A comparative site position study in a group of gamma-proteobacterial genomes. Co-occurrence of binding sites for pairs of TFs and its implication in regulation of transcription.

V. Espinosa[1], A. D. González[1], A. T. Ribeiro[2] and J. Collado-Vides[3]

[1] National Bioinformatics Center (BIOINFO), Industria y San José, Capitolio Nacional, CP. 10200, Habana Vieja, Habana, Cuba.

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

[3] Center of Genomics (CCG), UNAM, Mexico. AP 565-A Cuernavaca, CP 62100, Morelos, Mexico.

e-mail: vespinosa@bioinfo.cu

Abstract

A lot of work has been done to identify Transcription Factor (TFs) binding sites motifs in complete sequenced genomes using computational strategies, as a parallel to the experimental characterization approach. Here we present a comparative study of the specific location of binding sites for a group of TFs in the regulatory regions of genes in eight gamma-proteobacterial genomes using a set of predictions made by our group using a comparative genomic approach similar to the one described by Tan *et al.*, 2001. This study shows that exist similar patterns of preferred distances recognized by the TFs in organisms phylogenetically closed which in turn are different from the patterns in more distant related organisms. This must be related with similar regulatory mechanisms developed during the evolution and shared by closed related organism.

The second part of this study corresponds to the estimation of the statistical significance of the co-occurrence of pairs of TF's binding sites in the regulatory regions of groups of genes in the eight genomes and their possible implications in the regulation of transcription. The results obtained for a group of TFs is consistent with the results previously reported in the literature for TFs known to co-occur during the regulation of group of genes like CRP and FNR in *E. coli*. The results obtained for the other seven organisms, which regulation have not been studied so deeply, revealed patterns of co-occurrence very similar to the ones observed in *E. coli* which might be related with the conservation of the mechanisms mediating the regulation of transcription.

Methods

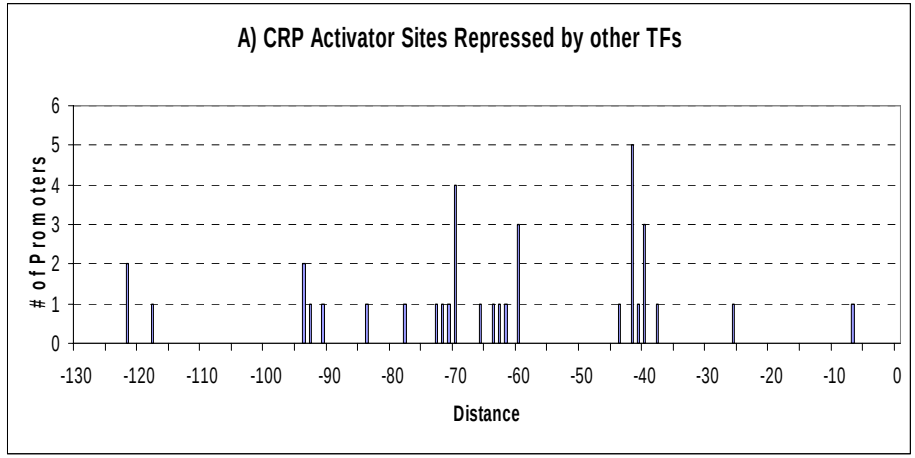
• Site-Promoter distance was calculated using the site predictions obtained by us in eight gamma-proteobacterial genomes (*Escherichia coli* K12 (NC_000913), *Haemophilus influenzae* (NC_000907), *Salmonella typhi* (NC_003198), *Salmonella typhimurium* LT2 (NC_003197), *Shewanella oneidensis* (NC_004347), *Shigella flexneri* 2a (NC_004337), *Vibrio cholerae* (NC_002505), *Yersinia pestis* KIM (NC_004088)) (González *et al.*, 2004) and the promoter predictions obtained using the methodology described by Huerta & Collado-Vides (Huerta & Collado-Vides, 2003) in the correspondent genomes. The study was restricted to those TFs for which we could rebuilt organism-specific models (48 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.

