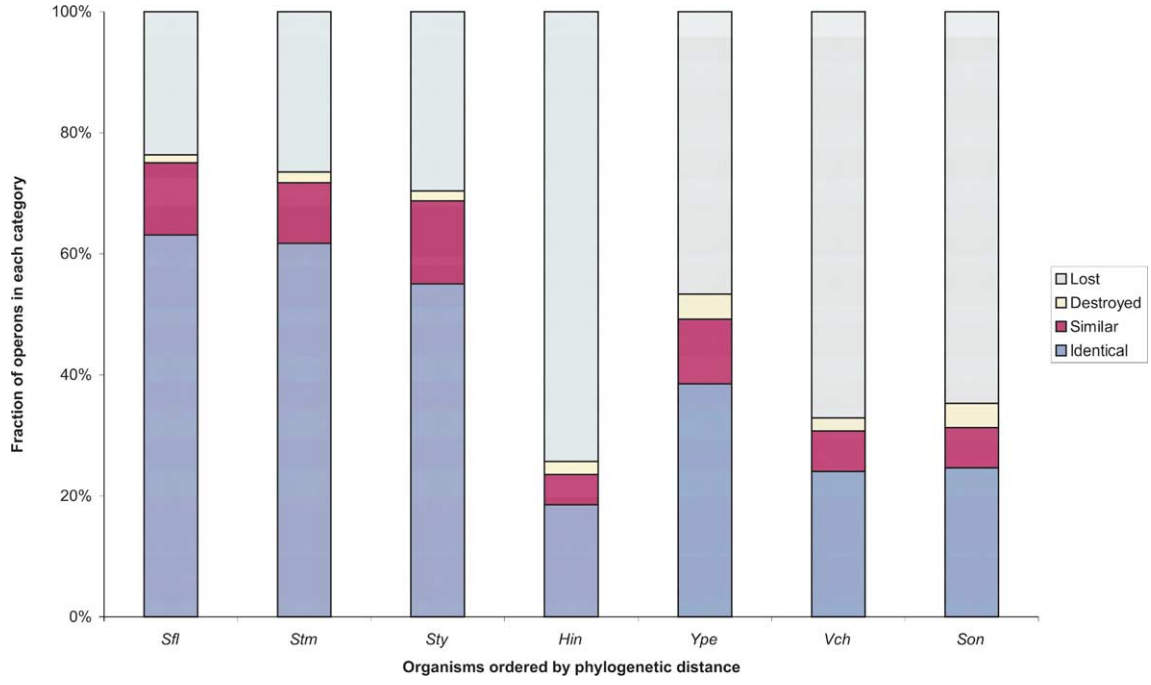

Figure 3. TU structure as a function of the phylogenetic distance with respect to *E. coli*. Relative number of operons belonging to each TU comparison category (*y* axis) in all the organisms ordered from left to right with respect to their phylogenetic proximity to *E. coli* (abscissa). The phylogenetic distance was estimated by calculating the fraction of genes in the given organism having orthologs in *E. coli*. In the Figure: *Sfl*, *Shigela flexneri*; *Sty*, *Salmonella typhi*; *Stm*, *Salmonella typhimurium*; *Hin*, *Haemophilus influenzae*; *Ype*, *Yersinia pestis*; *Vch*, *Vibrio cholerae*; *Son*, *Shewanella oneidensis*.

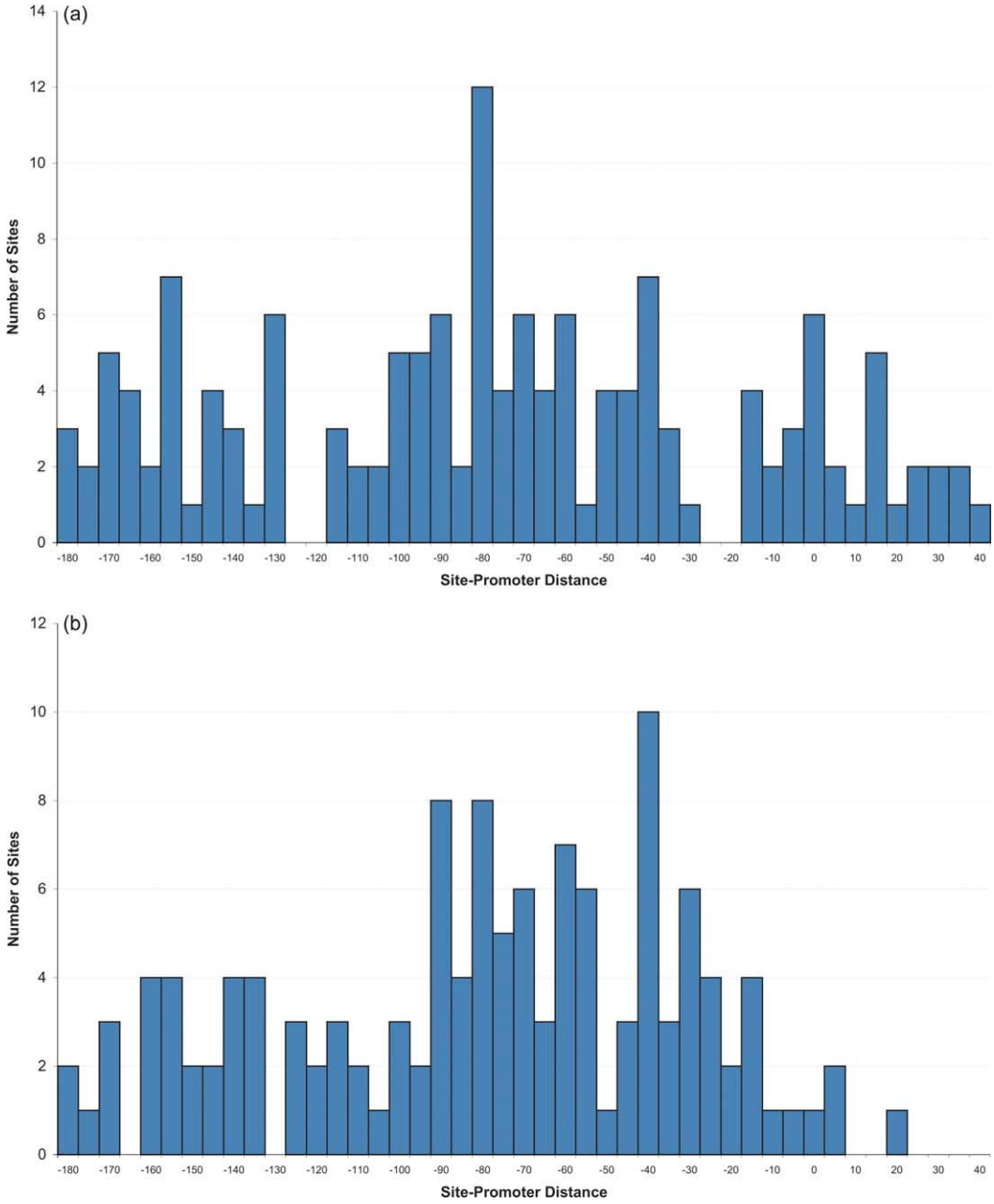

Figure 4 (*legend next page*)

category from left to right in the charts and a decrement in the Destroyed-Lost fraction.

