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

Natalia V. Sernova1, Mikhail S. Gelfand1,2*

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 Vibrioiales 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

Copyright: © 2012 Sernova, Gelfand. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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], *ppaA* [9], *nupC* [10], *udpG* [11], *udpD* [12], *tphH* [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]. 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<sub>CRP</sub>) per CytR-binding site (O<sub>CYTR</sub>) 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<sub>CRP-D</sub> [30], with an exception being *cddP*, where CRP binds stronger to its proximal operator O<sub>CRP-P</sub> [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<sub>CYTR</sub> operator being not symmetric relative to the flanking O<sub>CRP</sub> operators [17]. Fig. 1 shows a typical organization of the O<sub>CRP-D-O_CYTR-O<sub>CRP-P</sub> 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 *deoP2*, 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<sub>CRP</sub> and O<sub>CYTR</sub>, 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
genes were considered, in particular, in [17]. In the cytR, deoP and udh regulatory regions only one CRP-binding site participates in the activation; in the nupG, tss 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\textsubscript{CytR} (distal) and O\textsubscript{CytR} (proximal). Unlike the situation with many TFs, including repressors from the LacI-family, the length of spacers between parts of the O\textsubscript{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\textsubscript{CytR}D and O\textsubscript{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) interactions.

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

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
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_00149         |
| 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).

Figure 3. Sequence LOGOs of the CRP, CytR-distal, CytR-proximal operators. Horizontal axis: position in the binding site; vertical axis: information content 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\textsubscript{CRP} LOGO; B1) O\textsubscript{CytR distal} LOGO; B2) O\textsubscript{CytR proximal} LOGO.

doi:10.1371/journal.pone.0044194.g003
rather than protein-DNA (CytR-OCYTR) 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 OCYTR position was mapped precisely in few cases only, e.g. in the deoP2 promoter by point mutagenesis [36] or by exchange of udpP and deoP2 OCYTR 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 OCYTR was then predicted by the comparison with the consensus [11].

Figure 4. Alignments of upstream regions of gene orthologous to the E.coli CytR regulon members. OCYTR 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 grey – 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

Comparative Genomics of CytR
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 (4th) position of a 8 bp window. Blue – OCyTRD, magenta – OCyTRP, red – averaged positional information content.

doi:10.1371/journal.pone.0044194.g005

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
or 5'-AYGTGCAAC-Nx-GTTRCAT-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 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

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 [10] 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 OCRP positional weight matrix (PWM) using SignalX (see Data and Methods). Sequence LOGOs of the constructed motifs are shown in Fig. 3A. To construct the OCYTR PWMs, we considered five E. coli genes with clearly distinguishable OCYTR (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 OCRP 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 OCYTRD and OCYTRP were taken for deoC from ECO, SFL, ENT, CKO, SEH, STY, KPE; for wdp from SEH, STY, ECO, SFL, CKO; for pepD 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 OCRP sites for each E. coli gene listed above and its ortholog; (2) the same distance between OCYTR half-operators; (3) at most two mismatches in each OCRP site, and at most three total mismatches in the OCRP sites, compared to the E. coli OCRP sites; and (4) at most four mismatches in the OCYTR operator, and at most three mismatches in each OCYTR half-operator compared to the E. coli OCYTR boxes. The selected boxes were used to construct PWMs for the upstream (distal) and downstream (proximal) half-operators (OCYTRD, Fig. 3B1, and OCYTRP, 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 OCRP sites at a distance 10–40 bp, or (2) two candidate OCYTR sites at a distance not exceeding 20 bp, or (3) all four sites in the configuration OCRPD–N(−10, 20)–OCYTRD–N(0, 20)–OCYTRP–N(−10, 20)–OCRPP, 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 OCRP operators. As mentioned above, in many genomes there are no strong candidate OCYTR sites at positions corresponding to the E. coli OCYTR operators. To identify possible shifted OCYTR sites in the regions between pairs of OCRP 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 OCYTRD and, separately, OCYTRP using respective PWMs. Our assumption was that if the position of the CytR-binding site OCYTR, comprising both OCYTRD and OCYTRP and the length of the spacer between them, was conserved within the alignment, the SWAS plot would have two pronounced peaks. On the other
hand, if O_CYTRD and/or O_CYTRP 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 _old_ 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_CMRP 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_CYTRD and O_CYTRP, 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.
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\textsubscript{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, Photorehobacter luminosus and Vibrio fischeri. The distances between the candidate O\textsubscript{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\textsubscript{CYTRD} and O\textsubscript{CYTRP} 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\textsubscript{CYTR} sites, some overlapping with O\textsubscript{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\textsubscript{CYTR} sites: O\textsubscript{CYTRD} is immediately adjacent to O\textsubscript{CYTRP} (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 Photobacterium aestuarii, the 4-box cassettes had very low total weights (about 10 or even less) and variable distances between the O\textsubscript{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 spoH gene encodes the heat-shock sigma-factor (sigma-32 or \(\sigma^H\)). 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\textsubscript{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 spoH cassette belongs to type 1 in close E. coli relatives.

