OPEN ACCESS Freely available online

PLOS ONE

Comparative Genomics of CytR, an Unusual Member of the LacI Family of Transcription Factors

Natalia V. Sernova[1], Mikhail S. Gelfand[1,2]*

1 A.A.Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences (IITP RAS), Moscow, Russia, **2** Faculty of Bioengineering and Bioinformatics, M.V.Lomonosov Moscow State University, Moscow, Russia

Abstract

CytR is a transcription regulator from the LacI family, present in some gamma-proteobacteria including *Escherichia coli* and known not only for its cellular role, control of transport and utilization of nucleosides, but for a number of unusual structural properties. The present study addressed three related problems: structure of CytR-binding sites and motifs, their evolutionary conservation, and identification of new members of the CytR regulon. While the majority of CytR-binding sites are imperfect inverted repeats situated between binding sites for another transcription factor, CRP, other architectures were observed, in particular, direct repeats. While the similarity between sites for different genes in one genome is rather low, and hence the consensus motif is weak, there is high conservation of orthologous sites in different genomes (mainly in the Enterobacteriales) arguing for the presence of specific CytR-DNA contacts. On larger evolutionary distances candidate CytR sites may migrate but the approximate distance between flanking CRP sites tends to be conserved, which demonstrates that the overall structure of the CRP-CytR-DNA complex is gene-specific. The analysis yielded candidate CytR-binding sites for orthologs of known regulon members in less studied genomes of the Enterobacteriales and Vibrionales and identified a new candidate member of the CytR regulon, encoding a transporter named NupT (YcdZ).

Citation: Sernova NV, Gelfand MS (2012) Comparative Genomics of CytR, an Unusual Member of the LacI Family of Transcription Factors. PLoS ONE 7(9): e44194. doi:10.1371/journal.pone.0044194

Editor: Roy Martin Roop II, East Carolina University School of Medicine, United States of America

Received March 23, 2012; **Accepted** July 30, 2012; **Published** September 24, 2012

Funding: This study was partially supported by the Ministry of Education and Science of Russia via contract 07.514.11.4007, the Russian Foundation of Basic Research via grant 10-04-00431, and the Russian Academy of Sciences via the Program in Molecular and Cellular Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gelfand@iitp.ru

Introduction

CytR, the regulator of transport and utilization of nucleosides, was first mentioned in 1975 [1] and identified in 1985 [2]. The corresponding gene, *cytR*, was sequenced in 1986 [3].

Since then, nine *Escherichia coli* genes were experimentally shown to be regulated by CytR and form the CytR regulon: *deoC* [4,5], *cytR* itself [6], *tsx* [7], *cdd* [8], *ppiA* [9], *nupC* [10], *nupG* [11], *udp* [12], *rpoH* [13]. Later the corresponding binding sites in the upstream regions of these genes, except *nupC*, were proposed. In other species, CytR was shown to regulate its own gene, *cytR*, in *Salmonella typhimurium* [14] and *udp* in *Salmonella typhimurium*, *Yersinia pestis* and *Vibrio cholerae* [15]. Since among the gamma-proteobacteria CytR is present only in *E. coli* and its close relatives (up to *Vibrionales*), it has been suggested that *cytR* appeared in the Enterobacteriales due to horizontal transfer from the delta-Proteobacteria (*Geobacillus* sp.) or *Caulobacter* sp. [16].

The structural and functional features of CytR were reviewed in [17–20]. CytR is an atypical representative of the LacI-family [21]. Its affinity to its operators is rather weak [22] and because of that, in contrast to most prokaryotic repressors, CytR alone is not capable to repress transcription. CytR functions in a complex with a multifunctional transcription factor (TF), CRP [23,24].

The CRP protein is a dimer [25]. The subunit dimerization depends on the N-terminal domain, while the DNA recognition is performed by the C-terminal domain [26]. A possible regulatory mechanism was suggested, based on the crystal structure of CRP in complex with a DNA-fragment [27].

CytR protein is also dimeric [28]. The number of CRP-binding sites (O_{CRP}) per CytR-binding site (O_{CYTR}) varies from one to three [17]: one, as in the *cytR* promoter [6]; two, as in the majority of cases; or three, as in the *cdd* promoter [8,17,29]. This might indicate different structures of the CRP-CytR complex or repositioning of the CRP dimers upon interaction. In most promoters, CRP has a stronger affinity to the distal operator O_{CRP}D [30], with an exception being *cddP*, where CRP binds stronger to its proximal operator O_{CRP}P [29]. An important requirement is that at least one CRP-operator has to be situated at a distance not exceeding 5 nucleotides to the corresponding CytR-operator [20], with the position of the O_{CYTR} operator being not symmetric relative to the flanking O_{CRP} operators [17]. Fig. 1 shows a typical organization of the O_{CRP}D-O_{CYTR}-O_{CRP}P complex for five experimentally studied *E. coli* genes.

The mechanism of the CytR action is anti-activation rather than direct repression [17,31,32]. In particular, at the promoter *deo*P2, RNA polymerase and CytR compete for CRP that in this case acts as an activator [31]. CRP alone activates transcription, whereas the CRP-CytR cooperatively bound to O_{CRP} and O_{CYTR}, respectively, represses transcription. Cytidine binding to CytR releases the latter from the complex, hence the activation by CRP resumes and the gene is derepressed [33,34]. At that, the intrinsic CytR binding to DNA is not affected by cytidine binding [35]. The repression and activation of some other CytR-regulon

```
nupG aAaTGTtATCcAcATCACAaTT-04-TtTGCAAa-09-TTTGCAaT.....taTtTGccacAGgTaACAaAa
udp  TtaTGTGATtTgcATCACtTTT-12-TATGCAAc-03-TTTGCgTc.....ATgGTGATgagtATCACgaAa
ppiA TtTTGTGATCTgttTaAatgTT-02-atTGCAAT-10-aTTGCATT-03-agagGTGATtTtGATCACggAa
deoC TtaTtTGAaCcAGATCgCATTa-04-gATGCAAa-00-cTTGtAag-11-aATTGTGATgTgtATCgaAgTg
rpoH atTTcatcTCTAtgTCACATTT-02-TGcGtAAT-11-cTTGCATT-02-acTTGTGgataAaATCACggTc
```

Figure 1. Organization of upstream regions of five experimentally proven *E. coli* members of the CytR regulon. [13,19,37]. CytR-binding sites (O_{CYTR}) are highlighted in magenta, cores of CRP-binding sites (O_{CRP}) – in green. Numbers denote spacer lengths. Dots denote gaps in the alignment.
doi:10.1371/journal.pone.0044194.g001

genes were considered, in particular, in [17]. In the *cytR*, *deoP* and *udp* regulatory regions only one CRP-binding site participates in the activation; in the *nupG*, *tsx* and *cdd* promoters two CRP-binding sites are involved in the activation, and in all regulated genes, except *cytR*, two CRP-binding sites participate in the repression. Hence upstream regions of all CytR-regulated genes contain at least one CRP-binding site, either distal or proximal, that participates both in activation and repression, see [17] (Fig. 3, p.463).

The CytR-binding motif consists of two half-sites, denoted here as $O_{CYTR}D$ (distal) and $O_{CYTR}P$ (proximal). Unlike the situation with many TFs, including repressors from the LacI-family, the length of spacers between parts of the O_{CYTR} motif may vary in a wide interval from about zero to three DNA helical turns, with large spacers tending to comprise an integer number of turns, at most three [18].

In most studies, $O_{CYTR}D$ and $O_{CYTR}P$ were assumed to form degenerated inverted repeats [11,13,36], and the major role in specific binding was assigned to protein-protein (CytR-CRP)

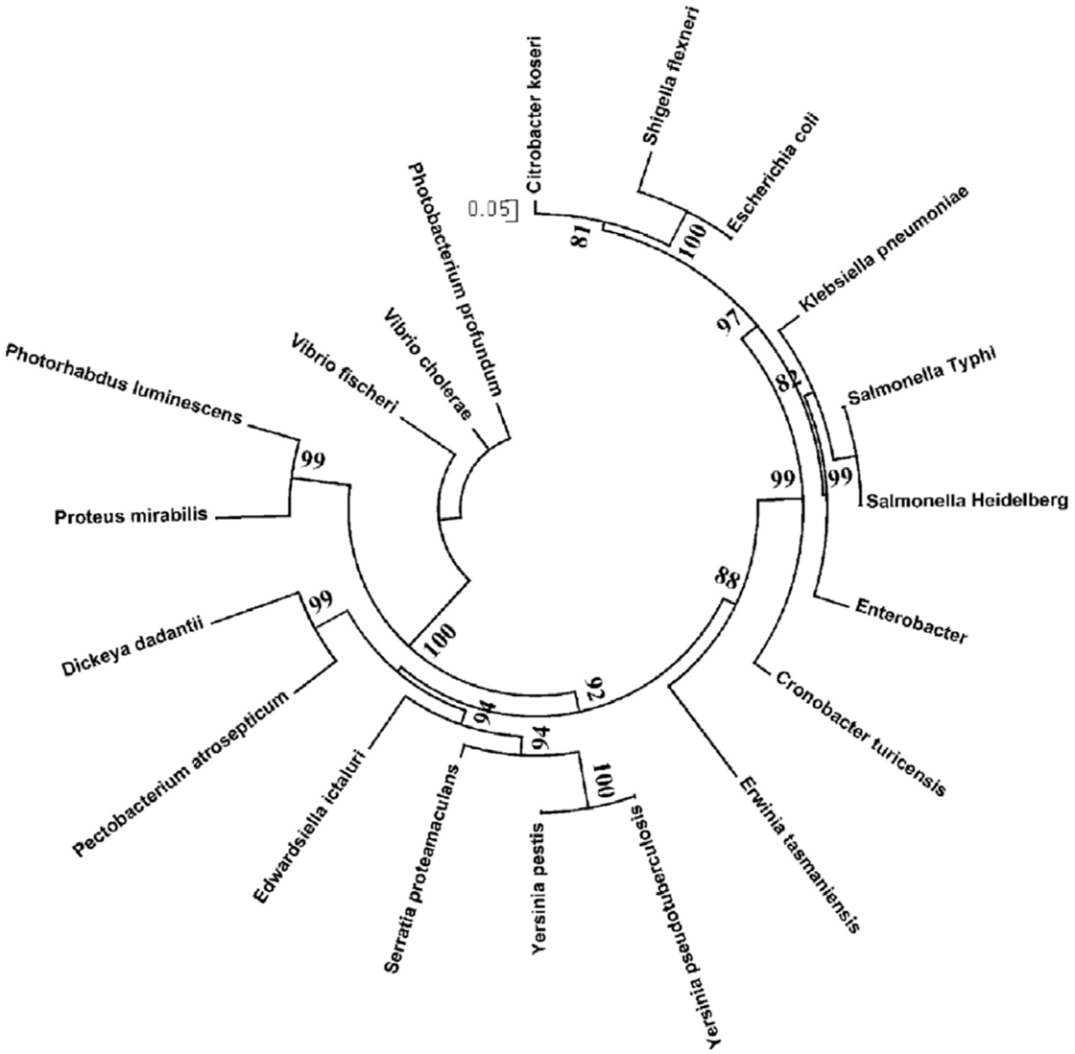

Figure 2. The ML-phylogenetic tree for the CytR proteins from the Enterobacteriales and Vibrionales. The tree defined the order of constructing alignments of upstream regions, see text for details.
doi:10.1371/journal.pone.0044194.g002

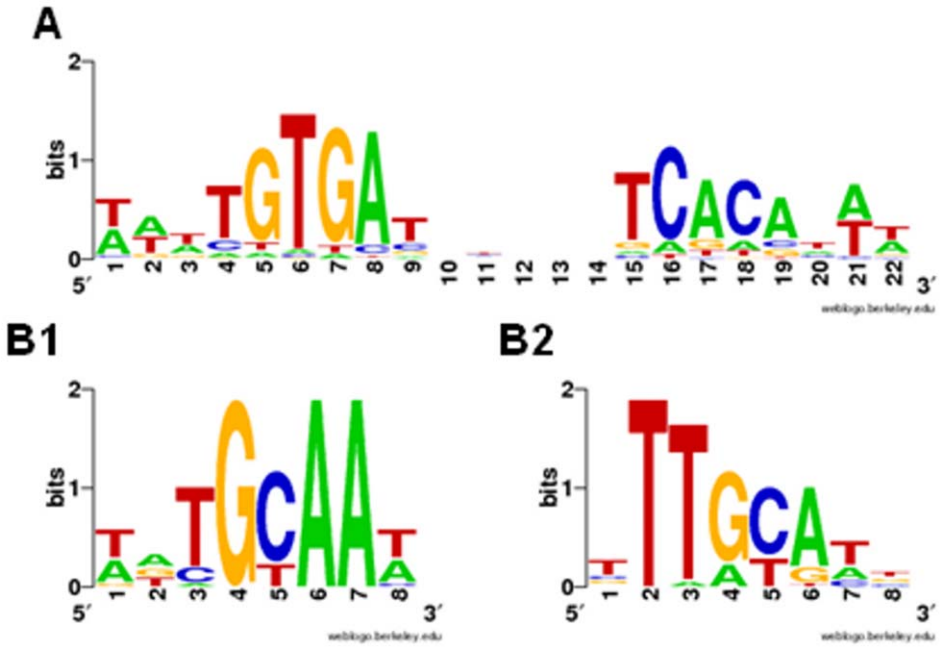

Figure 3. Sequence LOGOs of the CRP, CytR-distal, CytR-proximal operators. Horizontal axis: position in the binding site; vertical axis: informationcontent in bits. The height of each individual symbol reflects its prevalence at a given position, the height of each column is proportional to the positional information content in this position. A) O_{CRP} LOGO ; B1) O_{CYTR}D LOGO; B2) O_{CYTR}P LOGO.
doi:10.1371/journal.pone.0044194.g003

Table 1. The list of genomes with abbreviations.