We also plotted the fraction of TUs classified in each of the four aforementioned categories when compared to *E. coli*, ordered from left to right with respect to their phylogenetic proximity to *E. coli*. As can be seen in Figure 3, with the exception of *H. influenzae*, the percentage of Identical-Similar TUs decreases steadily as the phylogenetic distance increases, which is consistent with the results obtained by Itoh *et al*.[23] The different behavior observed for *H. influenzae* is mainly due to the size of its genome. We also counted the number of *E. coli* TUs whose corresponding genes are all present in other genomes but occur in more than one TU. Of 77 *E. coli* TUs whose structure is broken in at least one organism, only 16 regulatory sites (21%) are conserved upstream of all the TUs that result of the break event.

Site-promoter distance study

It has been shown that the structure of complex regulatory regions, the number and arrangements of sites, plays an important role in transcriptional regulation.[24,25] We explore the positional characteristics of the predictions found by calculating the site-promoter distance trying to find over-

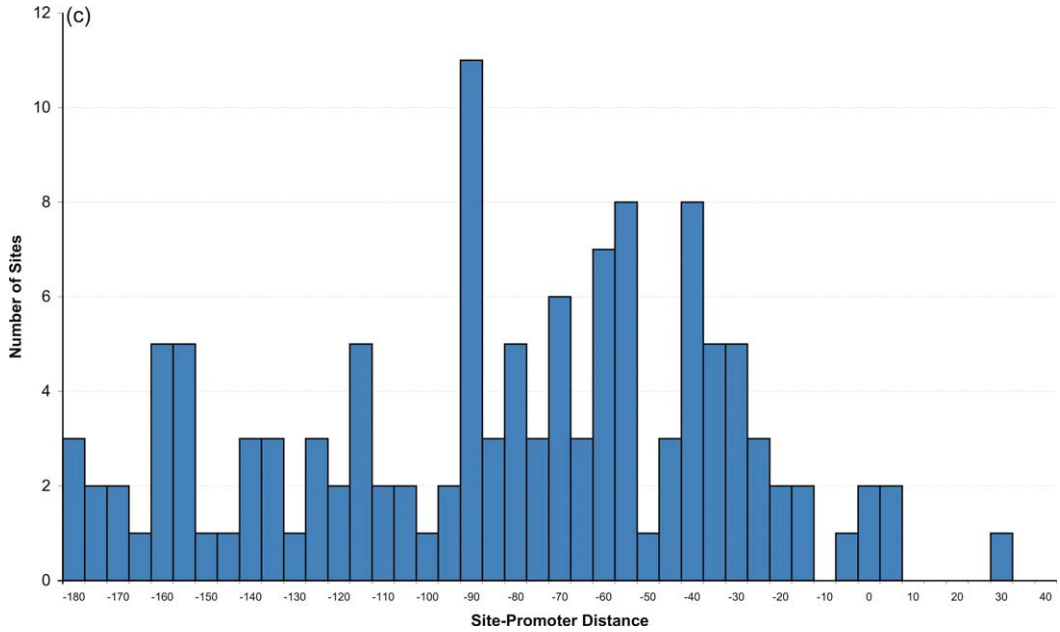

Figure 4. Site-promoter distance frequency histograms of the predictions in *S. flexneri*, *S. typhi* and *S. typhimurium*. The distance from the center of the putative site to the position 10 bp downstream from the center of the −10 box of the CRP predictions were plotted for (a) *S. flexneri*, (b) *S. typhi* and (c) *S. typhimurium* on the abscissa against the absolute number of predictions existing at a given distance from the promoter.

represented distance intervals in the regulatory regions of genes. All distances discussed here are computed relative to the initiation site of the corresponding promoter, based on a dataset of predicted promoters in several bacteria (A.M.H. *et al.*, unpublished results) generated by an extension of a former methodology developed by the same group (see Methods). The histograms obtained, depicted in Figure 4, show a clear conservation of distance ranges among more closely related organisms, such as *S. flexneri*, *Salmonella typhi* and *Salmonella typhimurium* for the CRP regulon. Conserved peaks are found at positions (−45→−41); (−65→−60) and (−85→−80). The conservation observed contrasts with the poor preference of sites for other positions. As expected, the histograms of the organisms more distant from *E. coli* differ considerably from what is obtained in closely related organisms, shown in Figure 4, when observed as a whole. However, the preferred distance ranges described above remain as conserved distance intervals in organism such as *S. oneidensis* and *V. cholerae*, despite having a smaller number of sites (data not shown).

We also generated the histograms of the experimentally characterized sites reported in RegulonDB[9] for the selected set of TFs used in this study (see Methods) in *E. coli* as an update of the previous report by Collado-Vides.[24] As can be seen in Figure 5, the distance intervals observed for the repressor and activator sites are different, and there is an almost complete separation between the peaks of both distributions, and even a similarity in the shape of the distributions obtained by us when compared to those reported by Collado-Vides. This result corresponds to the previous findings of dependence between the functionality and the position of the sites. The comparison between Figures 4 and 5 shows the correspondence of the distribution peaks obtained by using our predictions and those observed for the experimentally characterized sites, which opens a door to establishing a functional characterization of predicted sites from site-promoter distance information of known sites.

Conservation of the statistical significance of co-occurrence of putative binding sites

To examine the possible formation of higher-order regulatory arrangements of sites, we estimated the expected number and the probability of co-occurrences of sites in the regulatory regions of TUs in all the eight organisms and in all the defined bins (see Methods). Table 1 shows all the information of a group of the top-ranking pairs of TFs that we found co-occurring in the regulatory regions of genes in our predictions set (a complete list is available as Supplementary Data). As can be seen in Table 1, there is a correspondence between the values found and the phylogenetic distance among organisms, with a trend of co-occurrence of given pairs of TFs with similar statistical significance in closely related organisms.

The top-ranking pairs we found have been described for *E. coli* by Bulyk *et al.*,[35] as is the case for ArgR-ArgR, LexA-LexA, PhoB-PhoB and MetJ-MetJ. Those results were found despite the differences between the datasets used in both studies. In

