 Nucleic Acids Research, 1999, Vol. 27, No. 14 **2981–2989**

# Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes

**A. A. Mironov, E. V. Koonin[1], M. A. Roytberg[2] and M. S. Gelfand[3,*]**

State Center of Biotechnology, NIIGenetika, Moscow 113545, Russia, [1]National Center of Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Institute of Mathematical Problems of Biology, Pushchino 142292, Russia and [3]Institute of Protein Research, Russian Academy of Science, Pushchino 142292, Russia

Received November 18, 1998; Revised March 12, 1999; Accepted May 26, 1999

## ABSTRACT

**Recognition of transcription regulation sites (operators) is a hard problem in computational molecular biology. In most cases, small sample size and low degree of sequence conservation preclude the construction of reliable recognition rules. We suggest an approach to this problem based on simultaneous analysis of several related genomes. It appears that as long as a gene coding for a transcription regulator is conserved in the compared bacterial genomes, the regulation of the respective group of genes (regulons) also tends to be maintained. Thus a gene can be confidently predicted to belong to a particular regulon in case not only itself, but also its orthologs in other genomes have candidate operators in the regulatory regions. This provides for a greater sensitivity of operator identification as even relatively weak signals are likely to be functionally relevant when conserved. We use this approach to analyze the purine (PurR), arginine (ArgR) and aromatic amino acid (TrpR and TyrR) regulons of *Escherichia coli* and *Haemophilus influenzae*. Candidate binding sites in regulatory regions of the respective *H.influenzae* genes are identified, a new family of purine transport proteins predicted to belong to the PurR regulon is described, and probable regulation of arginine transport by ArgR is demonstrated. Differences in the regulation of some orthologous genes in *E.coli* and *H.influenzae*, in particular the apparent lack of the autoregulation of the purine repressor gene in *H.influenzae*, are demonstrated.**

## INTRODUCTION

With the sequencing of multiple complete bacterial and archaeal genomes, computational biology entered a new era. The availability of the sequences of all genes in several prokaryotic species created the opportunity of perceiving the relationships between prokaryotic genomes in a comprehensive and precise fashion, which was unattainable previously. Initially, the main efforts have been directed at large-scale comparison of proteomes with the aim of reconstructing the metabolism and other cellular functions in poorly characterized organisms and clarifying distant evolutionary relationships, particularly those between the three primary divisions of life—bacteria, archaea and eukaryotes (1–4). One unexpected result that has become immediately obvious was the lack of long-range conservation of the gene order in bacterial genomes, with the exception of species within the same genus (5–7). In fact, in distantly related bacteria, such as, for example, Proteobacteria and Cyanobacteria, there are only a few conserved operons that encode primarily, if not exclusively, genes whose products physically interact (8). At intermediate phylogenetic distances, however, for example in *Escherichia coli* and *Haemophilus influenzae*, a large number of operons are conserved, although their order is not (8,9).

An important further step in the functional annotation of genomes is the identification of regulatory signals, particularly binding sites for transcription factors. Although the problem of prediction of regulatory sites had been addressed for over 15 years (reviewed in 10), it is still far from being solved (11). One reason for this is that the learning sample rarely contains more than 20–30 sites. However, even for large samples, it proved to be extremely difficult to construct a good recognition rule. The physics of protein–DNA interaction is poorly understood, making it virtually impossible to derive a proper set of features for statistical or pattern recognition algorithms. Furthermore, the latter type of algorithms cannot take into account context effects, in particular, interactions between different regulatory sites, and structural properties of DNA. Nevertheless, in many cases, simple profile methods perform reasonably well, in the sense that they can correctly identify true sites if the number of alternatives is not too large (for benchmarking of several most popular algorithms; 12).

Good results in computer-assisted functional annotation of nucleotide sequences frequently have been obtained by combination of statistical analysis of DNA and comparative analysis of the protein sequences encoded by the respective genes. To a varying extent, this approach is used in the analysis of all genomic sequences. In more systematic efforts, it was employed in the construction of reliable gene recognition algorithms (13–15) and in the prediction of the specificity of

*To whom correspondence should be addressed at present address: State Center of Biotechnology, NIIGenetika, Moscow 113545, Russia. Tel: +7 095 948 82 19; Fax: +7 095 315 05 01; Email: misha@imb.imb.ac.ru

new restriction-modification systems (16). Here we apply this methodology to the analysis of bacterial transcription regulation in the context of a comparison of complete genomes.

The approach is based on the assumption that groups of genes subject to a specific mode of regulation (regulons) are at least partially conserved in evolution. This assumption generally seems to hold provided that the cognate regulatory factor is present in all compared genomes. Preliminary analyses have shown that in these cases, the regulatory signal also is conserved, and accordingly, a recognition rule derived for the most thoroughly studied genome can also be applied to other genomes (17). Under this approach, the assignment of a gene to a particular regulon is reinforced if not only this gene itself but also its orthologs in other genomes have candidate regulatory sites in the appropriate regions.

We applied this comparative approach to the analysis of purine, arginine, and aromatic amino acid regulons in *E.coli* and *H.influenzae*. Among the completely sequenced genomes, this is a natural choice for the first attempt of such a study since, first, *E.coli* gene regulation is by far the best understood among all bacteria, and second, *H.influenzae* is the only complete bacterial genome that is close enough to *E.coli* so that many operons are conserved but distant enough for significant differences to be apparent. Recognition rules derived from samples of known *E.coli* regulatory sites were used to predict sites in the *H.influenzae* genome and to detect likely new members of the three regulons in both species. We describe the general conservation of the three regulons in *E.coli* and *H.influenzae* along with differences in the regulation of some of the orthologous genes.

## MATERIALS AND METHODS

Complete genome sequences of *E.coli* (18) and *H.influenzae* (19) as well as partial sequences from other Proteobacteria were extracted from GenBank.

Three regulons were analyzed; the purine regulon (set of genes regulated by PurR) (20) and the arginine regulon (regulated by ArgR) (21) were considered separately, whereas the genes controlled by TrpR and TyrR were considered to comprise one aromatic amino acid regulon since some of them are subject to regulation by both factors (22). Known *E.coli* transcription factor binding sites were collected from the literature (20–23). Each site was considered in the orientation that corresponds to the coding strand of the regulated operon. Positional nucleotide weight matrices (profiles) were derived using the following formula for positional nucleotide weights:

$$W(b,k) = \log[N(b,k) + 0.5] - 0.25 \sum_{i = A,C,G,T} \log[N(i,k) + 0.5]$$

where $N(b,k)$ is the count of nucleotide $b$ in position $k$. The site score is the sum of the respective positional nucleotide weights. The base of the logarithm was chosen such that the standard deviation of the site score distribution on random oligomers equals 1. The site score defined by this formula is linearly related to the discrimination energy used in a number of other papers.

Candidate sites (PUR, ARG, TRP and TYR boxes) were identified in upstream regions of annotated *E.coli* and *H.influenzae* genes, including predicted ones. Thresholds and region boundaries in each case were selected so that none of the known sites were missed. Sets of potentially co-regulated genes were constructed from genes that have candidate regulatory sites in their upstream regions and genes that are located downstream of them if they are transcribed in the same direction and the intergenic distances do not exceed certain threshold (normally 100 nucleotides).

