D132–D136 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 6 November 2006*
doi:10.1093/nar/gkl800

Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes

Abel González Pérez, Vladimir Espinosa Angarica, Ana Tereza R. Vasconcelos[1],* and Julio Collado-Vides[2]

National Bioinformatics Center, Cuba, [1]National Laboratory for Scientific Computing, Brazil and [2]Center of Genomics, Mexico

Received August 16, 2006; Revised September 20, 2006; Accepted October 2, 2006

ABSTRACT

The version 2.0 of Tractor_DB is now accessible at its three international mirrors: www.bioinfo.cu/Tractor_DB, www.tractor.lncc.br and http://www.ccg.unam.mx/tractorDB. This database contains a collection of computationally predicted Transcription Factors' binding sites in gamma-proteobacterial genomes. These data should aid researchers in the design of microarray experiments and the interpretation of their results. They should also facilitate studies of Comparative Genomics of the regulatory networks of this group of organisms. In this paper we describe the main improvements incorporated to the database in the past year and a half which include incorporating information on the regulatory networks of 13—increasing to 30—new gamma-proteobacteria and developing a new computational strategy to complement the putative sites identified by the original weight matrix-based approach. We have also added dynamically generated navigation tabs to the navigation interfaces. Moreover, we developed a new interface that allows users to directly retrieve information on the conservation of regulatory interactions in the 30 genomes included in the database by navigating a map that represents a core of the known *Escherichia coli* regulatory network.

INTRODUCTION

The initiation of transcription in prokaryotic organisms is the most important stage in the regulation of gene expression in response to stimuli. The elucidation of the interactions that connect transcription factors (TFs) and their target genes is central to understand this regulatory mechanism. Several works in the past years have aimed at such elucidation, developing a variety of computational approaches to identify putative TFs' binding sites in organisms with completely sequenced genomes (1–6). The gamma-proteobacteria subclass has been widely employed in these works because the genomes of many (>30) of its members have been sequenced and it includes the organism with the best known regulatory network, *Escherichia coli*. In addition, many organisms of this subclass are pathogens of humans, animals or plants.

Two years ago, we developed a database (Tractor_DB) that contains information of computationally predicted regulatory interactions within the genomes of several organisms of this group. We presented its first version in the 2005 database issue (7). Tractor_DB is a relational database that uses the MySQL server with a web interface composed of several Perl scripts. The relational design of the database (i.e. the tables and the relations between them) has not changed with respect to the previous version (7).

In this paper, we describe the main modifications and improvements experienced by the database since. They have focused on the expansion of the biological information stored in the database and the improvement of the query and navigation interfaces.

CHANGES IN VERSION 2.0

Obtaining and preparing basic data

Genomic sequences of the gamma-proteobacteria included in Tractor_DB version 2.0 (see Table 1 for a list of organisms' names and genome sequences' accession numbers) were obtained from the GenBank database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria). Orthology relationships between gene pairs were determined using the BBH methodology (8). Transcription units (TUs) prediction (9) was then used to define orthologous TUs (those sharing at least a pair of orthologous genes). Regulatory regions (the targets for binding sites search) were defined as sequences stretching from −400 to +50 with respect to the first translated nucleotide of each TU, and orthologous regulatory regions as those upstream orthologous TUs. These orthology relationships were used in the prediction pipeline (see below). The sequences of TFs binding sites that have been identified

*To whom correspondence should be addressed at Bioinformatics Laboratory-LABINFO National Laboratory of Scientfic Computation Av. Getulio Vargas, 333, Quitandinha ZC: 25651-075 Petrópolis Rio de Janeiro, Brazil. Tel: +55 24 2233 6065; Fax: +55 24 2231 5595; Email: atrv@lncc.br

Table 1. Overview of the evolution of the number of regulatory interactions in Tractor_DB from version 1.0 to version 2.0, organized by organisms

