RESEARCH ARTICLE

GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution

Inna A. Suvorova[1]*, **Yuri D. Korostelev**[1], **Mikhail S. Gelfand**[1,2]

1 Research and Training Center on Bioinformatics, Institute for Information Transmission Problems RAS (The Kharkevich Institute), Moscow, Russia, **2** Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia

* inn1313@yandex.ru

OPEN ACCESS

Citation: Suvorova IA, Korostelev YD, Gelfand MS (2015) GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. PLoS ONE 10(7): e0132618. doi:10.1371/journal.pone.0132618

Editor: Igor B. Rogozin, National Center for Biotechnology Information, UNITED STATES

Received: March 31, 2015

Accepted: June 16, 2015

Published: July 7, 2015

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files. In addition, reconstructed regulon data are available in the RegPrecise database (http://regprecise.lbl.gov).

Funding: This research was supported by the Russian Science Foundation via grant 14-24-00155 (http://рнф.рф/, to MSG). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The GntR family of transcription factors (TFs) is a large group of proteins present in diverse bacteria and regulating various biological processes. Here we use the comparative genomics approach to reconstruct regulons and identify binding motifs of regulators from three subfamilies of the GntR family, FadR, HutC, and YtrA. Using these data, we attempt to predict DNA-protein contacts by analyzing correlations between binding motifs in DNA and amino acid sequences of TFs. We identify pairs of positions with high correlation between amino acids and nucleotides for FadR, HutC, and YtrA subfamilies and show that the most predicted DNA-protein interactions are quite similar in all subfamilies and conform well to the experimentally identified contacts formed by FadR from *E*. *coli* and AraR from *B*. *subtilis*. The most frequent predicted contacts in the analyzed subfamilies are Arg-G, Asn-A, Asp-C. We also analyze the divergon structure and preferred site positions relative to regulated genes in the FadR and HutC subfamilies. A single site in a divergon usually regulates both operons and is approximately in the middle of the intergenic area. Double sites are either involved in the co-operative regulation of both operons and then are in the center of the intergenic area, or each site in the pair independently regulates its own operon and tends to be near it. We also identify additional candidate TF-binding boxes near palindromic binding sites of TFs from the FadR, HutC, and YtrA subfamilies, which may play role in the binding of additional TF-subunits.

Introduction

Interactions between DNA and proteins lie at the heart of many biological processes including DNA recombination, replication, repair and transcription [1]. One of the main mechanisms of regulation of gene expression is specific binding of transcription factors (TFs) to DNA. While up to 10% of genes in genomes of free-living bacteria encode transcription factors [2, 3], their structure and DNA-binding specificity are usually unknown [1]. Understanding the recognition mechanism of protein-DNA interaction is one of the most important problems of

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

molecular and computational biology. Evolution of regulatory interactions in various organisms can be studied by comparative analysis of functional systems.

Empirical rules of the protein-DNA recognition reflect chemical and physical properties of the residues, such as partial charge interactions between amino acid side chains and bases, or amino acid side chain flexibility [4]. The contribution of the amino acid main chain to the specific interaction with DNA is minor compared to the amino-acid side-chain atoms [5], and important and favorable contacts are usually hydrogen bonds (because of their high specificity and directional character) and acid—base interactions [4, 6, 7]. However, they do not always dominate in determining the interaction, and other types of contacts are also important [8]. For example, though hydrophobic interactions are considered less important for DNA-binding, since there are relatively few non-polar atoms present in the DNA double helix grooves [4], and regions of protein-DNA contacts are rich in polar residues that are important for binding, as they are involved in the formation of electrostatic and hydrogen bonds [9], hydrophobic interactions can play a certain role in protein-DNA interaction. While hydrogen bonds are specific in recognizing purines, hydrophobic contacts are mainly involved in recognition of the pyrimidines, for example, protein side chains rely on hydrophobic interactions to differentiate thymine from cytosine [10].

However, these trends are not universal and do not explain all amino acid—base interactions that may depend on the structural context and, in particular, on the structural family of DNA-binding proteins [10, 11].

Conservation of base pairs in a motif is significantly correlated with the number of contacts they have with the bound TF [5, 8]. Base pairs that form more contacts tend to be more conserved in evolution, because some of these amino acid-base pair interactions stabilize the DNA-protein complex and changes in these positions are more deleterious [8]. Mutual information analysis can be used to predict amino acid—base contacts for particular transcription factor families, giving opportunity to yield structural insights from sequence information alone, which can be further experimentally verified [12–16].

Contacts between the protein and the DNA sugar-phosphate backbone are thought to play a minor role in determining the specificity [10], but they may impact it by positioning of TF recognition elements in an orientation allowing for proper interaction [10, 17].

GntR family

The GntR family of transcription factors, first described in 1991 and named after the gluconate-operon repressor in *Bacillus subtilis*, is a large group of proteins distributed among diverse bacteria and regulating various biological processes [18, 19, 20]. GntR-family regulatory proteins are comprised of a DNA-binding domain and a signaling domain, linked together [19, 20, 21]. All proteins from the GntR family share highly similar N-terminal HTH (helix—turn—helix) DNA-binding domains, but differ in the C-terminal effector-binding and oligomerization (E-O) domains [18, 20]. The HTH domain is widespread and detected in many TFs, being the most-studied and best-characterized DNA-binding motif in the prokaryotic world [18, 20, 21]. The N-terminal DNA-binding domain of GntR-family proteins comprises a central β-sheet cluster and three α-helices [20]. The HTH motif consists of the α-helix, the connecting loop, and the second α-helix, often referred to as the "recognition helix", as it directly interacts with the DNA [18, 20, 22]. Generally, HTH proteins bind as dimers to 2-fold symmetric DNA operator sequences so that each monomer recognizes a half-site [20, 22].

The C-terminal E-O domain does not bind to the DNA, but it can impose steric constraints on the DNA-binding domain, hence influencing the HTH motif, and thus plays an important role in regulation [21]. For example, E-O domain can restrict DNA-binding domain's mobility and thus reduce its ability to adapt to varying distances between the parts of a palindromic motif, which reflects on the binding motif structure [20]. Oligomerization and conformational changes due to binding of an inducer molecule allow for the correct HTH motif arrangement, modulating its orientation and presentation, and the subsequent DNA binding [20, 21], as shown for many diverse proteins [23, 24]. Thus, despite high conservation of the DNA-binding domain, the operator consensus sequences observed among GntR-family TFs may be different, likely due to the E-O domain variability and domain synergy [20].

