In Silico Biology

Vol. 5 no. 2 2005
Pages 1–6
(not for citation purposes)

Complementing computationally predicted regulatory sites in Tractor_DB using a pattern matching approach

Marylens Hernández Guía [1], Abel González Pérez [1], Vladimir Espinosa Angarica [1], Ana T. Vasconcelos [2] and Julio Collado-Vides [3*]

[1]National Bioinformatics Center, Industria y San José, Capitolio Nacional, CP. 10200, Habana Vieja, Habana, Cuba
[2]National Laboratory for Scientific Computing, Av. Getulio Vargas 333, Quitandinha, CEP 25651-075, Petropolis, Rio de Janeiro, Brazil
[3]Center of Genomics, UNAM, Cuernavaca, Mexico. AP 565-A Cuernavaca, CP 62100, Morelos, Mexico

Received on December 07, 2004; revised and accepted on December 23, 2004; published January 04, 2005

Edited by Edgar Wingender

ABSTRACT

Prokaryotic genomes annotation has focused on genes location and function. The lack of regulatory information has limited the knowledge on cellular transcriptional regulatory networks. However, as more phylogenetically close genomes are sequenced and annotated, the implementation of phylogenetic footprinting strategies for the recognition of regulators and their regulons becomes more important. In this paper we describe a comparative genomics approach to the prediction of new gamma-proteobacterial regulon members. We take advantage of the phylogenetic proximity of *Escherichia coli* and other 16 organisms of this subdivision and the intensive search of the space sequence provided by a pattern-matching strategy. Using this approach we complement predictions of regulatory sites made using statistical models currently stored in Tractor_DB, and increase the number of transcriptional regulators with predicted binding sites up to 86. All these computational predictions may be reached at Tractor_DB[†]. We also take a first step in this paper towards the assessment of the conservation of the architecture of the regulatory network in the gamma-proteobacteria through evaluating the conservation of the overall connectivity of the network.

Keywords: transcriptional regulatory networks, comparative genomics, gamma-proteobacterial regulons
Contact: collado@ccg.unam.mx
[†]**Availability:** www.ccg.unam.mx/Computational_Genomics/tractorDB; www.tractor.lncc.br; www.bioinfo.cu/Tractor_DB

INTRODUCTION

One of the major challenges in the post-genomic era is the identification of all the elements taking part in an organism's transcriptional regulatory network, in the quest to understand how the cell reacts to environmental stimuli at the level of transcription regulation. Intense research is being carried out in this direction (Buchler *et al.*, 2003; Cases *et al.*, 2003; Conant and Wagner, 2003; Salmon *et al.*, 2003). A first step towards this goal is the computational prediction of all the genes regulated by a Transcription Factor (TF) -*i.e.* its regulon. Although a significant amount of research has been dedicated to this issue in the past few years, it is far from resolved.

Recently, our group and others, have generated computational prediction of important regulatory elements in the *E. coli* genome: promoters (Huerta and Collado-Vides, 2003), operons (Salgado *et al.*, 2000), TFs (Perez-Rueda and Collado-Vides, 2000), TF binding sites (Thieffry *et al.*, 1998b). Since more bacterial genomes are being sequenced, it has become important to extend these works to other organisms, in the effort to deciphering their transcriptional regulatory networks and establishing comparative regulatory studies (Aguilar *et al.*, 2002; McCue *et al.*, 2001; Mirny and Gelfand, 2002; Moreno-Hagelsieb and Collado-Vides, 2002; Munch *et al.*, 2003; Perez-Rueda and Collado-Vides, 2001; Rajewsky *et al.*, 2002; Tan *et al.*, 2001).

Computational strategies to predicting new regulon members have thus increasingly included phylogenetic information in order both to extend predictions to more organisms, and as a way to discriminate false positive predicted sites. These strategies include the construction of a statistical model for the DNA binding site of each known *E. coli* TF which is used to predict putative binding sites in *E. coli* and other closely related bacteria (Tan *et al.*, 2001), and the construction of the models from the regions upstream orthologous genes in phylogenetically close organisms (phylogenetic footprinting, (McCue *et al.*, 2001; Thompson *et al.*, 2003)).