3-letter abbreviation*	The full name of the bacteria	Accession number
CKO	*Citrobacter koseri*	NC_009792
CTU	*Cronobacter turicensis*	NC_013282
DDC	*Dickeya dadantii*	NC_013592
ECO	*Escherichia coli*	NC_000913
EIC	*Edwardsiella ictaluri*	NC_012779
ENT	*Enterobacter*	NC_009425
ETA	*Erwinia tasmaniensis*	NC_010693
KPE	*Klebsiella pneumoniae*	NC_011283
PEB	*Pectobacterium atrosepticum*	NC_004547
PLU	*Photorhabdus luminescens*	NC_005126
PMR	*Proteus mirabilis*	NC_010554
PPR	*Photobacterium profundum*	NC_006370(NC_006371)**
SEH	*Salmonella Heidelberg*	NC_011083
STY	*SalmonellaTyphi*	NC_003198
SFL	*Shigella flexneri*	NC_008258
SPE	*Serratia proteamaculans*	NC_009832
VCH***	*Vibrio cholerae*	NC_002505(NC_002506)
VFM***	*Vibrio fischeri*	NC_011184(NC_011186)
VHA***	*Vibrio harveyi*	NC_009783(NC_009784)
VPA***	*Vibrio parahaemolyticus*	NC_004603(NC_004605)
VSP***	*Vibrio splendidus*	NC_011753(NC_011744)
VVU***	*Vibrio vulnificus*	NC_004459(NC_004460)
YPN	*Yersinia pestis*	NC_008149
YPS	*Yersinia pseudotuberculosis*	NC_006155

*Abbreviations in the left column were taken from KEGG database.
**The accession number for the second chromosomes is in parentheses.
***In the alignments, these genomes are denoted by the first two letters and a digit denoting the chromosome (1 or 2).
doi:10.1371/journal.pone.0044194.t001

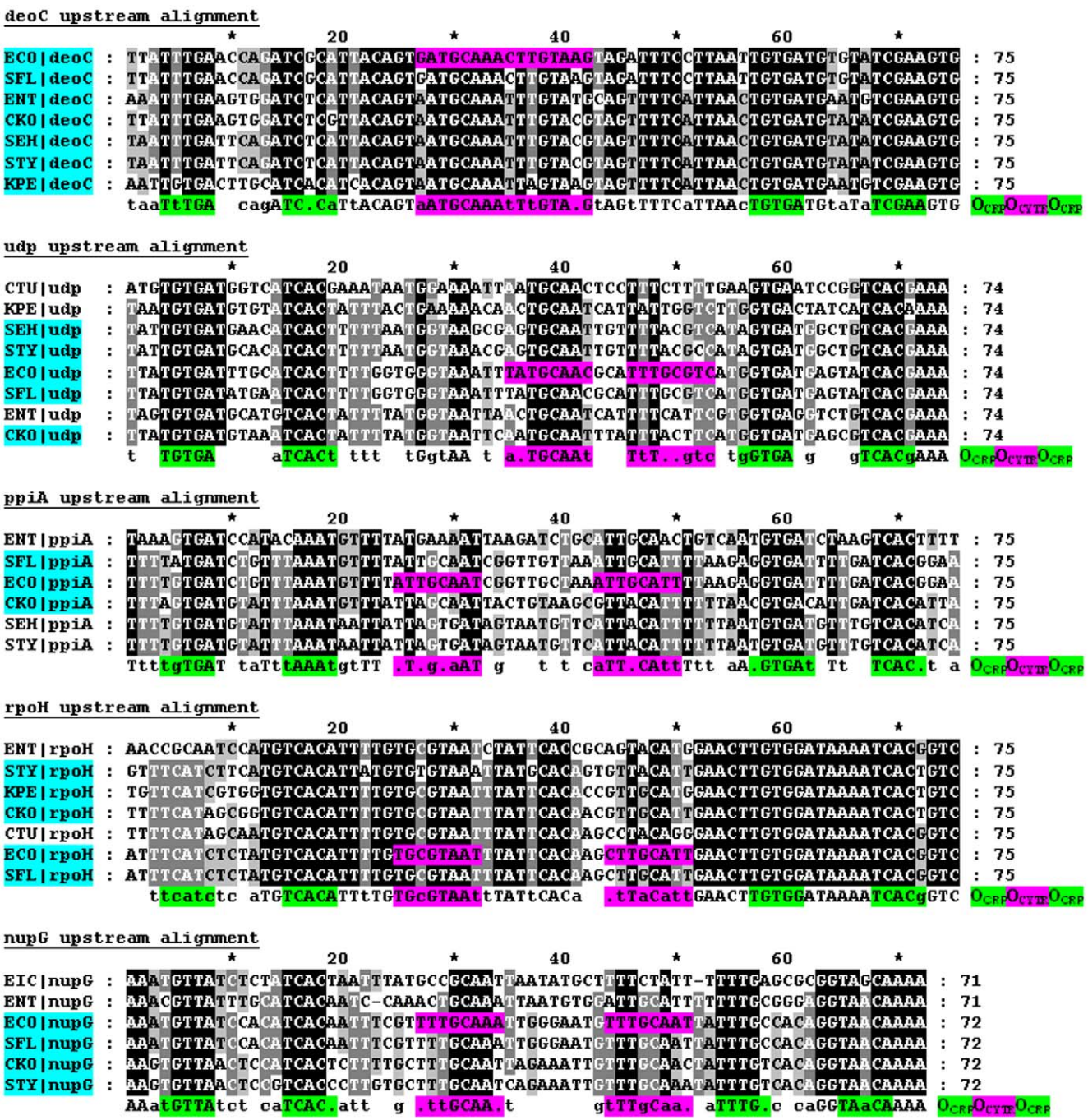

Figure 4. Alignments of upstream regions of gene orthologous to the *E.coli* CytR regulon members. O_{CYTR} boxes are highlighted in magenta. The consensus CytR and CRP motifs are shown at the bottom in magenta and green, respectively. Blue in the left column marks the genomes whose CytR-binding sites were used to construct the PWM. Shadows of grey denote the level of conservation, as set by GeneDoc: black – 100% conservation; dark gray – the consensus nucleotide frequency between 75% and 100%, light grey – the consensus nucleotide frequency between 50% and 75%; white – no conservation.
doi:10.1371/journal.pone.0044194.g004

rather than protein-DNA (CytR-O_{CYTR}) interactions. Still, at the physiological concentration of CytR, the CytR-DNA interactions are absolutely necessary for the repressor complex to be formed [37]. The exact O_{CYTR} position was mapped precisely in few cases only, e.g. in the *deo*P2 promoter by point mutagenesis [36] or by exchange of *udp*P and *deo*P2 O_{CYTR} operators [19]. In the majority of other cases, binding sites were located approximately by using the protein shift assay, protein footprinting with DNAase I or hydroxyl radical footprints, DMS-treatment, or cloning into a plasmid and measuring the level of CytR-repression [11,22]. The exact position of O_{CYTR} was then predicted by the comparison with the consensus [11].

The latter was described as an inverted pentameric repeat TGCAA-N_{2-3}-TTGCA [36], (where N denoted the number of nucleotides), a palindrome TTGCAA [38], or a pair of inverted octameric repeats (either 5′-AATGYCAAC-GC-GTTGCATT-3′

A

```
                    *         20         *         40         *         60         *
SEH|udp  :  TATTGTGATGAACATCACTTTTTAATGGTAAGCGAGTGCAATTGTTTTACGTCATAGTGATGGCTGTCACGAAA  :  74
STY|udp  :  TATTGTGATGCACATCACTTTTTAATGGTAAACGAGTGCAATTGTTTTACGCCATAGTGATGGCTGTCACGAAA  :  74
ECO|udp  :  TTATGTGATTTGCATCACTTTTGGTGGGTAAATTTATGCAACGCATTTGCGTCATGGTGATGAGTATCACGAAA  :  74
SFL|udp  :  TTATGTGATATGAATCACTTTTGGTGGGTAAATTTATGCAACGCATTTGCGTCATGGTGATGAGTATCACGAAA  :  74
CKO|udp  :  TTATGTGATGTAAATCACTATTTTATGGTAATTCAATGCAATTTATTTACTTCATGGTGATGAGCGTCACGAAA  :  74
            T  TGTGAT     TCACT ttt  tGGTAA   ..TGCAA.   TtT.CgtCAT.GTGATG  t TCACgAAA
```

B

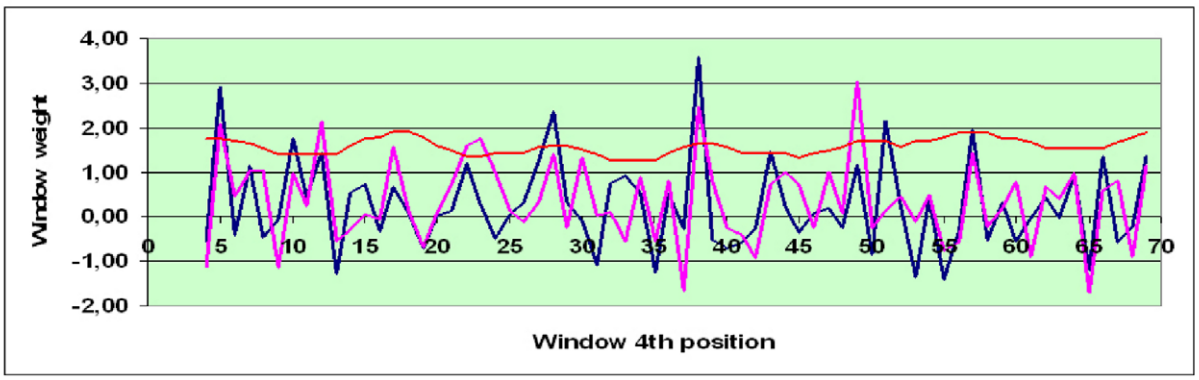

Figure 5. Alignment and SWAS plots of upstream regions of *udp* in close relatives of *E.coli*. The detected sites are highlighted in the consensus of alignment. A) Alignment of the upstream regions. Green – CRP-boxes, magenta – CytR-boxes. B) SWAS and information content plots. Scores are plotted corresponding to the middle (4^{th}) position of a 8 bp window. Blue – $O_{CYTR}D$, magenta – $O_{CYTR}P$, red – averaged positional information content.
doi:10.1371/journal.pone.0044194.g005

A

```
                    *         20         *         40         *         60         *
ENT|udp  :  TAGTGTGATGCATGTCACTATTTTATGGTAATTAACTGCAATCATTTTCATTCGTGGTGAGGTCTGTCACGAAA  :  74
CTU|udp  :  ATGTGTGATGGTCATCACGAAATAATGGAAAATTAATGCAACTCCTTTCTTTTGAAGTGAATCCGGTCACGAAA  :  74
KPE|udp  :  TAATGTGATGTGTATCACTATTTACTGAAAAACAACTGCAATCATTATTGGTCTTGGTGACTATCATCACAAAA  :  74
PMR|udp  :  ATTTGTGATAACTATCACGTTTTTATGAAATTGTTACTTTTTTTATTACCTGACAGGTGACGTCAATCACTAAA  :  74
ETA|udp  :  AAATGTGACATTTATCGCATTTTTATGTATTTACGATGTGGTTTTCATTTTGATAAATGATGTCTATCACATAA  :  74
DDC|udp  :  AAGTGTGATTAAAATCACGTTATTAATGTAAATAGATGTTCGTTACTTTCCAATGAGCGACGTGCTTCACTTAA  :  74
YPN|udp  :  ATGTGTGATTTAAATCACAATTAAAATGTATTTATGTAATCACTTTCTCTTTCAAGGTGACATAAATCACGAAA  :  74
YPS|udp  :  ATGTGTGATTTAAATCACAATTAAAATGTATTTATGTAATCACTTTCTCTTTCAAGGTGACATAAATCACGAAA  :  74
PEB|udp  :  AAAAGTGATTTAAGTCACATTTTTTATGTATGTTGATCTCCGTTTCTTTTTGGTAGGTGACTCTCATCACTTAA  :  74
SPE|udp  :  TTTTGTGACTTTAATCACAAAAGTTATGTAGATTGCATTCTGTTTCTTTATTCATAGTGATGGTTGTCACGAAA  :  74
            tGTGAt    aTCaCa t              t          t  t    .gtGAc     TCAC. AA
```