According to the type of the C-terminal domain, the GntR family is divided into four main (FadR, HutC, MocR, and YtrA) and two minor subfamilies (AraR and PlmA) [18, 20, 21, 25–29].

The FadR subfamily is the largest one, it comprises about 40% of known GntR-family TFs, with α-helical C-terminal domain, which is 150–170 amino acids in length, formed by either seven or six α-helices [18, 20]. TFs of the FadR subfamily bind effectors, small organic ligands, such as carboxylic acids, and then undergo conformational changes that affect DNA-binding [18]. Most FadR-subfamily proteins are involved in the regulation of oxidized substrates related to amino acids or emerging from the central metabolism, or at the crossroads of various metabolic pathways, such as glycolate (GlcC), galactonate (DgoR), pyruvate (PdhR), lactate (LldR), or gluconate (GntR) [18, 20].

The C-terminal domain of the second subfamily, HutC, is about 170 amino acids in length and contains both α-helical and β-sheet structures [20]. This subfamily comprises about 30% of GntR-family regulators [20]. The C-terminal E-O domain of HutC-subfamily transcription factors has the same fold as chorismate lyase (UbiC in *Escherichia coli*), which suggests that it may bind small effector molecules, such as histidine (HutC), fatty acids (FarR), sugars (TreR), and alkylphosphonates (PhnF), in a mode similar to chorismate lyase [19]. Some HutC-subfamily TFs are involved in the regulation of N-acetylglucosamine utilization (DasR, NagR, NagQ) and in conjugative plasmid transfer in various *Streptomyces* species (e.g., KorSA, KorA, and TraR) [20, 22, 30].

The third subfamily, MocR, is different, as proteins from this group have a large E-O domain, whose average length is about 350 amino acids [20]. This domain is homologous to class I aminotransferase proteins (TyrB of *E. coli*) [20, 31, 32]. The latter catalyze transamination of amino acids to α-keto acids and use pyridoxal 5'-phosphate (PLP) as a cofactor [20, 31, 32]. A similar requirement for PLP was shown for some MocR-subfamily proteins (TauR, GabR) [25, 31, 32]; moreover, PdxR in *Streptomyces* spp. is directly involved in the regulation of the PLP synthesis [20, 33]. Aminotransferases are known to form head-to-tail dimers and such dimerization likely occurs in MocR-subfamily proteins as well [20, 32].

Proteins from the fourth subfamily, YtrA, which is the smallest one among the main subfamilies (about 6%), have a reduced C-terminal domain with only two α-helices, of the average length about 50 amino acids [20]. This may seriously restrict effector-binding and dimerization abilities of the C-terminal domain, though the latter is obviously possible, since long palindromic binding motifs have been identified upstream of candidate regulated operons [20]. Most genes of the YtrA-subfamily TFs form operons with ATP-binding cassette (ABC) transporters [20].

The PlmA subfamily is composed exclusively of TFs from cyanobacterial species [26]. It is close to the YtrA and MocR subfamilies, and its likely ancestor arose from one of them [26]. PlmA (encoded by *all1076*) controls plasmid maintenance in *Anabaena* (*Nostoc*) sp. strain PCC 7120, but it is unclear whether it is a common function of PlmA-subfamily TFs, since there are no identified plasmids in several cyanobacterial species, all of which contain *plmA* orthologs [26].

AraR-subfamily TFs exhibit chimeric organization with two domains of different phylogenetic origin: its N-terminal DNA-binding region contains a winged HTH motif similar to that of the GntR family, while the C-terminal domain is homologous to the C-terminal domain of the GalR/LacI family [27, 28, 29]. AraR controls expression of genes encoding transporters and enzymes involved in uptake and utilization of L-arabinose and arabinose-containing polysaccharides, xylose and galactose in Firmicutes [27, 28, 29].

Structure of binding motifs

Different DNA-binding domain types recognize distinct motifs [34], whereas DNA-binding proteins from the same family tend to recognize sites of similar length, symmetry, and specificity [5]. Within each family, structure and fold of the DNA-binding domain and its mode of interaction with the binding motif are usually conserved, which results in a characteristic pattern of DNA-amino acid contacts [5]. However, even proteins with very high (up to 60–70%) amino acid sequence identity may bind to distinct DNA motifs [34].

As mentioned above, the HTH motifs are conserved in all the GntR family, although there are differences between consensus sequences for each subfamily [20]. The level of similarity between the HTH domains of the MocR and YtrA subfamilies is the highest. One of these two subfamilies has likely emerged from the other via replacement of the C-terminal domain [20].

Many experimentally identified and predicted binding motifs of GntR-family TFs match the palindromic $N_yGTN_xACN_y$ consensus sequence [20]. The motifs differ in the number (x, y) and the nature (N) of the nucleotides that surround the consensus GT and AC pairs [20]. This neighborhood often consists of A and T residues, and their number differs between subfamilies [20]. For example, the consensus for the FadR-subfamily TFs is $N_yGTM\text{-}N_{0-1}\text{-}KACN_y$, and for the HutC subfamily, $N_yGTMTAKACN_y$ [20,21]. The center of the palindrome is usually highly conserved, while the periphery varies [20].

Some TFs from the FadR and HutC subfamilies recognize unique motifs with different or no symmetry [20], for example, FarR (direct repeats TGTATTAWTT) [35], NagQ (direct repeats TGGTATT) [30], BioR (TTATMKATAA) [36, 37], NanR (direct repeats TGGTATAW) [38].

The distance between the half-sites is important for the correct presentation of a DNA site to a TF, and it varies weakly among the FadR and HutC subfamilies, but differs between these groups and the YtrA subfamily [20]. In the YtrA subfamily, the conserved GT and AC residues are located far from the center of the palindrome [20]. This feature of motifs may be due to short C-terminal domains of the YtrA-subfamily TFs, which may cause a specific mode of dimerization and DNA-binding, and hence yield an unusual motif structure [20].

Comparative studies of the MocR subfamily did not reveal any conserved palindromic sequence satisfying the GntR consensus or common to the whole subfamily [20]. For example, predicted binding motifs for some of the MocR-subfamily regulatory proteins include direct repeats of ATACCA for GabR [31], CTGGACYTAA for TauR [25] and AAAGTGGW(–/T) CTA for PdxR [39]. There are no obvious similarities in these structures and thus they may not be compared. Such organization of binding motifs could be due to the head-to-tail dimerization of MocR-type TFs, which yields direct repeats with sufficiently long spacers that allow for DNA looping [20].