In this work we use a comparative genomics approach based on regular expressions to find putative orthologous sites of *E. coli* regulons in other 16 gamma-proteobcteria, as a complement of the statistical models approach. Each TF known binding site in *E. coli* is treated as a regular expression, and its occurrence searched in orthologous non-coding regions (see Methods) allowing up to eight mismatches. The statistical significance of each predicted

*to whom correspondence should be addressed

orthologous site is then assessed using TFs' weight matrices. Using these two complementary approaches, we have predicted new members for 86 gamma-proteobacterial regulons. All these computational predictions have been included in the Tractor_DB database[†] (Gonzalez *et al.*, 2005).

METHODS

Selecting organisms and TFs

We worked with the organisms that are already included in Tractor_DB (with sites predicted using statistical models), which in time were selected to be a representative set of the gamma-proteobacteria whose genomes are sequenced. This set comprises: *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), *Yersinia pestis* KIM (NC_004088), *Buchnera aphidicola* (NC_004545), *Pseudomonas aeruginosa* (NC_002516), *Pseudomonas syringae* (NC_004578), *Pasteurella multocida* (NC_002663), *Pseudomonas putida* KT2440 (NC_002947), *Vibrio parahaemolyticus* (NC_004603), *Vibrio vulnificus* CMCP6 (NC_004459), *Xylella fastidiosa* (NC_002488), and *Xanthomonas axonopodis* (NC_003919). We selected all TFs with at least one known binding site annotated in RegulonDB v. 4.0: 105 (Salgado *et al.*, 2004). The lengths and sequences of all sites were obtained from this database.

Orthology searching

Orthology relationships were searched using the definition by (Huynen and Bork, 1998): two genes from different organisms are orthologs if they are the best bi-directional blast hits (BBHs). We used computationally predicted transcription units (TUs) by (Moreno-Hagelsieb and Collado-Vides, 2002), to find TUs upstream regions. Combining these two sources of information, we designated two TUs from different organisms as orthologous if they contain at least one pair of orthologous genes. Using the same reasoning, two non-coding regions from two different organisms are considered orthologous (or "r-orthologous" to denote regulatory orthology), if they occur upstream of two orthologous TUs. Following this idea we organized orthologous TUs into clusters centered in *E. coli*: each *E. coli* transcription unit was selected and all orthologous TUs in the other 16 organisms were recruited to form a cluster. Clearly, one organism may contribute to the cluster with more than one TU; the idea is that all the fragments of a broken TU may retain the same regulation of the *E. coli* TU.

Pattern matching

We converted each *E. coli* known binding site into a regular expression following these rules: if the site length was even and shorter than 14 nucleotides, the site sequence was expanded at both ends to reach 14 nucleotides (with the center between the two central base pairs); if the site was odd and shorter than 13 nucleotides, the site sequence was expanded at both ends until 13 nucleotides (with the center at the central base pair); on the other hand, the site length reported at RegulonDB was respected if the site was longer than 13 (odd) or 14 (even) nucleotides. The reason to take 13 or 14 nucleotides as the minimum length for odd and even sites respectively is that the length of most prokaryotic TF binding sites rages from 14 to 26 nucleotides. After this step, site lengths ranged from 13 nucleotides (DnaA, IHF, IciA, XapR, and CspA) to 61 nucleotides (ArcA).

To perform the pattern matching, we extracted the r-orthologous regions (from -400 to +50 with respect to the first base of the TU, the region where most TF binding sites are known to occur). We searched the occurrences of the *E. coli* regular expressions of each TF in all orthologous regulatory regions and allowed in the process of pattern matching the lowest number of mismatches that recovered a number of matches of the same order that the number of *E. coli* known sites of the TF, a cutoff selected to reduce the number of spurious matches. No gaps were allowed in the pattern matching, taking into account the high sequence identity between orthologous regulators in these organisms (Brown and Callan, 2004; Tan *et al.*, 2001). We allowed no more than 8 mismatches for any TF. (See Table 1 for details.)

