D98–D102 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki054

TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes

Abel D. González, Vladimir Espinosa, Ana T. Vasconcelos[1], Ernesto Pérez-Rueda[2] and Julio Collado-Vides[3,*]

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

Received August 6, 2004; Revised and Accepted October 1, 2004

ABSTRACT

Experimental data on the *Escherichia coli* transcriptional regulatory system has been used in the past years to predict new regulatory elements (promoters, transcription factors (TFs), TFs' binding sites and operons) within its genome. As more genomes of gamma-proteobacteria are being sequenced, the prediction of these elements in a growing number of organisms has become more feasible, as a step towards the study of how different bacteria respond to environmental changes at the level of transcriptional regulation. In this work, we present TRACTOR_DB (TRAnscription FaCTORs' predicted binding sites in prokaryotic genomes), a relational database that contains computational predictions of new members of 74 regulons in 17 gamma-proteobacterial genomes. For these predictions we used a comparative genomics approach regarding which several proof-of-principle articles for large regulons have been published. TRACTOR_DB may be currently accessed at http://www.bioinfo.cu/Tractor_DB, http://www.tractor.lncc.br/ or at http://www.cifn.unam.mx/Computational_Genomics/tractorDB. Contact Email id is tractor@cifn.unam.mx.

INTRODUCTION

One of the challenges of Functional Genomics is the identification of all the elements that take part in an organism's transcriptional regulatory network. This is necessary to understand how the cell reacts to environmental stimuli at the level of transcriptional regulation. Intense research is being carried out in this direction (1–3). The first step towards this goal is the recognition of all the genes regulated by a transcription factor (TF), i.e. its regulon.

Computational approaches to recognizing the location of regulatory sites in bacterial genomes include the use of weight matrices (4), phylogenetic footprinting (5), searching for statistical overrepresentation of oligonucleotides within a genome and clustering co-expressed genes in order to find conserved patterns in their upstream regions (6,7), among others (8,9). Recently, Tan *et al*. (10) proposed a new methodology which brings together the advantages of both, the weight matrices and the phylogenetic footprinting approaches.

In the past few years, a great amount of research has been dedicated to computational prediction of important regulatory elements in the *Escherichia coli* genome: promoters (11), operons (12), TFs (13) and TF binding sites (9). As more bacterial genomes are sequenced, it is becoming more important to extend these efforts to other organisms, and decipher their transcriptional regulatory networks by means of comparative regulatory studies (10,14–19).

Our two major goals in this work were the production of a reliable set of binding site predictions for as many gamma-proteobacterial TFs as possible in 17 organisms of this division [*E.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), *Xanthomonas axonopodis* (NC_003919) and *Xylella fastidiosa* (NC_002488)], and the construction of a database (TRACTOR_DB, accessible at http://www.bioinfo.cu/Tractor_DB, http://www.tractor.

*To whom correspondence should be addressed. Tel: +527 773 132063; Fax: +527 773 175581; Email: collado@ccg.unam.mx

lncc.br and http://www.cifn.unam.mx/Computational_Genomics/tractorDB) with a user-friendly navigation interface containing these computationally predicted sites. Several recent papers have addressed the first goal; however, to our knowledge, the present work is the most complete in terms of the number of regulons and organisms that it comprises. Most other studies have been limited to *E.coli* and *H.influenzae* (10) or to one or few regulons (18,19). In contrast, the work by McCue *et al*. (20) predicted a large number of putative regulatory sites in 10 gamma-proteobacterial genomes, but could not associate many of those sites to a given TF.

METHODS

Selecting organisms and regulons

We used the main ideas of the methodology proposed by Tan *et al*. (10) to predict new regulon members in 17 gamma-proteobacterial genomes. Orthology information is used in this methodology to assess the biological significance of putative regulon members. The methodology requires that the organisms selected be phylogenetically close to *E.coli*, since *E.coli* known binding sequences are used to build a statistical model of a TF's binding site, and this model is subsequently used to predict new members of that regulon in the other organisms within the study. Hence, we included in our study a group of organisms from different subdivisions of the gamma-proteobacteria subclass whose genomes were completely sequenced. We started working with all TFs with at least one binding site known in *E.coli*.

Outline of the predictive methodology

Original training sets contain *E.coli* TFs' binding sequences extracted from RegulonDB, version 4.0 (21). The first eight organism referenced in the list above (those sharing at least 30% of orthologous genes with *E.coli*) were used to construct the original training sets. A training set was built for each TF with at least one known binding sequence in *E.coli* (21) and it included also those of orthologous non-coding regions when less than 25 binding sequences were known in *E.coli*. We built weight matrices only for TFs with training sets larger than four sequences. Two statistical models were built for each TF: one using CONSENSUS, and the second using the Gibbs-SAMPLER. The training sets were filtered twice as proposed by Tan *et al*. (10) to eliminate possible weak binding sequences and cutoff values were selected in accordance with their methodology.