Several crystal structures of GntR-family proteins have been solved so far, for example, in the FadR subfamily, these are FadR from *Escherichia coli* (PDB code 1H9T, 1HW1, 1HW2), LldR from *Corynebacterium glutamicum* (2DI3), TM0439 from *Thermotoga maritime* (3SXK, 3SXY); in the HutC subfamily, YvoA (NagR) from *Bacillus subtilis* (2WV0), HutC from *Pseudomonas syringae pv. tomato* str. DC3000 (2PKH), AgaR from *Enterococcus faecalis* V583

Table 1. DNA-amino acid contacts in FadR from *E.coli* and AraR from *B. subtilis*.

Position in the HTH domain	Amino acid of FadR *E.coli*	Contact in FadR-DNA or other related function	Amino acid of AraR *B. subtilis*	Contact in AraR-DNA or other related function
0	Ser-7	Non-specific with sugar-phosphate backbone	Pro-25	-
1	Pro-8	Non-specific with sugar-phosphate backbone	Lys-26	Non-specific with sugar-phosphate backbone
2	Ala-9	Non-specific with sugar-phosphate backbone	Tyr-27	Non-specific with sugar-phosphate backbone
27	Glu-34	Non-specific with sugar-phosphate backbone; electrostatic bonds with Arg-35, Arg-45, Arg-49	Glu-52	Hydrogen bonds with Arg-63, Arg-67
28	Arg-35	Arg-G, specific	Asn-53	-
37	Thr-44	Non-specific with sugar-phosphate backbone; Thr-C and Thr-G, specific	Ser-62	-
38	Arg-45	Arg-G, specific	Arg-63	Arg-G, specific; Arg-A, water-mediated specific; Arg-A, acetate-mediated
39	Thr-46	Non-specific with sugar-phosphate backbone; Thr-C and Thr-G, specific	His-64	His-G and His-T, water-mediated specific
40	Thr-47	Non-specific with sugar-phosphate backbone	Thr-65	Non-specific with sugar-phosphate backbone
42	Arg-49	Non-specific with sugar-phosphate backbone	Arg-67	Non-specific with sugar-phosphate backbone
43	Glu-50	Electrostatic bonds with Arg-35, Arg-45, Arg-49	Lys-68	-
56	Ile-63	Non-specific with sugar-phosphate backbone	Ser-81	-
58	His-65	His-A and His-G, specific	Gln-83	Gln-A and Gln-T, specific
59	Gly-66	Non-specific with sugar-phosphate backbone; helps avoiding steric clash	Gly-84	Gly-T, specific; Gly-T and Gly-A, water-mediated specific; Gly-T and Gly-A, acetate-mediated; helps avoiding steric clash
60	Lys-67	Non-specific with sugar-phosphate backbone	Gly-85	-
62	Thr-69	Non-specific with sugar-phosphate backbone	Gly-86	-

doi:10.1371/journal.pone.0132618.t001

(3DDV); and in the AraR subfamily, AraR DNA-binding domain from *Bacillus subtilis* (4EGY, 4EGZ, 4H0E). However, only two of these TFs (FadR and AraR) are solved in a complex with DNA. The structural data (summarized in Table 1) shows that FadR from *E. coli* and AraR from *B. subtilis* form a number of non-specific interactions with the DNA sugar-phosphate backbone, but only few base pairs are specifically recognized within the complex [40, 41, 42]. Main base-specific interaction present in FadR-DNA complex and in all analyzed structures of AraR DNA-binding domain with DNA is a hydrogen bond formed by Arg, which is part of the recognition helix, with the base G [28, 40, 41, 42]. Thus, such recognition may be a conserved feature of the GntR family.

Goals

We use the comparative genomics approach to reconstruct regulons and predict binding motifs of the regulators from three subfamilies of the GntR family—FadR, HutC, and YtrA. We report correlations between the DNA binding motifs and amino acid sequences of TFs and predict the most favorable DNA-protein contacts.

Further, we analyze the divergon structure in the FadR and HutC subfamilies and characterize preferred site positions relative to regulated genes.

We also identify additional candidate TF-binding boxes near strong binding sites in a number of the orthologous groups of transcription factors from the FadR, HutC, and YtrA subfamilies.

Materials and Methods

All genomic sequences were obtained from GenBank [43]. Known GntR-family TFs were collected from the literature. New members of the family were found using exhaustive BLAST search [44]. Homologs of TFs were found by PSI-BLAST [44] searches (E-value cutoff, 10^{-20}), and orthologs were identified by construction of phylogenetic trees for identified homologs supplemented by analysis of gene neighborhoods on chromosomes (*e.g.*, co-localization with genes of a certain metabolic pathway in most genomes). Normally, an ortholog group contained one TF per genome. However, in some cases several TFs in one genome, resulting from recent duplications or close-range horizontal transfers, were assigned to the same ortholog group.

Amino acid and nucleotide sequence alignment was performed using the MUSCLE package (with default parameters) [45]. Phylogenetic trees were constructed with the PHYLIP package, using the protdist program for the calculation of distances and the maximum-likelihood method for the tree construction (with default parameters) [46].

Candidate binding sites were identified (or confirmed if they were previously predicted) by phylogenetic footprinting [2]. We manually analyzed alignments of upstream regions of orthologous genes and identified groups of consecutive conserved positions, relying on the assumption that binding sites are more conserved than surrounding intergenic regions. Nucleotide position weight matrices (PWMs, profiles) for each TF were then constructed by the SignalX program as previously described [47], using training sets of upstream regions of genes presumably belonging to the respective regulon (genes encoding TFs, as they are often auto-regulated, and genes co-localized with them). The profiles were then used to search for additional regulon members.

Computational search for candidate TF-binding sites in upstream gene regions (for all genes in genomes, 400 nucleotides (nt) upstream and 50 nt downstream relative to the annotated gene start) was performed using the GenomeExplorer program package [48] and the RegPredict web server [49]. Score thresholds for the identification of sites were selected so that candidate sites upstream of functionally relevant genes were accepted, while the fraction of genes preceded by candidate sites did not exceed 5% in each studied genome. Under these conditions, for some long, conserved motifs, the number of candidate sites per genome did not exceed 50. Weaker sites (with scores 10% less than the threshold) were also taken into account if their positions were similar to positions of stronger sites upstream of orthologous genes and there were no stronger competing sites in the same intergenic region. New candidate members were assigned to a regulon if they were preceded by candidate binding sites in several genomes, the exact number of genomes depending on the number of sequenced genomes in a taxonomy unit. The reconstructed regulons were extended to include all genes in putative operons, the latter defined as the strings of genes transcribed in the same direction, with intergenic distances not exceeding 200 nt, when such organization persisted in several genomes. Motif logos were constructed using WebLogo [50].