Table 1. Results of the pattern matching step

TF	Mm	Mt	Len	Reg	TF	Mm	Mt	Len	Reg
Ada	7	6	28	3	MarR	7	9	21	2
AraC	6	72	17	15	MelR	1	6	18	6
ArcA	7	21	61	21	MetJ	5	68	8	5
ArgR	5	85	16	12	MetR	5	9	24	3
BirA	2	5	40	2	MhpR	8	8	17	1
CRP	7	257	19	128	Mlc	7	22	26	6
CspA	0	3	5	3	ModE	8	16	27	6
CysB	8	6	42	7	Nac	6	99	15	4
CytR	8	10	40	7	NagC	8	21	26	8
DnaA	4	94	9	9	NarL	6	98	19	38
FadR	7	83	17	6	NarP	8	29	19	3
FarR	8	2	21	2	NtrC	4	64	15	11
FhlA	7	5	40	3	OmpR	4	76	10	16
FIS	1	78	16	56	OxyR	5	7	45	4
FlhD	6	3	40	2	PdhR	0	6	21	2
FNR	8	190	22	45	PhoB	6	63	17	8
FruR	5	84	14	8	PhoP	8	65	18	3
Fur	0	9	19	7	PurR	5	93	16	16
GadX	8	7	22	6	RcsB	8	3	25	3
GalR	5	8	17	4	RhaR	6	7	20	3
GalS	5	9	16	2	RhaS	3	7	17	3
GcvA	6	15	29	4	SoxR	8	7	19	2
GlcC	7	6	22	2	SoxS	7	47	18	8
GlpR	7	67	17	19	TdcA	5	6	15	1
GntR	2	7	16	2	TdcR	5	3	12	1
IciA	4	7	29	1	TorR	1	8	10	6
IclR	7	3	34	1	TreR	4	7	15	2
IHF	4	212	13	49	TrpR	8	9	27	3
KdpE	4	8	12	1	TyrR	8	88	22	19
LexA	8	87	20	11	UhpA	1	3	24	1
Lrp	2	78	40	41	XapR	4	7	13	2
MalI	4	5	12	4	XylR	4	9	16	4
MalT	4	59	12	9	YiaJ	2	3	35	1
MarA	8	30	10	7					

Mm: Maximum number of mismatches allowed for each regulon in the pattern matching process; **Mt**: Number of matches found with the regular expressions of each regulon; **Len**: Length of the regular expressions used to search matches for each regulon; **Reg**: Number of sequences in the *E. coli* regulon and number of regular expressions used to search matches in each regulon

Assessing the statistical significance of predicted sites

Using the CONSENSUS program (Hertz and Stormo, 1999) we built weight matrices for each TF aligning the sequences of original *E. coli* sites and their putative orthologous sites identified by pattern matching. These sequences were extended five nucleotides at each side before introducing them into the program. The idea is that pattern matching may locate a true orthologous site displaced several bases at the left or right: the CONSENSUS algorithm is able to correctly locate the sites by maximizing the *log*-likelihood of the alignment. The program was run including the reverse-complement sequences into the matrices in the case of dimeric binding TFs.

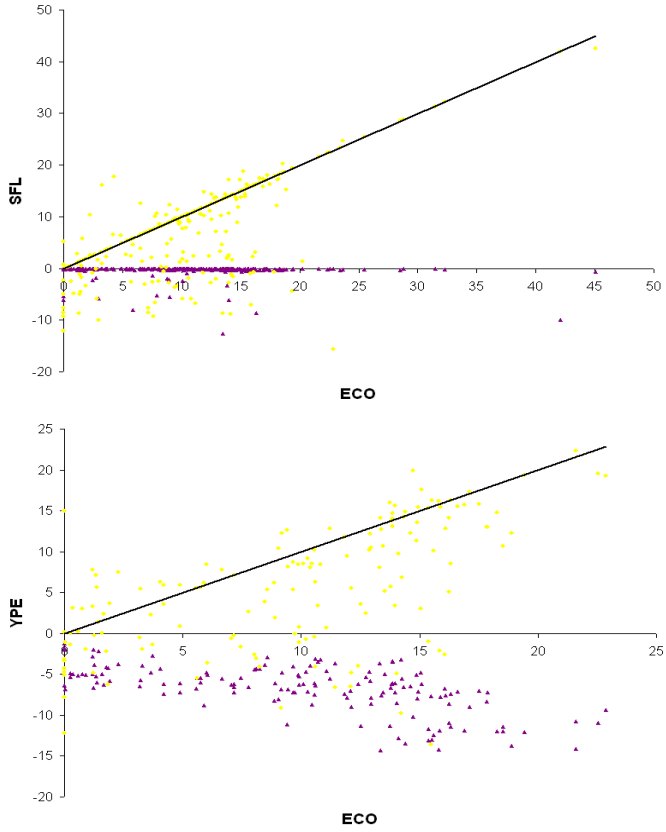

Fig. 1. Real (yellow) and expected (purple) score of orthologous sites predicted in *S. flexneri* (SFL; top) and *Y. pestis* (YPE; bottom)

