TRACTOR_DB as a source for comparative studies of transcriptional regulation mechanisms in gamma-proteobacterial genomes

Vladimir Espinosa[1], Abel D. González[1], Ana T. Ribeiro[2] Araceli M. Huerta[3] and Julio Collado-Vides[3].

[1] National Bioinformatics Center. Industria y San Jose, Capitolio Nacional, Havana 10200, Cuba. [2] National Laboratory for Scientific Computing. Av. Getulio Vargas, 333, Quitandinha CEP: 25651-075, Petrópolis - Rio de Janeiro. [3] Center for Genomic Sciences. CCG-UNAM, Av. Universidad s/n. Cuernavaca, Morelos 62100, Mexico.

contact: vespinosa@bioinfo.cu

Introduction

The great amount of genomic information generated by the international and private sequencing programs have made possible the study of transcription regulation from a genomic and global point of view. However, even for the best-studied free-living organism, the *enterobacterium Escherichia coli*, much still remains to be understood about how genes are regulated at the level of transcription initiation. Transcription Factors (TFs) bind to specific motif sequences in the regulatory regions of genes determining their expression by interacting in some way with the transcription machinery. A lot of work has been done to identify those sequence motifs in complete sequenced genomes using computational strategies as a parallel to the experimental characterization approaches.

We've developed a publicly accessible database known as TRACTOR_DB which stores well supported information about the transcriptional regulation systems in 17 gamma-protaobacteria, most of which are important pathogens and biological models (http://www.bioinfo.cu/Tractor_DB;http://www.tractor.lncc.br; http://www.cifn.unam.mx/Computational_Genomics/tractorDB). TRACTOR_DB was constructed following a methodology based on the conservation of TF-binding sites in the regulatory regions of closely related species and, at least to our knowledge, is the most complete resource with information of putative regulon members in gamma-proteobacteria. The principal aim of this study is to exploit the wealth of the information stored in TRACTOR_DB to conduct a group of comparative studies concerning the conservation of transcriptional regulation mechanisms in eight gamma-proteobacterial genomes.

Methods

•Site-Promoter distance was calculated using the site predictions reported in TRACTOR_DB [1] in eight gamma-proteobacterial genomes (*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)) and the promoter predictions obtained using the methodology described by [2] in the aforementioned genomes. The study was restricted to those TFs for which we could rebuild organism-specific models (38 TFs). The data generated in this study was used to construct the frequency histograms of site-promoter distance for a given TF in each organism under study.

•TF binding specificity (x) was estimated for each TF in each one of the eight genomes using an information-based approach which takes into account the score distribution of the TF recognizing its cognate sequences and a set of random sequences —represented by the entire genome of the organism analyzed. As described by [3] the formula used to calculate this estimator was the following:

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

•The relationship between genome organization and transcriptional regulation was estimated exploring the fate of each *E. coli*'s TU in the other 8 genomes and the resulting regulation of the fragments generated. We classified each pair of TUs as ***identical, similar, destroyed*** or ***lost*** as proposed by [4]. TF-binding site conservation was estimated using a site-orthology score (***SOS***) which accounts for the presence of conserved regulation in both TUs under analysis (a_i) and the evolutionary distance between the genomes compared (A_i), as defined in the following formula:

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

•We also study the co-occurrence of pairs of TF-binding sites in the regulatory regions of genes and the statistical significance of the number of co-occurrences in the organisms under study. Using the approximation described by [5] to estimate the statistical significance of co-occurrences of sites in a genome, we calculated the expected number of occurrences ($E(x)$) and the probability of obtaining by chance the observed number of occurrences ($P(bin)$) within regions of fixed base pair lengths ($bins$) in each genome for a given pair of TFs using the formula:

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

Results

a) The binding specificity of a regulator tends to be conserved across genomes, as proved by the linear behavior observed in the plots depicted in **Figure 1** which is related to the conservation of the signal-to-noise ratio of each *E. coli* regulon and all orthologous regulons in the other organisms.