Data on composition of the characterized GntR-family regulons, and respective binding sites are available in RegPrecise database (http://regprecise.lbl.gov/RegPrecise/).

Only TFs that had palindromic predicted binding motifs satisfying the GntR-family consensus were selected to analyze the correlation between amino acid sequences of TFs' DNA-binding HTH domains and nucleotides in the binding sites. The structural data of FadR from

Escherichia coli and AraR from *Bacillus subtilis* in complexes with DNA was taken as a reference model. At that, positions of known interacting amino acids [40, 41, 42] were re-numbered starting from the beginning of the HTH domain, counting from zero (Table 1).

Correlations were calculated for each subfamily using the Prot-DNA-Korr program package. The program calculates the correlation between each pair of columns, one from the amino acid alignment, and the other from the site alignment (dataset used in this work is given in S1 File). As a measure of correlation, the mutual information is used. The statistical significance value of the mutual information is calculated as the Z-score. Correlated pairs of positions are presented as a heatmap, where the pairs are colored according to the statistical significance, and as the contingency tables (given in S2 File) which contain expected and observed counts of amino acid-nucleotide pairs. For more detailed information concerning Prot-DNA-Korr program, see the link: http://bioinf.fbb.msu.ru/Prot-DNA-Korr/main.html.

Statistical analysis was performed using the STATISTICA program package [51].

The comparative genomic analysis relied on the following basic assumptions. It is known that the majority (~60–70%) of bacterial TFs are auto-regulated [52, 53]. Negative auto-regulation, in which a transcription factor represses its own gene expression, is the most common network motif (e.g., approximately 40% of the known TFs in *E. coli*) [54, 55]. Besides auto-regulation, in many cases genes encoding TFs and the genes they regulate are co-localized in the genome, since they tend to evolve concurrently, and this fact can be used when linking novel TFs with their DNA motifs and candidate regulon members [47,53].

Results and Discussion

Here we report the results of the analysis of transcription factors from three subfamilies of the GntR family—FadR, HutC, and YtrA. Candidate binding sites were predicted for 1252 GntR-family TFs from 307 genomes of bacterial species (S1 Table). The TFs were classified into 64 orthologous groups. The representation of the GntR-family TFs in individual genomes and among taxonomic groups varied (Table 2). For example, YtrA-subfamily TFs are common among *Firmicutes*, while FadR-subfamily TFs are more typical for *Proteobacteria*.

Protein-DNA correlations

FadR subfamily. The common consensus of all analyzed binding sites of FadR-subfamily TFs is an A/T-rich palindromic sequence with conserved TKGT/ACMA boxes (Fig 1), likely the most important for the DNA-protein interaction. As was mentioned earlier, the typical distance between GT and AC in most FadR-subfamily TF-binding sites is 3 nt (DgoR, ExuR, FadR, GlcC, LldR, PdhR *etc*), although in several orthologous groups it is 2 nt (*e.g*., GntR, HpxS, HypR, MdcY, PrpR, UxuR). The latter sites were included into the dataset (Spreadsheet FadR in S1 File) for further analysis after insertion of a single-nucleotide gap into the center of the motif.

Some FadR-subfamily TFs were excluded from the correlation analysis, since their binding motifs did not conform to the common consensus sequence and hence could not be aligned. The examples are NanR that binds direct repeats TGGTATAW [38], or BioR with the binding site consensus TTATMKATAA [36,37].

The correlation analysis (Fig 1, Spreadsheet FadR in S2 File) shows that, for FadR-subfamily TFs, significant amino acid positions correlated with the site positions and likely responsible for the binding specificity correspond well to those identified for the FadR (*E.coli*) and AraR (*B.subtilis*) protein-DNA structures.

Due to the symmetrical structure of the analyzed binding motifs and, consequently, of the obtained heat maps, correlations are usually shown for either G/C or A/T pair, while further

Table 2. General statistics of the analyzed GNTR-family TFs.

Number of \ Subfamily		FadR	HutC	YtrA
Orthologous groups		36	16	12
TFs analysed		634	389	229
Regulated operons, total		1740	975	283
Sites total (including divergent and multiple)		2396	1341	294
Taxonomy distribution of analyzed TFs				
Proteobacteria	Alpha	76	39	3
	Beta	151	64	0
	Gamma	308	112	25
	Delta	10	1	0
Firmicutes	Bacilli	18	97	89
	Clostridia	1	14	53
Actinobacteria		64	60	43
Thermotogae		0	0	14
Chloroflexi		6	0	1
Cyanobacteria		0	1	0
Bacteroidetes		0	1	0
Archaea		0	0	1

doi:10.1371/journal.pone.0132618.t002

disambiguation between G and C, or A and T is not always possible, since it requires additional consideration, such as comparing correlation data to the known contacts in the FadR-DNA and AraR-DNA complexes, taking into account donor-acceptor properties etc. It is known that, in general, hydrogen-bond donor residues (like Arg, His, Lys, Ser, Thr) prefer G, hydrogen bond acceptor residues (such as acidic Asp, Glu) prefer C, while Asn and Gln, that possess both donor and acceptor moieties, prefer A [6, 7].

Amino acids in position 28 of the HTH domain are correlated with nucleotides in site positions 6/14 (in TKGT/ACMA groups), known to form a contact in FadR-DNA complex