B

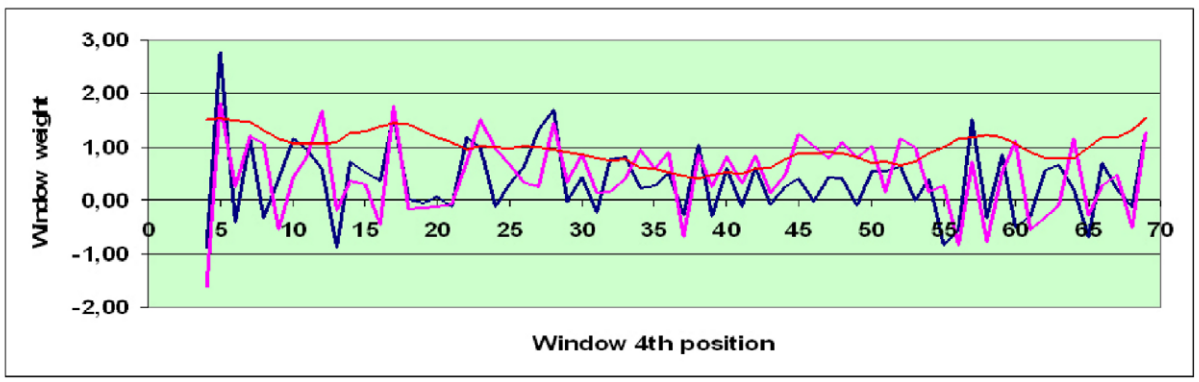

Figure 6. Alignment and SWAS plots of upstream regions of *udp* in distant Enterobacteriales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g006

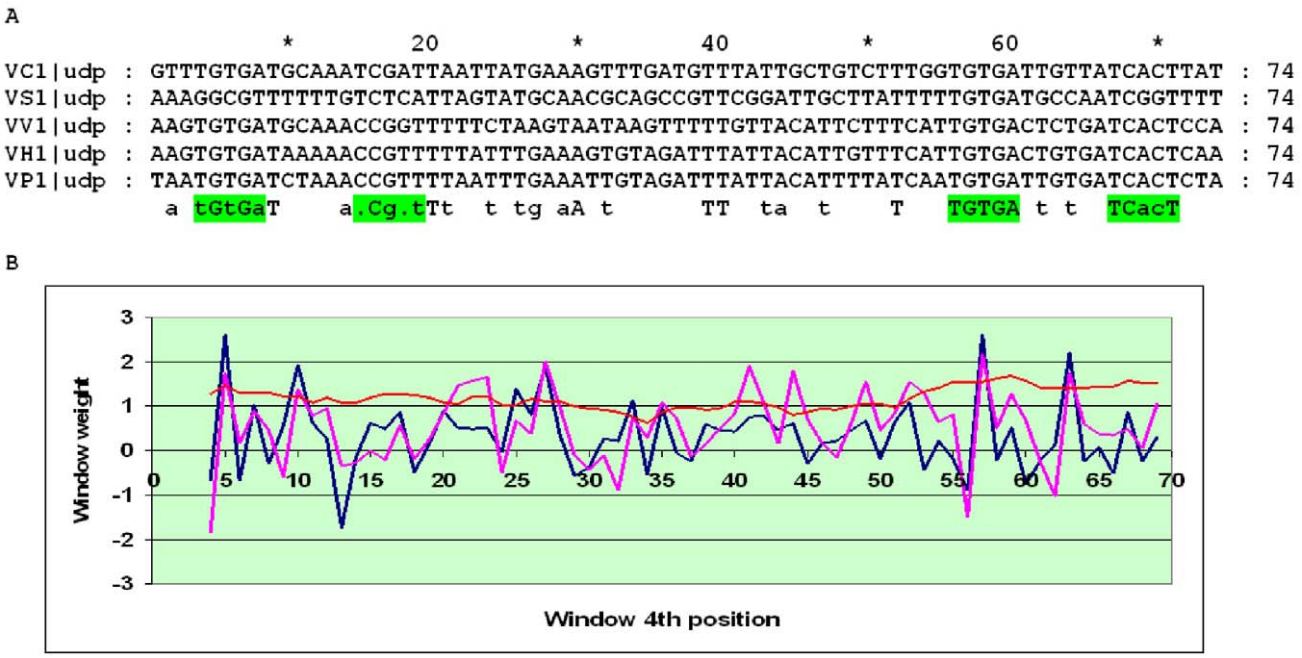

Figure 7. Alignment and SWAS plots of upstream regions of *udp* in the Vibrionales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g007

or 5′-AYGTGCAAC-N_x-GTTRCATT-3′, where Y = T or C, R = A or G, and x = 10, 11, 12 or 13) which are the optimal CytR-binding sites in the absence of CRP, or direct repeats of octamers in either orientation with a 1 bp spacer [37]. The most recent description implies only octameric repeats with a spacer allowing both of them to be situated on the same side of the DNA-helix, with the spacer being less than 4–5 nucleotides or roughly a helical turn, that is 10–11 nucleotides [18,20]. The current experimental data agree with this description [13,29,39]. The distances of about

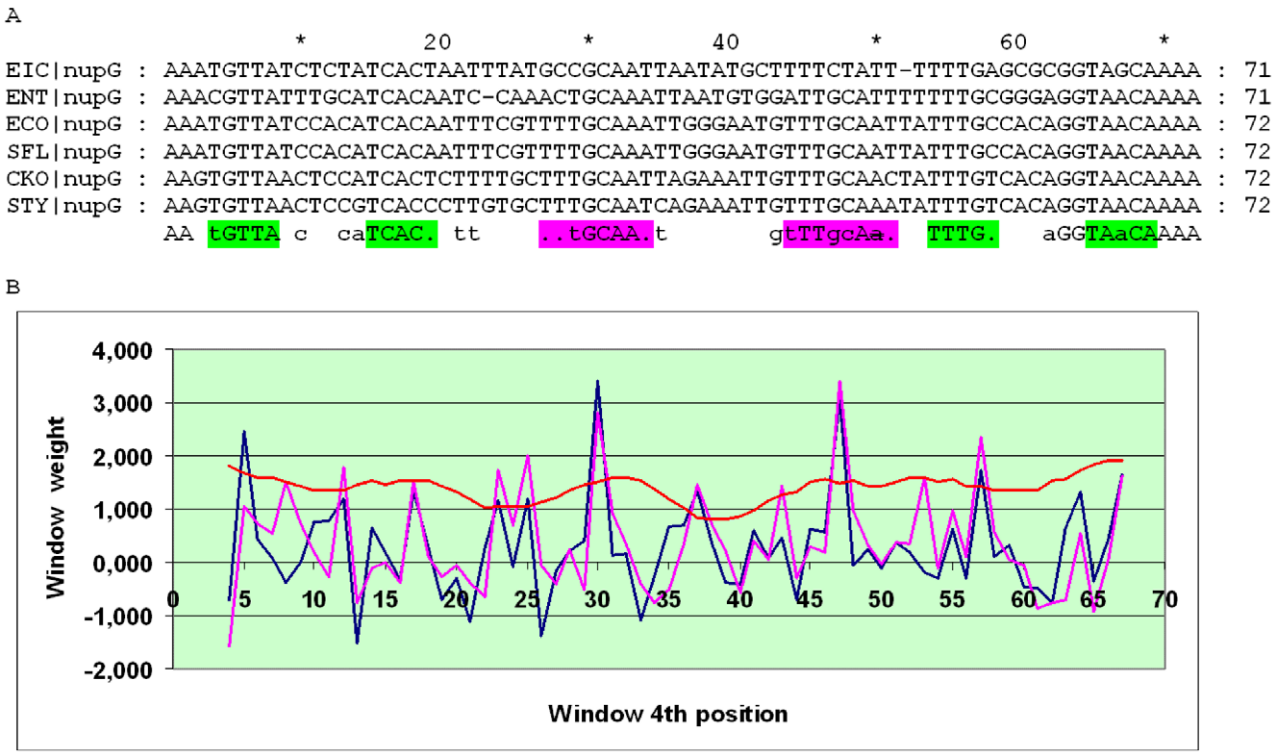

Figure 8. Alignment and SWAS plots of upstream regions of *nupG* in the Enterobacteriales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g008

two or three helical turns were experimentally proven to be possible [18] but so far have not been observed in nature.

Here, we study the evolution of CytR-binding sites, characterize their common features, and identify new candidate members of the CytR regulon in the Enterobacteriales and Vibrionales.

Results

Recognition rules

We compiled a list of gamma-proteobacterial genomes encoding orthologs of CytR. Orthologs were initially defined by the bidirectional best hit criterion and confirmed by construction of phylogenetic trees (Fig. 2). All these genomes belong to the Enterobacteriales and Vibrionales, the list is given in Table 1. We also identified orthologs of genes known to be regulated by CytR in *E. coli* (Table S1) (see Data and Methods).

We used 69 published CRP-binding sites [40] to construct the O_{CRP} positional weight matrix (PWM) using SignalX (see Data and Methods).). Sequence LOGOs of the constructed motifs are shown in Fig. 3A. To construct the O_{CYTR} PWMs, we considered five *E. coli* genes with clearly distinguishable O_{CYTR} (Fig. 1). We performed multiple alignment of the upstream regions of these genes and their orthologs. At that, we gradually increased the number of aligned sequences, starting with closest *E. coli* relatives and then including more distant ones, in the order given by the phylogenetic tree of the CytR proteins (Fig. 2), while the O_{CRP} sites could be reliably aligned and the distance between them remained approximately constant. Then we selected only the sequences that were conserved in the regions corresponding to the *E. coli* sites: both $O_{CYTR}D$ and $O_{CYTR}P$ were taken for *deoC* from ECO, SFL, ENT, CKO, SEH, STY, KPE; for *udp* from SEH, STY, ECO, SFL, CKO; for *ppiA* from SFL, ECO, CKO; for *rpoH* from STY, KPE, CKO, ECO, SFL; and for *nupG* from ECO, SFL, CKO, STY; see Table 1 for genome abbreviations; the selected genes are highlighted in blue in Fig. 4. Sites in other species were accepted for the matrix construction if they satisfied the following conservation conditions: (1) the same distance between O_{CRP} sites for each *E. coli* gene listed above and its orthologs; (2) the same distance between O_{CYTR} half-operators; (3) at most two mismatches in each O_{CRP} site, and at most three total mismatches in the O_{CRP} sites, compared to the *E. coli* O_{CRP} sites; and (4) at most four mismatches in the O_{CYTR} operator, and at most three mismatches in each O_{CYTR} half-operator compared to the *E. coli* O_{CYTR} boxes. The selected boxes were used to construct PWMs for the upstream (distal) and downstream (proximal) half-operators ($O_{CYTR}D$, Fig. 3B1, and $O_{CYTR}P$, Fig. 3B2, respectively).

To identify new candidate CytR regulon members, we used three recognition rules to select regions for construction and manual analysis of alignments, requiring either (1) two candidate O_{CRP} sites at a distance 10–40 bp, or (2) two candidate O_{CYTR} sites at a distance not exceeding 20 bp, or (3) all four sites in the configuration $O_{CRP}D$–$N_{(-10)–20}$–$O_{CYTR}D$–$N_{0–20}$–$O_{CYTR}P$–$N_{(-10)–20}$–$O_{CRP}P$, with negative numbers denoting overlapping sites.

For each set of orthologous genes, both known regulon members and new candidates, we performed multiple alignment anchored at pairs of the O_{CRP} operators. As mentioned above, in many genomes there are no strong candidate O_{CYTR} sites at positions corresponding to the *E. coli* O_{CYTR} operators. To identify possible shifted O_{CYTR} sites in the regions between pairs of O_{CRP} sites we used a variation of the sliding window technique, SWAS (sliding window average score) plots (see Data and Methods).