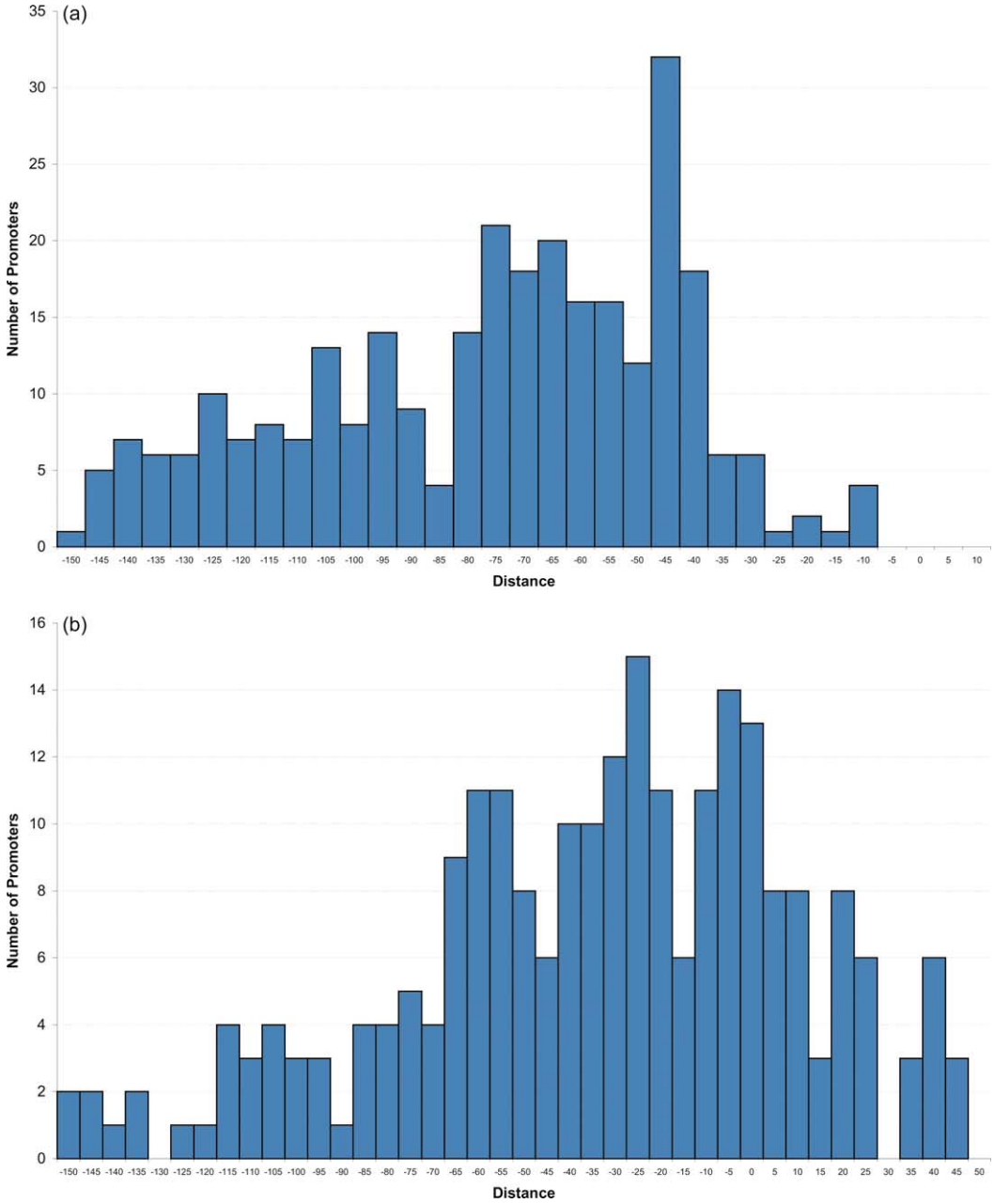

Figure 5. Frequency histograms of (a) the known activator and (b) repressor sites of the set of 38 TFs used to make the site-distance study. The information was downloaded from RegulonDB (v. 4.0)[9] and the distance from the center of each site with respect to the Transcription Start Site (TSS) was plotted against the absolute number of sites found at this position in the dataset.

addition, we found information about co-occurrence of those pairs in the other organisms studied.

Discussion

Although there is an obvious gap between the number of TUs regulated by global and specific TFs, this gap narrows dramatically when the analysis is centered in binding specificity. This binding-specificity measure can be regarded as the signal-to-noise ratio that the TF must discriminate when it binds its sites within a genome. The binding specificity is a better estimator of the signal-to-noise ratio recognized by the transcription factor than the absolute count of transcription units within the regulon, since it reflects how well represented is the space sequence recognized by the factor, and is independent of the number of known (or putative) binding sites. Our results indicate that the signal-to-noise ratio recognized by an *E. coli* TF is more or less equal to that recognized by the same TF in another genome. So, if a TF is identified as a global regulator in *E. coli*,[38] it is likely that it plays the same role in the transcriptional regulatory network in any other of the organisms included in this study. As can be

Table 1. Co-occurrence information in a group of TF–TF pairs

TFs	Bin	Number of occurrences	Organism	Expected occurrences	Probability
CRP-CRP	0–30	3	*Haemophilus influenzae*	1.211	0.12289
CRP-CRP	0–30	21	*Shigella flexneri* 2a	4.7304	3.26×10^{-8}
CRP-CRP	0–30	9	*Vibrio cholerae*	1.4904	2.64×10^{-5}
CRP-CRP	0–30	22	*Escherichia coli* K12	8.7975	0.00012807
CRP-CRP	0–30	31	*Salmonella typhimurium* LT2	6.9963	3.17×10^{-11}
CRP-CRP	0–30	40	*Salmonella typhi*	17.497	2.78×10^{-6}
CRP-CRP	0–30	12	*Yersinia pestis* KIM	1.7745	3.99×10^{-7}
CRP-CRP	0–30	35	*Shewanella oneidensis*	1.5733	2.45×10^{-12}
CRP-CRP	30–60	17	*Haemophilus influenzae*	1.0395	7.22×10^{-14}
CRP-CRP	30–60	48	*Shigella flexneri* 2a	4.3375	4.33×10^{-12}
CRP-CRP	30–60	11	*Vibrio cholerae*	1.2913	1.28×10^{-7}
CRP-CRP	30–60	50	*Escherichia coli* K12	7.6659	1.69×10^{-12}
CRP-CRP	30–60	55	*Salmonella typhimurium* LT2	6.1744	2.40×10^{-12}
CRP-CRP	30–60	58	*Salmonella typhi*	15.826	4.13×10^{-12}
CRP-CRP	30–60	29	*Yersinia pestis* KIM	1.6229	2.31×10^{-12}
CRP-CRP	30–60	31	*Shewanella oneidensis*	1.4428	4.91×10^{-13}
CRP-FNR	0–30	31	*Haemophilus influenzae*	0.85481	1.12×10^{-13}
CRP-FNR	0–30	168	*Shigella flexneri* 2a	4.3747	0
CRP-FNR	0–30	15	*Vibrio cholerae*	0.52512	5.44×10^{-15}
CRP-FNR	0–30	189	*Escherichia coli* K12	8.3963	0
CRP-FNR	0–30	189	*Salmonella typhimurium* LT2	5.2176	0
CRP-FNR	0–30	188	*Salmonella typhi*	14.161	0
CRP-FNR	0–30	74	*Yersinia pestis* KIM	1.183	0
CRP-FNR	0–30	20	*Shewanella oneidensis*	0.36014	5.61×10^{-13}
CRP-FNR	30–60	12	*Haemophilus influenzae*	0.73379	2.57×10^{-11}
CRP-FNR	30–60	56	*Shigella flexneri* 2a	4.0114	4.01×10^{-12}
CRP-FNR	30–60	8	*Vibrio cholerae*	0.45497	3.03×10^{-8}
CRP-FNR	30–60	56	*Escherichia coli* K12	7.3163	1.61×10^{-12}
CRP-FNR	30–60	48	*Salmonella typhimurium* LT2	4.6046	1.79×10^{-12}
CRP-FNR	30–60	57	*Salmonella typhi*	12.809	3.34×10^{-12}
CRP-FNR	30–60	19	*Yersinia pestis* KIM	1.0819	1.54×10^{-12}
CRP-FNR	30–60	7	*Shewanella oneidensis*	0.33028	6.37×10^{-8}
Fur-Fur	0–30	11	*Shigella flexneri* 2a	0.076004	4.91×10^{-14}
Fur-Fur	0–30	1	*Vibrio cholerae*	0.0083593	0.0083248
Fur-Fur	0–30	28	*Escherichia coli* K12	0.21695	3.11×10^{-14}
Fur-Fur	0–30	25	*Salmonella typhimurium* LT2	0.19467	2.69×10^{-13}
Fur-Fur	0–30	16	*Salmonella typhi*	0.42272	1.10×10^{-13}
Fur-Fur	0–30	27	*Yersinia pestis* KIM	0.045663	2.59×10^{-14}
ArgR-LexA	0–30	7	*Shigella flexneri* 2a	0.053717	2.76×10^{-13}
ArgR-LexA	0–30	8	*Escherichia coli* K12	0.099249	1.97×10^{-13}
LexA-LexA	0–30	3	*Shigella flexneri* 2a	0.0084412	9.84×10^{-8}
LexA-LexA	0–30	2	*Vibrio cholerae*	0.010827	5.78×10^{-5}
LexA-LexA	0–30	3	*Escherichia coli* K12	0.020269	1.36×10^{-6}
LexA-LexA	0–30	1	*Salmonella typhimurium* LT2	0.019904	0.019708
LexA-LexA	0–30	2	*Salmonella typhi*	0.059662	0.0017076
LexA-LexA	0–30	1	*Yersinia pestis* KIM	0.0043398	0.0043304
LexA-LexA	0–30	3	*Shewanella oneidensis*	0.0038707	9.43×10^{-9}