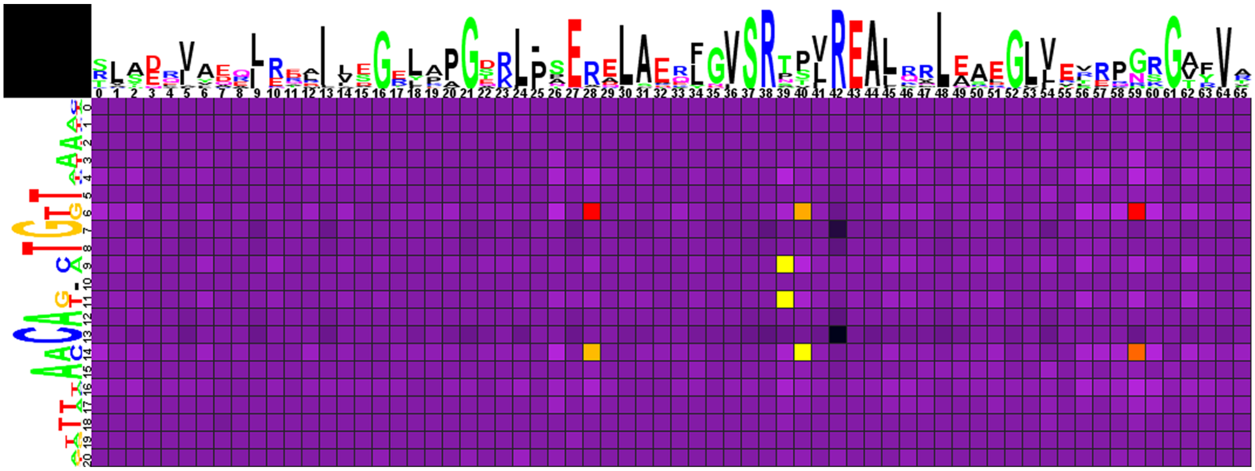

Fig 1. Heat map of correlations between amino acids and nucleotides for FadR-subfamily TFs and their binding sites. Sequence logos of HTH DNA-binding domains and corresponding binding sites are shown on the top and to the left of the heat map, respectively. The total height of the symbols in each position equals the positional information content, whereas the height of individual symbols is proportional to the positional amino acid or nucleotide frequency. The correlation scores are color ramped from yellow to red for amino acid-nucleotide pairs with statistical significance greater than an automatically defined threshold (with red assigned for the most correlated pair). The violet-black palette is used for other pairs.

doi:10.1371/journal.pone.0132618.g001

(Table 1) [40, 41]. Arg, the most frequent amino acid in this position, strongly prefers the G/C pair, while the A/T pair is significantly avoided. Asp, which is much rarer in position 28 than Arg, also significantly prefers the G/C pair. According to the electrochemical characteristics of these amino acids, we can conclude that the possible contacts in this position are Arg-G and Asp-C.

Amino acid residues in positions 40 and 59, which are known to be important for the FadR and AraR interactions with DNA (Table 1), are also correlated with nucleotide in positions 6/14. The most frequent amino acids in position 40 are Pro and Ser. Ser significantly prefers the G/C pair (possibly interacting with G), while Pro significantly avoids it.

Gly, that is most frequent in position 59, strongly prefers the G/C pair, while the A/T pair is significantly avoided. However, the Gly-G/C association might not be linked to a direct contact. In FadR-DNA complex, glycine occupying the same position does not form specific contacts, but due to the absence of the side chain allows for the interaction of the adjacent amino acid with DNA [40, 41]. Asn is also frequent in position 59 and exhibits a preference of the A/T pair, but it is not statistically significant.

Moreover, amino acids in position 39 of the HTH domain, also involved in binding to DNA by FadR and AraR (Table 1), are correlated with central nucleotides in positions 9/11. Asn here significantly prefers the A/T pair, possibly interacting with A according to the interaction trends described above. Thr is also frequent in this position and it shows a trend towards the preference of the A/T pair, but it is not statistically significant.

HutC subfamily. The consensus sequence of all analyzed binding motifs of HutC-subfamily TFs is very similar to the one of the FadR subfamily (Fig 2). The distance between GT and AC in most HutC-subfamily TF binding sites is 4 nt. Among the exceptions there are FarR (direct repeats TGTATTAWTT) [35], NagQ (direct repeats TGGTATT) [30], SdhR (palindrome with additional symmetry TCTTATGTCTTATATAAGACATAAGA) [56]. These TFs were excluded from the correlation analysis, as their binding sites could not be aligned and compared with the main group of sites (Spreadsheet HutC in S1 File).

The correlation analysis (Fig 2, Spreadsheet HutC in S2 File) shows that positions significant for binding specificity of HutC-subfamily TFs resemble the ones identified for FadR from

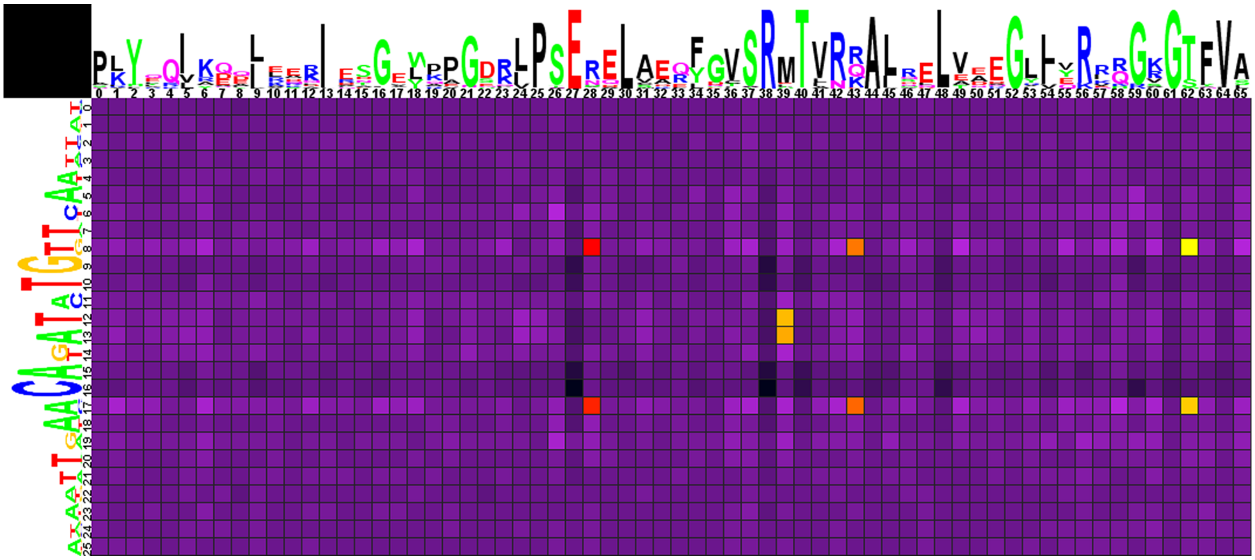

Fig 2. Heat map of correlations between amino acids and nucleotides for HutC-subfamily TFs and their binding sites. Notation as in Fig 1.

doi:10.1371/journal.pone.0132618.g002

E.coli and the FadR subfamily in general. In particular, amino acids in position 28 of the HTH domain correlate with nucleotides 8/17. As in the FadR subfamily, Arg, the most frequent amino acid in position 28, strongly prefers the G/C pair (according to the electrochemical characteristics, possible contact in this position is Arg-G), while the A/T pair is significantly avoided. Asn is also frequent in this position of the HTH domain, weakly preferring the A/T pair (no statistical significance).