Orthologous genes in *E.coli* and *H.influenzae* were identified by comparing the complete sets of protein sequences from the two species using the gapped BLASTP program or the Smith–Waterman algorithm as implemented in the GENOME program (A.A.Mironov, unpublished), selecting pairs of proteins with the greatest similarity to each other and checking for the conservation of domain architecture (6,24). The upstream regions of genes that are orthologous to genes containing regulatory sites were examined for candidate sites, even if these were not detected automatically. Site recognition was performed using the DNA-SUN (25) and GENOME programs (A.A.Mironov, unpublished). The non-redundant protein and nucleotide databases at the NCBI were searched using the gapped BLAST programs (26). Multiple sequence alignments were constructed using the CLUSTALX program (27). Phylogenetic trees were constructed using the PHYLIP package programs NEIGHBOR (the neighbor-joining method) and PROTPARS (maximum parsimony method) (28). Sequence logos were constructed using the MAKELOGO program (29) as implemented on the WorldWide Web by Stephen E. Brenner (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi ).

## RESULTS

### Identification of candidate regulator-binding sites

Sequence logos for the PUR, ARG, TRP and TYR boxes are shown in Figure 1. The boxes vary strongly in terms of information content, with the PUR and TRP boxes being stronger, and the ARG and TYR boxes being weaker. The latter sites are often present in a regulatory zone of a gene in several copies that are recognized co-operatively. The recognition weight matrices are shown in Table 1.

The distributions of candidate site scores for the four boxes are shown in Figure 2. Scores of the sites from the learning sample and their positions relative to the gene starts are given in Table 2. Comparison of this table with Figure 2 shows that, even for strong signals with a relatively large learning sample (the PUR box), the use of a statistical recognition rule is not sufficient to reliably predict operators.

We attempted to take into account co-operative binding of ArgR to tandemly repeated ARG boxes. A procedure that searched for pairs of ARG boxes performed quite well in the sense that it clearly separated all sites from the learning sample from all other sequences (data are not shown). However, since ArgR can bind to single ARG boxes, albeit with a low specificity (30), we used the single box recognizer for further analysis.

### Evolution of regulons

*The purine regulon. Haemophilus influenzae* retains the regulation of the PurR regulon genes directly responsible for purine biosynthesis, and the structure of the operons *purEK*, *cvpApurF*, *purC*, *purMN*, *purL* is the same in *E.coli* and *H.influenzae* (Table 3). Other genes of the core regulon also retain the regulation, although with some modifications (see

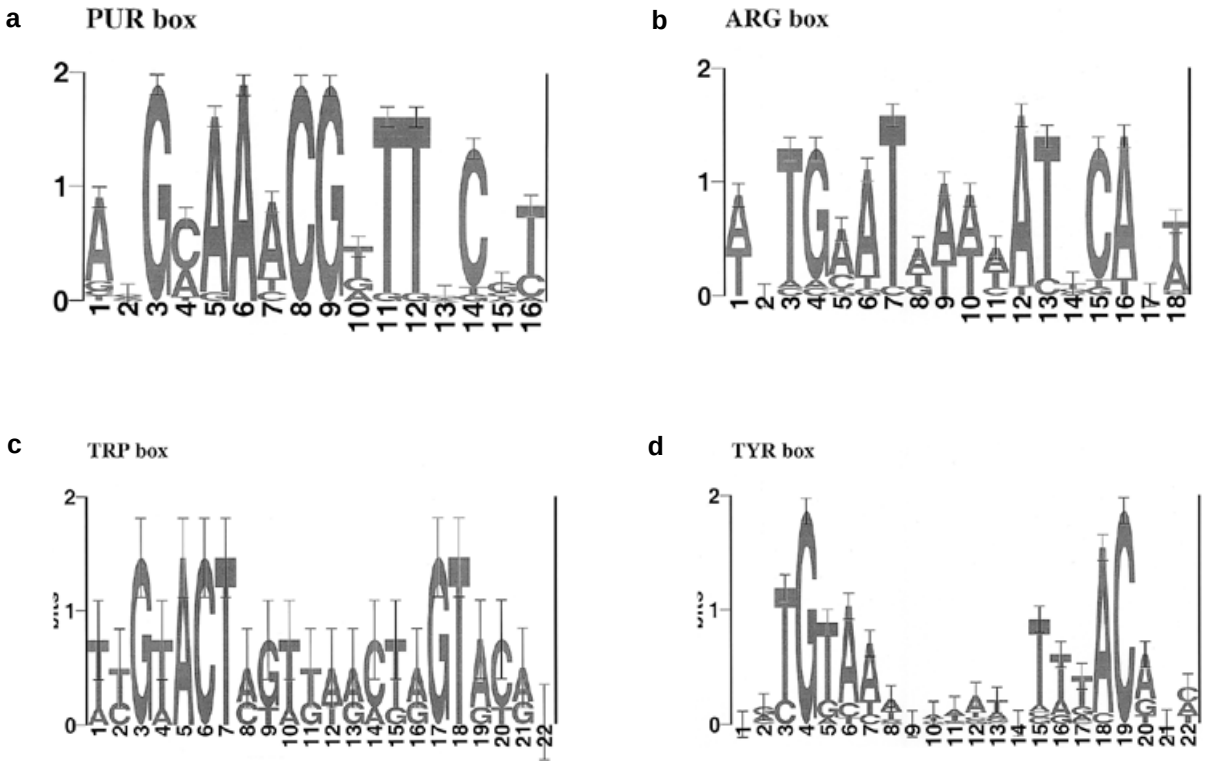


**Figure 1.** Sequence logos for the PUR, ARG, TRP and TYR boxes. Horizontal axis, position in the binding site; vertical axis, information in bits. The height of each stack of letters is proportional to the positional information content in the given position; the height of each individual letter reflects its prevalence in the given position. The logos were constructed from the aligned sequences of the known *E.coli* regulatory sites (Table 2).

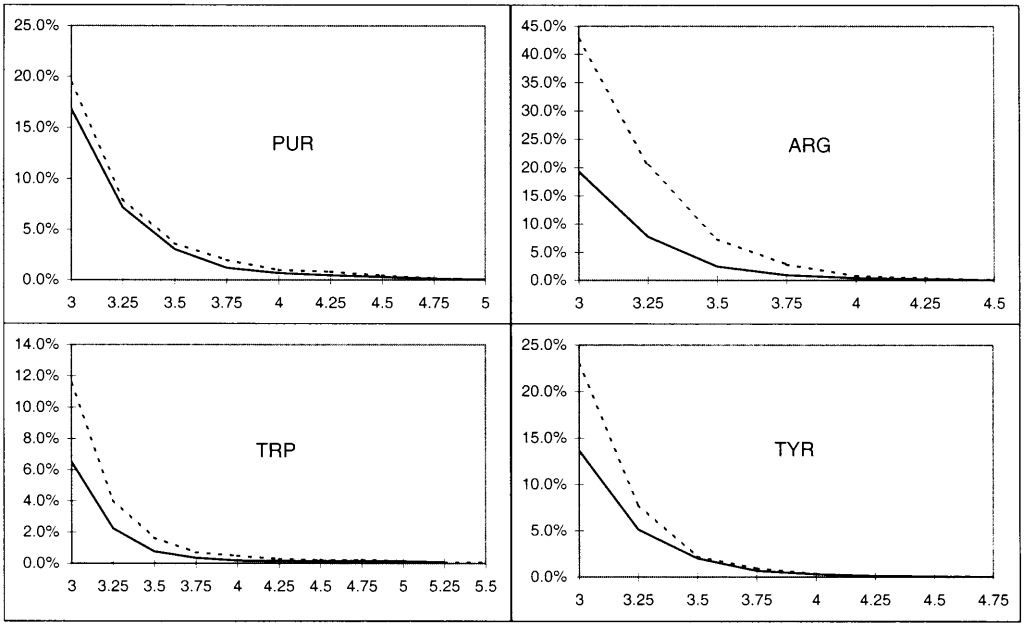


**Figure 2.** Histograms and distribution functions of candidate site scores for PUR, ARG, TRP and TYR boxes. Horizontal axis, score; vertical axis, percentage of genes whose candidate binding sites (highest scoring sites in upstream non-coding regions) for the given regulatory factor have a score greater than the respective value. Solid curves, *E.coli*; broken curves, *H.influenzae*.