To asses the statistical significance of putative orthologous sites we used the idea developed by (Brown and Callan, 2004), for CRP. Briefly, each predicted orthologous sequence is aligned against the weight matrix that represents the TF binding site and the score of this alignment is termed "real". On the other hand an "expected" score of the predicted sequence is calculated as a result of its alignment against the matrix provided that it mutated following a rate similar to the rest of the non-coding region where it is located. (Mutation rates are calculated from the alignment of the r-orthologous regions.) The idea here is that in an interspecies comparison, true orthologous regulatory sites should be more conserved than the rest of the non-coding regions. (For a thorough description of the statistics carried out in this calculations see (Brown and Callan, 2004).) To calculate the real score of putative orthologous sites we aligned them to the corresponding weight matrix using the PATSER program (Hertz and Stormo, 1999), and to estimate the score of the site had it mutated with the same rate that the surrounding non-coding region (expected score) we used the package of programs provided by Brown and Callan in their already cited work, located at www.princeton.edu/∼ccallan/binding.

Web logos of the TFs' binding sites were built using the WebLogo server at weblogo.berkeley.edu/logo.cgi (Schneider and Stephens, 1990).

RESULTS

Orthologous sites predicted by pattern matching

For 66 TFs we succeeded predicting orthologous sites using regular expressions obtained from *E. coli* known binding sites. For the other 39 TFs in the original set, we either failed finding an orthologous

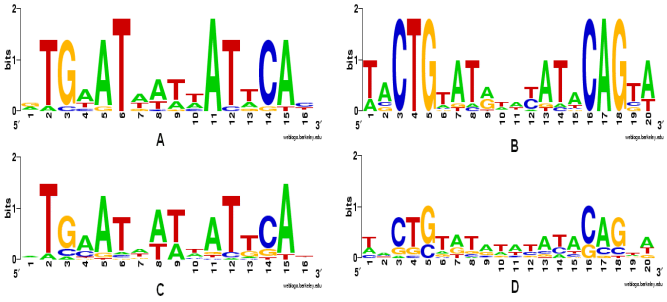

Fig. 2. Site logos for ArgR (A and C) and LexA (B and D) in *S. flexneri* (A and B) and *E. coli* (C and D)

TF in every other organism or the putative orthologous sites did not allow us to build a weight matrix with a p-value lower than 10^{-6} (a matrix whose probability of having emerged by chance is one in one million). For the set of 66 TFs with predictions, the p-values of the weight matrices range from 6.5×10^{-297} for PurR and LexA to 10^{-6} for FarR, and the number of putative orthologous sites ranges from 2 (FarR) to 257 (CRP). Table 1 summarizes the results of the pattern matching step.

Figure 1 shows the populations of real and expected scores of putative sites predicted in *Shigella flexneri* and *Yersinia pestis* by pattern matching using *E. coli* sites as regular expressions. (The analysis with other organisms follows the same trend.) The x coordinate of each dot in the graph corresponds to the score of an *E. coli* regulatory site, and its y coordinate corresponds to the real score of the orthologous site in *S. flexneri* or *Y. pestis* (yellow), or to its expected score (purple). The graphic shows that both populations tend to separate as the scores of the *E. coli* original sites increase. They separate completely for sites with score five or higher, which was taken as a cutoff to select good candidates to true regulatory sites. Brown and Callan obtained the same behavior for CRP binding sites comparing *E. coli* sites and their orthologs in *S. typhimurium*.

Comparing sequence logos is one way of visualizing the conservation of a TF motif in two organisms: Figure 2 illustrates this conservation through the logos of the LexA and ArgR motifs in *E. coli* and *S. flexneri*. *E. coli* logos were built from ArgR and LexA known binding sequences; for *S. flexneri* we used the sequence of putative orthologous sites that passed the cutoff. Both LexA logos exhibit the typical motif of this TF, although the relative importance of each position differs in both organisms. This is mainly due to the stringency applied in both the pattern matching and the assessment of the statistical significance of predicted sites, which determines that we rescue only those putative sites that resemble the most the *E. coli* LexA model. The same applies to the ArgR models.

All regulatory sites predicted by this methodology have been incorporated to Tractor_DB†. Table 2 summarizes the number of regulons and predicted sites per organism currently stored in Tractor_DB, grouped by prediction method: statistical models (Gonzalez *et al.*, 2005) or regular expressions.