The first column indicates the pair of TFs co-existing in the same regulatory region, the second column indicates the distance interval where the predictions were found to co-occur, the third column indicates the number of TUs having predictions for both TFs, the fourth column indicates the organism where the predictions where found, the fifth column gives the expected number of co-occurrences calculated from the total number of predictions for each TF (N_{TFa} N_{TFb}) in the complete genome of the given organism and the probability of two randomly chosen base-pairs are separated by a distance $x(\pi(x))$, see Methods. The last column gives the probability, given the expected occurrences, of obtaining at least the observed number of pairs in the given distance range (second column), providing a measure of the statistical significance of obtaining the observed number of co-occurrences.

seen in Figure 1, the regression coefficient of the lines ranges from 0.59 to 0.83, with the lower values corresponding to organisms that are evolutionarily more distant from *E. coli*.

Itoh *et al*. have pointed out that when an operon is destroyed and divided into two or more TUs, the latter units require new regulation, otherwise they would become pseudogenes.[23] Our finding that most of such fragments (approximately 75%) in our data set apparently do not retain original regulation suggests that transcription of the destroyed operon may be subject to a more complicated or a different regulation.

Much work has been dedicated to study the correlation of the evolution of TF binding sites, TF sequences and the sequences of the genes they regulate in gamma-proteobacteria.[6,39] However, the results we have obtained open the possibility that such correlation may be extended to TU structure conservation. The analysis presented in Figure 3 reveals that the fraction of operons in each TU comparison category follows a linear behavior,

decreasing the fraction of TUs with Identical or Similar structure and increasing the fraction of operons whose structure is partially or completely lost as the phylogenetic distance increases with respect to *E. coli*. Our results show that in all the organisms studied, the TUs with more conserved structure generally have putative binding sites supported by orthology in most of the organisms, those with higher orthology scores in Figure 2 located at the right in the chart. Such highly conserved TUs may be part of what Erill *et al.*[39] have named the regulon core in the LexA case.[40] This result shows a clear tendency for similar regulation of TUs, which are more frequently conserved in different gamma-proteobacteria, which is related also to the fact that the evolutionary history of operon structure and that of its upstream regulatory elements is closely related.

The histograms of site-promoter distance show that TF-binding sites tend to appear in preferred positions in closely related organisms, as is the case of those shown in Figure 4. Previous studies had reported the relationship between the position of the sites in the regulatory regions and the functionality of the sites and the arrangements they form.[18,24,25] Here, we found that the distances where TF-binding sites occur do not follow a normal distribution in *E. coli*, with the peculiarity that some distance ranges are more populated than others. It is interesting that the shape of the histograms is similar in organisms like *S. flexneri* and *S. typhimurium*, whose genome size and physiology are similar to that of *E. coli*. This finding reveals that orthologous TUs in related organisms tend to conserve the sites recognized by the same TFs and the distance characteristics of the sites that exist in the regulatory regions of *E. coli* TUs. The histograms of organisms like *H. influenzae* or *V. cholerae*, which are more distant from *E. coli*, show marked differences from those depicted in Figure 4. This is mainly due to the characteristics of the methodology used to obtain the predictions stored in TRACTOR_DB based on the possibility of establishing orthology relationships between TUs of different organisms.[8] The smaller number of predictions in those distantly related organisms determines that the corresponding histograms are less populated and the overall shape is lost. However, the preferred distance ranges remain in those histograms, although the peaks observed include fewer sites (data not shown).

We also constructed the frequency histograms of the known sites reported in RegulonDB[9] for the selected TFs. As can be seen in Figure 5, the shape of the distributions and the location of the peaks for activator and repressor distributions are similar to those reported by Collado-Vides *et al.* in a more limited dataset of known sites.[24] The comparison of the distributions of known sites, which are functionally characterized, and those of the predictions match in punctual intervals i.e. $-45 \rightarrow -41$. Despite the existence of this correlation, we could not establish a methodology to make a functional classification of the predictions found in *E. coli* and the other seven organisms based on the site-promoter distance observed for the known sites, due to a series of drawbacks. In the first place, and most important, there is evidence for different mechanisms of action depending on the position of the TF, as is the case of promoters of type I and II activated by CRP and FNR.[41–43] On the other hand, promoters regulated by multiple sites of the same TF and complex promoters generally show a great variety of combinations of sites situated at different positions,[44,45] which determines the final regulatory output.

A close examination of Table 1 shows some interesting aspects about the co-occurrence of sites in the organisms included in this study. We obtained co-occurrences of sites of CRP with up to 30 different TFs, some of which have been reported, as is the case of CRP-CytR[27] and CRP-AraC.[46] We also found a high number of co-occurrences of CRP sites with itself and with FNR, which is consistent with previous findings in *E. coli*. There is an interesting correlation of the observed number of co-occurrences and the statistical significance in the case of CRP-CRP and CRP-FNR with the phylogenetic distance between organisms, being clearly the trend of closely related organisms to have similar numbers of co-occurrences and probability values, as can be seen in Table 1 both for the 0–30 and 30–60 bins. It is worth noting the high number of CRP-FNR co-occurrences, related to lower, highly significant probability values, even higher than the number corresponding to CRP with itself. This is mainly due to the high level of similarity of the PWMs used to scan the genomes, which is related to previous experiments that allowed interconverting the site recognition specificities of CRP and FNR by replacing a single base at the hemisites,[47] or by substituting the appropriate amino acid residues at the recognition helices of both factors.[48]