The model built using CONSENSUS was aligned against TUs promoter regions (from −400 to +50) using PATSER (4), setting the lower threshold of the program to the weak cutoff value; the DSCAN software was used to scan these regions with the model built using the Gibbs-SAMPLER (22). The sets of putative sites thus produced by CONSENSUS and Gibbs-SAMPLER were merged and subsequently filtered using orthology information and score (10).

New training sets were built rescuing putative sites predicted for each TF in all organisms. For those TFs that produced more than four putative sites in *E.coli* (after the orthology filtering) and in at least one other organism, a training set was built for each separate organism with more than four putative sites. These sets were then used to build a specific model for each organism using CONSENSUS, which were then used to re-calculate the cutoffs and to re-scan the regulatory regions of each organism.

Summing up, the main features included in the approach proposed by Tan *et al*. (10) in order to extend it to as many TFs as possible were as follows: (i) the use of two different algorithms to build the statistical models of each TF binding site: the CONSENSUS (4) and the Gibbs-SAMPLER (22); (ii) the inclusion, within the training set used to build the model of each TF, of non-coding regions upstream the TUs of the organisms other than *E.coli* that are orthologous to those that in *E.coli* are known as regulon members (orthologous non-coding regions), along with *E.coli* known binding sequences of the TF; and (iii) the reconstruction of the models for each TF in each organism after the prediction process, which allows the refinement of the search for new members of the regulon within each genome.

All the sites found for a given TF—in the eight initial organisms—after the second scanning using the rebuilt models were aligned, using CONSENSUS, to produce a Positional Weight Matrix (PWM). Those PWMs were then used to scan the genomes of the other nine organisms for new putative binding sites. The orthology filtering process was done as described in the Supplementary Material, Section I.5, using only the first eight organisms at the centre of the analysis.

For a thorough description of the predictive methodology see Section I.(1–6) and Figure 1 of the Supplementary Material.

Designing and building the database

The database was designed and built following the relational model, and installed on a MySQL server. The web interface is managed by a cgi PERL script that does all the work, from querying the database at the user's request, to generating the dynamic web pages that form the interface. The design that we have adopted makes it very easy to incorporate new instances to the database, which may be very easily accommodated into the interface. Figure 1 shows the main relationships between the tables that compose TRACTOR_DB and the most important queries carried out by the interface program.

Several links in the dynamically generated web pages that form the interface make the navigation easy and user-friendly. The dynamic pages are linked to other databases such as RegulonDB (21), EcoCyc (23) and the NCBI database. All the data stored in the database may be downloaded in the form of flat files and the complete system will be available for installation upon request in the near future.

RESULTS

Putative new regulon members by organism

Table 1 of the Supplementary Material shows the number of members (TUs) found in all organisms for each of the 74 regulons for which a statistical model was built. (In those cases for which the model of the binding site could be rebuilt, the results shown correspond to the search done with that rebuilt model.) For those regulon members found in *E.coli*, predictions are compared to regulon members (TUs) annotated