Comparing network connectivity

One of the goals of Tractor_DB is to serve as a starting point in theoretical studies of the evolution of the gamma-proteobacterial

Table 2. Computationally predicted sites and TFs in the seventeen organisms grouped in Tractor_DB

Organism	Reg Exp		Models		Total	
	S	Reg	S	Reg	S	Reg
Buchnera aphidicola	2	1	7	3	9	4
Escherichia coli	138	45	842	72	873	86
Haemophilus influenzae	17	9	144	5	149	10
Pseudomonas aeruginosa	11	9	22	11	29	16
Pseudomonas putida	12	9	17	9	28	16
Pseudomonas syringae	12	6	16	9	27	14
Salmonella typhi	88	36	710	44	735	58
Salmonella typhimurium	96	37	762	45	782	60
Shewanella oneidensis	16	8	287	9	294	12
Shigella flexneri	104	40	667	56	693	69
Vibrio cholerae	14	11	242	11	246	18
Vibrio parahaemolyticus	0	0	141	28	141	28
Vibrio vulnificus	0	0	112	23	112	23
Xanthomonas axonopodis	11	9	3	2	14	11
Xanthomonas campestris	9	7	5	4	14	10
Xylella fastidiosa	11	12	7	4	12	6
Yersinia pestis	43	18	364	16	373	25

Reg Exp: sites predicted using pattern matching; **Models**: sites predicted using statistical models; **S**: number of predicted binding sites; **Reg**: number of genes in the regulon

transcriptional regulatory network, since it is the largest collection of computationally predicted regulatory interactions in gamma-proteobacterial genomes. One important architectural feature of cellular networks is connectivity (Thieffry *et al.*, 1998a). We compared the connectivity of the regulatory network of five of the organisms included in Tractor_DB (*E. coli*, *S. flexneri*, *S. typhi*, *S. typhimurium* and *Y. pestis*). Figure 3 presents the results of this comparison.

The *E. coli* regulatory network connectivity (incoming connections) that results from integrating all the data in Tractor_DB (a set of 873 TUs) is very similar to that obtained by (Thieffry *et al.*, 1998a), using a study set of approximately 200 TUs. The comparison with the networks of the four closest organisms with regulatory sites in Tractor_DB (excluding *H. influenzae*, given the smaller size of its genome) revealed a significant conservation of the overall connectivity of the regulatory network (in outgoing as in incoming regulatory interactions). The fact that the regulatory predicted sites stored in Tractor_DB for the organisms used in this comparison were ultimately produced using experimental information from *E. coli* (either as statistical models or as regular expressions) does not influence this result. In the process of prediction of new regulon members, separate statistical models were rebuilt for each organism to tune the search for new regulatory sites and the orthology filtering was iterated using each organism at the center of the analysis each time, see (Gonzalez *et al.*, 2005). This resulted in the prediction of regulatory sites in these organisms with no ortholog in *E. coli*. For example, 48% of the TUs that are members of the CRP regulon in *S. typhimurium* do not have orthologs in *E. coli*.

DISCUSSION

Tractor_DB stores information on computationally predicted gamma-proteobacterial regulon members (Gonzalez *et al.*, 2005) in 17 organisms. Its main goal is to aid experimentalist researchers in the interpretation of microarray data, and any other situation in which a piece of information regarding gene expression regulation may be of help. But it also intends to serve as a first step in theoretical studies of the evolution of the gamma-proteobacterial transcriptional regulatory network. In the beginning, it only stored sites predicted using statistical models built from *E. coli* sites, see (Gonzalez *et al.*, 2005), Supplementary Material. This approach permitted predicting sites for a large number of TFs (72) in 17 gamma-proteobacterial genomes (including *E. coli*). However, for some TF/organism combinations, we failed predicting regulon members. These combinations corresponded mainly to small regulons, whose binding motifs are expected to vary more from organism to organism than those of large regulons (Rajewsky *et al.*, 2002). Thus, the aim of the present work was to complement the former approach using another strategy that allowed discovering sites mainly for those small regulons.

