**COMPUTER ANALYSIS OF TRANSCRIPTION REGULATORY PATTERNS IN COMPLETELY SEQUENCED BACTERIAL GENOMES.**

GELFAND M.S.1\*, MIRONOV A.A.2

1Institute of Protein Research, Russian Acad. Sci., Pushchino, 142292, Russia; misha@imb.imb.ac.ru

2State Center of Biotechnology NIIGenetika, Moscow, 113545, Russia; mir@vnigen.msk.su

\*Corresponding author

Keywords: transcription regulatory patterns, bacterial genomes, site recognition, escherichia coli, haemophilus influenzae, purine and arginine regulons

Recognition of transcription regulation sites is one of the most difficult problems of computational molecular biology. In most cases small sample size and low degree of sequence conservation do not allow for construction of reliable recognition rules. We suggest a new approach to this problem based on simultaneous analysis of several related genomes. At that, we assume that groups of genes subject to some specific regulation ("regulons") are evolutionary stable. Thus, in each genome we select genes that have candidate sites in regulatory regions. Then all comparisons between the selected genes are performed and groups of homologous genes are determined. In order to distinguish between paralogs and orthologs, the selected genes from each set are compared with the the total gene complements of the other genomes.

We applied this technique to analysis of purine (PurR), arginine (ArgR) and aromatic amino acid (TrpR and TyrR) regulons of *Escherichia coli* and *Haemophilus influenzae*. Candidate binding sites in regulatory sites of *H.influenzae* were found, a new family of purine transport proteins subject to PurR regulation was described, ArgR regulation of arginine transport was demonstrated, and differences in regulation of some *E.coli* and *H.influenzae* genes were discovered.

**1. Data and Algorithms**

Three regulons were analyzed (the purine/PurR and arginine/ArgR regulons were considered separately, the TrpR and TyrR regulons were combined, as some genes from the aromatic amino acid regulon are subject to regulation by both of these factors). Genes belonging to the *E.coli* regulons were collected from the literature [1] and their orthologs in *H.influenzae* were identified. Known *E.coli* transcription factor binding sites were collected and positional nucleotide weight matrices (profiles) were derived. The positional nucleotide weights are defined by

where N(b,k) is the number of occurences of nucleotide b at position k. Site score is the sum of the respective positional nucleotide weights. The base of the logarithm is chosen so as the standard distribution of the site score on random Bernoulli sequences equals 1.

Candidate sites (PUR, ARG, TRP and TYR boxes) were selected in regions upstream of annotated genes of *E.coli* and *H.influenzae*. Thresholds and region boundaries in each case were selected so as to lose none of the known sites. Sets of potentially co-regulated genes were constructed. They consisted of genes having candidate sites in the upstream regions and genes downstream of those, if they were transcribed in the same direction and the intergenic distances did not exceed some threshold (usually 100 nucleotides).

Pairwise alignment of all genes from the *E.coli* and *H.influenzae* was performed. Pairs of genes having strong similarity were retained for further analysis. This included comparison of the genes with the total gene complements of *E.coli* and *H.influenzae* in order to distinguish orthologs (that can be assumed to have the same role in the cell) from paralogs. Some genes with strong sites, that had known potentially relevant function, were also compared with GenBank and their close homologs were analyzed for the presence of candiadte sites in their upstream regions.

All analysis was performed using the programs DNA-SUN [2] and GENOME (A.M., unpublished).

**2. Results**

**2.1. Transport proteins in the purine and arginine regulons**

Analysis of the PurR regulon resulted in identification of a family of transport proteins that has representatives in *E.coli* and *H.influenzae*, as well as a number of other bacteria. The family consists of two subfamilies. The known members of one subfamily are uracyl and xanthine transporters [3], whereas the other subfamily has no proteins with known specificity. *E.coli* has representativs in both subfamilies, and they happen to form pairs of close paralogs (*yicO* and *yieG*, *yjcD* and *ygfQ/R*, *yicE* and *ygfO*). In each case the first member of a pair has a strong PUR box and thus is likely to be regulated by PurR, whereas the second member has no PUR boxes. All close relatives of the *yicE-ygfO* pair and one more gene with a PUR box, *ygfU*, are H+/purine(xanthine) symporters, and thus purine transport is a very likely function for these genes. The two other pairs, *yicO-yieG* and *yjcD-ygfQ/R*, as well as the *H.influenzae* gene *HI0125*, which is an ortholog of the latter pair, can be ascribed only an unspecified transport function.