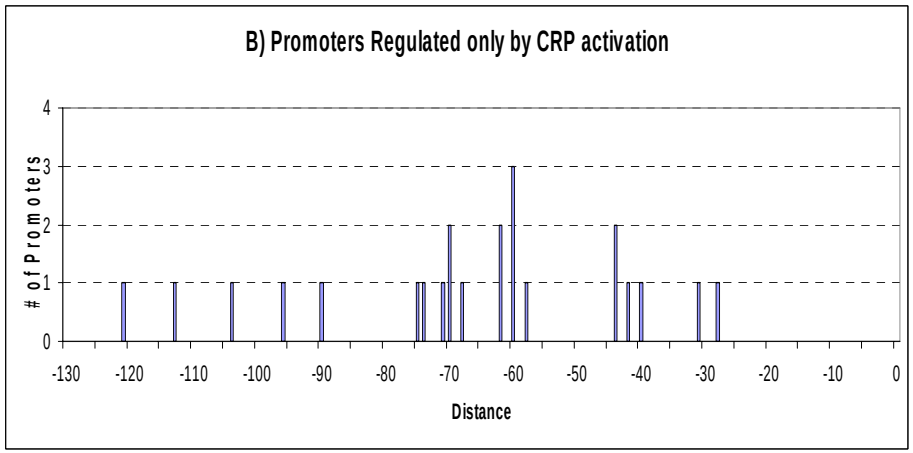
• We correlated the binding strength -using the site score as an estimator- of all sites in complex regulatory regions -regions with more than one site- found in a given organism and the characteristics of the surroundings of the site. The scores were obtained when scanning the genomes with PATSER using the rebuilt models generated using CONSENSUS. For each site in a complex regulatory region we calculated the average distance from the site to the other sites found in that region. The site scores were then plotted against the average distance of each site producing the correlation graphics depicted below.

• We also study the co-occurrence of pairs of sites within 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 Bulyk *et al.* to estimate the statistical significance of co-ocurrence of sites in a genome (Bulyk *et al.*, 2004) 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.

Results

The figures correspond to the site-promoter distance frequency histograms of CRP regulated promoters. These results obtained from the total known sites reported in RegulonDB (release v4.0) are consistent with the results obtained by Collado-Vides (Collado-Vides, 1992) using a quite lower number of sites. These outcomes corroborate the relationship between the function of a given site and the characteristics of the surroundings of the site. This means that the distance from site to promoter determines in some way the activator-repressor function of the site and the coexistence of sites from the same or different TFs affects the regulatory meaning of the regulatory region because of the coordination that may exist between those sites to exert their function.





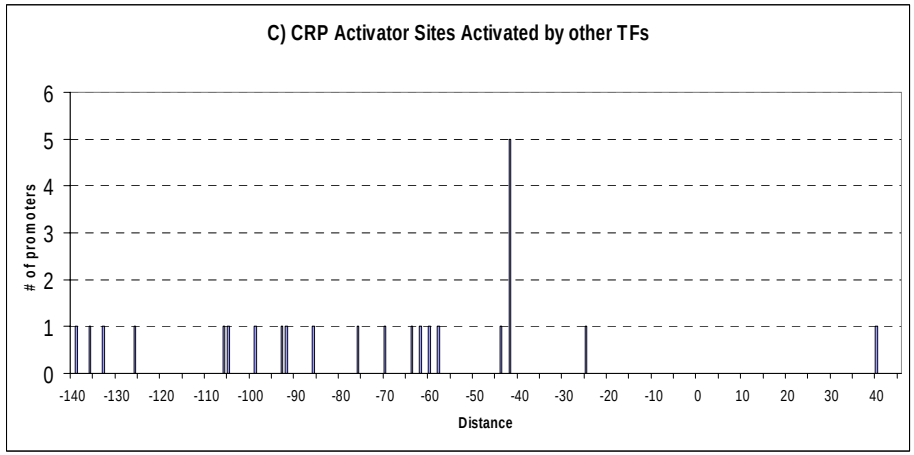
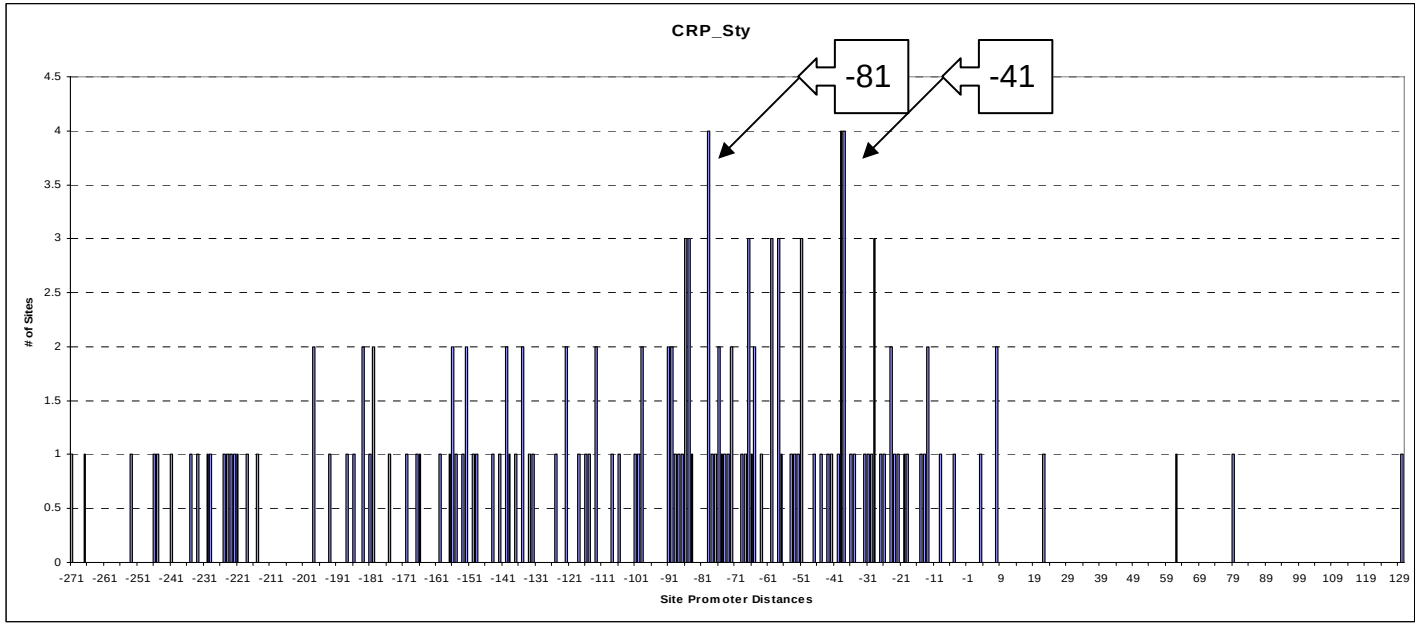