The results obtained for Fur-Fur co-occurrences are also significant, due to their high statistical significance in all the organisms with the exception of *V. cholerae*, with a single case of co-occurrence, and *H. influenzae* where the factor is missing. Fur is said to be a global regulator related to iron uptake, oxidative response and acid tolerance response in *E. coli*, *S. oneidensis* and *S. typhimurium*.[49–51] Some clues have been found about the possibility of formation of an arrangement of two close positioned dimers, each recognizing one site to form a stable TF-DNA complex.[49] Our results show that there is a conserved pattern of co-occurrences in the 0–30 bin in almost all the organisms, which might be related to a similar mechanism of DNA recognition and way of action for Fur. Other interesting results are those of the ArgR-LexA and LexA-LexA co-occurrences. They correspond quite well to those obtained by Bulyk *et al.*[35] in the case of LexA in *E. coli* and we also found a conservation of the co-occurrence pattern in *S. flexneri* and *S. oneidensis*. In the case of ArgR-LexA coexistence,

we have not found any previous report of LexA co-occurrence with any other TF. The high significance found by us is interesting because the two phylogenetically closest organisms, *S. flexneri* and *E. coli*, have similar number of co-occurrences in the 0–30 bin which may be a signal of a possible interaction not yet described.

The fact that new regulon members in all organisms other than *E. coli* were predicted using information that comes exclusively from *E. coli*, or is biased towards this organism, see above, poses the question of whether the results of the comparisons conducted in this study have arisen due to circularity. The term circularity in this context means that regulatory sites, and regulons as a whole, might show the same behavior as those of *E. coli* as a result of orthology rather than due to the conservation of the basic mechanisms of transcription. However, two steps of the methodology used to predict new regulon members in the organisms included in this study prevent the occurrence of this circularity: the reconstruction of TF binding sites weight matrices to adjust them to each organism, and the use of each organism as the center of the orthology filtering.[36]

The combination of these two steps allows recovering putative regulatory sites in each organism, even if they do not have an orthologous site in *E. coli*, either because they have an ortholog in a third organism or because they score above the strong cutoff in that organism using the rebuilt matrix.[36] The reconstruction of matrices actually changes the structure of the binding site model searched for in each organism. This change allows recovering new putative sites that do not score above the threshold when the original matrices, enriched in *E. coli* sequences, are used.

As a result of the combination of the two aforementioned steps, the regulons reconstructed in all seven organisms other than *E. coli* are not a subset of those reconstructed in that enterobacterium. For example, almost half of all members of the CRP regulon in *S. typhimurium* either do not have an ortholog in *E. coli* or their orthologs do not belong to the CRP regulon in this organism (data not shown). This implies that the list of promoters under CRP regulation is qualitatively different in *E. coli* and *S. typhimurium*. Nevertheless, the distribution of promoter-regulatory sites distances is conserved in these two organisms, as can be seen in Figure 4.

Figure 6 of the Supplementary Data illustrates the diversity in composition of the FUR (A) and Lrp (B) regulons in *E. coli* (Eco), *S. typhi* (Sty), and *S. typhimurium* (Stm) by two Venn diagrams. The comparison of the composition of all regulons included in Tractor_DB[36,37] reveals the same variability. Yet, the TF binding specificity or the signal-to-noise recognition ratio is conserved across these three organisms, despite the variability in regulon sizes among them. These findings suggest that the influence of circularity in the results presented here may be dismissed.

Another interesting aspect that reinforces the validity of the predictions used in this study is the one regarding the conservation of gene function across the regulons and organisms studied. It has been reported that regulons are very well conserved structures across related species and despite its members are usually susceptible to Lateral Gene Transfer (LGT) the regulon as a whole, and the regulatory proteins related to it, tend to be quite stable from the evolutionary point of view.[40] The evolutionary stability of a regulon can be correlated with its gene contents[6] and self-regulation.[38] It seems evident that, in the case of a large and self-regulated gene network, regulon structure (i.e. regulatory protein, regulon functional core genes and regulatory motifs) will tend to be preserved because a mutation either in the gene encoding the regulatory protein or its operator region, will often lead to severe deregulation and, thus, to a substantial disruption in cellular equilibrium.[39] Regulon conservation has been confirmed in bacterial genomes for some specific regulons, as is the case of ArgR[52] and FUR.[53]

Recent research of the FUR[53] and LexA[39,54] regulons have given some insights into the characteristics of these regulons across alpha and gamma-protaobacteria. In contrast to the initial ideas of regulon structure, it has been proposed that regulons present double information content: the regulon core with an evolutionary stable structure and the global gene set, or the periphery, more prone to variation among organisms.[39] The results obtained by us for the 38 regulons that are the subject of this study (and depicted in Table 2 of the Supplementary Data) are consistent with these previous findings. Table 2 of the Supplementary Data summarizes the variability of gene content across five of the eight organisms for each one of the 38 regulons studied when compared to their *E. coli* counterparts. *V. cholerae* and *H. influenzae* were excluded from this Table because the number of TF-binding sites found in these organisms for the regulons studied is very low. For each organism, there is a description of the percentage of genes that leave the regulon: those for which no orthologs were found (column GP) and those for which an ortholog is found but not regulated by the orthologous TF in that organism (column IP) as well as those genes entering the regulon: those genes that are regulated by the orthologous TF in a specific organism but not in *E. coli* (column NC). The number of conserved interactions, genes with orthology relationships in all organisms and regulated by the same TF in all organisms, are summarized in column C.

As can be seen in Table 2 of the Supplementary Data, there is a variability in the composition of regulons across the genomes studied, with a group of genes regulated by the same TFs that have orthology relationships in all the organisms (regulon core genes), column C of each organism-specific section, and another group of genes that either are lost in the subject organism, no ortho-

logous genes were found there, or for which orthologous genes were found but belonging to different regulons, summarized in the first (GP) and second (IP) columns of each organism-specific section, respectively. In the same way there is a group of new genes within the regulon in each specific organism, fourth column (NC) of each organism-specific section, that also accounts for the variability of this regulon when compared to its counterpart in *E. coli*.

The characteristics of regulons of global regulators like CRP and FNR, having a large number of target operons, are very similar to those of other global regulators regulating a smaller number of genes like the case of FUR. As can be seen in the aforementioned Table, the percentage of genes belonging to the regulon core, conserved throughout the organisms, decreases as the phylogenetic distance among organisms increases,[54] with a corresponding increase of the variability of the genes forming the regulon's periphery, sumarized in the first (GP), second (IP) and fourth (NC) columns of Table 2 of the Supplementary Data. This increase in the variability of regulon's periphery as the phylogenetic distance increases is expected, because the number of phenomena like LGT, deletions and mutations related to the adaptation of the organisms to their specific habitats are more important. However, this variability is far less evident in the regulon core set, formed by the genes related to the central functions of the regulon.

In addition to this, the variability of both regulon core and regulon periphery content is higher in regulons of global regulators than in small regulons corresponding to specific regulators like LexA or NtrC. Certainly, global regulators are specialized to recognize a wide group of motifs with different affinities governing the expression of multiple operons, frequently involved in different functional classes. This high-level organization of multiple functions within a global regulator seems to be more variable across species. It is reasonable that global regulators, having lower values of binding specificity, as can be seen in Figure 1, are more susceptible to having organism-specific genes entering the regulon as well as genes that belong to the regulon in *E. coli* changing of regulation in other organisms, because their binding domains can adapt more easily to changes that mutations can cause in TF-binding sites as well to regulate new genes acquired by LGT. In this respect, it is important to note that global regulators generally exert their function in conjunction with other specific TFs that may contribute to the adaptation of the variability of regulon gene content.