Amino acid residues in positions 43 and 62 also show correlations with nucleotides in positions 8/17. The most frequent amino acids are Arg, Gln, Lys in position 43, and Thr and Ser in position 62, but neither of them shows significant preference of any base pair, while less frequent in position 62 Trp significantly prefers the G/C pair (possibly interacting with C).

Moreover, amino acids in position 39 of the HTH domain are correlated with central nucleotides 12/13 (as it has been shown for the FadR subfamily). The most frequent amino acid in this position is Met, with non-significant preference of the A/T pair, while less frequent Asp here significantly prefers the G/C pair, where it likely interacts with C, since it is a hydrogen bond acceptor amino acid.

YtrA subfamily. This subfamily, its binding motifs and regulons have many features different from those of other studied GntR subfamilies. The divergent organization of regulated operons, frequently observed for FadR- and HutC-subfamilies TFs (see below), is very rare in YtrA-subfamily regulons. Consistent with previous observations [20], most YtrA-subfamily regulons consist of a single operon comprised of genes encoding ATP-binding cassette (ABC) transporters. Moreover, most genes regulated by YtrA-subfamily TFs are preceded by single binding sites, and very few double or triple binding sites (quite common for the FadR and HutC subfamilies) have been identified (Spreadsheet YtrA in S1 File).

Binding motifs of TFs from the YtrA subfamily are significantly longer than motifs of other GntR-family TFs (Fig 3). Still, due to the conserved HTH domain structure in the GntR family, YtrA-type DNA-binding domains can be aligned accurately with domains from the other subfamilies, and our analysis has shown that amino acid positions that determine the binding specificity in the YtrA subfamily are mostly similar to those of FadR and HutC.

As was already mentioned, consensuses of the GntR-family binding motifs are generally palindromic, but each particular site can deviate from the consensus, being not strictly symmetric. In the case of FadR and HutC subfamilies these deviations are averaged by the large number of studied sites (Table 2), and thus the corresponding correlation data heat maps are symmetric. The YtrA subfamily is the smallest one, with the number of analyzed sites being an order of magnitude less than in other subfamilies (Table 2), which leads to some asymmetries in the corresponding correlation data heat map (Fig 3). Moreover, asymmetry can be caused by the lack of the divergently regulated operons in YtrA subfamily, unlike the FadR and HutC subfamilies.

The correlations (Fig 3, Spreadsheet YtrA in S2 File) show that nucleotides in positions 12-13/29-30 may specifically interact with amino acids in positions 27 and 28. As in the case of FadR and HutC subfamilies, Arg and G/C is the most frequent amino acid/nucleotide pair in position 28, though they do not show significant correlation here, while Asn and Tyr, rarer in this position, are significantly associated with the A/T pair (both likely interacting with A, according to the interaction trends described above). Val is the most frequent amino acid in position 27, but it shows no statistically significant base-pair preferences, while Thr, that is also frequent here, significantly prefers the A/T pair (possibly interacting with A, being a polar uncharged amino acid) in nucleotide positions 12,13 and 30.

Correlations are also observed for nucleotides 16-17/25-26 and amino acids 37 and 39, and that conforms well to the structural data for FadR, where these positions are important for the interaction with DNA (Table 1). Asn is the most frequent amino acid residue in positions 37

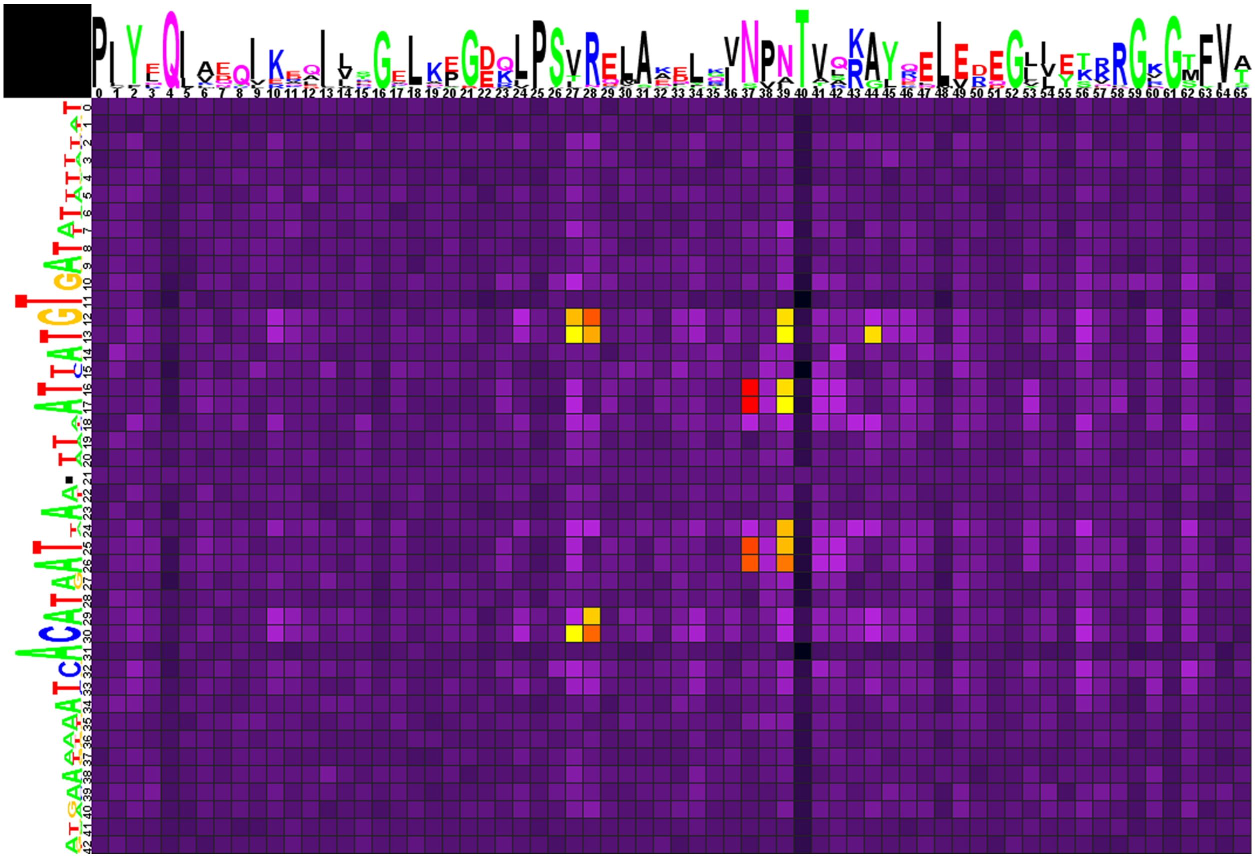

Fig 3. Heat map of correlations between amino acids and nucleotides for YtrA-subfamily TFs and their binding sites. Notation as in Fig 1.