Organism	Accession no.	Version 1.0		Version 2.0	
		TFs	TUs	TFs	TUs
Acinetobacter sp ADP1	NC_005966	—	—	9	41
Buchnera aphidicola Bp	NC_004545	3	7	8	22
Erwinia carotovora SCRI1043	NC_004547	—	—	7	18
Escherichia coli K12	NC_000913	74	853	87	938
Escherichia coli O157H7	NC_002655	—	—	46	411
Haemophilus ducreyi 35000HP	NC_002940	—	—	12	60
Haemophilus influenzae Rd KW20	NC_000907	5	140	19	193
Legionella pneumophila Lens	NC_006369	—	—	6	9
Legionella pneumophila Paris	NC_006368	—	—	8	11
Legionella pneumophila Philadelphia	NC_002942	—	—	8	8
Methylococcus capsulatus Bath	NC_002977	—	—	3	3
Photobacterium profundum	NC_006370, NC_006371	—	—	21	104
Photorhabdus luminescens TT01	NC_005126	—	—	23	136
Pseudomonas aeruginosa PA01	NC_002516	11	22	11	24
Pseudomonas putida KT2440	NC_002947	9	17	9	17
Pseudomonas syringae DC3000	NC_004578	9	16	11	18
Salmonella typhi CT18	NC_003198	23	725	68	812
Salmonella typhimurium LT2	NC_003197	26	752	69	832
Shewanella oneidensis MR-1	NC_004347	6	285	20	345
Shigella flexneri 2a 301	NC_004337	32	658	62	718
Shigella flexneri 2a 2457T	NC_004741	—	—	47	367
Vibrio cholerae N16961	NC_002505, NC_002506	7	234	22	267
Vibrio parahaemolyticus RIMD 2210633	NC_004603, NC_004605	28	141	29	182
Vibrio vulnificus CMCP6	NC_004459, NC_004460	23	112	26	157
Xanthomonas axonopodis 306	NC_003919	2	3	3	4
Xanthomonas campestris ATC 33913	NC_003902	4	5	5	6
Xylella fastidiosa 9a5c	NC_002488	4	7	6	9
Yersinia pestis Mediaevails 91001	NC_005810	—	—	25	183
Yersinia pestis KIM	NC_004088	11	354	28	405
Yersinia pseudotuberculosis IP32953	NC_006155	—	—	26	185

Number of TFs with binding sites (TFs) and number of transcription units with regulatory inputs (TUs) included in both versions.