**Table 1.** Positional nucleotide weight matrices (profiles) for PUR, ARG, TRP and TYR boxes[a]

| A | C | G | T | Cns[b] | A | C | G | T | Cns[b] |
|---|---|---|---|---|---|---|---|---|---|
| | | **PUR** | | | | | **ARG** | | |
| | | | | | 0.23 | -0.24 | -0.24 | 0.25 | W |
| 0.28 | -0.12 | -0.04 | -0.12 | A | 0.05 | -0.17 | 0.08 | 0.05 | N |
| 0.01 | 0.10 | 0.01 | -0.12 | C | -0.05 | -0.05 | -0.23 | 0.34 | T |
| -0.16 | -0.16 | 0.47 | -0.16 | G | -0.05 | -0.05 | 0.34 | -0.23 | G |
| 0.15 | 0.25 | -0.29 | -0.11 | C | 0.29 | 0.06 | -0.08 | -0.26 | A |
| 0.41 | -0.20 | -0.02 | -0.20 | A | 0.31 | -0.25 | -0.07 | 0.02 | A |
| 0.47 | -0.16 | -0.16 | -0.16 | A | -0.19 | -0.01 | -0.19 | 0.39 | T |
| 0.29 | -0.03 | -0.29 | 0.03 | A | 0.21 | -0.30 | 0.02 | 0.07 | A |
| -0.16 | 0.47 | -0.16 | -0.16 | C | 0.30 | -0.23 | -0.23 | 0.17 | A |
| -0.16 | -0.16 | 0.47 | -0.16 | G | 0.25 | -0.24 | -0.24 | 0.23 | W |
| 0.04 | -0.32 | 0.07 | 0.21 | T | 0.19 | 0.02 | -0.30 | 0.10 | A |
| -0.20 | -0.20 | -0.02 | 0.41 | T | 0.39 | -0.19 | -0.19 | -0.01 | A |
| -0.20 | -0.20 | -0.02 | 0.41 | T | -0.21 | 0.06 | -0.21 | 0.36 | T |
| -0.02 | -0.12 | 0.08 | 0.06 | G | -0.04 | -0.09 | 0.04 | 0.09 | N |
| -0.24 | 0.36 | -0.06 | -0.06 | C | -0.19 | 0.39 | -0.19 | -0.01 | C |
| -0.11 | -0.01 | 0.14 | -0.01 | G | 0.36 | -0.21 | -0.21 | 0.06 | A |
| -0.10 | 0.11 | -0.29 | 0.28 | T | -0.05 | 0.00 | 0.00 | 0.06 | N |
| | | | | | 0.15 | -0.10 | -0.28 | 0.23 | T |
| | | **TRP** | | | | | **TYR** | | |
| 0.05 | -0.16 | -0.16 | 0.27 | T | 0.07 | -0.11 | -0.04 | 0.07 | N |
| -0.18 | 0.14 | -0.18 | 0.21 | T | 0.06 | -0.02 | 0.14 | -0.18 | G |
| -0.12 | -0.12 | 0.36 | -0.12 | G | -0.25 | 0.12 | -0.25 | 0.38 | T |
| 0.05 | -0.16 | -0.16 | 0.27 | T | -0.17 | -0.17 | 0.50 | -0.17 | G |
| 0.36 | -0.12 | -0.12 | -0.12 | A | -0.09 | -0.30 | 0.07 | 0.32 | T |
| -0.12 | 0.36 | -0.12 | -0.12 | C | 0.35 | 0.02 | -0.29 | -0.08 | A |
| -0.12 | -0.12 | -0.12 | 0.36 | T | 0.29 | -0.02 | -0.32 | 0.05 | A |
| 0.21 | 0.14 | -0.18 | -0.18 | A | 0.15 | -0.08 | -0.17 | 0.10 | A |
| -0.16 | -0.16 | 0.27 | 0.05 | G | -0.05 | -0.05 | 0.00 | 0.10 | T |
| 0.05 | -0.16 | -0.16 | 0.27 | T | 0.11 | -0.10 | -0.10 | 0.08 | A |
| -0.18 | -0.18 | 0.14 | 0.21 | T | 0.12 | -0.03 | -0.18 | 0.09 | A |
| 0.21 | -0.18 | -0.18 | 0.14 | A | 0.19 | -0.08 | -0.08 | -0.02 | A |
| 0.21 | -0.18 | 0.14 | -0.18 | A | 0.03 | -0.18 | -0.02 | 0.17 | T |
| 0.05 | 0.27 | -0.16 | -0.16 | C | 0.07 | 0.03 | -0.05 | -0.05 | N |
| -0.16 | -0.16 | 0.05 | 0.27 | T | -0.11 | -0.11 | -0.11 | 0.32 | T |
| 0.21 | -0.18 | 0.14 | -0.18 | A | 0.09 | -0.33 | -0.02 | 0.26 | T |
| -0.12 | -0.12 | 0.36 | -0.12 | G | -0.06 | -0.16 | 0.00 | 0.22 | T |
| -0.12 | -0.12 | -0.12 | 0.36 | T | 0.44 | -0.01 | -0.22 | -0.22 | A |
| 0.27 | -0.16 | 0.05 | -0.16 | A | -0.17 | 0.50 | -0.17 | -0.17 | C |
| -0.16 | 0.27 | -0.16 | 0.05 | C | 0.26 | -0.33 | 0.09 | -0.02 | A |
| 0.21 | -0.18 | 0.14 | -0.18 | A | -0.01 | 0.06 | -0.01 | -0.05 | N |
| 0.08 | -0.03 | -0.03 | -0.03 | N | 0.13 | 0.16 | -0.35 | 0.06 | C |

[a]Each column shows the weights of the given nucleotide in the consecutive positions of the respective binding site.
[b]Cns, Consensus derived for each position by majority rule.