doi:10.1371/journal.pone.0132618.g003

and 39, weakly preferring the A/T pair in 16-17/25-26 positions (no statistical significance). Ser, less common in position 37, and Ile in position 39, both strongly prefer A/T in positions 16-17/25-26 of the motif; while His in position 39 is significantly correlated with the G/C pair in positions 25 and 26. In the latter case the contact is likely His-G, conforming to the interaction trends of the hydrogen-bond donor residues, as well as to the His-G contact in the same position in the AraR-DNA complex (Table 1).

Moreover, significant correlations are also identified for Ala in position 39 with A/T in nucleotide positions 12, 13 and with G/C in position 24; and Gly in position 44 is correlated with A/T in position 13.

Overall, despite significant differences in the binding motifs, DNA-protein interactions in the YtrA subfamily seem to be at least partly similar to that of the FadR and HutC subfamilies.

Overview of protein-DNA correlations. Our data shows that predicted protein-DNA interactions for all three analyzed subfamilies of the GntR family correspond well to known nucleotide-amino acid contacts of FadR and AraR [40, 41, 42].

It has been shown in the literature that Arg, Asn, Lys, Gln, Thr, Ser, Asp and Gly account for more than 70% of contacts, with Arg alone accounting for 23% [7]. This trend was demonstrated in our study as well: majority of the predicted interactions involved exactly these amino acids.

Arg-G, Asn-A, Asp-C, Gln-A, Glu-C, Lys-G, and to a lesser extent His-G and Ser-G, appear to be the most relevant, strongest and highly specific contacts [4, 7]. Preferences are also known for Ala-C, Cys-G, Gly-G, Leu-A, Thr-G, and Trp-C [7].

Though there is some controversial data (for example, both Ser-A/T and Ser-G/C correlations), the majority of favorable contacts (Arg-G, Asn-A, Asp-C, Gly-G, His-G, Trp-C), predicted by the correlation analysis of the GntR-family TFs and their binding sites in all analyzed subfamilies, conforms to the general interaction trends described in the literature [6, 7].

Divergons

Many genes regulated by FadR—and HutC-subfamily TFs are organized in two divergently transcribed operons (divergons), and it is not immediately clear what is the relationship between the intergenic sites and each of the operons. The YtrA subfamily has not been represented in this analysis, as TFs from this subfamily almost never have sites between divergently transcribed operons.

The divergons were divided into two groups: divergons that consist of structural genes only (the control group), and divergons comprising a TF gene. Divergons with single or double intergenic sites were studied separately.

For divergons with a single binding site, we analyzed the length of the intergenic region and the distance between the center of the binding site and the starts of both genes that form the divergon.

For divergons with double binding sites we calculated the length of the intergenic region, the distance between the center of the proximal binding site and the start of a gene, and the distance between the two binding-sites' centers.

In the case of divergons with single sites, our aim was to determine whether these sites regulate both divergent operons, or one operon only (e.g., divergon could comprise a regulated operon containing structural genes and a divergent, not auto-regulated TF gene).

It is known that TFs, for example AraR, can cooperatively bind to several adjacent sites, allowing for a more flexible and tight control of the expression [42, 57]. In the case of divergons with double sites, we aimed to distinguish between the following alternatives: the site pair could be involved in regulation of the both divergent operons (or one particular operon), hence essentially being a single complex site, or each site in the pair could separately regulate its own operon.

Divergons with a single site. For both FadR- (n = 96) and HutC-subfamily (n = 94) divergons comprising a TF gene in one of the operons we observed an approximately linear increase of the distance between the site and the start of each gene in the divergon, as the intergenic distance increased (Fig 4A and 4B). The same tendency was also observed for the control divergons (FadR, n = 33; HutC, n = 23) (Fig 4C; due to the complete match only one regression line

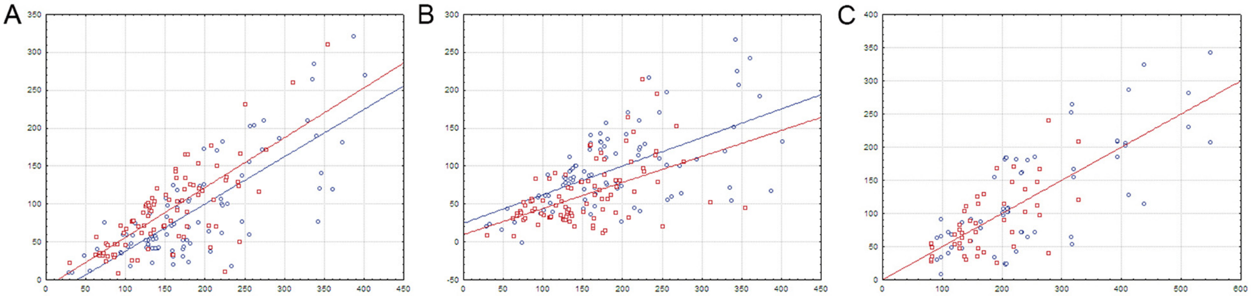

Fig 4. Distances between regulated genes and TF-binding sites in divergons with single sites. A—operons with a TF gene; B—operons with structural genes only; C—the control group (includes divergons without TF genes). The vertical axis is the distance between the site center and the start codon. The horizontal axis is the intergenic distance. Each dot corresponds to one site. The regression lines are shown. Blue color denotes the FadR subfamily; red color, the HutC subfamily.

doi:10.1371/journal.pone.0132618.g004

Table 3. Interdependence of the intergenic distance and the distance to a single site.