In addition, PUR boxes were found upstream of the gene *tsx* encoding outer membrane nucleoside-specific channel in *E.coli*, *Enterobacter aerogenes*, *Klebsiella pneumoniae*, and *Salmonella typhimurium* [4, 4A]*.*

Analysis of the ArgR regulon allowed us to identify ARG boxes upstream of operons encoding arginine-specific ABC transport systems (*artPIQM* and *artJ* from *E.coli*, *HI1180-HI1177* from *H.influenzae*) and thus to place these operons in the arginine regulon.

**2.2. Changes in operon structure with retained regulation**

There are two main types of differences between *E.coli* and *H.influenzae* operons subject to the same regulation. First, genes can be absent in an operon. The gene *HI0811*, which is a candidate member of the *H.influenzae* ArgR regulon, is an ortholog of the last gene of the *E.coli* *argCBH* operon, whereas the first two genes have no orthologs in *H.influenzae*. Similarly, presumably TyrR-regulated gene *HI1290* is an ortholog of *tyrA*, whereas the first gene of the *aroFtyrA* operon of *E.coli* has no orthologs in *H.influenzae*. Finally, *purB* of *E.coli* is an ortholog of *H.influenzae* gene *HI0639*, whereas the PUR box is upstream of the first gene in the operon-like gene string *HI0638-HI0639*.

The second type of changes is breaking of an operon into two parts with retained regulation. Two *E.coli*operons *purHD* and *glyA*, both regulated by PurR, correspond to a single *H.influenzae* gene string *HI0887-HI0889*, and a PUR box is found upstream of *HI0887*.

Both types of differences occur in the tryptophan operon(s) regulated by TrpR and having TRP boxes in upstream regions. There is a single operon on interobacteria (*trpLEDCBA* in *E.coli*, *trpEGDC/FB* in *Vibrio parahaemoliticus*) and two operons in *H.influenzae*: *HI1387-1389.1* (*trpEDDC*) and *HI1430-HI1432* (*ydfGtrpBA*). On the other hand, *HI1430* (*ydfG*) is a hypothetical oxidoreductase that is absent in the *trpBA* operon of *Pasteurella multocida*, a close relative of *H.influenzae*.

**2.3. Changes of regulation**

In some cases regulation patterns seem to be changed. The simplest case is the loss of regulation; the most interesting example of this type is the absence of PUR boxes in the upstream region of *HI1632* (*H.influenzae*ortholog of *purR*). This means that unlike its *E.coli* counterpart [5,6], this gene is not autoregulated. A more subtle case is the change of the regulation mechanism: *purB* is regulated by PurR via the roadblock mechanism [7], which explains an unusual location of the PUR box in the coding region of this gene (around codon 60), whereas the position of PUR box in the corresponding *H.influenzae* operon *HI0638-HI0639* is similar to position of PUR boxes in other operons.

The most interesting situation seems to be that of the unique *H.influenzae* DAPH-synthase (there are three DAPH-synthases in *E.coli* encoded by *aroH*, *aroG* and *aroF* and feedback repressed by tryptophan, phenylalanine and tyrosine, respectively [1]). The gene *HI1547* is an ortholog of *aroG* and thus encodes DAPH-synthase-PHE (E.Koonin, personal communication). However, unlike *aroG*, regulated by TyrR (with phenyalanine and tryptophan acting as co-repressors), it has a TRP box, but no TYR boxes, similarly to the tryptophan-regulated gene *aroH* coding for DAPH-synthase-TRP. Thus either the regulation of this gene has changed, or a very subtle non-orthologous displacement has taken place [8]. There seem to be no computational way for resolving this ambiguity, that thus should be subject to experimental analysis.

**3. Discussion**

Computer analysis was used for prediction of bacterial transcription signals for more than 15 years (reviewed, in particular, in [9]), and in many cases it served as a basis for further experimental work (e.g. [10]). However, this study represents the first attempt to completely characterize regulons in newly sequenced genomes using large-scale genomic comparison.