**Table 2.** Scores and positions relative to the gene start of sites from the learning samples

| Operon | Regulator | binding site(s) | score | pos. |
|---|---|---|---|---|
| Purine regulon | | | | |
| *purR* | PurR | AgGCAAACGTTTaCcT | 4.76 | -61 |
| | | gaGCAAACGTTTcCac | 4.07 | +27 |
| *purEK* | PurR | ACGCAAcCGTTTTCcT | 4.61 | -86 |
| *cvpApurF* | PurR | ACGCAAACGTTTTCtT | 4.93 | -71 |
| *purC* | PurR | ACGCAAACGTgTGCGT | 4.67 | -168 |
| *purMN* | PurR | tCGCAAACGTTTGCtT | 4.55 | -80 |
| *purL* | PurR | ACGCAAACGgTTTCGT | 4.94 | -91 |
| *purB* | PurR | ACGCAAtCGgTTaCcT | 4.45 | +185 |
| *guaBA* | PurR | AtGCAAtCGgTTaCGc | 4.21 | -68 |
| *purHD* | PurR | gCGCAAACGTTTTCGT | 4.76 | -122 |
| *glyA* | PurR | AgGtAAtCGTTTGCGT | 4.30 | -133 |
| *pyrD* | PurR | cgGaAAACGTTTGCGT | 4.51 | -103 |
| *prsA* | PurR | AaGaAAACGTTTTCGc | 4.72 | -357 |
| *glnB* | PurR | AtGCAAACGaTTTCaa | 4.06 | -82 |
| *purA* | PurR | AaGCAAACGgTgattT | 3.77 | -23 |
| | | AgGaAAACGaTTGgcT | 4.17 | -122 |
| *codBA* | PurR | ACGaAAACGaTTGCtT | 4.68 | -83 |
| *pyrC* | PurR | AgGaAAACGTTTcCGc | 4.54 | -66 |
| *purT* | PurR | ACGCAAACGTTTTCGT | 5.08 | -54 |
| *gcvTHP* | PurR | AaGagAACGaTTGCGT | 4.31 | -105 |
| *speAB* | PurR | AaGaAAcCGgTTGCGc | 4.28 | -133 |
| Arginine regulon | | | | |
| *argR* | ArgR | TtTGcATAAAAATTCATc | 4.24 | -63 |
| | | TaTGcAcAAtAATgttgT | 3.34 | -43 |
| *argA* | ArgR | AcaGAATAAAAATaCAcT | 3.98 | -50 |
| | | TtcGAATAAtcATgCAaa | 3.98 | -39 |
| *argCBH* | ArgR | TaTcAATAttcATgCAgT | 4.61 | -128 |
| | | TaTGAATAAAAATaCAcT | 4.61 | -109 |
| *argD* | ArgR | AGTGAtTttttATgCATa | 4.01 | -68 |
| | | TGTGgtTAtAAtTTCAca | 3.50 | -47 |
| *argE* | ArgR | AGTGtATttttATTCATa | 3.80 | -64 |
| | | AcTGcATgAAtATTgATa | 3.39 | -43 |
| *argF* | ArgR | AaTGAATAAttAcaCATa | 4.16 | -65 |
| | | AGTGAATtttAATTCAaT | 4.41 | -44 |
| *argG* | ArgR | TGTGAATgAAtATcCAgT | 4.31 | -210 |
| | | AtTaAATgAAAAcTCATT | 3.90 | -91 |
| | | TtTGcATAAAAATTCAgT | 4.51 | -70 |
| *argI* | ArgR | AaTGAATAAtcATcCATa | 4.33 | -63 |
| | | AtTGAATtttAATTCATT | 4.49 | -42 |
| *carAB* | ArgR | TGTGAATtAAtATgCAaa | 4.36 | -50 |
| | | AGTGAgTgAAtATTCtcT | 3.79 | -39 |
| Aromatic amino acid regulon | | | | |
| *trpR* | TrpR | TcGTACTctTTAgCgAGTACAA | 5.22 | -68 |
| *trpLEDCBA* | TrpR | TcGaACTAGTTAACTAGTACgc | 5.40 | -47 |
| *aroH* | TrpR | aTGTACTAGagAACTAGTGCAt | 5.03 | -166 |
| *mtr* | TrpR | TTGTACTcGTgtACTgGTACAg | 5.48 | -72 |
| | TyrR | TcTGTAAAatAATATaTACAgC | 4.60 | -129 |
| | TyrR | TGcGTAAtcAtcgcTgaACAgC | 3.56 | -159 |
| *aroLM* | TrpR | TTGTACTAGTTtgaTgGTAtgA | 5.15 | -81 |
| | TyrR | AGTGgAAtTtttTcTTTACAat | 4.40 | -108 |
| | TyrR | gGTGTAttgAgATTTTcACttt | 3.39 | -131 |
| | TyrR | AaTGTAAtTtAtTATTTACACt | 4.71 | -184 |
| *tyrR* | TyrR | TGTGTcAATgAtTgTTgACAga | 3.99 | -91 |
| | | gcTGaccggAtATcTTTACgCc | 3.20 | -135 |
| *tyrB* | TyrR | cGTGTttcaAAAagTTgACgCC | 3.30 | -2 |
| | | cccGTAAAcctggAgaacCAtC | <3 | |
| *aroG* | TyrR | AGTGTAAAaccccgTTTACACa | 3.90 | -91 |
| *aroFtyrA* | TyrR | TGTGTAAATAAAaATgTACgaa | 4.81 | -165 |
| | | AGTGTAAATTtATcTaTACAga | 4.49 | -113 |
| | | TaTGgAttgAAAacTTTACttt | 3.79 | -90 |
| *aroP* | TyrR | AacGgAAtTgcAaacTTACACa | 3.58 | -66 |
| | | gaTGTAAAcAAATtaaTACAaC | 4.05 | -89 |
| *tyrP* | TyrR | TaTGTAAcgtcggTTTgACgaa | 3.52 | -89 |
| | | AtTGTAcATttATATTTACACC | 4.49 | -112 |

below). Orthologs of several genes that in *E.coli* belong to the purine regulon, namely *codBA*, *pyrC*, *gcvTHP*, *speAB*, *purT* are missing in *H.influenzae*. Of these genes, only *purT* is directly involved in purine synthesis, but its function is redundant with that of *purN* (20). Finally and most interestingly, orthologs of some genes of the *E.coli* PurR regulon, namely *pyrD*, *prsA*, *glnB*, *purA* and *purR* itself, are present in *H.influenzae* but apparently have lost the PurR regulation. [The regulation of *E.coli purA* by PurR binding to the two rather weak PUR boxes in its upstream regions is in fact questionable (31,32).]. The *E.coli purR* gene is autoregulated through two PUR boxes. However, no sequence resembling a PUR box can be found upstream of *purR* in *H.influenzae*, and it seems that direct autoregulation in this case can be ruled out.

Several operons of the purine regulon have different gene organization and/or mode of regulation in *E.coli* and *H.influenzae*. 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* (Fig. 3a). Thus these three *H.influenzae* genes are confidently predicted to form a single PurR-regulated operon. The *E.coli purB* gene is the ortholog of the *H.influenzae* gene *HI0639*. In *E.coli*, this gene is regulated by PurR via the roadblock mechanism (33),

**Table 3.** *Haemophilus influenzae* operons predicted to belong to the purine regulon

| Operon | PurR-binding site | score | pos. |
|---|---|---|---|
| *HI1615-HI1616 (purEK)* | taGCAAACGTTTGCcT | 4.46 | -77 |
| *HI1206-HI1207 (cvpApurF)* | ACGCAAACGTTTTCtT | 4.93 | -84 |
| *HI1726 (purC)* | taGCAAACGTTTGCtT | 4.46 | -31 |
| *HI1429-HI1428 (purMN)* | tCGCAAACGTTTGCtT | 4.55 | -62 |
| *HI0752 (purL)* | AtGCAAACGTTTGCtT | 4.73 | -1 |
| *HI0638-9 (ycfCpurB)* | ACGgAAACGTTTTCcT | 4.39 | -28 |
| *HI0887-9 (purHDglyA)* | AaGCAAACGTTTGCGT | 5.01 | -65 |