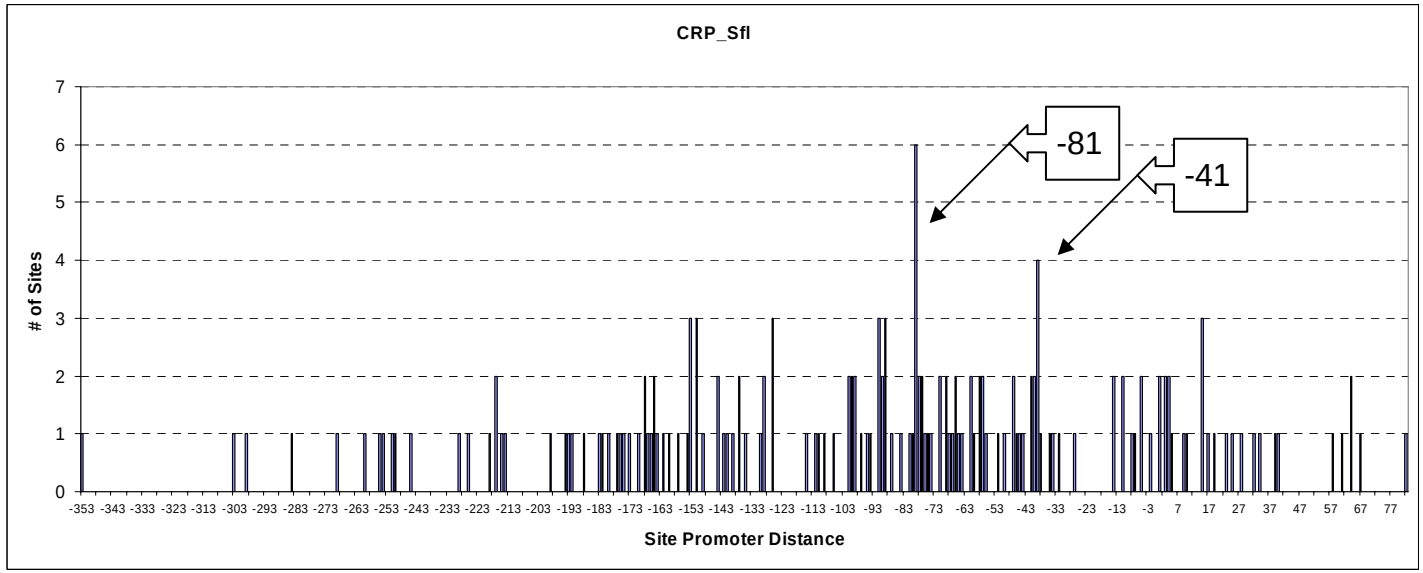
Fig 1: Frecuency histograms (site-promoter distance) of *E. coli* promoters: A) Containing CRP sites and repressor sites of another TF; B) Only subject to CRP activation and C) Containing CRP sites and activator sites of another TF