At each window position, we calculated the average weight of the $O_{CYTR}D$ and, separately, $O_{CYTR}P$ using respective PWMs. Our assumption was that if the position of the CytR-binding site O_{CYTR}, comprising both $O_{CYTR}D$ and $O_{CYTR}P$ and the length of the spacer between them, was conserved within the alignment, the SWAS plot would have two pronounced peaks. On the other

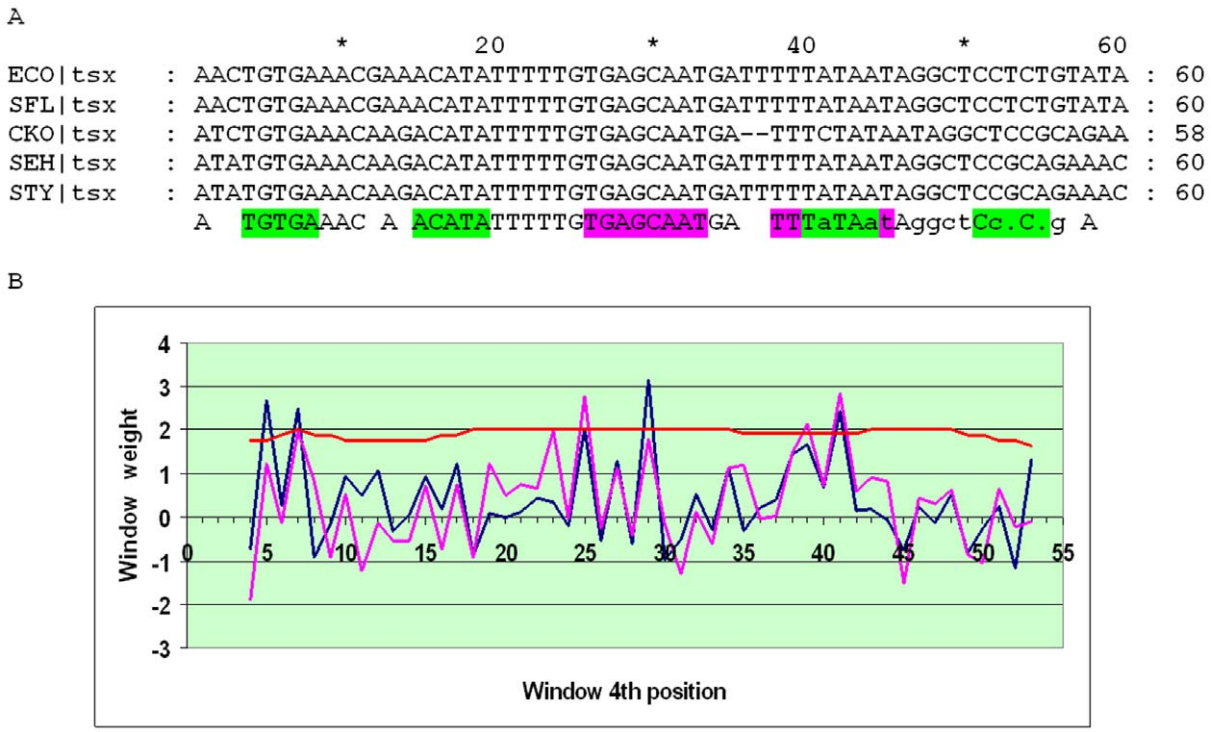

Figure 9. Alignment and SWAS plots of upstream regions of *tsx* in close relatives of *E. coli*. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g009

hand, if $O_{CYTR}D$ and/or $O_{CYTR}P$ shifted in a fraction of genomes, each new position would be represented by a new smaller peak. We accepted a peak if the average score within a window exceeded 3, or a single prominent peak with the score slightly below the average (e.g. about 2.7 for O_{CYTR} of *cdd* or *cytR*, see below). The positional conservation was also assessed using the plots of the information content (see Data and Methods). A SWAS peak was assumed to be more reliable if it was observed in a region of a more or less constant positional conservation.

Evolution of CytR-binding sites

To characterize the conservation of CytR-binding sites, we constructed three groups of alignment of gene upstream regions for closest relatives of *E. coli*, for other Enterobacteriales (in some cases, for all available Enterobacteriales including *E .coli*), and for the Vibrionales, and analyzed these alignments using the SWAS plots. It should be noted that the representation of gene orthologs in genomes varied and, further, in some genomes the intergenic regions diverged beyond recognition. The criteria for the inclusion of upstream regions to alignments were based on the scores of the O_{CRP} sites and the conservation of the distance between them.

The operator cassettes may be classified into four main types by the pattern of conservation observed in the SWAS plots. The first type has two clear peaks that correspond to $O_{CYTR}D$ and $O_{CYTR}P$, yielding conservation of both sites and the distance between them. The second rare type has one clear peak and a diffuse group of scattered minor peaks, reflecting conservation of one O_{CYTR} site and absence or shift of the other one. The third type is characterized by the absence of clear peaks. Finally, the fourth type is two peaks of the same type, reflecting direct rather than inverted repeats. There were few such cassettes, but they also could be conserved to some extent. Note that the above definitions may depend on the number and similarity of sequences in an alignment: the closer they are, the more likely the respective gene would belong to type 1 rather than to type 3.

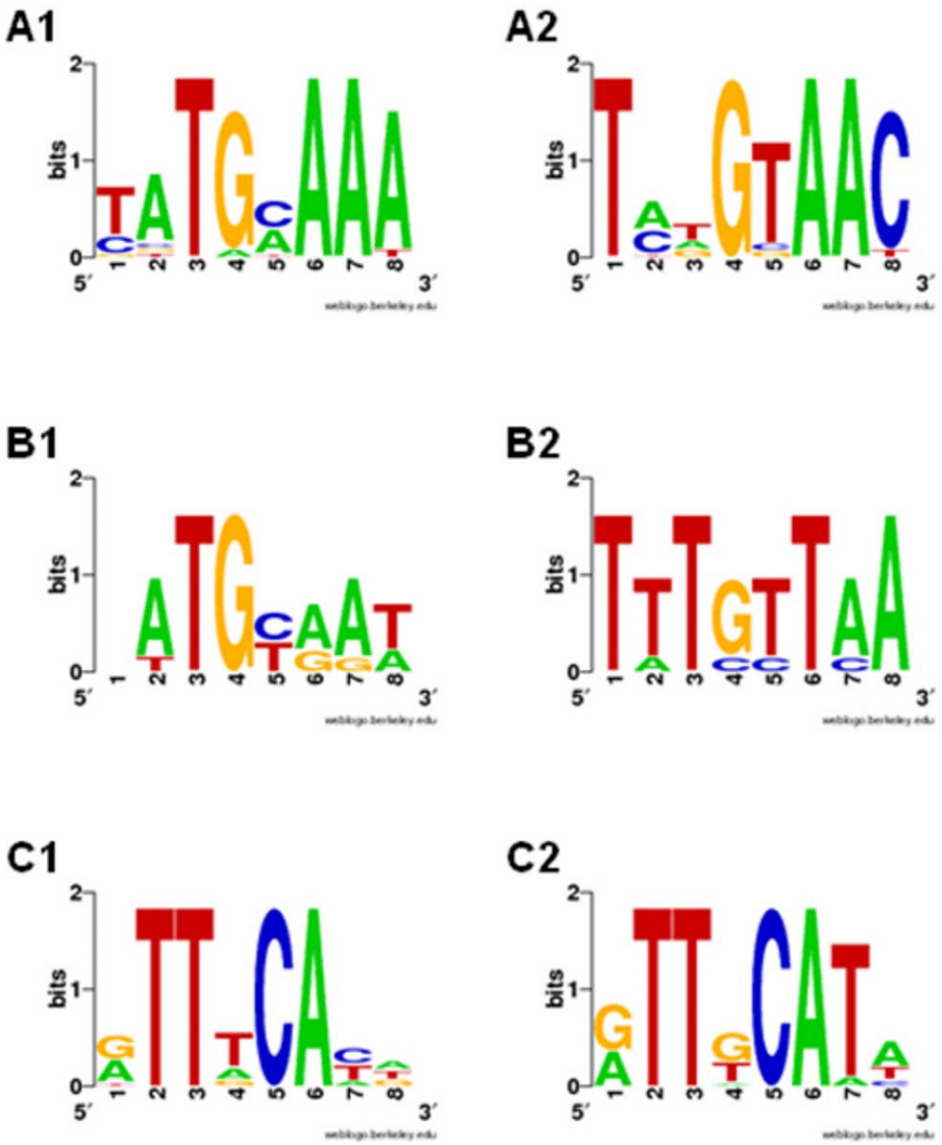

Figure 10. Sequence LOGOs of CytR-binding motifs, direct repeat type. Notation as in Fig. 3. A) $O_{CYTR}D$ LOGO for *cytR* from 16 Enterobacteriales; B) $O_{CYTR}D$ LOGO for *cytR* from 6 Vibrionales; C) $O_{CYTR}P$ LOGO for *cdd* from 14 Enterobacteriales.
doi:10.1371/journal.pone.0044194.g010

The *udp* gene encodes uridine phosphorylase in many Enterobacteriales and Vibrionales. The detailed structure of the *udp* cassette in *E. coli* was studied in [29]. The distance between the O_{CRP} sites in the *udp* promoter is conserved (30 or 31 bp) in almost all Enterobacteriales and Vibrionales; the only exceptions with non-conserved intersite distances and an overall low score of the cassette are *Photobacterium profundum*, *Photorhabdus luminescens* and *Vibrio fischeri*. The distances between the candidate O_{CYTR} sites are not constant, and the alignment may be divided into three subalignments. In the SWAS plot of close relatives of *E. coli*, two pronounced peaks corresponding to O_{CYTR}D and O_{CYTR}P are visible (Fig. 5). In more distant Enterobacteriales and in the Vibrionales, no clear peaks are seen, and there are many genome-specific non-conserved candidate O_{CYTR} sites, some overlapping with O_{CRP}, that cannot be confidently predicted based on the sequence analysis alone (Fig. 6 and Fig. 7, respectively). Hence, the *udp* cassette is of type 1 at close distances and of type 3 at more distant ones.

The *deoC* gene encodes NAD(P)-linked 2-deoxyribose-5-phosphate aldolase. Prominent SWAS-plot peaks are observed in close relatives of *E. coli* where, unusually, there is no spacer between the O_{CYTR} sites: O_{CYTR}D is immediately adjacent to O_{CYTR}P (Fig. S1 and [29]). In more distant Enterobacteriales (*Edwardsiella ictaluri*, *Dickeya dadantii*, *Erwinia tasmaniensis* and *Klebsiella pneumoniae*) no clear peaks are seen (Fig. S2). In all Vibrionales including *P. profundum*, two *Yersinia* species (*Y. pestis* and *Y. pseudotuberculosis*) and *Pectobacterium atrosepticum*, the 4-box cassettes had very low total weights (about 10 or even less) and variable distances between the O_{CRP} sites, and hence the respective regions were not included into the alignments. Thus again we have type 1 behavior at close, and type 3 at distant Enterobacteriales.

Very low scores of 4-site cassettes were observed for most *ppiA* (peptidyl-prolyl cis-trans isomerase A) gene orthologs. Nevertheless, for six closest *E. coli* relatives, two peaks in the SWAS plot are clearly seen (Fig. S3), therefore the *ppiA* cassette is of type 1 at close Enterobacteriales, similar to previously characterized *ppiA* cassette in *E. coli*, see [9] (Fig. 1.B, p. 990).

The *rpoH* gene encodes the heat-shock sigma-factor (sigma-32 or σ^H). The *E. coli* cassette was described in [13]. The 4-site cassette scores for this gene are rather low, mainly because of the low scores of the O_{CRP} sites. Moreover, the scores in most Enterobacteriales are lower than those in *E. coli*. However, the SWAS plot features two clear peaks (Fig. S4), thus the *rpoH* cassette belongs to type 1 in close *E. coli* relatives.