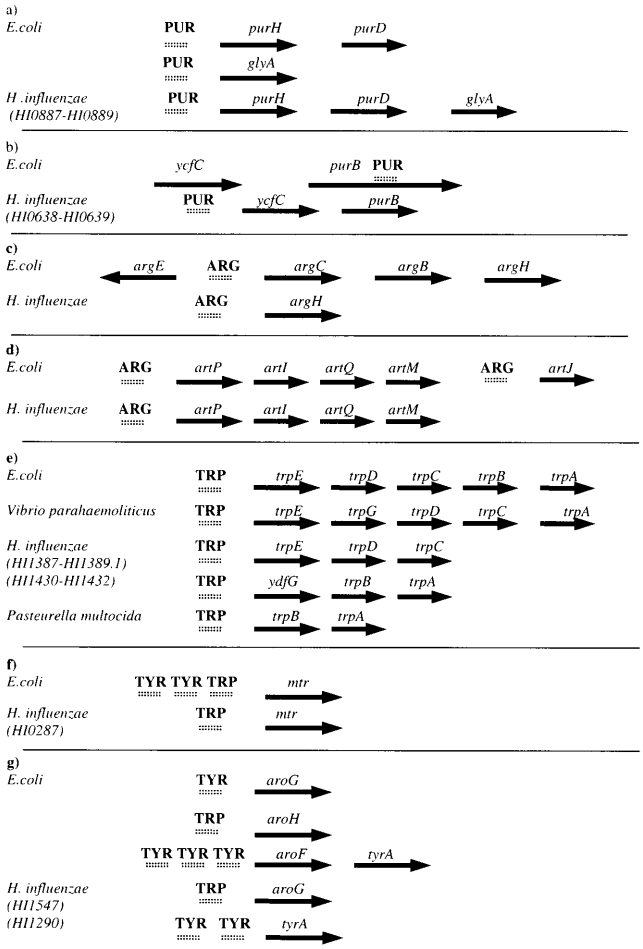


**Figure 3.** Some Proteobacterial operons with variations in gene organization and/or mode of regulation. (**a**) The purine regulon, the *purHD* operon. (**b**) The purine regulon, the *purB* operon. (**c**) The arginine regulon, the *argECDH* operon. (**d**) The arginine regulon, the art operon. (**e**) The aromatic amino acid regulon, the *trp* operon. (**f**) The aromatic amino acid regulon, the *mtr* operon. (**g**) The aromatic amino acid regulon, the *aroF,G,H* operons The candidate binding sites are indicated by a double dotted line.

which explains an unusual location of the PUR box within the coding region (around codon 60). In *H.influenzae* the PUR box is found upstream of the first gene in the operon-like gene string *HI0638–HI0639*. Notably, *HI0638* is the ortholog of the uncharacterized *E.coli* gene *ycfC*, which is located upstream of *purB* (Fig. 3b).

**Table 4.** *Haemophilus influenzae* operons predicted to belong to the arginine regulon

| Operon | ArgR-binding site | score | pos. |
|---|---|---|---|
| *HI1209 (argR)* | AGTGAATttttATgCAaT | 4.27 | -50 |
| *HI0811 (argH)* | TaTGAATAAAtATgCAca | 4.52 | -54 |
| *HI1727 (argG)* | AtaGAATttttATTCAca | 3.87 | -64 |
| | AtcGAtTAtttATTCAaT | 3.75 | -43 |

The regulation status of the *guaBA* (*HI0221–HI0222*) operon of *H.influenzae* is unclear since the only candidate PUR box is within the second gene of the operon in position (+260) and is weak (score = 3.90). Although it could be another case of a distinct regulation mechanism, it is more likely that this operon is not regulated by PurR.

*The arginine regulon.* Of this *E.coli* regulon, *H.influenzae* retains only the repressor and two genes, namely *argG* and *argH*, which encode enzymes that catalyze the conversion of citrulline into arginine (Table 4). Orthologs of the other genes of the *argCBH* operon, as well as the single-gene operon *argE* (that in *E.coli* is transcribed in the opposite direction and is regulated by the same operator), are all missing in *H.influenzae* (Fig. 3c). The *argR* and *argH* genes of *H.influenzae* have single ARG boxes, and thus the regulatory effect is predicted to be weak.

*The aromatic amino acid regulon.* This case is the most complicated, and the analysis has been supplemented by consideration of the available genome fragments from other Proteobacteria. The autoregulation is conserved for the orthologs of *trpR* and *tyrR* genes in *H.influenzae*, as well as for *trpR* of *Enterobacter cloacae* and *Salmonella typhimurium* (Table 5) and *tyrR* of *Citrobacter braakii* (Table 6). The main tryptophan operon *trpLEDCBA* is conserved in the enterobacterium *Vibrio parahaemoliticus* but is broken into two parts in *H.influenzae*. The first part, which includes the *HI1430–HI1432* genes (orthologs of *E.coli ydfG-trpBA*), contains an additional gene *ydfG*, which encodes a predicted oxidoreductase. This gene may be a relatively recent addition to the operon since it is not present in the *trpBA* operon of the closely related species *Pasteurella multocida* (Table 5; Fig. 3e). In *Pseudomonas aeruginosa*, the *trpBA* operon is regulated by an unrelated transcription factor *trpI*, and accordingly, no TRP boxes are found upstream of this operon.

In *E.coli*, the *aroLM* and *mtr* operons are regulated by both TrpR and TyrR. There are no orthologs of *aroL* and *aroM* genes in *H.influenzae*; the ortholog of the *mtr* gene has only the TRP box (Fig. 3f). Other operons that have no orthologs in *H.influenzae* are *tyrB* and *aroP*. By contrast, *H.influenzae* has two paralogous *tyrP* genes (*HI0477* and *HI0528*). The former has three candidate TYR boxes, whereas the latter has only one; the single *E.coli tyrP* gene has two binding sites for TyrR.

The most interesting case is that of the unique *H.influenzae* 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAPH) synthase. There are three DAPH-synthases in *E.coli*, which are encoded by *aroH*, *aroG* and *aroF* and feedback-inhibited by tryptophan, phenylalanine and tyrosine, respectively (34). The gene *HI1547* is confidently identified as the ortholog of *aroG* (data not shown) and thus is predicted to encode DAPH-synthase-

**Table 5.** Operons of various bacteria predicted to be regulated by TrpR

| Operon | Sp[a] | TrpR-binding site | score | pos. |
|---|---|---|---|---|
| *HI0830 (trpR)* | Hin | aTGcACTAGTTtAaTAGTgtAA | 4.49 | -38 |
| *trpR* | Ecl | cTGTACTcGTTAAtgAGTACAA | 4.72 | -36 |
| *trpR* | Sty | TcGTACTctTTAgCgAGTACAA | 5.22 | -68 |
| *HI1387-89.1 (trpEDDC)* | Hin | TTGcACTAGTTtAaTAGTACAA | 5.15 | -47 |
| *HI1430-32 (ydfGtrpBA)* | Hin | TTGTACTAcTTtAaTAGTACAA | 5.15 | -48 |
| *trpEGDC/FB* | Vpa | TcGcACTAGTTAACTAGTACAc | 5.26 | -53 |
| *trpBA* | Pmu | cTGaACTAGTTtAaTAGTtCAA | 4.50 | -95 |
| *HI1547 (aroG)* | Hin | TcGaACTAGTTtACTAGTACAA | 5.51 | -81 |
| *HI0287 (mtr)* | Hin | gTGTACTAcTatAaTAGTgCAA | 4.11 | -43 |

[a]Species: Hin, *H.influenzae*; Ecl, *E.cloacae*; Sty, *S.typhimurium*; Vpa, *V.parahaemoliticus*; Pmu, *P.multocida*.