Using pattern matching, we succeeded predicting regulatory sites for 133 TF/organism combinations for which statistical models found no sites. These correspond mainly to small regulons (less than 10 sites in *E. coli*). This strategy thus is more suitable for predicting putative regulatory sites when the TF binding motif is less well conserved; pattern matching allows exploring the sequence space more intensively than statistical models in these cases. (A statistical model such as a weight matrix, built from *E. coli* sequences will fail recognizing orthologous sites unless they are very similar to the original sites, because the information content of the matrix will be limited by the number of *E. coli* sites.) A pattern matching strategy, on the other hand, "relaxes" the search criterion by permitting mismatches, and hence orthologous sites with more divergent sequences may be detected. For this same reason, pattern matching may help overcoming difficulties associated with the extrapolation of statistical models built using *E. coli* TFs binding sites to other organisms with background base compositions (for instance, different GC content in their non-coding regions). For large regulons the case is the opposite: since the strategy we implemented here is limited to recognizing orthologous sites, the upper limit is the number of *E. coli* known sites. A statistical models approach explores the sequence space more extensively, because the significance of sites orthologous to *E. coli* predicted sites is assessed as well (Gonzalez *et al.*, 2005).

A central problem to all computational strategies for TF binding sites prediction is the determination of cutoff values to discriminate significant matches from spurious results. In the approaches that use statistical models to predict new TF binding sites this problem relates to the selection of a score cutoff for the alignment of predicted sequences to the model. In the pattern matching approach that we describe here, we addressed this problem in various ways at different steps of the methodology. In order to limit the number of spurious putative sites in the pattern matching step, we allowed only the maximum number of mismatches that permitted rescuing a number of matches of the same order that the number of known sequences within the *E. coli* regulon. The point here is that since we are looking only at putative orthologous sites, the number of *E. coli* known sites in a regulon sets a maximum limit to the number of sites that can be found. However, the number of mismatches was never allowed to be greater than eight, given the length of the regular expressions used to search for putative matches (between 13 and 61 nucleotides). Since we did not use a position-dependent model

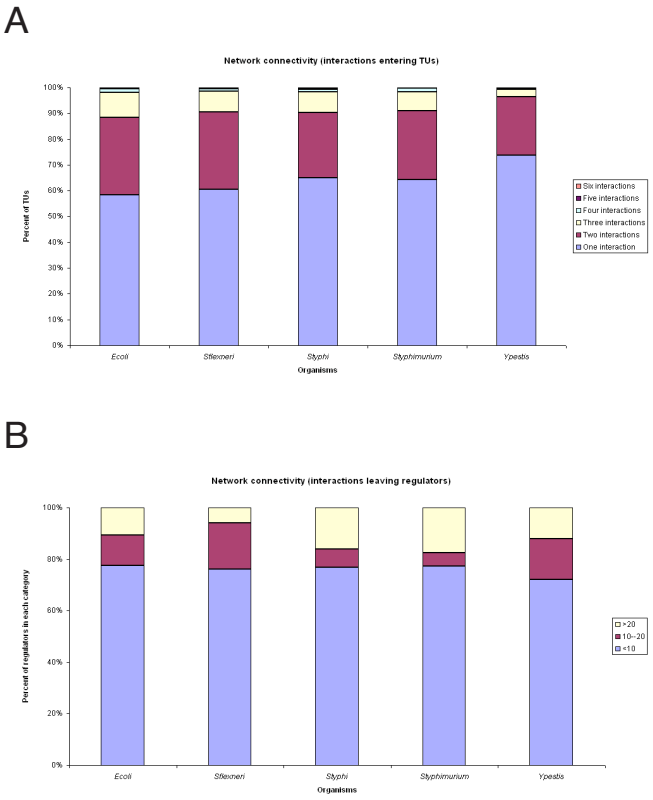

Fig. 3. Comparison of regulatory network connectivity in five gamma-proteobacteria. A: Fraction of TUs with different number of incoming interactions (less than 0.5% with five and six interactions in each organism); B: Fraction of TFs with different number of outgoing interactions

of mismatches at the pattern matching process, weight matrices (which are position-dependent models) were used in order to filter out possible spurious sites.

The inclusion of these computationally predicted regulon members hence complements our previous work using statistical models and contributes to the completeness of Tractor_DB.

One significant difference between our work and that of Brown and Callan (beside the number of TFs and organisms involved in the study) is that they identified orthologous sites aligning non-coding regions with ClustalW. Although this method may suit very well the comparison between *E. coli* and *S. typhimurium*, it may not work so well for more distant organisms. Pattern matching helps bridging this gap, although for some of the most distant organisms we had difficulties too identifying orthologous sites (*i.e.*, one regulon for *B. aphidicola* and three regulons for *X. fastidiosa*).