A closer look into some specific regulons reveals that some important genes described previously as part of the regulon core in the LexA regulon,[39,54] such as *recA*, *recN*, *lexA*, *uvrAB*, *ssb* and *feoA*, *fhuAF* and *sodA* in the FUR regulon,[53] are part of the regulon core according to our results (Table 3 of the Supplementary Data). In the case of the LexA regulon, we also identified the gene *sulA* as one of the more conserved across the closely related organisms, with the exception of *H. influenzae*, *S. oneidensis* and *V. cholerae*, which is consistent with the proposed relevance of this gene in the control of gene variability in the LexA regulon and the recent appearance of this cell division inhibitor in the evolution of this regulon.[39] Another interesting case is the conservation of the genes belonging to the *glnALG* operon of the NtrC regulon. All these genes are related to the central functions of this important regulon, taking part in nitrogen assimilation and the adaptation of bacteria to different nitrogen source media.[55] The three gene products of this regulon, the proteins glutamine synthetase and proteins NR_I and NR_{II}, which respond coordinately to the exterior signals and regulate all the genes belonging to this regulon,[55] were identified as conserved across all the organisms studied, with the exceptions of *H. influenzae* and *V. cholerae*, where we did not find predictions for this TF. The results obtained by us for the 38 regulons studied show that this is a general trend in the regulatory network: regulon content varies from one organism to the next, with the exception of the regulon core, formed by genes that also tend to be located in transcription units with more conserved structures, which is more stable in terms of gene content (Table 3 of the Supplementary Data).

This result of the conservation of gene function across regulons and genomes and its relation with the results obtained in previous studies, as well as the specificities of the predictive methodology followed to avoid circularity in the predictions stored in TRACTOR_DB,[36] gives additional support to the data used in this study and validates our conclusions about the conservation of the basic mechanisms of transcription regulation in closely related organisms.

Concluding Remarks

We found some preliminary clues that give insights into the conservation of the mechanisms of transcription regulation in eight closely related gamma-proteobacteria and 38 regulons. We were able to correlate the characteristics of transcription regulation and genome organization in a group of eight gamma-proteobacteria, all of which have been studied poorly from this point of view. Our results show that TFs play similar roles in the regulatory networks of closely related organisms, which is consistent with similar binding specificity values across the genomes studied for global and specific regulators. There is also a correspondence between the strength of the orthology support of a TF-binding site, the number of organisms having orthologous TUs with sites for the same TF, and the structure of these TUs. This suggests the existence of important links between transcriptional regulation and genome organization, which have

not been described yet in such a wide group of related organisms.

We found that the behavior observed in *E. coli* for TF-binding sites located at specific distances from the corresponding promoter is found also in the other gamma-proteobacteria studied. Besides, our findings show a trend of predictions to appear in specific distance ranges in all the organisms, a trend that is more marked in closely related organisms. Exploring the possibility of functional interactions that might exist among TFs in our set of predictions for *E. coli*, we found that some TFs are more likely to appear in the same regulatory regions (co-occurrence of TF-binding sites), at specific distances from others TFs. In this case, we found again that this behavior was conserved in the other gamma-proteobacteria related to *E. coli*. All these results point to the similarity of transcriptional regulatory mechanisms in the organisms studied. To our knowledge, this is the first comparative study that deals with the information on the mechanisms and organization of transcriptional regulation in *E. coli* to unravel important structural and functional similarities of these mechanisms in other related organisms on such a large scale.

Methods

Selecting organisms and data

TRACTOR_DB[36] contains information on transcriptional regulation in 17 closely related gamma-proteobacteria and we selected eight out of those 17 on the basis of their phylogenetic proximity to *E. coli*. Hence, we included in our study those gamma-proteobacteria with at least 30% of their genes having orthologs in *E. coli*, given the fact that the predictive methodology used to obtain the predictions stored in TRACTOR_DB relies on orthology relationships found between organisms, and for those more distant from *E. coli* only few predictions were found (for a thorough description see González *et al*.[36]).The organisms selected for the present study were: *Escherichia coli* K12 (NC_000913), *Haemophilus influenzae* (NC_000907), *Salmonella typhi* (NC_003198), *Salmonella typhimurium* LT2 (NC_003197), *Shewanella oneidensis* (NC_004347), *Shigella flexneri* 2a (NC_004337), *Vibrio cholerae* (NC_002505) and *Yersinia pestis* KIM (NC_004088).

In order to calculate site-promoter distances in our predictions, we used the promoter predictions generated by means of the methodology described by Huerta & Collado-Vides in each of the organisms mentioned above.[56] These data were kindly delivered to us by Dr Araceli M. Huerta (personal communication).

TFs binding specificity estimation

For each TF in *E. coli*, we apply the formula described by Rajewsky *et al*.[6] to estimate the real size of the corresponding regulon instead of just counting the total predictions rescued during the search:

$$x = \frac{m_{\text{experimental}} - m_{\text{random}}}{\sqrt{\sigma^2_{\text{experimental}} + \sigma^2_{\text{random}}}}$$

In this formula $m_{\text{experimental}}$ is the mean of the distribution of scores of putative sites' sequences and m_{random} is the mean of the distribution of scores of the whole genomic sequence. The denominator of the expression is the combined variance of these two distributions. The site scores were calculated by scanning the regulatory regions with PATSER using the matrix corresponding to a TF in the case of $m_{\text{experimental}}$ and the site scores obtained when scanning the complete genome for the same TF matrix in the case of m_{random}. After calculating the x value for the given TF in *E. coli*, the procedure was repeated in each organism and the results obtained were plotted to construct Figure 1.

Transcription unit classification and orthology score calculation

Orthologous TUs were defined as those containing at least one orthologous gene between one given organism and *E. coli*, and the orthology relationships were established following the definition given by Huynen & Bork,[57] using the best bi-directional BLAST hits (BBHs). The classification of TU structure was done as proposed by Itoh *et al*.[23] according to that, two TUs are classified as Identical if their structures, number of genes and order, are identical. The TUs may be classified as Similar if the structure was conserved in part, with translocations, deletions and two insertions allowed (this means that two TUs are similar even if the number of genes in the operon vary due to insertions of, at most two, new genes and deletions and/or the structure vary due to translocations of one or more genes within the operon). The term Destroyed is used when at least two orthologous genes were found and the structure of the *E. coli*'s TU was not conserved (two TUs were classified within this category when two or more ortholog were found in the other genome belonging to independent transcription units) and the term Lost is used when one or no orthologs are found, so that the comparison of the structure cannot be estimated. For a more complete description of the operon categories, please see Itoh *et al*.[23]

The conservation of a regulatory site in a collection of n genomes was measured by means of the site-orthology score:

$$SOS = \sum_{\text{organism}=1}^{n} a_i A_i$$

In this formula a_i is a binary coefficient with value 1 if a site for the same TF is found both in the regulatory regions of an *E. coli* TU and the corresponding orthologous TU in organism i and zero if it is not; and A_i is a weight that reflects the evolutionary distance between organism i and *E. coli* (we defined A_i as 1 minus the fraction of orthologous genes shared by the two organisms, which means that A_i values are closer to zero for closely related organisms). Scores were normalized to 1 by dividing by the highest value. According to this, each TU with a putative site found within its regulatory region has an orthology score associated with it on the basis of those orthologous TUs that have a regulatory site for the same TF in other organisms.