**Table 6.** Operons of various bacteria predicted to be regulated by TyrR

| Operon | Sp[a] | TyrR-binding site | score | pos. |
|---|---|---|---|---|
| *HI0410 (tyrR)* | Hin | taTGTAAAatAATATTTACACt | 4.82 | -76 |
| *tyrR* | Cbr | gcTGTcAATAttTgTTgACAga | 3.90 | -91 |
| *HI1290 (tyrA)* | Hin | ctTGTAAtTAAtTtTTTACAat | 4.13 | -45 |
| | | taTGTAAgatAAaAaTTACAgg | 3.45 | -22 |
| *aroF* | Sty | NGtgtaaagtttttgatacgaa | --- | -167 |
| | | gGTGTAAAgtttatTTTACgaa | 4.17 | -115 |
| | | taTGgAttgAAATcTTTACttt | 3.93 | -92 |
| *HI0477 (tyrP)* | Hin | acTGTAAATtAtacaaTACAat | 3.83 | -60 |
| | | atcGTAAATtttTtaTTACAtC | 3.67 | -37 |
| | | taaGgAAAcAAAaATgaACAaa | 3.27 | -14 |
| *HI0528 (tyrP)* | Hin | taaGgAAAcAAAaATgaACAaa | 3.31 | -42 |

[a]Species: Hin, *H.influenzae*; Cbr, *C.braakii*; Sty, *S.typhimurium*.

PHE. However, unlike *aroG*, which is regulated by TyrR (with phenylalanine and tryptophan acting as co-repressors), this *H.influenzae* gene has a TRP box, but no TYR boxes, similarly to the *E.coli* tryptophan-regulated gene *aroH*, which encodes the DAPH-synthase-TRP (Fig. 3g). Two alternative explanations of this evolutionary conundrum seem possible: (i) the *H.influenzae* DAPH-synthase-PHE is regulated by tryptophan at the transcriptional level, the functional implications of which are unclear, and (ii) the *H.influenzae* DAPH-synthase, although in phylogenetic terms orthologous to *aroG*, has changed the specificity of allosteric inhibition and is feedback-inhibited by tryptophan. A final solution can be reached only by experimental analysis of the *H.influenzae* enzyme.

Finally, catabolic operons *tutBA* in *Erwinia herbicola* and *tpl* in *Citrobacter freundii* also are regulated by TyrR and their regulatory regions contain multiple TYR boxes (data are not shown).

*Transport proteins: new members of known regulons.* Our analysis of the PurR regulon resulted in the identification of a family of transport proteins that is represented in *E.coli* and *H.influenzae*, as well as a number of other bacteria (Fig. 4). The family consists of two subfamilies. The known members of one subfamily are uracyl and xanthine transporters (35), whereas the other subfamily does not include any transporters with a known specificity. *Escherichia coli* has representatives in both subfamilies, and notably, they form pairs of closely related 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 (Table 7 and Fig. 4). All close relatives of the *yicE–ygfO* pair and one additional gene with a PUR box, *ygfU*, encode $H^+$/purine(xanthine) symporters, and thus purine transport is the most likely function for these genes. The other two pairs, *yicO–yieG* and *yjcD–ygfQ/R*, as well as the *H.influenzae* gene *HI0125*, which is the ortholog of the latter pair, can be assigned only an unspecified transport function.

**Table 7.** Transport operons predicted to belong to the purine and arginine regulons

| Operon | Sp[a] | Regulator | binding site(s) | score | pos. |
|---|---|---|---|---|---|
| *yjcD* | Eco | PurR | AgctAAACGTTTGCtT | 3.87 | -131 |
| | | | ACGatAACGTTTGCGc | 4.22 | -273 |
| *HI0125 (yjcD)* | Hin | PurR | ACGCAAACGTTTaCtT | 4.85 | -85 |
| *yieG* | Eco | PurR | ACGCAAtCGTTgcCGT | 4.21 | -114 |
| *ygfU* | Eco | PurR | ACGtAAACGgTTGCtT | 4.45 | +21 |
| *yicE* | Eco | PurR | tgGCAAACGTTTGCtT | 4.46 | -69 |
| | | | gCGCAAcCGgTTGCGc | 4.15 | +29 |
| *tsx* | Eco | PurR | ACGCAAtCGaTTaCGT | 4.57 | -153 |
| *tsx* | Eae | PurR | ACGCAAtCGaTTaCGc | 4.40 | -143 |
| *tsx* | Kpn | PurR | ACGCAAtCGaTTaCGT | 4.57 | -157 |
| *tsx* | Sty | PurR | ACGCAAtCGaTTaCGT | 4.57 | -156 |
| *artPIQM* | Eco | ArgR | AtTGcATAAttATTCtgT | 4.08 | -72 |
| *artJ* | Eco | ArgR | AtTGcATAtAAATTCAcT | 4.36 | -86 |
| *HI1180-77 (artPIQM)* | Hin | ArgR | TaTGcATAAAAATgtAaT | 4.01 | -50 |

[a]Species: Eco, *E.coli*; Hin, *H.influenzae*; Eae, *E.aerogenes*; Kpn, *K.pneumoniae*; Sty, *S.typhimurium*.

In addition, PUR boxes were found upstream of the *tsx* gene, which encodes an outer membrane nucleoside-specific channel in *E.coli*, *Enterobacter aerogenes*, *Klebsiella pneumoniae* and *S.typhimurium* (36,37).

The analysis of the ArgR regulon resulted in the identification of ARG boxes upstream of the operons that encode arginine-specific ABC transport systems (*artPIQM* and *artJ* from *E.coli*; *HI1180–HI1177* from *H.influenzae*); thus these operons belong to the arginine regulon (Table 7). In this system, ArtP is the ATPase, ArtQ and ArtM are transmembrane proteins, and ArtI and ArtJ are periplasmic arginine-binding proteins. The orthologous operon of *H.influenzae* has the same gene order. The *E.coli artJ* gene is located immediately downstream of the *artPIQM* operon, but is transcribed independently and has its own ARG box (Fig. 3d); *H.influenzae* has no ortholog for this gene. The regulatory regions of each of the transport operons contain a single ARG box, which suggests that the regulatory effect caused by ArgR binding is likely to be low.

## DISCUSSION

Computer analysis had been used for prediction of bacterial transcription signals for more than 15 years (10,38–44) and on many occasions the results have served as the basis for further experimental work (e.g. 43). Co-evolution of regulons and regulators also was examined (45). However, to the best of our knowledge, this study is the first attempt to systematically characterize regulatory sites in two or more genomes by comparing the respective complete gene sets.