In this paper we also take a first step in assessing the conservation of the regulatory network architecture through the gamma-proteobacteria class. The finding that the overall connectivity of the regulatory network is conserved (at least in the enterobacteria included in Tractor_DB), taken together with other results obtained by our group using this same data set (manuscript in preparation) regarding the conservation of regulon dispersion, *i.e.* the signal-to-noise ratio that a TF discriminates when recognizing its true binding sites from the background, the conservation of TU structure in connection with the conservation of its regulatory site, and the probable conservation of higher order regulatory complexes (co-occurring pairs of TFs and the distances between their binding sites, as well as the distances between TFs binding sites and promoters) suggest the possibility of the existence of a highly conserved architecture in the regulatory network in the gamma-proteobacteria class.

ACKNOWLEDGEMENTS

We thank Fernanda Mendonça (LNCC), Roger Paixão (LNCC), Heladia Salgado (CCG), and César Bonavides (CCG) for maintenance of Tractor_DB.

REFERENCES

Aguilar, D., Oliva, B., Aviles, F. X., and Querol, E. (2002). Transcout: prediction of gene expression regulatory proteins from their sequences. *Bioinformatics*, **18**(4), 597–607.

Brown, C. T. and Callan, C. G. J. (2004). Evolutionary comparisons suggest many novel camp response protein binding sites in escherichia coli. *Proc Natl Acad Sci U S A*, **101**(8), 2404–2409.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, **100**(9), 5136–5141.

Cases, I., de Lorenzo, V., and Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol*, **11**(6), 248–253.

Conant, G. C. and Wagner, A. (2003). Convergent evolution of gene circuits. *Nat Genet*, **34**(3), 264–266.

Gonzalez, A. D., Espinosa, V., Vasconcelos, A. T., Perez-Rueda, E., and Collado-Vides, J. (2005). Tractor_db: a database of regulatory networks in gamma-proteobacterial genomes. *Nucleic Acids Res*, **33**(Database issue), D98–102.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**(7-8), 563–577.

Huerta, A. M. and Collado-Vides, J. (2003). Sigma70 promoters in escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. *J Mol Biol*, **333**(2), 261–278.

Huynen, M. A. and Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A*, **95**(11), 5849–5856.

McCue, L., Thompson, W., Carmack, C., Ryan, M. P., Liu, J. S., Derbyshire, V., and Lawrence, C. E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, **29**(3), 774–782.

Mirny, L. A. and Gelfand, M. S. (2002). Structural analysis of conserved base pairs in protein-dna complexes. *Nucleic Acids Res*, **30**(7), 1704–1711.

Moreno-Hagelsieb, G. and Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18 Suppl 1**, S329–36.

Munch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003). Prodoric: prokaryotic database of gene regulation. *Nucleic Acids Res*, **31**(1), 266–269.

Perez-Rueda, E. and Collado-Vides, J. (2000). The repertoire of dna-binding transcriptional regulators in escherichia coli k-12. *Nucleic Acids Res*, **28**(8), 1838–1847.

Perez-Rueda, E. and Collado-Vides, J. (2001). Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol*, **53**(3), 172–179.

Rajewsky, N., Socci, N. D., Zapotocky, M., and Siggia, E. D. (2002). The evolution of dna regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*, **12**(2), 298–308.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. (2000). Operons in escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A*, **97**(12), 6652–6657.

Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., and Collado-Vides, J. (2004). Regulondb

(version 4.0): transcriptional regulation, operon organization and growth conditions in escherichia coli k-12. *Nucleic Acids Res*, **32**(Database issue), D303–6.

Salmon, K., Hung, S.-p., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003). Global gene expression profiling in escherichia coli k12. the effects of oxygen availability and fnr. *J Biol Chem*, **278**(32), 29837–29855.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**(20), 6097–6100.

Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J., and Stormo, G. D. (2001). A comparative genomics approach to prediction of new members of regulons. *Genome Res*, **11**(4), 566–584.

Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998a). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, **20**(5), 433–440.

Thieffry, D., Salgado, H., Huerta, A. M., and Collado-Vides, J. (1998b). Prediction of transcriptional regulatory sites in the complete genome sequence of escherichia coli k-12. *Bioinformatics*, **14**(5), 391–400.

Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003). Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res*, **31**(13), 3580–3585.