A

```
                    *         20         *         40         *         60
ECO|cytR    : AAATTCAATATTCATCACACTTTTCATGAAAATTCTGTAACCGTTTTCACGCGCTATCTGCTAAAAATG : 69
SFL|cytR    : AAATTCAATATTCATCACACTTTTCATGAAAATTCTGTAACCGTTTTCACGCGCTATCTGCTAAAAATG : 69
CKO|cytR    : AAATTAAATATCCATCACACTTTTTATGAAAATTCTGTAACCGCTTCCATTAACTATCTGATAAAAATG : 69
SEH|cytR    : AAATTAAAGATCCATCACACTTTTTATGAAAATTCTGTAACCGCTTCCACGCACTATCTGATAAAAATG : 69
STY|cytR    : AAATTAAAGATCCATCACACTTTTTATGAAAATTCTGTAACCGCTTCCACGCACTATCTGATAAAAATG : 69
EIC|cytR    : AAAATCAATTCATATCACACTTTCCCTGCAAATTAAGTAACTGTCACCCCTGATGTTCCGTTTAATATG : 69
DDC|cytR    : AAAACCGATAATCCAAACGGCACAGGCGTAAATTATTCGATAAATCCTGGGCGCAATCTGTTAAGAGTG : 69
PEB|cytR    : GAACACGATATTCATCACACTTTCTGTGCAAATTAAGCAACAAAACCTGCGTTCAATCTGTTATCACTG : 69
SPE|cytR    : AAAACCTTCCCAGGTCACACTTTTTATGCAAATTAGGTAACAAAGGCAGGACGTATTCTGTTTAGAATA : 69
YPN|cytR    : AAAACCTACTTACGTCACATTTTTTATGCAAATTAGGTAACAAATTACCCCCACTATCTGTATAGAATG : 69
YPS|cytR    : AAAACCTACTTACGTCACATTTTTTATGCAAATTAGGTAACAAATTACCCCCACTATCTGTATAGAATG : 69
PLU|cytR    : ATAAATGATTTATATCACACTTT-TTTACAAACTAAGGAATAAATGAATTTGCCAAACTGTTTAGAATG : 68
PMR|cytR    : AAAAACGATTGGTATCACACTTTTTATGCAAACTAGGTAACAGTATCTATTCAAAAAGTGGGTAAAATA : 69
ETA|cytR    : AAAAGTCGCACAAGTCACAAATTTTATGCAAATTAAGTAAC-ATCTTGCCGCTATATTTGTTAACAATG : 68
CTU|cytR    : AAATTTAATATCCGTCACATCTTCGATGTAATTTTTGTAACTGGCG-CGCCCAATATCTGCTAAGAATG : 68
ENT|cytR    : AAATCAAACGCTCCTCACACTTTTTATGAAAATTCTGTAACCGCTTCCATTCACTATCTG-TAAAAATG : 68
KPE|cytR    : AAATTCAACAAGCATCACGTTTTTCATGAAAATTCTGTAACTCACTGCCGCGACTATCTGATAAAAATG : 69
              aaA           tcACa       ..tg.AA.tT..g.aAc
```

B

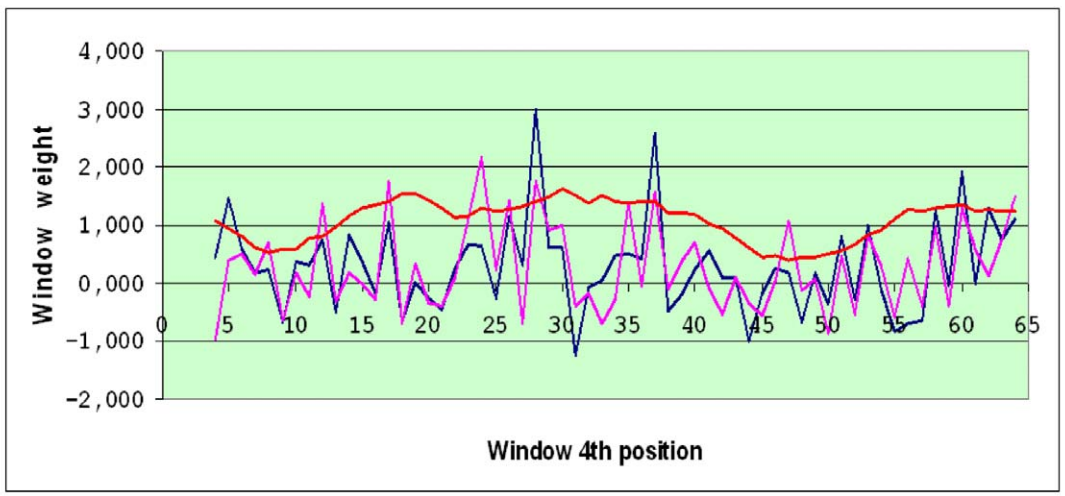

Figure 11. Alignment and SWAS plots of upstream regions of *cytR* in the Enterobacteriales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g011

A

```
                    *         20         *         40         *         60
PP1|cytR  :  CAATGCGATGTAAGTCACGTATTGAATGTAATTTATGCTCATAAAGCCATGTTTAATCAACAA  :  63
VC1|cytR  :  TATCGTGATGAAAATCACAGTGACTATGCAAAATTTGTTAATCTCGCCGTAACTTTCTGTTCT  :  63
VH1|cytR  :  TAATGTGATGAGATTCACATTAGGATTGCAATTTTTGTTAAACAGACCGTAACTTTCCCTTTA  :  63
VV1|cytR  :  TTACGTGATGAACATCACGAATGAAATGCGGATTTTGTTAATTCGCCCGTAACTTTAAATAAA  :  63
VF1|cytR  :  TCATGTGAGAGAAATCACAATACAGATGTAATTTTTGTTAAACGAGCCGTATTTAAATCTTTA  :  63
VS1|cytR  :  TCTCGTGAGCAACTTCACACTAGTCATGTGATTTTTGTTAATCTAGCCGTAACTTTACATGGG  :  63
               .GtGA       TCACa       .aTG..a.tTtTGtTaA      CCgTa  T     t
```

B

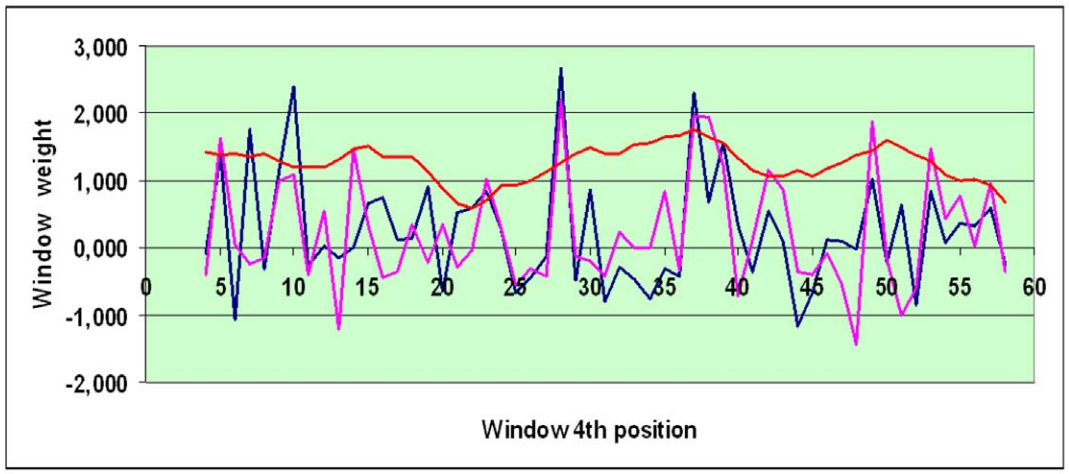

Figure 12. Alignment and SWAS plots of upstream regions of *cytR* in the Vibrionales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g012

The *nupG* gene encodes one of two high affinity nucleoside transporters in *E. coli*. It is present in seven genomes, the fewest among all considered genes. Further, the gene annotated as *nupG* in *Salmonella enterica* Heidelberg is in fact *xapB*, encoding xanthosine MFS transporter [41], as demonstrated by the analysis of phylogenetic trees (not shown) and co-localisation with *xapA*, the latter encoding a subunit of xanthosine phosphorylase. In *K. pneumoniae*, the total score of the best O_{CRP} pair is too low (about 7.1), and although the distance between them (30–31 nucleotides) is not sharply different from that in other genomes (27–28 bp), it is

A

```
                    *         20         *         40         *         60         *
CK0|nupC  :  AAATTAAACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGTCCTGGCCCTCGCTGTT  :  70
EC0|nupC  :  AAATTTAACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGTACTGGCACTTGCCGTT  :  70
SFL|nupC  :  AAATTTAACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGTACTGGCACTTGCCGTT  :  70
SEH|nupC  :  AAATTAAACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGTCCTGGCGCTTGCCGTT  :  70
STY|nupC  :  AAATTAAACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGTCCTGGCGCTTGCCGTT  :  70
             AAATT.AACCTGCCTCATATTTGGAGCAAATATGGACCGCGTCCTTCATTTTGT.CTGGC CTtGCcGTT
```

B

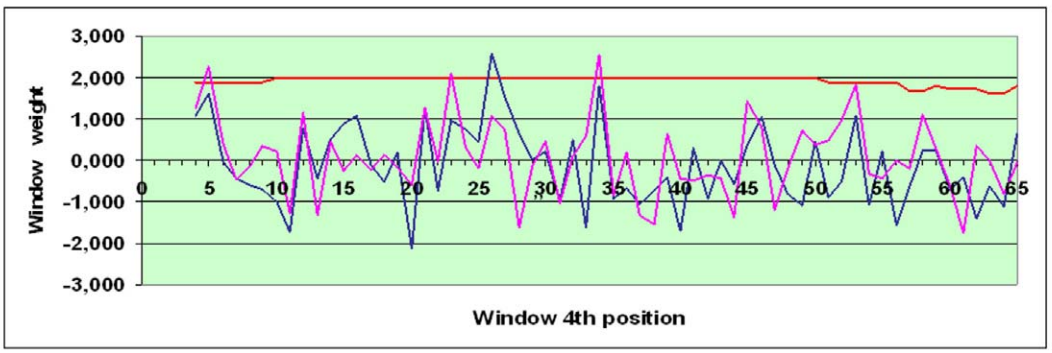

Figure 13. Alignment and SWAS plots of upstream regions of *nupC* in close relatives of *E. coli* (inverted O_{CYTR} repeats). Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g013

A

```
                    *         20         *         40         *         60         *
CKO|ycdZ    : AAATGTGACCGTTGTCACTTTGTTGTGCAACGTGATTAATTGTCTTGCATAAAATGTGTGCGTGATCTCATTC : 73
SEH|ycdZ    : TAATGTGATGTAAATCACCATTAACTGCAACGAGTTTTACATCATTGCATAAAATGTGTGCGCGATCTCATTA : 73
STY|ycdZ    : TAATGTGATGTAAATCACCATTAACTGCAACGGGTTTTGCATCATTGCATAAAATGTGTGCGTGATCTCATTA : 73
ECO|ycdZ    : AAGTGTGATCTACGTCACTCATAACTGCAACGGATAATTTGTTGTTGCATAAAATGTGTGCTCGATCTCATTC : 73
SFL|ycdZ    : AAATGTGATCATCGTCACTCATAACTGCAACAGATAATTTGTTGTTGCATAAAATGTGTGCTCGATCTCATTC : 73
                a TGTGA       tCACN   taacTGCAAC    t   t    t  .TTGCATAAAATGTGTGC  GATCTCATT
```

B

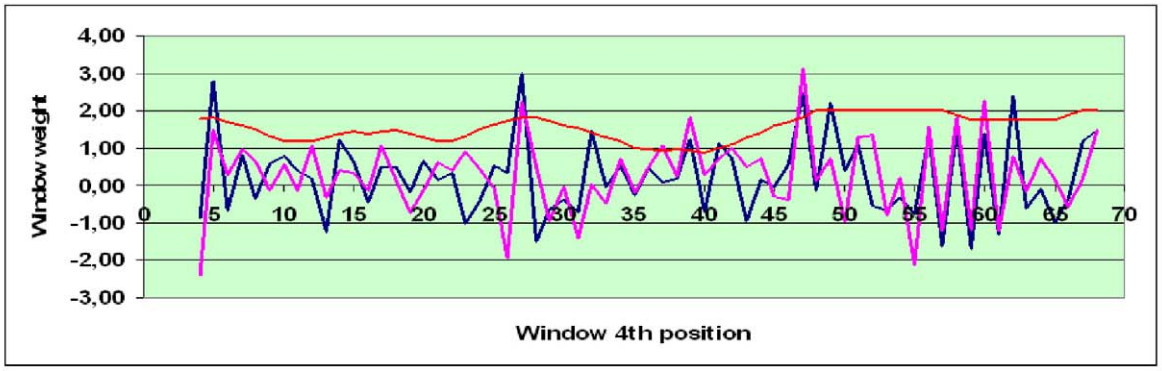

Figure 14. Alignment and SWAS plots of upstream regions of *ycdZ* in close relatives of *E. coli*. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g014

likely that the regulation of *nupG* in *K. pneumoniae* has been lost. The SWAS plot has two pronounced peaks corresponding to the $O_{CYTR}D$ and $O_{CYTR}P$ sites with the conserved distance of 9 bp between them and an overlap between the latter and the proximal O_{CRP} site (Fig. 8 and [39]). Hence the *nupG* cassette belongs to type 1.