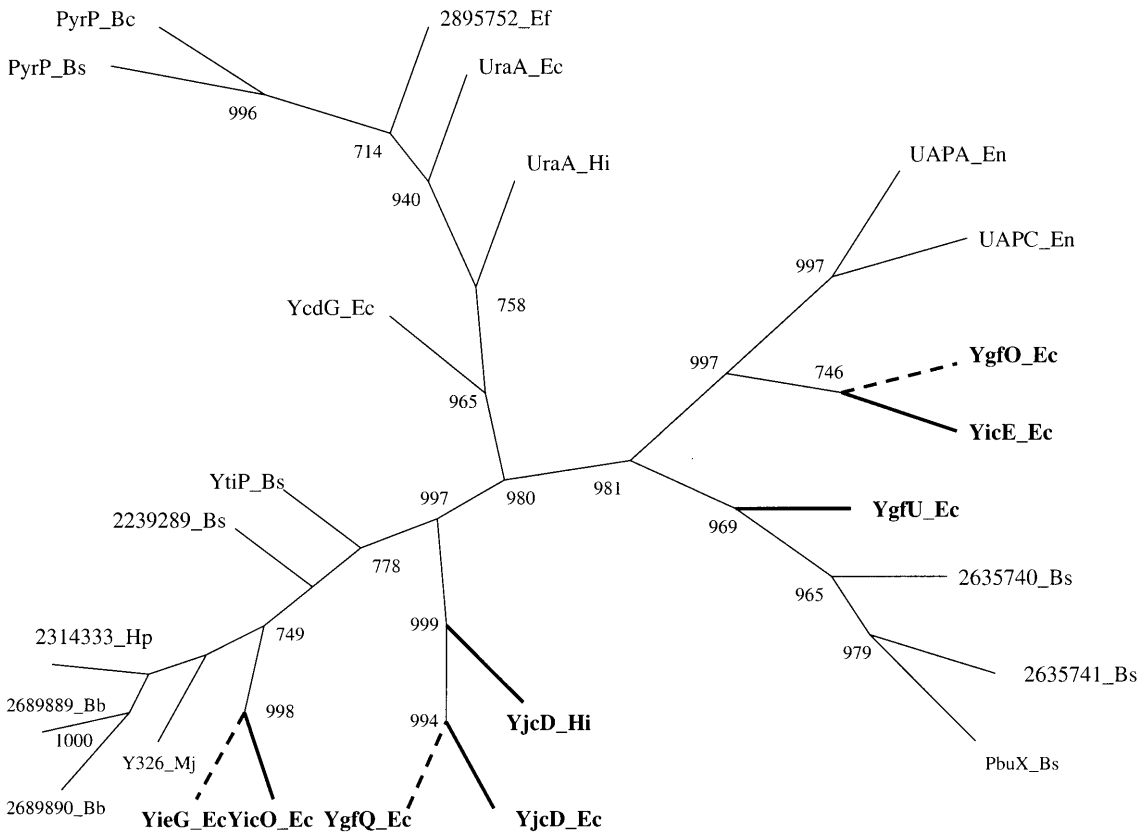


**Figure 4.** A phylogenetic tree of purine and uracyl permeases. The tree was constructed using the neighbor-joining method. The numbers at forks indicate the percentage of bootstrap replications (out of 1000), in which the given grouping was observed. The putative permeases from *E.coli* and *H.influenzae* that were predicted to belong to the *pur* regulon as a result of our analysis and their apparently unregulated paralogs (broken lines) are shown by bold type. PyrP, UraA, uracil permeases; UapA, uric acid-xanthine permease; UapC, broad specificity purine permease; PbuX, xanthine permease. The remaining proteins are functionally uncharacterized gene products that are indicated either by provisional gene name (starting with the letter Y) or by Gene Identification number. Species abbreviations: Bb, *Borrelia burgdorferi*; Bc, *Bacillus caldoliticus*; Bs, *B.subtilis*; Ec, *E.coli*; Ef, *Enterococcus faecalis*; En, *Emericella nidulans*; Hi, *H.influenzae*; Hp, *H.pylori*; Mj, *Methanococcus jannaschii*.

This comparative approach involves three main components: (i) prediction of transcription factor binding sites, (ii) delineation of orthologous relationship between genes by comparing their protein products and (iii) comparison and, when necessary, prediction of protein functions. The use of complete genomes facilitates the identification of orthologs and thus increases the reliability of inferences regarding identical or similar cellular roles of proteins. However, in spite of potential uncertainty in terms of orthology, identification of homologous genes in all bacterial species, including those whose genome sequences have not been completed yet, using similarity search in GenBank is a useful supplement to this analysis.

All sites considered in this paper are approximately palindromic. However, we used the sites in the orientation corresponding to the direction of transcription and did not symmetrize the profiles. There were two reasons for this. First, we were interested in designing a general procedure for site recognition, rather than one that is applicable to symmetrical sites only. Second, it is not guaranteed that even the dimeric factors bind their operators in the symmetric manner. This possibility has been raised in the case of TrpR based on the crystallographic data (46) and chemical modification of natural sites (47), and in the case of AraC based on mutational analysis (48). The Lrp binding signal derived from the SELEX data is not symmetrical either (49).

The comparative analysis of the *E.coli* and *H.influenzae* genomes revealed three principal types of differences between operons that are subject to the same mode of regulation. The differences of the first type are limited to the presence or absence of individual genes in otherwise conserved operons. The examples in *H.influenzae* are operons *ycfCpurB* (*purB* in *E.coli*, Fig. 3b), *argH* (*argCBH* in *E.coli*, Fig. 3c), *ydfGtrpBA* (*trpBA* in *P.multocida*, Fig. 3e) and *tyrA* (*aroFtyrA* in *E.coli*, Fig. 3g).

The second type of changes involves breaking of an operon into two parts, both of which retain the regulation. Two *E.coli* operons, *purHD* and *glyA*, both regulated by PurR, correspond, in *H.influenzae*, to the gene string *HI0887–HI0889* with a PUR box upstream of *HI0887* (Fig. 3a). Similarly, the tryptophan operon is broken in *H.influenzae* into two parts, *trpEDC* and *trpBA*, both of which have strong TRP boxes in the regulatory regions.

Finally, some operons lose or switch regulation. The most interesting case in this category is the elimination of *purR*

autoregulation in *H.influenzae*. The loss of 'regulation of regulators' appears to be a more general phenomenon: in *E.coli*, the repressor IlvY regulates both its own gene *ilvY* and the adjacent *ilvC* gene, which are transcribed from divergent promoters. By contrast, in *H.influenzae*, although the overall location of these genes is the same, the distance between them is much larger, and a candidate binding site is close to *ilvC*, but too distant from *ilvY* to expect autoregulation (M.Gelfand, unpublished observation). The elimination of this higher level of regulation may be linked to the evolution of the parasitic lifestyle of *H.influenzae* that requires much less versatility in the response of the bacterium to environmental changes than its free-living relatives, such as *E.coli*. Another clear case of simplification in regulation includes the loss of the TYR box by the *H.influenzae mtr* operon, which in *E.coli* is regulated by both TrpR and TyrR. The roadblock mechanism of repression of *purB* by *purR* in *E.coli* is not conserved in *H.influenzae*, although the repression itself seems to exist. Finally, it is possible that the gene *aroG* of *H.influenzae* has switched its regulation from TyrR to TrpR.

The conservation of a regulatory DNA-binding protein in an uncharacterized bacterial genome seems to be a reliable predictor of the conservation of the binding sites in at least some operons, even if most of the regulon is missing. For example, there are only three known genes in the arginine regulon of *H.influenzae*, including the repressor ArgR itself (but not counting the transport proteins predicted to belong to the arginine regulon in this work), but the ARG boxes are conserved. The *E.coli* ARG box recognition matrix seems capable of detecting the relevant signals even in the distantly related *Bacillus subtilis* genome, which also encodes an ortholog of ArgR (A.A.Mironov and M.S.Gelfand, unpublished observations). Conversely, there are no strong PUR boxes in the *Helicobacter pylori* genome that does not encode a PurR ortholog. Similarly, although there is a purine repressor in *B.subtilis*, it is unrelated to the *E.coli* PurR, and indeed, the type of regulation (mostly by attenuation) and regulatory sites (in a few genes regulated at the transcription level) of the *B.subtilis* purine regulon differ from those of *E.coli*. The *P.aeruginosa* operon *trpBA* is regulated by the repressor TrpI, which is unrelated to TrpR of *E.coli* and *H.influenzae*, and predictably, there are no TRP boxes in the region upstream of this operon.