Calculation of the site-promoter distance in our predictions

We restrict the site-promoter distance study to a group

of 38 of the 74 TFs for which we reported predictions in TRACTOR_DB. The selected TFs were those for which we could reconstruct the models in each organism (see Gonzalez *et al*.[36] Supplementary Data for a thorough description), producing an organism-specific PWM used to rescue the predictions that formed the dataset used in this study (LexA, CRP, FNR, MetJ, PurR, PhoB, OmpR, CadC, GcvA, AppY, FruR, MelR, NtrC, GalS, ArgR, DnaA, Lrp, TrpR, GlcC, Fur, LysR, FadR, ArcA, NarL, ModE, PhoP, Ada, AraC, Mlc, FabR, FarR, TyrR, Nac, GntR, CsgD, BirA, CaiF, NadR).

Using all the predictions supported by orthology, or those that exceeded the strong cutoff, we calculated the corresponding distance to each promoter predicted in the same operon. The predictive methodology described by Huerta & Collado-Vides[56] finds sometimes multiple promoter signals in the regulatory region of the same TU. In those cases, we calculated the distance from the center of each site lying in a TU regulatory region to each promoter predicted in the same regulatory region.

The methodology used to predict the promoters returns, in each case, the corresponding positions of the −10 and −35 boxes, the distance between them and other information, such as the scores of the alignment of each box (A.M.H. *et al*., unpublished results). In the case of the promoters that are experimentally characterized and reported in RegulonDB,[9] there is information about their transcription start site (TSS), the transcription +1 position. However, the methodology used to generate the promoter predictions used in this study does not provide any insight about the position of the TSS. To make our calculations, we had to consider the position of the TSS exactly ten bases downstream with respect to the center of the −10 box, even when it has been reported that the distance from the −10 hexamer to the +1 might vary from four to 12 base-pairs.[58]

Calculation of the statistical significance of co-occurrence of pairs of sites in the regulatory regions

For all the TFs coexisting in the same regulatory region with other TFs, either different or identical, the estimation of the statistical significance of the co-occurrence of putative sites in the regulatory regions of the TUs was done in each organism following the methodology described by Bulyk *et al*. in *E. coli*.[35] We established eight spacing bins, each corresponding to a variable distance range between 0 bp and 450 bp (0–30 bp, 30–60 bp, 60–90 bp, 0–100 bp, 100–200 bp, 200–300 bp, 300–400 bp and 0–450 bp). The rank of each pair of sites was based on the probability of obtaining the observed number of hits for the most over-represented bin,[35] and was calculated using the following formula:

$$P(\text{bin}) = 1 - \sum_{S=0}^{\text{obs(bin)}-1} \frac{Na \cdot Nb}{S} \cdot \prod^{S} \cdot (1 - \prod)^{Na \cdot Nb - S}$$

where N_a and N_b are the total number of putative sites in the genome for *TFa* and *TFb* respectively, S is an index variable in the summation running from 0 to the observed number of co-occurrences minus 1, obs(bin) is the observed number of co-occurrences in the corresponding bin and Π is the sum of the π values across the complete length of the spacing bin, which was calculated using the formula:

$$\prod = \sum_{x=0}^{\text{binsize}} \pi(x)$$

In this formula $\pi(x)$ is the probability that two randomly chosen non-coding base-pairs are separated by a distance x.

$\pi(x)$ was computed by tabulating the number of occurrences of all possible pairwise combinations of non-coding bases in a set of pure non-coding regions extracted from the genome of each organism. The $\pi(x)$ *versus* distance correlations obtained were used to calculate the values of Π needed in the formulas described above.

Acknowledgements

This work was funded, in part, by a collaboration project on Bioinformatics between Cuba and Brazil supported by CNPq/MCT. We acknowledge NIH-NIGMS grant GM071962-01 to J.C.-V. and the Iberoamerican Bioinformatics Network (Red Iberoamericana de Bioinformatica-RIBIO VII.L, CYTED) for partial support for this project.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.09.037

References

1. Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E. & Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucl. Acids Res.* **31**, 266–269.
2. Aguilar, D., Oliva, B., Aviles, F. X. & Querol, E. (2002). TranScout: prediction of gene expression regulatory proteins from their sequences. *Bioinformatics*, **18**, 597–607.
3. McCue, L., Thompson, W., Carmack, C., Ryan, M. P., Liu, J. S., Derbyshire, V. & Lawrence, C. E. (2001). Phylogenetic footprinting of TF binding sites in proteobacterial genomes. *Nucl. Acids Res.* **29**, 774–782.
4. Mirny, L. A. & Gelfand, M. S. (2002). Structural analysis of conserved base pairs in protein–DNA complexes. *Nucl. Acids Res.* **30**, 1704–1711.
5. Moreno-Hagelsieb, G. & Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
6. Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002). The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res.* **12**, 298–308.
7. Pérez-Rueda, E. & Collado-Vides, J. (2001). A common origin of transcriptional repression by helix-turn-helix proteins in the context of the evolution of regulatory families in Archea and Eubacteria. *J. Mol. Biol.* **53**, 172–179.
8. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. & Stormo, G. (2001). A comparative genomics approach to prediction of new members of regulons. *Genome Res.* **11**, 566–584.
9. Salgado, H., Gama-Castro, S., Martínez-Antonio, A., Díaz-Peredo, E., Sánchez-Solano, F., Peralta-Gil, M.