The *tsx* gene, encoding the nucleoside channel, is present in many bacteria up to the Vibrionales. While the distance between the O_{CRP} sites flanking O_{CytR} is mostly conserved for the *tsx* orthologs in the Enterobacteriales, about 14 bp, the score of the 4-box cassette even in close relatives of *E. coli* is rather low, due to low scores of the O_{CRP} sites (about 3). However, the SWAS plot features two pronounced peaks, and hence the cassette is of type 1, although $O_{CYTR}P$ overlaps $O_{CRP}P$ (Fig. 9). The predicted sites in *E.coli* differ slightly from those suggested earlier (Fig. 9) here and [38] (Fig. 10, p.33253). Out of four other Enterobacteriales with *tsx* orthologs (*E. ictaluri*, *Enterobacter 638*, *K. pneumoniae*, *Serratia proteamaculans*) only three yield a relatively satisfactory alignment with the corresponding SWAS plot of type 2 (Fig. S5). Only four of the Vibrionales have *tsx* orthologs (*Vibrio harvey*, *Vibrio parahaemolyticus*, *Vibrio splendidus*, *Vibrio vulnificus*) (Fig. S6). Since a high-scoring $O_{CYTR}P$ peak is situated relatively close to $O_{CRP}D$ and all other peaks have very low scores (about 2), this case is assigned to type 3, as neither $O_{CYTR}D$ nor $O_{CRP}P$ can be reliably identified.

A

```
                    *         20         *         40         *         60         *
SPE|ycdZ    : TTTTGTGACAAATTCCACATC---TCTGTAATTATCTTGTTTCTCATTACATATTTTATGTGAAATACATCACATGA : 74
CTU|ycdZ    : AAATGTGATATTTGTCACATTTAAAATGTAACTCACT-GTACGTTATTTCATA--TGACGTGCGGCAGATCGCATTT : 74
ENT|ycdZ    : AAGAGTGACACTCGTCACATTTCAAATGTAACAGTGTA-TTTTGGTTTGCATA--AAATGTGCTCATGATCCCATTT : 74
KPE|ycdZ    : ATATGTGATGTTAATCACTATTTGAATGCAACTGAAAT-TAGGTTCATTCATA--TACTGTGTTTTAGATCTCATTC : 74
                tGTGA       tCACa      aaTGtAAc                 T CATA      tGTG.       TC.CA
```

B

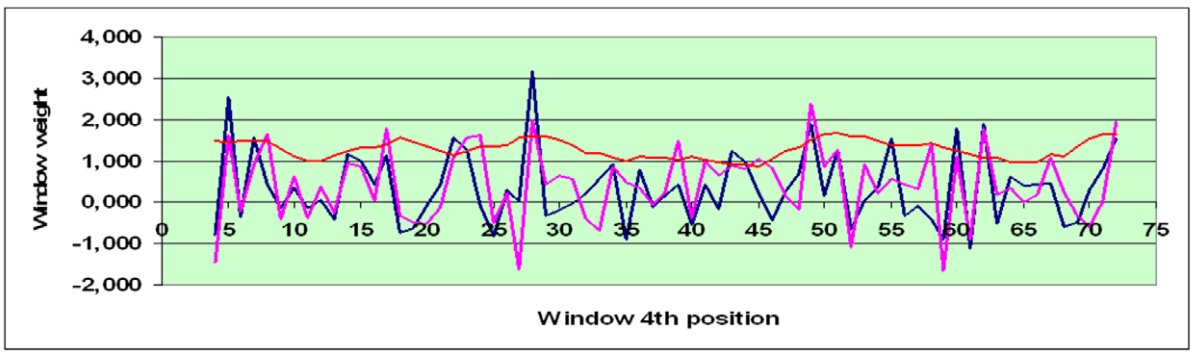

Figure 15. Alignment and SWAS plots of upstream regions of *ycdZ* in distant Enterobacteriales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g015

The *cytR* gene itself has only the distal O_{CRP} site. On the other hand, the bound complex has been observed in *E. coli K-12* [6] and *S. typhimurium* [14]. The O_{CYTR} site is often assumed to be an imperfect inverted repeat [36], but the alignment of operator cassettes from sixteen Enterobacteriales and, separately, six Vibrionales shows that the O_{CYTR} site is an imperfect direct repeat (Fig. 10). At that, the Enterobacteriales and Vibrionales seem to have conserved organization of the O_{CRP}D-O_{CYTR}D-O_{CYTR}P recognition site, but slightly different sequences of O_{CYTR}D and O_{CYTR}P. The unusual properties of this cassette may explain the fact that the scores of the O_{CYTR} sites are low, less than 3. However, the conservation of these sites in the alignment provides the evidence for their functional relevance (Fig. 11 and Fig. 12).

Another atypical cassette is that of *cdd*, cytidine/deoxycytidine deaminase. It contains O_{CRP} sites flanking direct repeats of O_{CYTR} [8,29]. Two O_{CRP}D sites denoted in the literature and here, in this particular case, O_{CRP}2 and O_{CRP}3 overlap by 20 bp, that is, they are shifted relative to each other by 2 bp. The arrangement of the sites is conserved in 14 genomes (Fig. S7 and Fig. S8).

To analyze direct repeats in the two latter cases, belonging to type 4, we applied the standard matrices for O_{CYTR}D and O_{CYTR}P (Fig. 3B1 and Fig. 3B2, respectively) and selected the matrix providing two highest SWAS-plot peaks. Both site cassettes have 1 bp spacers.

In the two cases of direct repeats, for the *cytR* cassette, pronounced SWAS-plot peaks were observed for the O_{CYTR}D PWM both for the Enterobacteriales and Vibrionales (Fig. 11 and Fig. 12, respectively), whether for the *cdd* cassette visible peaks are produced by O_{CYTR}P PWM for the Enterobacteriales, whereas none of the two matrices provides anything definite for the Vibrionales (Fig. S7 and Fig. S8, respectively).

NupC is a nucleoside transporter. It is unrelated to NupG and shows somewhat different specificity: unlike universal NupG, it does not transport guanosine and deoxyguanosine [42]. The *nupC* gene was proposed to be regulated by CytR based on its function in the nucleoside transport, similar to some other genes from the CytR regulon [43], and the location of candidate pentameric binding sites [10]. While the alignment of the *nupC* upstream regions of five closest relatives of *E. coli* contains conserved regions, they have very low O_{CRP} scores. Candidate O_{CYTR} sites are seen in the alignment as inverted repeats at a zero distance (Fig. 13). The corresponding peaks at the SWAS plot are weak (~2.6) but clearly visible. The score of O_{CRP}D is about 4, which is consistent with a usual model of regulation of promoters with two CRP-binding sites. An alternative is O_{CYTR} being a direct repeat with a 3 bp spacer, close to the one observed in a SELEX experiment for direct repeats [37]. In this case the score of one of the peaks is larger than 3 and the score of the second peak is about 2.5, both assessed by the O_{CYTR}P PWM (Fig. S9). Finally, there is a possibility that weaker O_{CRP} sites, in particular the one overlapping the transcription start site, also participate in formation of the regulatory complex. An experiment is needed to validate the predicted site and to select between the alternative descriptions of the cassette structure.

New candidate member of the CytR regulon

As the initial criterion for identification of new possible operator cassettes, we relied on conservation of the distance between candidate O_{CRP} sites and the presence of peaks in the SWAS plots, demonstrating conservation of the O_{CYTR} positions. We started with identification of *E. coli* genes preceded by high-scoring O_{CRP}D-O_{CYTR}D–O_{CYTR}P–O_{CRP}P cassettes. We required that the score of each cassette exceeded the minimal observed score for the known genes (cut-off 12.6) and that the distance between O_{CRP} sites was in the interval (10–40). As expected, the initial four genes used to construct the PWMs (*deoC*, *nupG*, *ppiA*, *udp*) had high total scores and were among the leaders in the list ordered by decrease of the total O_{CRP}-O_{cytR}D-O_{cytR}P-O_{CRP} score. We selected 37 *E.coli* genes satisfying these criteria, listed in Table S2.

Then we identified orthologs of these genes and checked the presence of a pair of O_{CRP} sites at approximately the same distance in at least five genomes. After that we aligned the promoter regions, anchored at O_{CRP}D and O_{CRP}P, and applied

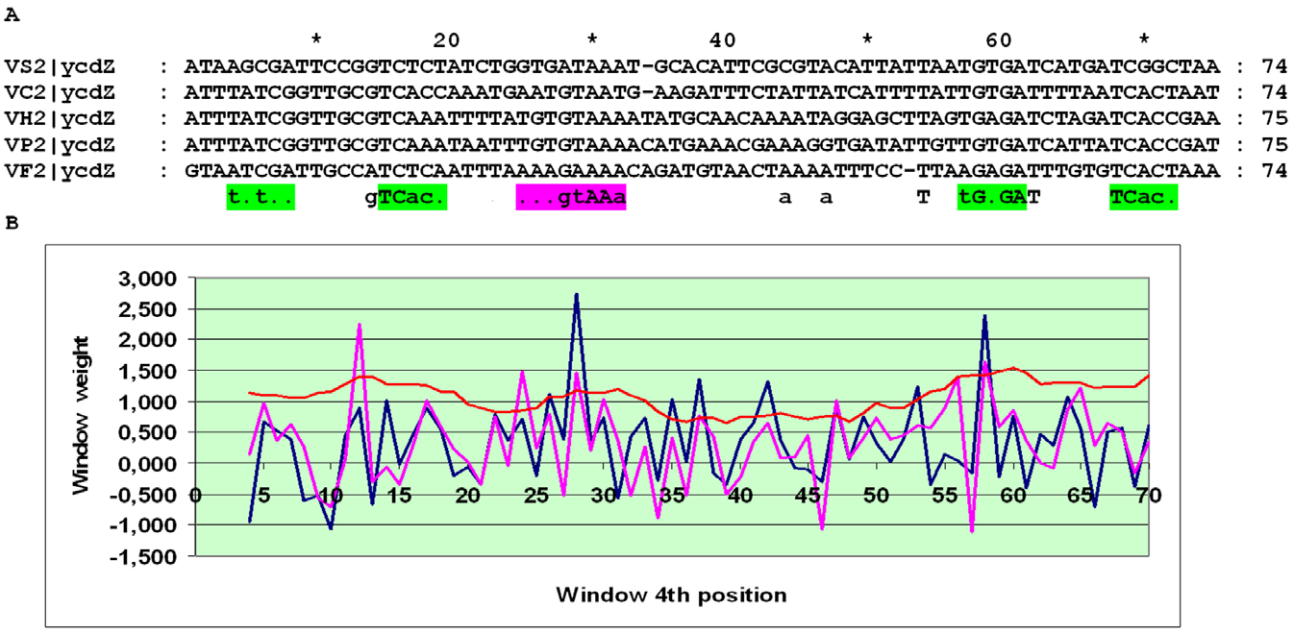

Figure 16. Alignment and SWAS plots of upstream regions of *ycdZ* in the Vibrionales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g016

the O_{CYTR}D and O_{CYTR}P PWMs, constructing SWAS plots for the spacer between O_{CRP}D and O_{CRP}P.

One strong candidate emerged from this analysis. The *ycdZ* gene of *E .coli* is preceded by a cassette formed by two O_{CRP} sites at a conserved distance (29–31 bp) and O_{CYTR} sites in the correct arrangement, and this cassette is conserved in 17 related genomes, that is, in almost all Enterobacteriales and Vibrionales. The exceptions were *D. dadantii*, *E. tasmaniensis*, *P. atrosepticum*, *Y. pestis* and *Yersinia pseudotuberculosis*, where this gene is simply absent, and *V. vulnificus* that has an atypical O_{CRP}–O_{CRP} distance.

The alignment of the *ycdZ* upstream regions may be divided into three subalignments. In close relatives of *E. coli*, two pronounced peaks in the SWAS plots, corresponding to O_{CYTR}D and O_{CYTR}P, are visible (Fig. 14). In more distant Enterobacteriales, one clear peak is seen (Fig. 15). In the Vibrionales, one peak is visible, but its average score is less than 3 (Fig. 16). Thus, the *ycdZ* cassette belongs to type 1 at close distances and to type 2 in more distant Enterobacteriales and in the Vibrionales. Hence we predict that *ycdZ* is a member of the CytR regulon.

The encoded protein YcdZ is an inner-membrane protein from the DUF1097 family. According to TMHMM (see Data and Methods) it has five transmembrane domains (Fig. 17). Hence YcdZ is likely to be a transporter. We suggest naming it NupT.

Discussion

We have observed that the distances between candidate O_{CRP} sites are conserved in upstream regions of orthologous genes regulated by the CRP-CytR complex. On the other hand, positions of the O_{CYTR} sites seem to be conserved only at close evolutionary distances, as the highest-scoring candidate sites may occupy different positions in distant Enterobacteriales and in the Vibrionales (e.g. Fig. 18). One possible explanation for that, discussed in the literature, is that the binding of CytR to DNA has very low specificity, and the regulation is based on the formation of multimetric CRP-CytR complexes stabilized by the CytR-DNA interaction [44]. However, the existence of the CytR-binding motif, albeit weak, as well as the intergenome conservation (higher than background) of O_{CYTR} sites argues against this explanation. This is represented by peaks in the SWAS plots.