There are three main components in our approach: prediction of transription factor sites, analysis of protein homologies, and consideration of protein function. The use of complete genomes allows us to identify orthologs, and thus to use sequence similarity to make conclusions about similar cellular role of proteins. However, a good supplement to our technique is analysis of homologous genes in all related bacterial species using similarity search in GenBank. Thus the approach is flexible, yet sufficiently robust to make non-trivial predictions even when the operon structure and regulatory interactions are not stable (cf. [11]).

An important prerequisite for this type of analysis is conservation of the regulatory protein itself. Thus, there are no strong PUR boxes in the *Helicobacter pylori* genome that does not contain a gene for PurR. Similarly, although there is a purine repressor in *Bacillus subtilis*, it is unrelated to PurR of *E.coli* and indeed, the type of regulation (mostly by attenuation) and regulatory sites (in a few genes regulated on the transcription level) of the *B.subtilis* purine regulon differ from those of *E.coli*. On the other hand, if the regulatory protein is conserved, the regulatory signals tend to be conserved as well. There are only three known genes in the arginine regulon of *H.influenzae*, including the repressor ArgR itself (not counting the transport proteins predicted to belong to the arginine regulon in this work), but the ARG boxes are conserved. Preliminary results show that *E.coli* ARG box recognition matrix can recover the relevant signals even in the distantly related *B.subtilis* genome.

This study allowed us to make a number of interesting predictions that can be checked by rather simple experimental techniques. One group of such predictions includes inferences about change of regulation patterns: loss of autoregulation in the *H.influenzae* homolog of *PurR*, different mode of repression of *purB*, and change of regulation of *aroG*. The second group is formed by predictions that extend the existing purine and arginine regulons both in *E.coli* and *H.influenzae* by inclusion of transport proteins (purine and arginine transporters). It is somewhat surprizing that these systems, especially the large family of H+/purine symporters, were not detected by genetic analysis. A possible explanation is that all PurR-regulated genes from this family have close non-regulated paralogs, and thus the influence of mutation in these genes would be weak or expressed in very specific conditions.

Our further plans involve analysis of global regulatory systems (SOS, CRP, Fur, Fnr regulons) and multiple interacting systems (e.g. interaction of purine/pyrimidine regulation), comparisons of with more distant genomes (in particular, *E.coli* and *B.subtilis*), development of multiple local alignment / signal definition algorithms that would allow to analyze functionally related regulons with non-homologous regulators, more detailed analysis of interaction between proteins and their binding sites from both structural and evolutionary point of view, and, as a distant goal, development of techniques for automated characterization of regulatory pathways in newly sequenced genomes.

**Acknowledgements**

This work was partially supported by grants from the Russian Fund of Fundamental Research and the US Department of Energy. We are grateful to Mikhail Roytberg and Eugene Koonin for discussions and assistance.

**References**

1. F.C. Neidhardt, Ed. “Escherichia coli and Salmonella. Cellular and Molecular Biology“ (ASM Press, Washington, 1996)
2. A.A. Mironov et al., Comput. Appl. Biosci. 11 (1995)
3. G. Diallinas, L. Gorfinkel, H.N. Arst Jr., G. Cecchetto, C. Scazzocchio. J. Biol. Chem. 270, 8610-8622 (1995)
4. E. Bremer, A. Middendorf, J. Martinussen., P. Valentin-Hansen. Gene 96, 59-65 (1990)
5. 4A. A. Nieweg, E. Bremer. Microbiology 143, 603-615 (1997)
6. R.F. Rolfes, H. Zalkin. J. Bacteriol. 172, 57585766 (1990)
7. L.M. Meng, P. Nygaard. Mol. Microbiol. 4, 2187-2192 (1990)
8. B. He, H. Zalkin. J. Bacteriol. 174, 7121-2127 (1992)
9. E.V. Koonin, A.R. Mushegian, P. Bork. Trends in Genetics 12, 334-336 (1996)
10. M.S. Gelfand. J. Comput. Biol. 2, 87-117 (1995)
11. B. He, K.Y. Choi, H. Zalkin. J. Bacteriol. 175, 3598-3606 (1993)
12. M.Y. Galperin, E.V. Koonin. In Silico Biol. 1, 0007 (1998) <<http://www.bioinfo.de/isb/1998/01/007/>>