This study allowed us to make several predictions that appear to be readily experimentally testable. One group of such predictions includes inferences about changes in regulation patterns, namely the loss of autoregulation in the *H.influenzae* ortholog of PurR, different mode of repression of *purB*, and the apparent change in the regulation of *aroG*. The second group of predictions extends the purine and arginine regulons both in *E.coli* and *H.influenzae* by inclusion of transport proteins (purine and arginine transporters). It is somewhat surprising that these transport systems, especially the large family of $H^+$/purine symporters, have not been identified as part of the purine regulon by genetic analysis. A possible explanation is that all genes from this family that are predicted to be under the PurR regulation have close non-regulated paralogs, and thus the effect of mutations in the regulated genes might be manifest only under very specific conditions.

Further research directions will include analysis of global regulatory systems, such as SOS, CRP, Fur and Fnr regulons, and multiple interacting systems, for example the interaction between purine and pyrimidine regulation or the interaction between the regulation by repression and by attenuation in the aromatic amino acid regulon, as well as comparisons between more distant genomes, such as *E.coli* and *B.subtilis*. As a more distant goal, we envisage development of techniques for systematic characterization of regulatory pathways in newly sequenced genomes.

## ACKNOWLEDGEMENTS

This work was partially supported by grants from the Russian Fund of Basic Research and the US Department of Energy (FG02-94ER61919).

## REFERENCES

1. Koonin,E.V. and Galperin,M.Y. (1997) *Curr. Opin. Genet. Dev.*, **7**, 757–763.
2. Koonin,E.V., Tatusov,R.L. and Galperin,M.Y. (1998) *Curr. Opin. Struct. Biol.*, **8**, 355–363.
3. Ouzounis,C., Casari,G., Sander,C., Tamames,J. and Valencia,A. (1996) *Trends Biotechnol.*, **14**, 280–285.
4. Huynen,M.A. and Bork,P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
5. Koonin,E.V., Mushegian,A.R. and Rudd,K.E. (1996) *Curr. Biol.*, **6**, 404–416.
6. Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) *Curr. Biol.*, **6**, 279–291.
7. Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) *Nucleic Acids Res.*, **25**, 701–712.
8. Mushegian,A.R. and Koonin,E.V. (1996) *Trends Genet.*, **12**, 289–290.
9. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) *Trends Biochem. Sci.*, **23**, 324–328.
10. Gelfand,M.S. (1995) *J. Comput. Biol.*, **2**, 87–115.
11. Thieffry,D., Huerta,A.M., Perez-Rueda,E. and Collado-Vides,J. (1998) *Bioessays*, **20**, 433–440.
12. Frech,K., Quandt,K. and Werner,T. (1997) *Comput. Appl. Biosci.*, **13**, 89–97.
13. Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
14. Mironov,A.A., Roytberg,M.A., Pevzner,P.A. and Gelfand,M.S. (1998) *Genomics*, **51**, 332–339.
15. Frishman,D., Mironov,A., Mewes,H.W. and Gelfand,M. (1998) *Nucleic Acids Res.*, **26**, 2941–2947.
16. Gelfand,M.S. and Koonin,E.V. (1997) *Nucleic Acids Res.*, **25**, 2430–2439.
17. Mironov,A.A. and Gelfand,M.S. (1999) *Mol. Biol.*, **33**, 127–132 (in Russian).
18. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1474.
19. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
20. Zalkin,H. and Nygaard,P. (1996) in Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella. Cellular and Molecular Biology*. ASM Press, Washington, DC, Vol. 1, pp. 561–579.
21. Glansdorff,N. (1996) in Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella. Cellular and Molecular Biology*. ASM Press, Washington, DC, Vol. 1, pp. 408–433.
22. Pittard,A.J. (1996) in Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella. Cellular and Molecular Biology*. ASM Press, Washington, DC, Vol. 1, pp. 458–484.
23. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) *Nucleic Acids Res.*, **26**, 55–59.
24. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
25. Mironov,A.A., Alexandrov,N.N., Bogodarova,N., Grigorjev,A., Lebedev,V.F., Lunovskaya,L.V., Truchan,M.E. and Pevzner,P.A. (1995) *Comput. Appl. Biosci.*, **11**, 331–335.

26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
27. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
28. Felsenstein,J. (1996) *Methods Enzymol.*, **266**, 418–427.
29. Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
30. Tian,G., Lim,D., Carey,J. and Maas,W.K. (1992) *J. Mol. Biol.*, **226**, 387–397.
31. Meng,L.M., Kilstrup,M. and Nygaard,P. (1990) *Eur. J. Biochem.*, **187**, 373–379.
32. He,B. and Zalkin,H. (1994) *J. Bacteriol*., **176**, 1009–1013.
33. He,B. and Zalkin,H. (1992) *J. Bacteriol*., **174**, 7121–7127.
34. Ahmad,S., Rightmire,B. and Jensen,R.A. (1986) *J. Bacteriol*., **165**, 146–154.
35. Diallinas,G., Gorfinkiel,L., Arst,H.N., Jr, Cecchetto,G. and Scazzocchio,C. (1995) *J. Biol. Chem.*, **270**, 8610–8622.
36. Bremer,E., Middendorf,A., Martinussen,J. and Valentin-Hansen,P. (1990) *Gene*, **96**, 59–65.
37. Nieweg,A. and Bremer,E. (1997) *Microbiology*, **143**, 603–615.
38. Berg,O.G. and Hippel,P.H.V. (1988) *Trends Biochem. Sci.*, **13**, 207–211.
39. Stormo,G.D. and Hartzell,G.W.D. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
40. Goodrich,J.A., Schwartz,M.L. and Mcclure,W.R. (1990) *Nucleic Acids Res.*, **18**, 4993–5000.
41. Rosenblueth,D.A., Thieffry,D., Huerta,A.M., Salgado,H. and Collado-Vides,J. (1996) *Comput. Appl. Biosci.*, **12**, 415–422.
42. Thieffry,D., Salgado,H., Huerta,A.M. and Collado-Vides,J. (1998) *Bioinformatics*, **14**, 391–400.
43. He,B., Choi,K.Y. and Zalkin,H. (1993) *J. Bacteriol*., **175**, 3598–3606.
44. Robison,K., McGuire,A.M. and Church G.M. (1998) *J. Mol. Biol*., **284**, 241–254.
45. Otsuka,J., Watanabe,H. and Mori,K.T. (1996) *J. Theor. Biol*., **178**, 183–204.
46. Niland,P., Huhne,R. and Muller-Hill,B. (1996) *J. Mol. Biol*., **264**, 667–674.
47. Saacke,D., Walter,B., Kisters-Woike,B., von Wilcken-Bergmann,B. and Muller-Hill,B. (1990) *EMBO J.*, **9**, 1963–1967.
48. Liu,Y.C. and Matthews,K.S. (1993) *Biochemistry*, **32**, 10532–10542.
49. Niland,P., Huhne,R. and Muller-Hill,B. (1996) *J. Mol. Biol*., **264**, 667–674.
50. Cui,Y., Wang,Q., Stormo,G.D. and Calvo,J.M. (1995) *J. Bacteriol*., **177**, 4872–4880.