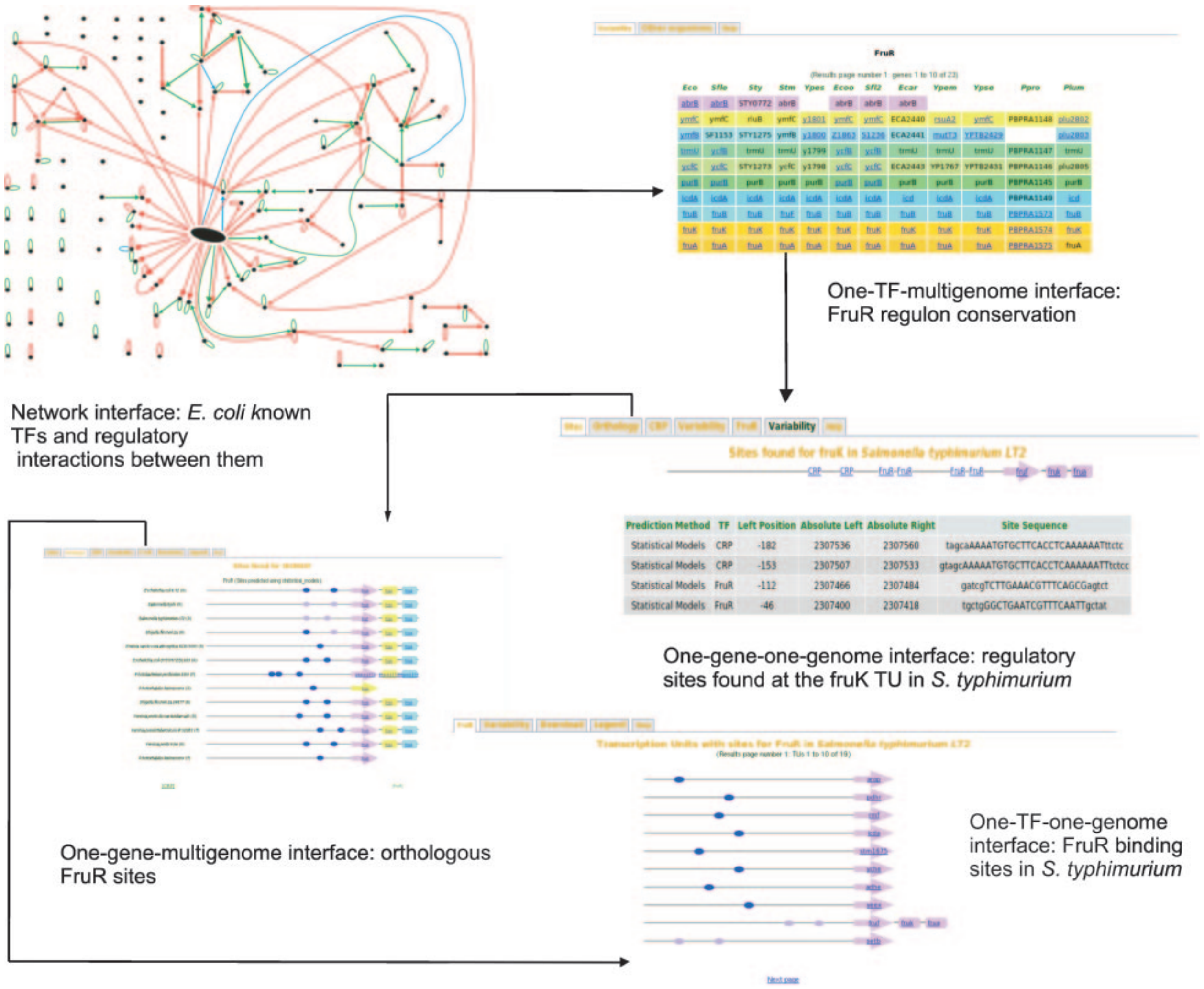
experimentally in *E.coli* were obtained from RegulonDB version 5.0 (10).

Expansion of the biological information in the database

Two main steps were taken aimed at the expansion of the information included in the database. First, thirteen new organisms were added to the pipeline of the weight matrix-based approach, used to predict regulatory interactions in the first version. The number of organisms of the gamma-proteobacteria subclass with information on regulatory interactions in the database was thus expanded to 30. Figure 1 of the Supplementary Data presents a flowchart of this approach (7).

Briefly, this strategy starts by building positional weight matrices from training sets constituted by the binding sites of each TF that are known in *E.coli* and orthologous regulatory regions in other seven organisms (those phylogenetically closer to *E.coli*, excluding *E.coli* O157H7 and *Shigella flexneri* 2a 2457T). Then, these training sets are filtered to eliminate possible weak binding sequences and two cutoff values for each TF are calculated. The regulatory regions of all genomes are then scanned for putative binding sites using each TF's matrix. The sites thus obtained are filtered using orthology information (an *E.coli* site without at least one ortholog in at least one of the other 29 genomes is discarded). Finally, a separate matrix is built for each organism from the putative binding sequences retrieved by the first matrix and the scanning and filtering steps are repeated. In this second filtering process, all possible inter-genome orthology relationships are included in the analysis. For instance, a putative site identified in *Salmonella typhi* is rescued if an orthologous site is identified in *S.typhi*, even though it does not have an orthologous site in *E.coli*. For details on the implementation of this approach, which shares many features with known phylogenetic footprinting strategies (3,5,6), please refer to the Supplementary Data of the 2005 database issue publication (7).