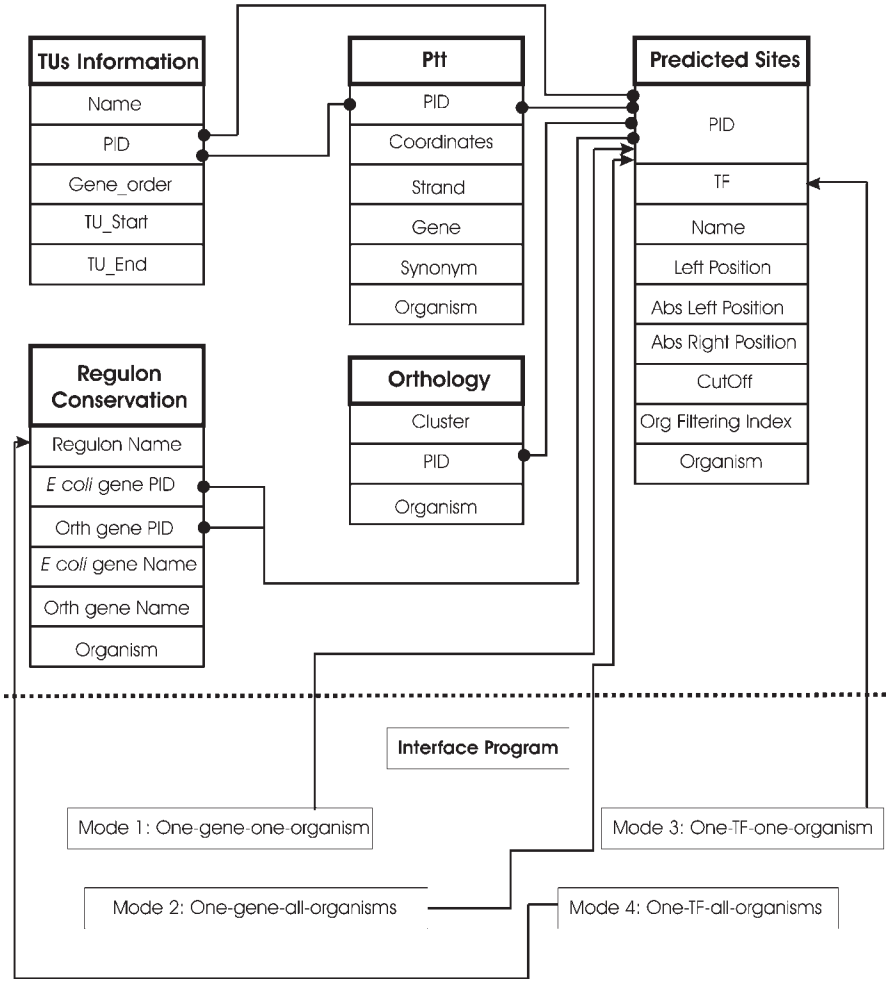

Figure 1. TRACTOR_DB design. Main relationships between tables (oval arrows), and main queries performed by the interface program (stealth arrows).

in RegulonDB (shown in column one), in order to assess the sensitivity of our search for new regulon members. Regulons may be grouped in three different clusters based on this sensitivity. First, there is a set of six regulons for which 50–80% of the known TUs are recognized by the methodology. This group contains several global regulators (24), and its most prominent members are CRP (75% of member TUs recognized) and FNR (65%). This improves the results obtained by Tan *et al*. (10) (44.6 and 46.2%, respectively), which are mostly due to the incorporation of new organisms into the study. The failure in detecting the remaining quarter of CRP, and one-third of FNR regulated TUs is mostly due to the restriction imposed by the weak cutoff. Most of the sites that the methodology does not recognize score below this cutoff. The second group comprises 63 regulons, with few (<10) TU members known in *E.coli*. For these, most TU members are recognized by the methodology (from one-half to all). Finally, there is a group of five regulons for which less than that one-half of the members are identified, in most cases due to the failure in producing a good model of the binding site. Our results recognizing *E.coli* known regulon members are comparable to those of McCue *et al*. (20) for small regulons and are better for large regulons.

For some regulon–organism combinations no putative sites are found although the TF is found in that particular organism. This happens mainly for small regulons, those for which there is a greater variation in the sequence of the recognition helix from genome to genome (16), and thus, for which a greater variation in the recognition sequence from one organism to the next is expected.

The number of regulons for which our methodology could predict sites is significantly low in the organisms more distant from *E.coli* (i.e. *Xanthomonas axonopodis* and *Buchnera aphidicola*, Table 1). This is mainly due to the limitations that impose the methodology which rely on orthology relationships between the organisms under study.

The database

We designed the web interface TRACTOR_DB so that it may be accessed and browsed in four different ways or modes. The first mode may be called one-gene-one-genome; it displays all the sites found for a TU of any organism. The user selects this mode by entering a gene name or gi and ticking an organism's name. In the second mode (one-gene-multigenome), the user enters a gene name or gi and ticks the *all_organisms*

Table 1. Total number of predictions—supported by orthology or scoring above the strong cutoff—and TFs found in each organism

Organism	Total predictions	Number of TFs
E.coli K12	853	74
S.typhi	725	23
S.typhimurium LT2	752	26
S.flexneri 2a	658	32
Y.pestis KIM	354	11
H.influenzae	140	5
V.cholerae	234	7
S.oneidensis	285	6
B.aphidicola	7	3
P.aeruginosa	22	11
P.putida KT2440	17	9
P.syringae	16	9
V.parahaemolyticus	141	28
V.vulnificus CMCP6	112	23
X.axonopodis	3	2
X.campestris	5	4
X.fastidiosa	7	4

option: the interface then displays the orthology information that supports all sites predicted upstream the *E.coli* TU where that gene is located. The third mode (one-TF-one-genome) displays all predicted members of the TF regulon in one organism. To access this mode, the user ticks an organism's name and selects a TF name from the menu next to it. The last mode, which may be called one-TF-multigenome, displays the information regarding the conservation of a simple or complex *E.coli* regulon (a set of genes regulated by the combination of two or more TFs) across all the other 16 genomes.