The rpoH gene encodes the heat-shock sigma-factor (sigma-32 or \(\sigma^H\)). 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\textsubscript{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 spoH cassette belongs to type 1 in close E. coli relatives.\(^\text{[29]}\)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure11.png}
\caption{Alignment and SWAS plots of upstream regions of cytR in the Enterobacteriales. Notation as in Fig. 5. doi:10.1371/journal.pone.0044194.g011}
\end{figure}
The \textit{nupG} gene encodes one of two high affinity nucleoside transporters in \textit{E. coli}. It is present in seven genomes, the fewest among all considered genes. Further, the gene annotated as \textit{nupG} in \textit{Salmonella enterica} Heidelberg is in fact \textit{xapB}, encoding xanthosine MFS transporter [41], as demonstrated by the analysis of phylogenetic trees (not shown) and co-localisation with \textit{xapA}, the latter encoding a subunit of xanthosine phosphorylase. In \textit{K. pneumoniae}, the total score of the best OCRP 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
likely that the regulation of \( nupG \) in \( K. \) pneumoniae has been lost. The SWAS plot has two pronounced peaks corresponding to the \( OCRP \) and \( OCYTRP \) sites with the conserved distance of 9 bp between them and an overlap between the latter and the proximal \( OCRP \) 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 \( OCRP \) sites flanking \( OCyTR \) 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 \( OCRP \) sites (about 3). However, the SWAS plot features two pronounced peaks, and hence the cassette is of type 1, although \( OCYTRP \) overlaps \( OCRP \) (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 (\( E. \) icat, \( E. \) ictaluri, \( K. \) pneumoniae, \( S. \) 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 (\( V. \) harveyi, \( V. \) parahaemolyticus, \( V. \) splendidus, \( V. \) vulnificus) (Fig. S6). Since a high-scoring \( OCYTRP \) peak is situated relatively close to \( OCRP \) and all other peaks have very low scores (about 2), this case is assigned to type 3, as neither \( OCYTRD \) nor \( OCRP \) can be reliably identified.
The cytR gene itself has only the distal OCRP site. On the other hand, the bound complex has been observed in E. coli K-12 [6] and S. typhimurium [14]. The OCRYTR 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 OCRYTR site is an imperfect direct repeat (Fig. 10). At that, the Enterobacteriales and Vibrionales seem to have conserved organization of the OCRPD-OCRYPD-OCRYPTRP recognition site, but slightly different sequences of OCRYPD and OCRYPTRP. The unusual properties of this cassette may explain the fact that the scores of the OCRYTR 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 OCRP sites flanking direct repeats of OCRYTR [8,29]. Two OCRP sites denoted in the literature and here, in this particular case, OCRP2 and OCRP3 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 OCRYPD and OCRYPTRP (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 OCRYPD PWM both for the Enterobacteriales and Vibrionales (Fig. 11 and Fig. 12, respectively), whether for the cdd cassette visible peaks are produced by OCRYP TR 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 OCRP scores. Candidate OCRYTR 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 OCRYPD is about 4, which is consistent with a usual model of regulation of promoters with two CRP-binding sites. An alternative is OCRYTR 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 5 and the score of the second peak is about 2.5, both assessed by the OCRYPTRPWM (Fig. S9). Finally, there is a possibility that weaker OCRP 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 OCRP sites and the presence of peaks in the SWAS plots, demonstrating conservation of the OCRYTR positions. We started with identification of E. coli genes preceded by high-scoring OCRPD-OCRYPD-OCRYPTRP-OCRPP 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 OCRP 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 OCRP-OCRYPD-OCRYPTRP-OCRPP 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 OCRP sites at approximately the same distance in at least five genomes. After that we aligned the promoter regions, anchored at OCRPD and OCRPP, and applied...
the $\text{OCYTRD}$ and $\text{OCYTRP}$ PWMs, constructing SWAS plots for the spacer between $\text{OCRPD}$ and $\text{OCRIP}$.