The inclusion of the genomes of 13 new organisms to the prediction pipeline of this methodology eventually allowed extending the identification of putative binding sites for the 17 organisms, already contained in version 1.0. The main reason for these new findings was the identification of new orthology relationships, and not the discovery of new sites previously overlooked by the weight matrices. As stated above, regulatory sequences from *E.coli* O157H7 and *S.flexneri* 2a 2457T strains were not included in the construction of original matrices since their similarity to their orthologous regulatory sequences in *E.coli* K12 would have biased the training sets. The matrices produced from these training sets would have been expected to work well in those organisms closer to *E.coli* (or increase the rate of false positive sites). However, these biased matrices would have probably failed identifying many true sites in more distant organisms. Such bias did not occur, as shown by the specificity values (1) calculated for each *E.coli* regulon, which ranged from 96.2 to 100% (except for CRP and FNR that showed 79.6 and 84.4%, respectively, a rate of false positives that may be attributed at least in part to site cross recognition). Forty-four regulons


Network interface: E. coli known TFs and regulatory interactions between them
One-TF-multigenome interface: FruR regulon conservation
Variability
Sites found for fruK in Salmonella typhimurium LT2
One-gene-one-genome interface: regulatory sites found at the fruK TU in S. typhimurium
One-gene-multigenome interface: orthologous FruR sites
One-TF-one-genome interface: FruR binding sites in S. typhimurium
Next page

FruR

(Results page number 1: genes 1 to 10 of 23)

Eco	Sfle	Sty	Stm	Ypes	Ecoo	Sfl2	Ecar	Ypem	Ypse	Ppro	Plum
abrB	abrB	STY0772	abrB		abrB	abrB	abrB				
ymfC	ymfC	rluB	ymfC	y1801	ymfC	ymfC	ECA2440	rsuA2	ymfC	PBPRA1148	plu2802
ymfB	SF1153	STY1275	ymfB	y1800	Z1863	S1236	ECA2441	mutT3	YPTB2429		plu2803
trmU	ycfB	trmU	trmU	y1799	ycfB	ycfB	trmU	trmU	trmU	PBPRA1147	trmU
ycfC	ycfC	STY1273	ycfC	y1798	ycfC	ycfC	ECA2443	YP1767	YPTB2431	PBPRA1146	plu2805
purB	purB	purB	purB	purB	purB	purB	purB	purB	purB	PBPRA1145	purB
icdA	icdA	icdA	icdA	icdA	icdA	icdA	icd	icdA	icdA	PBPRA1149	icd
fruB	fruB	fruB	fruF	fruB	fruB	fruB	fruB	fruB	fruB	PBPRA1573	fruB
fruK	fruK	fruK	fruK	fruK	fruK	fruK	fruK	fruK	fruK	PBPRA1574	fruK
fruA	fruA	fruA	fruA	fruA	fruA	fruA	fruA	fruA	fruA	PBPRA1575	fruA

Prediction Method	TF	Left Position	Absolute Left	Absolute Right	Site Sequence
Statistical Models	CRP	-182	2307536	2307560	tagcaAAAATGTGCTTCACCTCAAAAAATttctc
Statistical Models	CRP	-153	2307507	2307533	gtagcAAAAATGTGCTTCACCTCAAAAAATTtctcc
Statistical Models	FruR	-112	2307466	2307484	gatcgTCTTGAAACGTTTCAGCGagtct
Statistical Models	FruR	-46	2307400	2307418	tgctgGGCTGAATCGTTTCAATTgctat

Figure 1. Links between the five query and navigation interfaces included in Tractor_DB version 2.0, illustrated using the FruR regulon.

showed 100% of specificity in the identification of putative regulatory sites. On the other hand, sensitivity values behaved roughly as reported in the previous version of the database with >40 regulons for which 100% of known TUs were correctly identified (7).