et al. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucl. Acids Res.* **32**, 303–306.
10. Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucl. Acids Res.* **28**, 60–64.
11. Karp, P. D., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I. T. & Saier, M. H., Jr (2004). The *E. coli* EcoCyc database: no longer just a metabolic pathway database. *ASM News*, **70**, 25–30.
12. Serres, M. H., Goswami, S. & Riley, M. (2004). GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucl. Acids Res.* **32**, D300–D302.
13. Collado-Vides, J. (1991). The search for a grammatical theory of gene regulation is formally justified by showing the inadequacy of context free grammars. *CABIOS*, **7**, 321–326.
14. Collado-Vides, J. (1996). Towards a unified grammatical model of σ70 and σ54 bacterial promoters. *Biochimie*, **78**, 351–363.
15. Guo, Y., Lew, C. M. & Gralla, J. D. (2000). Promoter opening by sigma 54 and sigma 70 RNA polymerases: sigma factor-directed alterations in the mechanism and tightness of control. *Genes Dev.* **14**, 2242–2255.
16. Burrows, P. C., Severinov, K., Ishihama, A., Buck, M. & Wigneshweraraj, S. R. (2003). Mapping σ54-RNA polymerase interactions at the -24 Consensus Promoter Element. *J. Biol. Chem.* **278**, 29728–29743.
17. Wigneshweraraj, S. V., Kuznedelov, K., Severinov, K. & Buck, M. (2003). Multiple roles of the RNA polymerase beta subunit flap domain in sigma 54-dependent transcription. *J. Biol. Chem.* **278**, 3455–3465.
18. Rosenblueth, D. A., Thieffry, D., Huerta, A. M., Salgado, H. & Collado-Vides, J. (1996). Syntactic recognition of regulatory regions in *Escherichia coli*. *Comput. Appl. Biosci.* **12**, 415–422.
19. Mushegian, A. R. & Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**, 289–290.
20. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayse, W. S., Borodovsky, M. *et al*. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.
21. Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. (1997). Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucl. Acids Res.* **25**, 701–712.
22. Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. (1997). Genome plasticity as a paradigm of eubacterial evolution. *J. Mol. Evol.* **44**, S57–S64.
23. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16**, 332–346.
24. Collado-Vides, J., Magasanik, B. & Gralla, J. D. (1991). Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.* **55**, 371–394.
25. Collado-Vides, J. (1992). Grammatical model of the regulation of gene expression. *Proc. Natl Acad. Sci. USA*, **89**, 9405–9409.
26. GuhaThakurta, D. & Stormo, G. D. (2001). Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
27. Pedersen, H. & Valentin-Hansen, P. (1997). Protein-induced fit: the CRP activator protein changes sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. *EMBO J.* **16**, 2108–2118.
28. Sudarsanam, P., Pilpel, Y. & Church, G. M. (2002). Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res.* **12**, 1723–1731.
29. Hannenhalli, S. & Levy, S. (2002). Predicting transcription factor synergism. *Nucl. Acids Res.* **30**, 4278–4284.
30. Banerjee, N. & Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucl. Acids Res.* **31**, 7024–7031.
31. Makeev, V. J., Lifanov, A. P., Nazina, A. G. & Papatsenko, D. A. (2003). Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucl. Acids Res.* **31**, 6016–6026.
32. Kreiman, G. (2004). Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucl. Acids Res.* **32**, 2889–2900.
33. Liu, X., Brutlag, D. L. & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* **6**, 127–138.
34. Collado-Vides, J. (1993). A linguistic representation of the regulation of transcription initiation. I. An ordered array of complex symbols with distinctive features. *BioSystems*, **29**, 87–104.
35. Bulyk, M. L., McGuire, A. M., Masuda, N. & Church, G. M. (2004). A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res.* **14**, 201–208.
36. Gonzalez, A. D., Espinosa, V., Vasconcelos, A. T., Pérez-Rueda, E. & Collado-Vides, J. (2005). TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucl. Acids Res.* **33**, D98–D102.
37. Hernandez, M., Gonzalez, A. D., Espinosa, V., Vasconcelos, A. T. & Collado-Vides, J. (2004). Complementing computationally predicted regulatory sites in TRACTOR_DB using a pattern matching approach. *In Silico Biol.* **4**, 0020.
38. Martínez-Antonio, A. & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* **6**, 482–489.
39. Erill, I., Escribano, M., Campoy, S. & Barbé, J. (2003). *In silico* analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA-regulon. *Bioinformatics*, **19**, 2225–2236.
40. Gelfand, M. S., Novichkov, P. S., Novichkova, E. S. & Mironov, A. A. (2000). Comparative analysis of regulatory patterns in bacterial genomes. *Brief. Bioinfor.* **1**, 357–371.
41. Williams, R. M., Rhodius, V. A., Bell, A. I., Kolb, A. & Busby, S. J. W. (1996). Orientation of functional activating regions in the *Escherichia coli* CRP protein during transcription activation at class II promoters. *Nucl. Acids Res.* **24**, 1112–1118.
42. Rhodius, V. A., West, D. M., Webster, C. L., Busby, S. J. W. & Savery, N. J. (1997). Transcription activation at class II CRP-dependent promoters: the role of different activating regions. *Nucl. Acids Res.* **25**, 326–332.
43. Williams, S. M., Savery, N. J., Busby, S. J. W. & Wing, H. J. (1997). Transcription activation at Class I FNR-dependent promoters: identification of the activating

surface of FNR and the corresponding contact site in the C-terminal domain of the RNA polymerase σ subunit. *Nucl. Acids Res.* **25**, 4028–4034.
44. Wing, H. J., Williams, S. M. & Busby, S. J. W. (1995). Spacing requirements for transcription activation by *Escherichia coli* FNR Protein. *J. Bacteriol.* **177**, 6704–6710.
45. Barnard, A. M. L., Green, J. & Busby, S. J. (2003). Transcription regulation by tandem-bound FNR at *Escherichia coli* promoters. *J. Bacteriol.* **185**, 5993–6004.
46. Lobell, R. B. & Schleif, R. F. (1990). Looping and unlooping by AraC protein. *Science*, **250**, 528–532.
47. Spiro, S. & Guest, J. R. (1990). FNR and its role in oxygen-regulated gene expression in *Escherichia coli*. *FEMS Microbiol. Rev.* **75**, 399–428.
48. Spiro, S., Gaston, K. L., Bell, A. I., Roberts, R. E., Busby, S. J. & Guest, J. R. (1990). Interconversion of the DNA-binding specificities of two related transcription regulators, CRP and FNR. *Mol. Microbiol.* **4**, 1831–1838.
49. Lavrrar, J. L. & McIntosh, M. A. (2003). Architecture of a Fur binding site: a comparative analysis. *J. Bacteriol.* **185**, 2194–2202.
50. Thompson, D. K., Beliaev, A. S., Giometti, C. S., Tollaksen, S. L., Khare, T., Lies, D. P. *et al*. (2002). Transcriptional and proteomic analysis of a ferric uptake regulator (Fur) mutant of *Shewanella oneidensis*: possible involvement of Fur in energy metabolism, transcriptional regulation, and oxidative stress. *Appl. Environ. Microbiol.* **68**, 881–892.
51. Hall, H. K. & Foster, J. W. (1996). The role of Fur in the acid tolerance response of *Salmonella typhimurium* is physiologically and genetically separable from its role in iron acquisition. *J. Bacteriol.* **178**, 5683–5691.
52. Makarova, K. S., Mironov, A. A. & Gelfand, M. S. (2001). Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol.* **2**, research0013.1–0013.8.
53. Panina, E. M., Mironov, A. A. & Gelfand, M. S. (2001). Comparative analysis of FUR regulon in gamma-protaobacteria. *Nucl. Acids. Res.* **29**, 5195–5206.
54. Erill, I., Jara, M., Salvador, N., Escribano, M., Campoy, S. & Barbe, J. (2004). Differences in LexA regulon structure among proteobacteria through *in vivo* assisted comparative genomics. *Nucl. Acids Res.* **32**, 6617–6626.
55. Magasanik, B. (1988). Reversible phosphorylation of an enhancer binding protein regulates the transcription of bacterial nitrogen utilization genes. *Trends Biochem. Sci.* **13**, 475–479.
56. Huerta, A. M. & Collado-Vides, J. (2003). Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* **333**, 261–278.
57. Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
58. Harley, C. B. & Reynolds, R. P. (1987). Analysis of *E. coli* promoter sequences. *Nucl. Acids Res.* **15**, 2343–2361.

Edited by R. Ebright

(Received 7 June 2005; received in revised form 12 September 2005; accepted 13 September 2005)
Available online 4 October 2005