Predictions's histograms

Fig 2: Frequency histograms of the predictions found for CRP in the *S. typhi* and *S. flexneri* genomes

The histograms inspection shows the conservation of preferred distances from sites to promoters. Similar results were found for other TFs and other organisms. The closer the phylogenetic relation between the organisms the more similar the preferred distance ranges –as can be seen in the two figures for the -41 and -81 peaks-

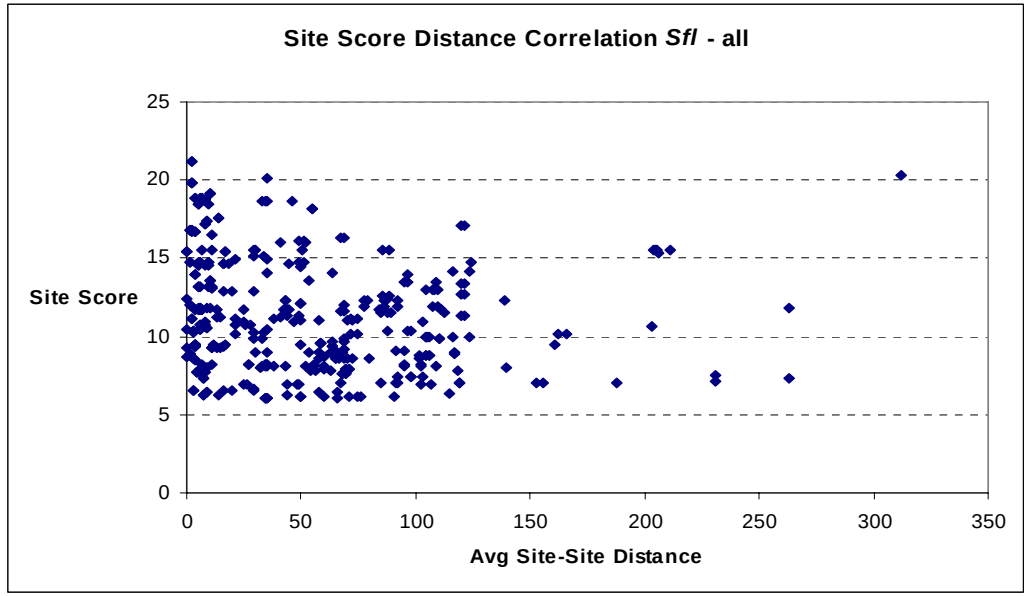


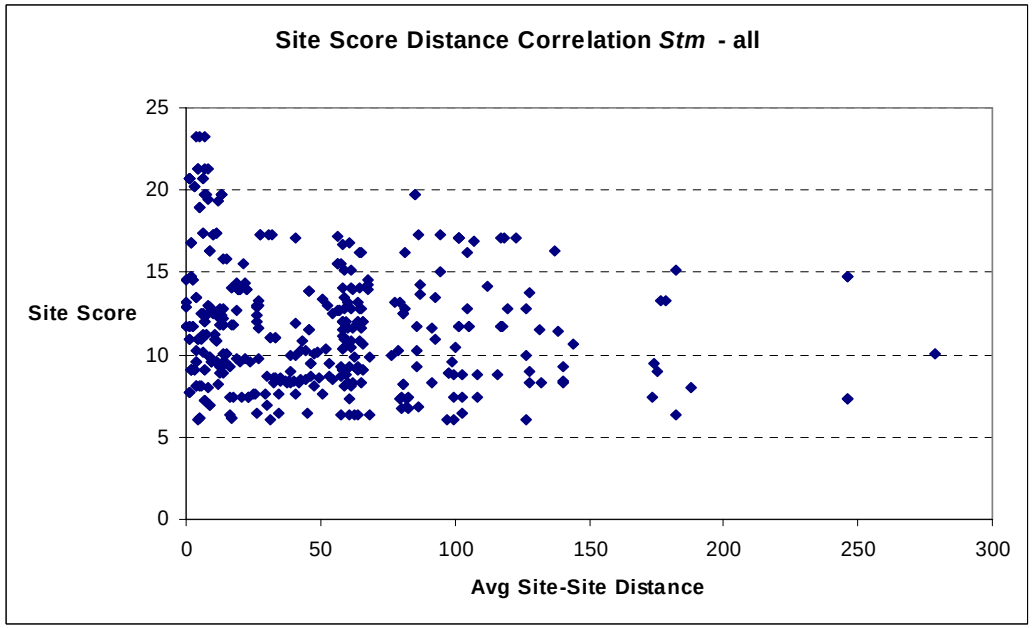


Binding energy-site position correlation

Fig 3: Correlation between the score -binding strength- of each site and the average distance to other sites co-occurring in complex regulatory regions in *S. typhimurium* and *S. flexneri*.