Further, a second computational strategy was added to the prediction pipeline, based on the use of regular expressions to identify putative orthologous regulatory sites to those that have been identified experimentally in *E.coli*. Figure 2 of the Supplementary Data presents a flowchart of this approach (11).

Briefly, this methodology uses *E.coli* regulatory sites, obtained from RegulonDB (10) to build regular expressions that are used to scan orthologous regulatory regions in the other 29 genomes. This scanning is conducted as a pattern matching, in which every position of the site is allowed to change with equal probability, thus permitting a more intensive exploration of the space of sequences recognized by the orthologous TF than do positional weight matrices. Each putative orthologous binding site is then assessed for its statistical significance. To do this, the score of the putative orthologous site identified by the pattern matching is calculated using a weight matrix for the TF that putatively binds to the site. This score is then compared to the score that the site would present if its sequence had changed (with respect to the matrix) at the same rate than the regulatory sequence where it is located has changed with respect to the *E.coli* orthologous regulatory sequence from which the original regular expression was derived (12). (For details regarding this second approach, please see ref. 11.)

The combination of these two computational strategies based on different principles resulted in a more complete reconstruction of the transcriptional regulatory network of the 30 gamma-proteobacteria included in the present version of the database. The weight matrix-based approach identified a greater number of regulatory links, mostly due to the reconstruction of a matrix ‘adapted’ to each organism, and the orthology filtering based on each separate organism. On the other hand, the regular expression-based approach allowed the identification of putative sites for TF-organisms combinations with few or no sites identified by the first approach. This complementation may be explained because the pattern matching-based approach indeed accomplished a more intensive exploration of the sequence space of orthologous regulons. An alternative explanation is that the results of the positional weight matrix-based approach may be affected by differences observed in GC contents among the genomes included in the prediction pipeline, since nucleotides background frequencies used to build the original matrices are calculated from the *E.coli* genome (7,11–13). The pattern matching-based approach identified putative binding sites for 133 TF-organism combinations for which the weight matrix-based approach failed to identify any sites.

Table 1 summarizes the data included in Tractor_DB version 2.0 compared to version 1.0. It presents the number of TFs' binding sites, and TUs under their regulatory control identified by the combination of the two approaches in each organism. *S.typhi* and *S.typhimurium* were the organisms with bigger increments both in the number of TFs (45 and 43) with regulatory outputs and the number of TUs (87 and 80) with regulatory inputs identified by either approach.

Improvements to the query and navigation interface

A new query interface was added to the four already implemented in version 1.0 (7) that allows the user to directly retrieve the data regarding the conservation of regulatory interactions within a given regulon (with respect to *E.coli*) from a map that contains all known *E.coli* TFs and the regulatory interactions that interconnect them. Each node in the map represents a TF, and it gives access to the information on the degree of conservation of each regulatory output (to individual structural genes) identified in *E.coli* across the genomes of all the other organisms included in the database.

Navigation tabs have been added to the dynamic pages generated by the Perl scripts in response to queries launched at any of the five interfaces. These tabs considerably ease the navigation between dynamic pages. Other minor improvements to the database interface comprise the inclusion of a download page, which allows direct access to flat files that contain the information stored in the database for each organism, and the segmentation of dynamic pages generated by the orthology view (one-gene-multigenome interface), the TF list view (one-TF-one-organism interface), and the Regulon conservation view (one-TF-multigenome) resulting in a speedup of the generation of dynamic pages by the Perl scripts. Figure 1 illustrates the new query interface, and the improvements that the tabs introduce to the navigation between pages.

Dynamic pages containing query results are linked to knowledge bases such as RegulonDB (10) and EcoCyc (14).