On the other hand, the intergenome conservation of the distances between O_{CRP} sites in promoters of specific genes together with intragenome differences between genes and relatively low conser-

TMHMM result

```
# YcdZ_ECO Length: 163
# YcdZ_ECO Number of predicted TMHs:  5
# YcdZ_ECO Exp number of AAs in TMHs: 105.864
# YcdZ_ECO Exp number, first 60 AAs:  37.34668
# YcdZ_ECO Total prob of N-in:        0.30359
# YcdZ_ECO POSSIBLE N-term signal sequence
YcdZ_ECO TMHMM2.0 outside      1    19
YcdZ_ECO TMHMM2.0 TMhelix     20    42
YcdZ_ECO TMHMM2.0 inside      43    48
YcdZ_ECO TMHMM2.0 TMhelix     49    68
YcdZ_ECO TMHMM2.0 outside     69    71
YcdZ_ECO TMHMM2.0 TMhelix     72    94
YcdZ_ECO TMHMM2.0 inside      95    98
YcdZ_ECO TMHMM2.0 TMhelix     99   118
YcdZ_ECO TMHMM2.0 outside    119   122
YcdZ_ECO TMHMM2.0 TMhelix    123   145
YcdZ_ECO TMHMM2.0 inside     146   163
```

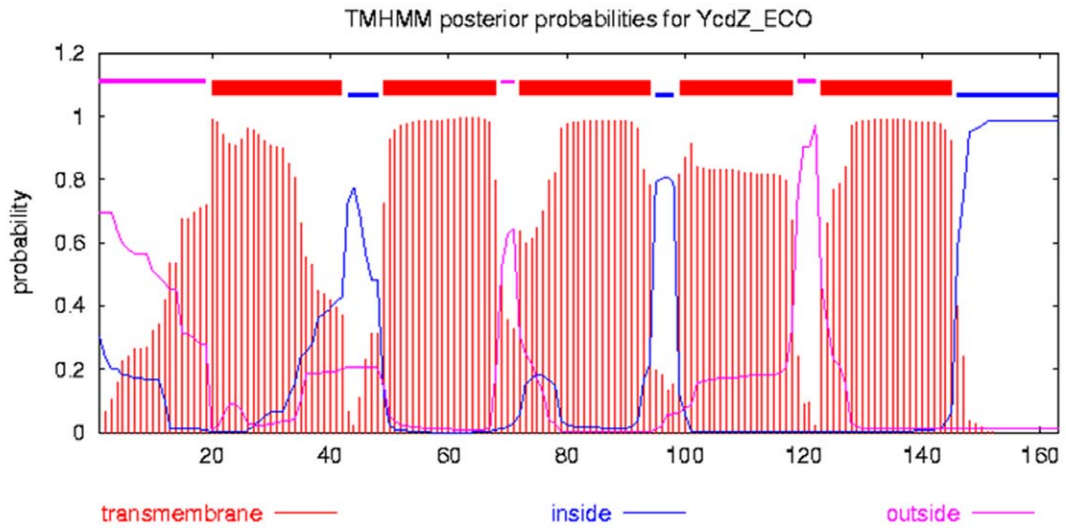

Figure 17. TMHMM predictions for YcdZ. Five transmembrane domains are predicted by TMHMM.
doi:10.1371/journal.pone.0044194.g017

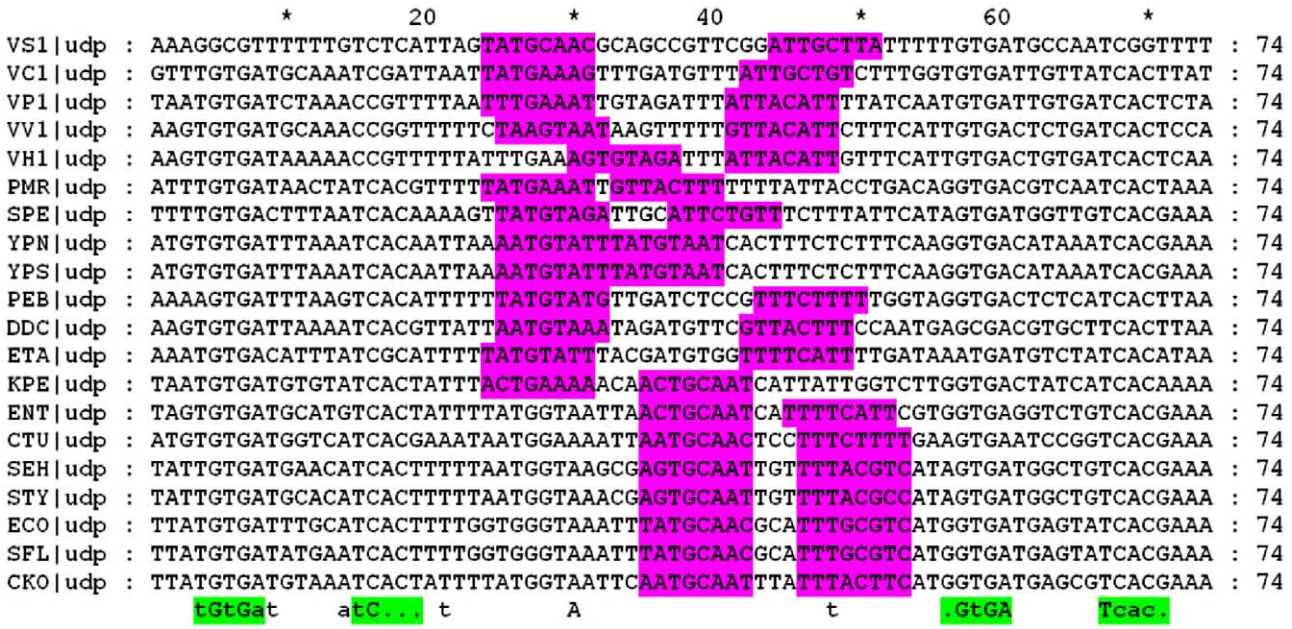

Figure 18. Alignment of upstream regions of *udp* with predicted O_{CYTR} sites, flanked by O_{CRP} sites in the Enterobacteriales and Vibrionales. Notation as in Fig. 5.
doi:10.1371/journal.pone.0044194.g018

vation of positions of candidate O_{CYTR} sites demonstrate that the structure of the complex is dictated by CRP molecules.

The problem of identification of the CytR-binding motif is not trivial either. Indeed, the experimental data do not define the binding sites up to nucleotide: the most commonly used method, DNA footprinting, leaves some uncertainty about the site extent and location [45]. When the motif is strong, it is simple to align the footprinted regions and identify the common core. However, for weak motifs this is far from being straightforward, and we believe that evolutionary considerations yielding the phylogenetic footprinting techniques also deserve attention.

Finally, the overall structure of the CytR-binding site may vary. In most cases, it is an inverted repeat with a variable spacer. However, as shown in a SELEX experiment with the *deo* operator, both inverted repeats of O_{CYTR} boxes with a large spacer (10 to 13 bp) and direct repeats in either direction (O_{CYTR}D or O_{CYTR}P) with a short spacer (1 bp) may be bound by CytR [37]. Direct repeats in the O_{CYTR}D orientation were observed to be conserved in the operators of *cytR* and *cdd*, again, with short spacers.

The comparative analysis also enables to identify new regulon members even for regulators with weak motifs. Of course, the predicted CytR regulation of the *nupT (ycdZ)* gene requires experimental verification.

Data and Methods

Complete genome sequences of the Enterobacteriales and Vibrionales in the gbk format were downloaded from GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) [46].

Multiple alignments of protein and DNA sequences were constructed using Muscle 3.6 (http://www.drive5.com/muscle/) [47] and visualized and manually edited using GeneDoc Editor version 2.6.002 (http://www.nrbsc.org/gfx/genedoc/; Nicholas, Karl B and Nicholas, Hugh B. Jr. 1997, GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors). Protein sequence database searches were performed using the latest version of BLASTP (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.25) [48], . All searches were run against the non-redundant protein sequence database at the NCBI. Positional information content of nucleotide alignments was calculated using ad hoc programs. Transmembrane segments were identified by TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) [49] and candidate N-terminal signal sequences were analyzed with SWMSignal (http://bio-cluster.iis.sinica.edu.tw/SVMSignal) [50].

The maximum-likelihood (ML) tree was constructed using morePHYML, (http://mobyle.pasteur.fr/cgi-bin/portal.py?%23forms::phyml#forms::morePhyML) [51]. The circular tree was built by MEGA 5.1 [52].

Motif identification, construction of recognition profiles, identification of candidate sites in genome sequences and protein similarity searches using the Smith–Waterman algorithm were performed using Genome Explorer, version 3 [53], modified by L.V. Lunovskaya and A. Shpilman. Candidate sites were identified in the interval (−300,+200) relative to the start codon. Positional weight matrices were constructed using SignalX (http://bioinf.fbb.msu.ru/SignalX) [54]. Sequence logos were constructed by WebLOGO (http://weblogo.berkeley.edu/logo.cgi) [55] .

The Positional Weight Matrix (PWM) was defined via:

$$W(\beta,\kappa)=\log[N(\beta,\kappa)+0{,}5]-0{,}25\sum_{\alpha=A,C,G,T}\log[N(\alpha,\kappa)+0{,}5],$$

where $W(\beta,\kappa)$ is the positional weight of nucleotide β at position κ of the PWM, $N(\beta,\kappa)$ is the count of nucleotide β at position κ in the training sample.

The sum of the positional weights for a site yields the site score:

$$Z(\beta_1\ldots\beta_k)=\sum_{\kappa=1}^{K}W(\beta_\kappa,\kappa).$$

Sliding window average score (SWAS) plots were constructed as follows. The upstream regions of each gene from the CytR regulon

were aligned in the bacterial groups with approximately constant CRP-CRP distance. Within a sliding window of size 8 nt, the average score of sites within the window was calculated using respective PWMs. Scores of sequences containing gaps in a given window position were set to 0, but these sequences were counted for averaging; hence, positions with gaps were penalized.

The positional information content was calculated as

$$I_i = 2 + \sum_{\alpha \in \{A,C,G,T\}} f(\alpha,i)\log_2(f(\alpha,i)),$$

where $f(\alpha, i)$ is the frequency of nucleotide α in the alignment position i.

Supporting Information

Figure S1 Alignment and SWAS plots of upstream regions of *deoC* in close relatives of *E. coli*. Notation as in Fig. 5.
(TIF)

Figure S2 Alignment and SWAS plots of upstream regions of *deoC* in distant Enterobacteriales. Notation as in Fig. 5.
(TIF)

Figure S3 Alignment and SWAS plots of upstream regions of *ppiA* in close relatives of *E. coli*. Notation as in Fig. 5.
(TIF)

Figure S4 Alignment and SWAS plots of upstream regions of *rpoH* in close relatives of *E. coli*. Notation as in Fig. 5.
(TIF)

Figure S5 Alignment and SWAS plots of upstream regions of *tsx* in distant Enterobacteriales. Notation as in Fig. 5.
(TIF)

Figure S6 Alignment and SWAS plots of upstream regions of *tsx* in the Vibrionales. Notation as in Fig. 5.
(TIF)

Figure S7 Alignment and SWAS plots of upstream regions of *cdd* in distant Enterobacteriales (direct O_{CYTR} repeats). Notation as in Fig. 5.
(TIF)

Figure S8 Alignment and SWAS plots of upstream regions of *cdd* in the Vibrionales. Notation as in Fig. 5.
(TIF)

Figure S9 Alignment and SWAS plots of upstream regions of *nupC*-in close relatives of *E. coli* (direct O_{CYTR} repeats). Notation as in Fig. 5.
(TIF)

Table S1 The list of orthologs of CytR-regulated genes from *E.coli* that have nearly constant O_{CRP}-O_{CRP} distances. & – "yes(value)": the corresponding ortholog exists and the value in parentheses is the O_{CRP}-O_{CRP} distance && – 0: no ortholog. # – "no(abbreviation)", the reason why the upstream region was not considered: 1 – (single) for *cytR* means only one O_{CRP}; 2 – (pstn), atypically distant start position for the group; 3 - (wght): weak maximal weight of the O_{CRP}-O_{CRP} operator pair; 4 - (dist): distance larger and smaller than ±2 nucleotides compared with the average distance for the group; 5 - (mist): probable misannotation (*nupG* in *Salmonella enteric* Heidelberg and *cytR* in *Edwardsiella ictaluri*) 6 - (triple): no triple O_{CRP}-O_{CRP}-O_{CRP} in the specified region. @ – IIc: the second chromosomes for all *Vibrio* – spp. and *Photobacterium profundum* * – candidate member of the CytR regulon ** – exceptional genes added as known from the literature co – cut-off, the smallest score of known cassettes for the respective gene.
(TIF)

Table S2 O_{CRP}D-O_{CYTR}D-O_{CYTR}P-O_{CRP}P cassettes of probable CytR regulon members in *E. coli*. 1 – O_{CRP}D is the distal CRP-operator with respect to the transcription start. 2 – O_{CYTR}D is the distal CytR-operator with respect to the transcription start. 3 – O_{CYTR}P is the proximal CytR-operator with respect to the transcription start. 4 – O_{CRP}P is the proximal CRP-operator with respect to the transcription start. 5 – *spr* is spacer length. 6 – Σ*score* is total score of the cassette. 7 – site score is in parentheses. $ – *start pos* is the start position of the cassette in the respective upsteam region. @ – no direct repeats for O_{CyTR}D-O_{CRP}P, only inverted ones; * – known regulon member with experimentally determined cassette; # – predicted regulon member, predicted cassette; #* – known regulon member, predicted cassette.
(TIF)

Acknowledgments

We are grateful to L.V. Lunovskaya and A. Shpilman for improvements in the Genome Explorer program, to A.A. Mironov for fruitful discussions, and to L.N. Serebrennikova for editing the manuscript.