One strong candidate emerged from this analysis. The $\text{ycdZ}$ gene of $E. \text{coli}$ is preceded by a cassette formed by two $\text{OCRIP}$ sites at a conserved distance (29–31 bp) and $\text{OCYTR}$ 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 $D. \text{dadantii}$, $E. \text{tasmaniensis}$, $P. \text{atrosepticum}$, $Y. \text{pestis}$ and $Y. \text{pseudotuberculosis}$, where this gene is simply absent, and $V. \text{vulnificus}$ that has an atypical $\text{OCRIP}$–$\text{OCRIP}$ distance.

The alignment of the $\text{ycdZ}$ upstream regions may be divided into three subalignments. In close relatives of $E. \text{coli}$, two pronounced peaks in the SWAS plots, corresponding to $\text{OCYTRD}$ and $\text{OCYTRP}$, 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 $\text{ycdZ}$ cassette belongs to type 1 at close distances and to type 2 in more distant Enterobacteriales and in the Vibrionales. Hence we predict that $\text{ycdZ}$ is a member of the $\text{CytR}$ 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 $\text{OCRIP}$ sites are conserved in upstream regions of orthologous genes regulated by the $\text{CRP-CytR}$ complex. On the other hand, positions of the $\text{OCYTR}$ 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 $\text{OCYTR}$ 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 $\text{OCRIP}$ 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 AA in TMHs: 105.864
# YcdZ ECO Exp number, first 60 AA: 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 TMH 20 42
YcdZ ECO TMHMM2.0 inside 43 48
YcdZ ECO TMHMM2.0 TMH 49 68
YcdZ ECO TMHMM2.0 TMH 69 71
YcdZ ECO TMHMM2.0 TMH 72 94
YcdZ ECO TMHMM2.0 TMH 95 98
YcdZ ECO TMHMM2.0 TMH 99 118
YcdZ ECO TMHMM2.0 outside 119 122
YcdZ ECO TMHMM2.0 inside 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
vation of positions of candidate OCYTR 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 OCYTR boxes with a large spacer (10 to 13 bp) and direct repeats in either direction (OCYTRDo rOCYTRP) with a short spacer (1 bp) may be bound by CytR [37]. Direct repeats in the OCYTRD 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:

\[ W(\beta, \kappa) = \log |N(\beta, \kappa) + 0.5| - 0.25 \sum_{z \in A,C,G,T} \log |N(\zeta, \kappa) + 0.5|, \]

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

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

\[ Z(\beta_1 \ldots \beta_k) = \sum_{k=1}^{K} W(\beta_k, \kappa), \]

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 \(0 \text{ nt}\), 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
\[
I_i = 2 + \sum_{x \in \{A,C,G,T\}} f(x,i) \log_2 (f(x,i)),
\]
where \(f(x,i)\) is the frequency of nucleotide \(x\) 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 \(OCyTR\) 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 \(OCyTR\) repeats). Notation as in Fig. 5. [TIF]

Table S1 The list of orthologs of CytR-regulated genes from \(E. coli\) that have nearly constant \(OCyTR-OCyRP\) distances. & – “yes(value)”:

| **Genes** | **Distances** | **Remarks** |
|-----------|--------------|------------|
| \(deoR\)  | \(1\)          | exceptional genes added as known from the literature co – cut-off, the smallest score of known cassettes for the respective gene. |

**Acknowledgments**

We are grateful to L.V. Lanovskaya 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.
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, Dandanel G, Valentin-Hansen P (1995) Gene-regulatory modules in Escherichia coli: nucleoprotein complexes formed by CRP-CRP and CytR at the nupC promoter. Mol Microbiol 17: 841–853.

12. Brikum I, Suziedelis K, Stemmman O, Zhong R, Alhakamian L, et al. (1996) Analysis of CRP-CyR interactions at the Escherichia coli udp promoter. J Bacteriol 176: 1614–1622.

13. Kallipolitis BH, Valentin-Hansen P (1998) Transcription of spoH, encoding the Escherichia coli heat-shock regulator sigma32, is negatively controlled by the cAMP-CRP/CyR nucleoprotein complex. Mol Microbiol 29:1091–9.

14. Thomsen LE, Pedersen M, Norregaard-Madsen M, Valentin-Hansen P, Kallipolitis BH (1999) Protein-ligand interaction: grafting of the uridine-specific determinants from the CyR regulator of Salmonella typhimurium to Escherichia coli CyR. J Mol Biol 288: 165–175.

15. Zohothnikina M, Ovcharova I, Eremina S, Erkais Lopez I, 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: e173."