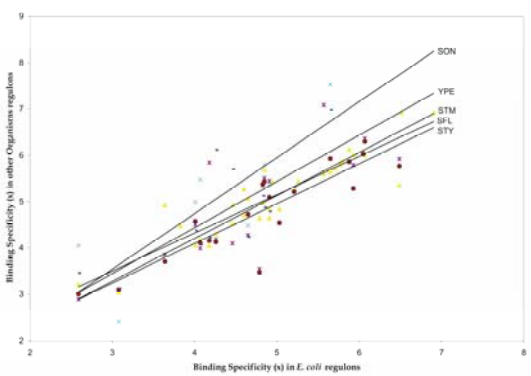

Figure 1: TFs' binding specificities in different gamma-proteobacteria relative to *E. coli* regulons. In the graph: STY, *Salmonella typhi* (***y = 0.8592x + 0.6598; R² = 0.7983***); STM, *Salmonella typhimurium* (***y = 0.9271x + 0.4952; R² = 0.683***); SON, *Shewanella oneidensis* (***y = 1.2062x - 0.0816; R² = 0.6783***); SFL, *Shigella flexneri* (***y = 0.8273x + 1.0126; R² = 0.834***); YPE, *Yersinia pestis* (***y = 0.9959x + 0.4547; R² = 0.5998***).

b) There is a clear conservation of distance ranges among more closely related organisms as can be seen in the histograms obtained, depicted in Figure 2 for the CRP regulon in *S. flexneri* and *S. typhi*. Conserved peaks are found at positions (-45 → -41); (-65 → -60) and (-85 → -80).

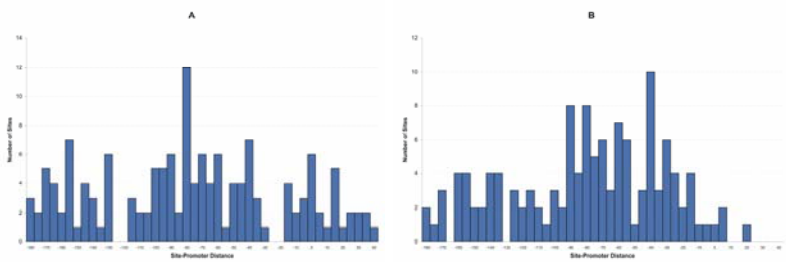

Figure 2: Site-promoter distance frequency histograms of the predictions in *S. flexneri* and *S. typhi*. The distance from the center of the putative site to the position 10bp downstream from the center of the -10 box of the CRP predictions were plotted in *S. flexneri* (**A**) and *S. typhi* (**B**) in the abscissa against the absolute number of predictions existing at a given distance-*bin* from the promoter.

c) As can be seen in Figure 3 A and B there is a tendency for predictions more conserved across organisms —those with higher site-orthology scores— to occur in the regulatory regions of TUs with more conserved structure.

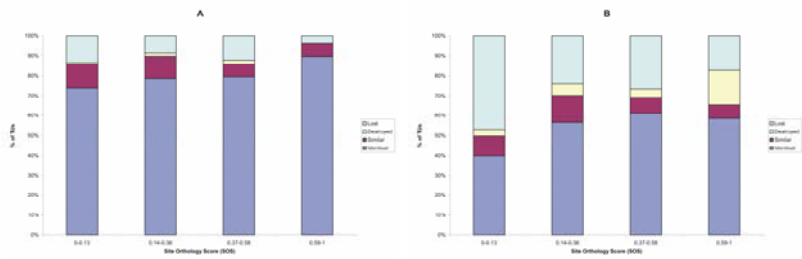

Figure 3: Conservation of TU structure and TU regulation in *S. flexneri* and *Y. pestis*. Fraction of operons in *S. flexneri* (A) and *Y. pestis* (B) resulting in each one of the four TU-comparison-categories when compared to *E. coli*'s TUs and the conservation of the sites existing in the regulatory regions of those TUs in the organisms studied estimated through the site orthology score (*SOS*). Each bar represents the percent of TUs in the given organism belonging to each category containing sites in their regulatory regions within the given site orthology score range.