The current release of TRACTOR_DB (1.0) contains computational predictions of new members of 74 regulons in the 17 gamma-proteobacterial genomes referenced above. Future releases will include transcription binding site predictions for new gamma-proteobacteria whose genomes are already sequenced and others which are currently in progress. The update process will be carried out as new genomes are available approximately once every six months. Table 1 summarizes the information available in the current release of TRACTOR_DB.

DISCUSSION

To fulfill the identification of new members of as many known gamma-proteobacterial regulons as possible in eight organisms of this group, we employed the main ideas developed by Tan *et al*. (10) in their approach to predict new members of regulons, as an effective way to reduce the high false positives rate inherent to all computational approaches used to predict regulatory signals. Moreover, we incorporated several extensions into that methodology in order to include as many regulons as possible to the study, including all with at least one known binding sequence in *E.coli*. Finally, we found new putative binding sites in *E.coli* for 74 out of 102 TFs with binding sites annotated in RegulonDB, and for 42 of them, we were able to rebuild the model of the binding site and tune the search in each genome. The sensitivity of this extended methodology in *E.coli* proved very good (Table 1 of Supplementary Material). In addition, we predicted new members of regulons in the other 16 gamma-proteobacterial genomes under study, in numbers ranging from two regulons in *X.axonopodis* to 32 in *S.flexneri*.

The most important result of this work is the establishment of TRACTOR_DB, a relational database that gives access, through a user-friendly browsing interface, to our computationally predicted members of proteobacterial regulons. The design of the database makes it very easy to include new computationally predicted regulon members in other gamma-proteobacterial genomes and/or using different predictive approaches. TRACTOR_DB should aid experimentalist researchers in the analysis of microarray results or in any other situation when a piece of information regarding the possible regulation of a gene or set of genes may be of help (orienting researchers about which genes are more likely to be regulated by a given TF or set of TFs).

TRACTOR_DB may also be regarded as a starting point in the understanding of the organization of the transcriptional regulatory network of bacteria other than *E.coli*, given the fact that—to our knowledge—this is the largest set of putative regulon members in gamma-proteobacterial genomes from the point of view of number of regulons and organisms.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Marcelo T. dos Santos (LNCC) for his important contribution to this work, Dr Gabriel Moreno-Hagelsieb for fruitful discussions, and Fernanda Mendonça (LNCC), Roger Paixão (LNCC) and Heladia Salgado (CCG) for their contribution to the database interface. This work was partly funded by a collaboration project on Bioinformatics between Cuba and Brazil supported by CNPq/MCT. J.C.-V. acknowledges support from grants 0028 from Conacyt and GM62205 from NIH.

REFERENCES

1. Buhler,N.E., Gerland,U. and Hwa,T. (2003) On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA*, **100**, 5136–5141.
2. Cases,I., de Lorenzo,V. and Ouzounis,C.A. (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol*., **11**, 248–253.
3. Thieffry,D., Huerta,A.M., Pérez-Rueda,E. and Collado-Vides,J. (1998) From specific gene regulation to genomics networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays*, **20**, 433–440.
4. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
5. Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs recursive sampler: finding TF binding sites. *Nucleic Acids Res*., **31**, 3580–3585.
6. Qian,J., Lin,J., Luscombe,N.M., Yu,H. and Gerstein,M. (2003) Prediction of regulatory networks: genome-wide identification of TF targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.
7. Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**, I273–I282.
8. Li,H., Rhodius,V., Gross,C. and Siggia,E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial gnomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.

9. Thieffry,D., Salgado,H., Huerta,A.M. and Collado-Vides,J. (1998) Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K12. *Bioinformatics*, **14**, 391–400.
10. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.
11. 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**, 261–278.
12. 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. USA*, **97**, 6652–6657.
13. Pérez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K12. *Nucleic Acids Res.*, **28**, 1838–1847.
14. Mirny,L.A. and Gelfand,M.S. (2002) Structural analysis of conserved base pairs in protein–DNA complexes. *Nucleic Acids Res.*, **30**, 1704–1711.
15. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
16. 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**, 298–308.
17. Pérez-Rueda,E. and 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.
18. 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.
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. 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.
21. Salgado,H., Gama-Castro,S., Martínez-Antonio,A., Díaz-Peredo,E., Sánchez- Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jiménez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martínez,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**, 303–306.
22. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
23. Karp,P.D., Arnaud,M., Collado-Vides,J., Ingraham,J., Paulsen,I.T. and Saier,M.H.,Jr (2004) The *E. coli* EcoCyc Database: no longer just a metabolic pathway database. *ASM News*, **70**, 25–30.
24. Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.