As in the case of distance range location of sites, the binding energy - Avg site-site location is also conserved between close related organisms. Fig 3 shows that the score of sites are tightly grouped in a confined section of the chart, corresponding to multiple complex regulatory regions in a genome having sites with the same surroundings characteristics –in average similarly distant from other sites- and different binding strengths. This behavior is also observed in the distal and proximal subpopulations (data not shown) which means that, with few exceptions, this behavior is general to all sites no matter their positional classification.





Co_Ocurrence of pairs of sites

Table 1: Co-occurrences of sites found in a selected group of TFs and its statistical significance in all the organism 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.

The results shown in Table 1 demonstrates that the co-occurrence of different pairs of sites is conserved among some of the organisms in each case –excluding those organisms in which a given TF doesn't exist-. These results are also consistent with the findings of other authors for co-existence of sites in the regulatory regions of genes in *E. coli* (Bulyk et al., 2004; Pedersen and Valentin-Hansen, 1997).

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

Discusion

The results obtained in this study are important extensions to the findings reported in the bibliography for *E. coli* TF binding sites. As reported by (Collado-Vides, 1992; Collado-Vides *et al.*, 1991) there is a relation between the distance from sites to promoters and their function and the preferred distances observed for a given TF and the characteristics of the surroundings of each site. We present here an update of the second study with the new data reported in the latest version of RegulonDB which demonstrates that this is a general behavior, at least for CRP. We couldn't find the similar results in other TFs starting from their known sites, which is related with different distance patterns depending on the characteristics of the TF instead of a general behavior for any TF.

Even we couldn't find a way of relating confidently site distance with its functionality, we find clear conservations of preferred distance in the frequency histograms of close related organisms (Fig 2) and marked differences between the distant ones. This is related with a conserved structure of TUs and their regulation which is more evident in closer related organisms.

Fig 3 shows that not only the distant patterns are conserved among organisms, but also the correlation of the binding energy with the position of the sites. Those charts show that within an organism multiple complex TUs containing sites with similar surroundings have different scores. This result are contrary to our idea that sites farther in average from other sites should have higher scores assuring an effective independent DNA recognition. Our results may be related with site's different surroundings depending not in average distance to other sites but with different TFs binding sites interacting with the given site.

Finally we could also find some correlation between co-occurrence of pairs of sites in the regulatory regions in the organisms under study. This is the first study –at least to our knowledge- reporting these kind of data supported by an statistical analysis. As can be seen in Table 1 closer related organisms have similar co-occurrence patterns in observed number of sites and probability values. This finding, along with the distance study, shows the possibility of making functional extrapolations of knowledge between closer related organisms in a comparative genomic study of transcription regulation.

References

1. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J., Stormo, G. A Comparative Genomics Approach to Prediction of New Members of Regulons. (2001) Genome Research, 11, 566-584.
2. González, AD., Espinosa, V., Vasconcelos, AT., Pérez-Rueda, E., Collado-Vides, J. TRACTOR_DB: a Database of Regulatory Networks in Gamma-Proteobacterial Genomes. (2005) Nucleic Acids Res. **33,** D989-102.
3. Huerta, AM., 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.
4. Bulyk, ML., McGuire, AM., Masuda, N., Church, GM. A Motif Co-Occurrence Approach for Genome-Wide Prediction of Transcription-Factor-Binding Sites in *Escherichia coli.* (2004) Genome Res. 14, 201-208.
5. Collado-Vides, J. Grammatical model of the regulation of gene expression. (1992) PNAS. 89, 9405-9409.
6. Pedersen, H. and Valentin-Hansen, P. Protein-induced fit: the CRP activator protein changes sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. (1997) EMBO Journal. 16, 2108-2118.
7. Collado-Vides, J. Magasanik, B., Gralla, JD. Control Site Location and Transcriptional Regulation in *Escherichia coli.* (1991) Microbiological Reviews. 55, 371-394.