Author Contributions

Conceived and designed the experiments: MSG. Performed the experiments: NVS. Analyzed the data: NVS MSG. Contributed reagents/materials/analysis tools: NVS MSG. Wrote the paper: NVS MSG .

References

1. Hammer-Jespersen K, Munch-Petersen A (1975) Multiple regulation of nucleoside catabolizing enzymes: regulation of the *deo* operon by the *cytR* and *deoR* gene products. Mol Gen Genet 137: 327–335.
2. Singer JT, Barbier CS, Short SA (1985) Identification of the *Escherichia coli deoR* and *cytR* gene products. J Bacteriol 163: 1095–1100.
3. Valentin-Hansen P, Larsen JE, Højrup P, Short SA, Barbier CS (1986) Nucleotide sequence of the *cytR* regulatory gene of *E. coli* K-12. Nucleic Acids Res 14: 2215–2228.
4. Søgaard-Andersen L, Martinussen J, Møllegaard NE, Douthwaite SR, Valentin-Hansen P (1990a) The CytR repressor antagonizes cyclic AMP-dependent AMP receptor protein activation of the *deoC*p2 promoter of *Escherichia coli* K-12. J Bacteriol 172: 5706–5713.
5. Søgaard-Andersen L, Møllegaard NE, Douthwaite SR, Valentin-Hansen P (1990b) Tandem DNA-bound cAMP-CRP complexes are required for transcriptional repression of the *deo*P2 promoter by the CytR repressor in *Escherichia coli*. Mol Microbiol 4: 1595–1601.
6. Gerlach P, Valentin-Hansen P, Bremer E (1990) Transcriptional regulation of the *cytR* repressor gene of *Escherichia coli*: autoregulation and positive control by the cAMP/CAP complex. Mol Microbiol 4: 479–488.
7. Gerlach P, Søgaard-Andersen L, Pedersen H, Martinussen J, Valentin-Hansen P, et al. (1991) The cyclic AMP (cAMP)-cAMP receptor protein complex functions both as an activator and as a corepressor at the *tsx*-p2 promoter of *Escherichia coli* K-12. J Bacteriol 173: 5419–5430.
8. Holst B, Søgaard-Andersen L, Pedersen H, Valentin-Hansen P (1992) The cAMP CRP/CytR nucleoprotein complex in *Escherichia coli*: two pairs of closely linked binding sites for the cAMP-CRP activator complex are involved in combinatorial regulation of the *cdd* promoter. EMBO J 11: 3635–3643.
9. Nørregaard-Madsen M, Mygind B, Pedersen R, Valentin-Hansen P, Søgaard-Andersen L (1994) The gene encoding the periplasmic cyclophilin homologue, PPIase A, in *Escherichia coli*, is expressed from four promoters, three of which are activated by the cAMP-CRP complex and negatively regulated by the CytR repressor. Mol Microbiol 14: 989–997.

10. Craig JE, Zhang Y, Gallagher MP (1994) Cloning of the *nupC* gene of *Escherichia coli* encoding a nucleoside transport system, and identification of an adjacent insertion element, IS 186. Mol Microbiol 11: 1159–1168.
11. Pedersen H, Dall J, Dandanell G, Valentin-Hansen P (1995) Gene-regulatory modules in *Escherichia coli*: nucleoprotein complexes formed by cAMP-CRP and CytR at the *nupG* promoter. Mol Microbiol 17: 843–853.
12. Brikun I, Suziedelis K, Stemmann O, Zhong R, Alikhanian L, et al. (1996) Analysis of CRP-CytR interactions at the *Escherichia coli udp* promoter. J Bacteriol 178: 1614–1622.
13. Kallipolitis BH, Valentin-Hansen P (1998) Transcription of *rpoH*, encoding the *Escherichia coli* heat-shock regulator sigma32, is negatively controlled by the cAMP-CRP/CytR nucleoprotein complex. Mol Microbiol 29:1091–9.
14. Thomsen LE, Pedersen M, Nørregaard-Madsen M, Valentin-Hansen P, Kallipolitis BH (1999) Protein-ligand interaction: grafting of the uridine-specific determinants from the CytR regulator of *Salmonella typhimurium* to *Escherichia coli* CytR. J Mol Biol 288: 165–175.
15. Zolotukhina M, Ovcharova I, Eremina S, Errais Lopes L, Mironov AS (2003) Comparison of the structure and regulation of the *udp* gene of *Vibrio cholerae*, *Yersinia pseudotuberculosis*, *Salmonella typhimurium*, and *Escherichia coli*. Res Microbiol 154: 510–520.
16. Price MN, Dehal PS, Arkin AP (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 3: 1739–1750.
17. Valentin-Hansen P, Søgaard-Andersen L, Pedersen H (1996) A flexible partnership: the CytR anti-activator and the cAMP-CRP activator protein, comrades in transcription control. Mol Microbiol 20: 461–466.
18. Jørgensen CI, Kallipolitis BH, Valentin-Hansen P (1998) DNA-binding characteristics of the *Escherichia coli* CytR regulator: a relaxed spacing requirement between operator half-sites is provided by a flexible, unstructured interdomain linker. Mol Microbiol 27: 41–50.
19. Tretyachenko-Ladokhina V, Ross JB, Senear DF (2002) Thermodynamics of *E. coli* cytidine repressor interactions with DNA: distinct modes of binding to different operators suggests a role in differential gene regulation. J Mol Biol 316:531–46.
20. Kallipolitis BH, Valentin-Hansen P (2004) A role for the interdomain linker region of the *Escherichia coli* CytR regulator in repression complex formation. J Mol Biol 342: 1–7.
21. Weickert MJ, Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. J Biol Chem 267:15869–15874.
22. Pedersen H, Søgaard-Andersen L, Holst B, Valentin-Hansen P (1991) Heterologous cooperativity in *Escherichia coli*. The CytR repressor both contacts DNA and the cAMP receptor protein when binding to the *deo*P2 promoter. J Biol Chem 266:17804–17808.
23. Søgaard-Andersen L, Mironov A S, Pedersen H, Sukhodelets V V, Valentin-Hansen P (1991a) Single amino acid substitutions in the cAMP receptor protein specifically abolish regulation by the CytR repressor in *Escherichia coli*. Proc Natl Acad Sci USA 88: 4921–4925.
24. Søgaard-Andersen L, Pedersen H, Holst B, Valentin-Hansen P (1991b) A novel function of the cAMP-CRP complex in *Escherichia coli*: cAMP-CRP functions as an adaptor for the CytR repressor in the deo operon. Mol Microbiol 5: 969–975.
25. Anderson WB, Schneider AB, Emmer M, Perlman RL, Pastan I (1971) Purification of and properties of the cyclic adenosine 3′,5′-monophosphate receptor protein which mediates cyclic adenosine 3′,5′-monophosphate-dependent gene transcription in *Escherichia coli*. J Biol Chem 246: 5929–5937.
26. Aiba H, Krakow JS (1981) Isolation and characterization of the amino and carboxyl proximal fragments of the adenosine cyclic 3′ ,5′-phosphate receptor protein of *Escherichia coli*. Biochemistry 20: 4774–4780.
27. Schultz SC, Shields GC, Steitz TA (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. Science 253:1001–1007.
28. Barbier CS, Short SA (1993) Characterization of *cytR* mutations that influence oligomerization of mutant repressor subunits. J Bacteriol 175: 4625–4630.
29. Holt AK, Senear DF (2010) An unusual pattern of CytR and CRP binding energetics at *Escherichia coli cdd*P suggests a unique blend of class I and class II mediated activation. Biochemistry 49: 432–442.
30. Meibom KL, Kallipolitis BH, Ebright RH, Valentin-Hansen P (2000) Identification of the subunit of cAMP receptor protein (CRP) that functionally interacts with CytR in CRP-CytR-mediated transcriptional repression. J Biol Chem 275: 11951–11956.
31. Møllegaard NE, Rasmussen PB, Valentin-Hansen P, Nielsen PE (1993) Characterization of promoter recognition complexes formed by CRP and CytR for repression and by CRP and RNA polymerase for activation of transcription on the *Escherichia coli deo*P2 promoter. J Biol Chem 268:17471–17477.
32. Kallipolitis BH, Nørregaard-Madsen M, Valentin-Hansen P (1997) Protein-protein communication: structural model of the repression complex formed by CytR and the global regulator CRP. Cell 89: 1101–11019.
33. Rasmussen PB, Holst B, Valentin-Hansen P (1996) Dual-function regulators: the cAMP receptor protein and the CytR regulator can act either to repress or to activate transcription depending on the context. Proc Natl Acad Sci U S A 93: 10151–10155.
34. Barbier CS, Short SA, Senear DF (1997) Allosteric mechanism of induction of CytR-regulated gene expression. CytR repressor-cytidine interaction. J Biol Chem 272: 16962–16971.
35. Busby S, Ebright RH (1999) Transcription activation by catabolite activator protein (CAP). J Mol Biol 293: 199–213.
36. Rasmussen PB, Søgaard-Andersen L, Valentin-Hansen P (1993) Identification of the nucleotide sequence recognized by the cAMP-CRP dependent CytR repressor protein in the *deo*P2 promoter in *E. coli*. Nucleic Acids Res 21: 879–885.
37. 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 J 16: 2108–2118.
38. Perini LT, Doherty EA, Werner E, Senear DF (1996) Multiple specific CytR binding sites at the *Escherichia coli deo*P2 promoter mediate both cooperative and competitive interactions between CytR and cAMP receptor protein. J Biol Chem 271: 33242–33255.
39. Tretyachenko-Ladokhina V, Cocco MJ, Senear DF (2006) Flexibility and adaptability in binding of *E. coli* cytidine repressor to different operators suggests a role in differential gene regulation. J Mol Biol 362: 271–286.
40. Zheng D, Constantinidou C, Hobman JL, Minchin SD (2004) Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. Nucleic Acids Res 32: 5874–5893.
41. Seeger C, Poulsen C, Dandanell G (1995) Identification and characterization of genes (*xapA*, *xapB*, and *xapR*) involved in xanthosine catabolism in *Escherichia coli*. J Bacteriol 177: 5506–5516.
42. Zhang Y, Craig JE, Gallagher MP (1992) Location of the *nupC* gene on the physical map of *Escherichia coli* K-12. J Bacteriol 174: 5758–5759.
43. Patching SG, Baldwin SA, Baldwin AD, Young JD, Gallagher MP, et al. (2005) The nucleoside transport proteins, NupC and NupG, from *Escherichia coli*: specific structural motifs necessary for the binding of ligands. Org Biomol Chem 7: 462–470.
44. Søgaard-Andersen L, Valentin-Hansen P (1993) Protein-protein interactions in gene regulation: the cAMP-CRP complex sets the specificity of a second DNA-binding protein, the CytR repressor. Cell 75: 557–566.
45. Hampshire AJ, Rusling DA, Broughton-Head VJ, Fox KR (2007) Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. Methods 42: 128–140.
46. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Res 39: D32–7.
47. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–97.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2008) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
49. Möller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17: 646–653.
50. Lai J-S, Cheng C-W, Sung T-Y, Hsu W-L (2012) Computational Comparative Study of Tuberculosis Proteomes Using a Model Learned from Signal Peptide Structures. PLoS ONE 7(4): e35018.
51. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307–321.
52. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
53. Mironov AA, Vinokurova NP, Gelfand MS (2000) Software for analyzing bacterial genomes. Mol Biol (Mosk) 34: 253–262.
54. Gelfand MS, Koonin EV, Mironov AA (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucl Acids Res 28: 695–705.
55. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.