COMPARATIVE GENOMICS AND THE REGULATION OF GENE EXPRESSION

The availability of experimental information on the regulation of gene expression in *E.coli* and the development of methodologies for the identification of putative regulatory sites in a number of other gamma-proteobacteria have driven comparative studies regarding the organization of one or several regulons (15–19). The information stored in Tractor_DB should aid such studies in the future. Recently, using these data, we have conducted a study regarding the conservation of general regulatory mechanisms in six organisms of this subclass (20).

AVAILABILITY

Tractor_DB version 2.0 may be accessed at any of its three mirrors: the National Bioinfomatics Center (Cuba) mirror (www.bioinfo.cu/Tractor_DB); the National Laboratory for Scientific Computing (Brazil) mirror (www.tractor.lncc.br); and the Genomics Center (Mexico) mirror (http://www.ccg.unam.mx/tractorDB).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Luiz Gonzaga Paula de Almeida (LNCC) and César Bonavides (CCG) for their contribution to the database interface, and for maintaining the LNCC and CCG mirrors. We are thankful to two anonymous referees for their comments. This work was partly funded by a collaboration project on Bioinformatics between Cuba and Brazil supported by CNPq/MCT. V.E.A. would also like to acknowledge the support given to this project by Red Iberoamericana de Bioinformática (RIBIO rt VIIL), CYTED. Funding to pay the Open Access publication charges for this article was provided by the National Laboratory for Computation of Brazil.

Conflict of interest statement. None declared.

REFERENCES

1. Benítez-Bellón,E., Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA. *Genome Biol*., **3**, 0013.1–0013.6.
2. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*., **12**, 739–748.
3. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of TF binding sites in proteobacterial genomes. *Nucleic Acids Res*., **29**, 774–782.
4. Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*., **30**, 5549–5560.
5. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.A. (2001) Comparative Genomics Approach to Prediction of New Members of Regulons. *Genome Res*., **11**, 566–584.
6. Tan,K., McCue,L.A. and Stormo,G. (2004) Making connections between novel transcription factors and their DNA motifs. *Genome Res*., **15**, 312–320.
7. González,A., Espinosa,V., Vasconcelos,A.T., Pérez-Rueda,E. and Collado-Vides,J. (2004) TRACTOR_DB: a Database of Regulatory Networks in Gamma-Proteobacterial Genomes. *Nucleic Acids Res*., **33**, D98–D102.
8. Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *PNAS*, **95**, 5849–5856.
9. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
10. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al.* (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*., **34**, D394–D397.
11. Hernández,M., González,A., Espinosa,V., Vasconcelos,A.T. and Collado-Vides,J. (2004) Complementing computationally predicted

regulatory sites in Tractor_DB using a pattern matching approach. *In Silico Biol*., **5**, 0020.

12. Brown,C.T. and Callan,C.G., Jr (2004) Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli*. *PNAS*, **101**, 2404–2409.
13. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
14. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta–Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.
15. Makarova,K.S., Mironov,A.A. and Gelfand,M.S. (2001) Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol*., **2**, research0013.1–0013.8.
16. Panina,E.M., Mironov,A.A. and Gelfand,M.S. (2003) Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *PNAS*, **100**, 9912–9917.
17. Panina,E.M., Mironov,A.A. and Gelfand,M.S. (2001) Comparative analysis of FUR regulons in gamma-proteobacteria. *Nucleic Acids Res*., **29**, 5195–5206.
18. Erill,I., Jara,M., Slavador,N., Escribano,M., Campoy,S. and Barbé,J. (2004) Differences in LexA regulon structure among Proteobacteria through *in vivo* assisted comparative genomics. *Nucleic Acids Res*., **32**, 6617–6626.
19. Erill,I., Escribano,M., Campoy,S. and Barbé,J. (2003) *In silico* analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA-regulon. *Bioinformatics*, **19**, 2225–2236.
20. Espinosa,V., González,A., Huerta,A.M., Vasconcelos,A.T. and Collado-Vides,J. (2005) Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. *J. Mol. Biol*., **354**, 184–199.