TFs	obs(x)	Organism	E(x)	P(x)
CRP-CRP	24	*Haemophilus influenzae*	3.313	6.55E-13
CRP-CRP	93	*Shigella flexneri* 2a	13.956	0
CRP-CRP	29	*Vibrio cholerae*	4.1255	6.54E-13
CRP-CRP	98	*Escherichia coli* K12	24.453	0
CRP-CRP	112	*Salmonella typhimurium* LT2	19.722	0
CRP-CRP	120	*Salmonella typhi*	50.767	0
CRP-CRP	56	*Yersinia_pestis* KIM	5.2178	7.88E-13
CRP-CRP	87	*Shewanella oneidensis*	4.6382	0
FNR-FNR	6	*Haemophilus influenzae*	0.40114	4.06E-06
FNR-FNR	34	*Shigella flexneri* 2a	2.9663	1.53E-12
FNR-FNR	3	*Vibrio cholerae*	0.12363	0.00029
FNR-FNR	47	*Escherichia coli* K12	5.5404	3.73E-13
FNR-FNR	26	*Salmonella typhimurium* LT2	2.7251	1.19E-12
FNR-FNR	30	*Salmonella typhi*	8.262	4.26E-09
FNR-FNR	18	*Yersinia pestis* KIM	0.57281	8.66E-14
FNR-FNR	3	*Shewanella oneidensis*	0.058987	3.26E-05
Fur-Fur	13	*Shigella flexneri* 2a	0.22423	1.15E-13
Fur-Fur	3	*Vibrio cholerae*	0.023139	1.97E-06
Fur-Fur	31	*Escherichia coli* K12	0.60302	4.06E-14
Fur-Fur	25	*Salmonella typhimurium* LT2	0.54875	2.39E-13
Fur-Fur	16	*Salmonella typhi*	1.2265	5.88E-13
Fur-Fur	27	*Yersinia pestis* KIM	0.13427	2.02E-14
MetJ-MetJ	5	*Haemophilus influenzae*	0.037341	5.33E-10
MetJ-MetJ	3	*Shigella flexneri* 2a	0.01506	5.52E-07
MetJ-MetJ	11	*Vibrio cholerae*	0.066111	1.04E-14
MetJ-MetJ	13	*Escherichia coli* K12	0.077061	5.00E-15
MetJ-MetJ	18	*Salmonella typhimurium* LT2	0.079524	3.46E-14
MetJ-MetJ	10	*Salmonella typhi*	0.14262	2.42E-14
MetJ-MetJ	15	*Yersinia pestis* KIM	0.021674	3.33E-15
MetJ-MetJ	37	*Shewanella oneidensis*	0.047072	2.28E-14

d) As can be seen in Table 1 there is a correspondence between the values obtained for the expected number and the probability of co-occurrences of sites in the regulatory regions of TUs, with a trend of co-occurrence of given pairs of TFs with similar statistical significance in closely related organisms.

Table 1: Co-occurrences of sites found in a selected group of TFs and its statistical significance in all the organisms under study. The first column shows the TF-TF pair, the second one the observed number of predictions found by our search, the third one shows the Organism and the fourth and fifth ones show the number of co-occurrences expected by chance for the given pair within the genome and the probability of finding the observed number of co-occurrences in the 0-100 bin.

Discussion

- The TF-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. 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* [8] 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 seen in **Figure 1** the regression coefficients of the lines range from 0.59 to 0.83, with the lower values corresponding to organisms that are evolutionarily more distant from *E. coli*.
- 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 2**. 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 [6, 7]. 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's interesting that this behavior is conserved in other related organisms and the shape of the histograms is similar in organisms like *S. flexneri* and *S. typhimurium* whose genome sizes 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, which means that not only the sites are conserved during evolution in this group of organisms, but also their positional characteristics
- The results depicted in **Figure 3** show that in all the organisms studied those TUs with more conserved structure generally have putative binding sites supported by orthology in most of the organisms —those with higher orthology scores and located right in this chart. Such highly conserved TUs may be part of what have been named as the regulon core in the LexA case [9]. These results show a clear tendency for similar regulation of TUs which are more frequently conserved in different gamma-proteobacteria, which is also related to the fact that the evolutionary history of operon structure and that of its upstream regulatory elements is closely related.
- A close examination of **Table 1** shows some interesting aspects concerning the co-occurrence of sites in the organisms included in this study. The results shown in **Table 1** demonstrates that the co-occurrence of different pairs of TF-binding sites is conserved among some of the organisms in each case –excluding those organisms where a given TF doesn't exist. These results are also consistent with previous findings for co-existence of sites in the regulatory regions of genes in *E. coli* [5, 10] and prove that also this property of transcription regulation is conserved in closely related organisms which might be related with the conservation of higher order regulatory complexes during evolution.

References

1. 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.
2. 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. doi:10.1016/j.jmb.2003.07.017.
3. 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.
4. 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.
5. 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.
6. Collado-Vides, J., Magasanik, B. & Gralla, J. D. (1991). Control Site Location and Transcriptional Regulation in *Escherichia coli*. *Microbiol. Rev.* **55(3)**, 371-394.
7. Collado-Vides, J. (1992). Grammatical model of the regulation of gene expression. *Proc. Natl. Acad. Sci. USA.* **89**, 9405-9409.
8. Martínez-Antonio, A. & Collado-Vides, J. (2003). Identifying Global Regulators in Transcriptional Regulatory Networks in Bacteria. *Current Opinion in Microbiol.* **6**, 482-489.
9. 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.
10. 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 Journal.* **16(8)**, 2108-2118.

We also acknowledge the Iberoamerican Bioinformatics Network ***(Red Iberoamericana de Bioinformatica-RIBIO VII.I., CYTED)*** *for partial support to this project.*