	Linear regression coefficient (R^2)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,60 (0,55)	0,62 (0,60)	0,66 (0,56)
Operons with structural genes only	0,40 (0,35)	0,38 (0,35)	0,34 (0,26)
Control divergons	0,50 (0,58)	0,50 (0,58)	0,50 (0,42)
	Pearson correlation coefficient (p-value)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,74 ($p<1\cdot10^{-7}$)	0,77 ($p<1\cdot10^{-7}$)	0,75 ($p<1\cdot10^{-7}$)
Operons with structural genes only	0,59 ($p<1\cdot10^{-7}$)	0,59 ($p<1\cdot10^{-7}$)	0,51 ($p = 2\cdot10^{-7}$)
Control divergons	0,76 ($p<1\cdot10^{-7}$)	0,76 ($p<1\cdot10^{-7}$)	0,65 ($p = 1\cdot10^{-6}$)

doi:10.1371/journal.pone.0132618.t003

is visible). Thus, single sites usually tend to be localized approximately in the middle of the intergenic spacer, although in the divergons with TF genes they usually are slightly closer to the structural operon (Table 3, Fig 4).

Divergons with double sites. As mentioned earlier, there are three possible variants of regulation in the case of divergons with double sites. If each site in a pair regulates its own operon, the distance between the sites should be positively correlated with the size of the intergenic region, as each site would tend to be closer to its regulated operon. Vice versa, if the sites are co-operatively involved in the regulation of the both operons (or one particular operon from a pair), the distance between the sites would likely be approximately constant and hence would not correlate with the size of the intergenic region. In this case, similarly to the single-site one, if the common site pair is involved in the regulation of both divergent operons, these sites would tend to be situated in the central part of intergenic region, otherwise, the site pair would be positioned near the regulated operon.

In both FadR (n = 100) and HutC (n = 60) subfamilies, we observe two fractions of divergons with a TF gene (Table 4, Fig 5A). The first group includes divergons (FadR, n = 29; HutC, n = 32) where the distance between double sites is relatively constant (Fig 6A). In this group, the distance to the proximal binding site tends to be higher for longer intergenic regions in both structural operons and operons with a TF gene (Table 5, Fig 7A and 7B). Thus, double sites in this group of divergons usually tend to be localized near the center of the intergenic area, and we may conclude that they form a pair involved in the co-operative control of both operons.

The second group (FadR, n = 71; HutC, n = 28) consists of divergons where the distance between the sites in a pair linearly increases with the size of the intergenic region (Fig 6B). Thus, these sites are presumably independent, and each of them controls its own operon. There is also a trend towards increasing of the distance to the proximal site for both structural operons and operons with a TF gene, as the intergenic region grows, but this trend is not as

Table 4. Two fractions of divergons with double sites, interdependence of the intergenic distance and the distance between double sites.

	Linear regression coefficient (R^2)		
	Both subfamilies	FadR	HutC
Divergons with common sites (constant inter-site distance)	0,06 (0,26)	0,05 (0,41)	0,06 (0,11)
Divergons with separate sites (increasing inter-site distance)	0,50 (0,53)	0,49 (0,61)	0,49 (0,43)
	Pearson correlation coefficient (p-value)		
	Both subfamilies	FadR	HutC
Divergons with common sites (constant inter-site distance)	0,51 ($p = 3\cdot10^{-5}$)	0,64 ($p = 2\cdot10^{-4}$)	0,32 (p = 0,07)
Divergons with separate sites (increasing inter-site distance)	0,73 ($p<1\cdot10^{-7}$)	0,78 ($p<1\cdot10^{-7}$)	0,65 ($p = 2\cdot10^{-4}$)

doi:10.1371/journal.pone.0132618.t004

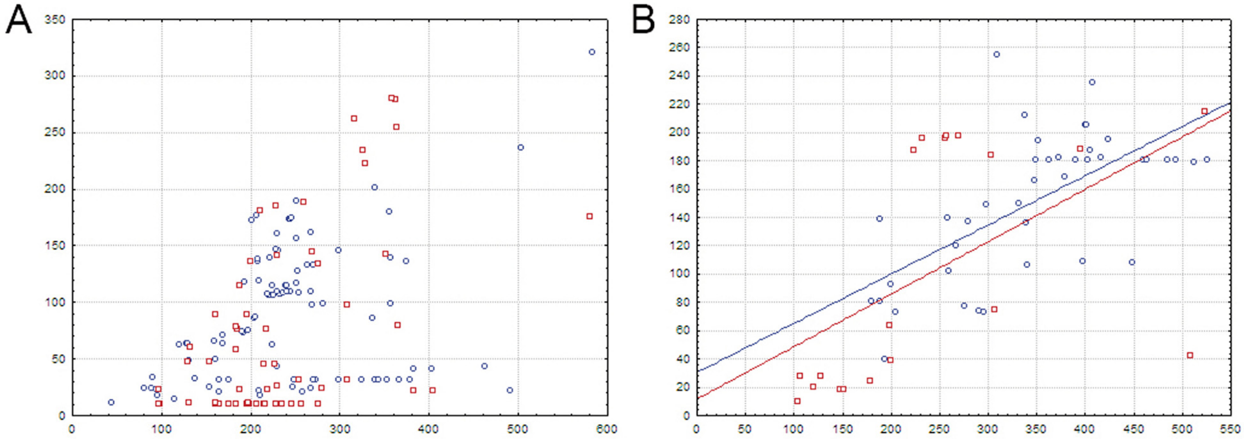

Fig 5. Distances between double sites in divergons. A—divergons with a TF gene; B—the control group. The vertical axis is the inter-site distance. The horizontal axis is the intergenic distance. Notation as in Fig 4.

doi:10.1371/journal.pone.0132618.g005

prominent as in case of divergons with relatively constant inter-site distance (the first group) (Table 6, Fig 8A and 8B). The same tendencies were also observed in the control group in both FadR (n = 46) and HutC (n = 19) subfamilies (Table 6, Figs 5B and 8C). Thus, in the control group there is only one type of divergons, where the sites do not act co-operatively, and each likely regulates the adjacent operon.

Additional half-sites near binding sites of the GntR-family TFs

A typical GntR-family binding motif is a palindrome, but we have found that considerable number of identified palindromic binding sites is accompanied by a weaker adjacent half-site (box) at a distance of 7–12 nt. For a more quantitative analysis, regions flanking candidate binding sites of all studied TFs were considered. In 23 analyzed orthologous groups (13 groups, 170 TFs and 450 binding sites in the FadR subfamily; 4 groups, 186 TFs and 514 sites in the HutC subfamily; and 6 groups, 120 TFs and 167 binding sites in the YtrA subfamily; data not shown) weaker boxes were found at the 7–12 nt distance from the binding site center (on one

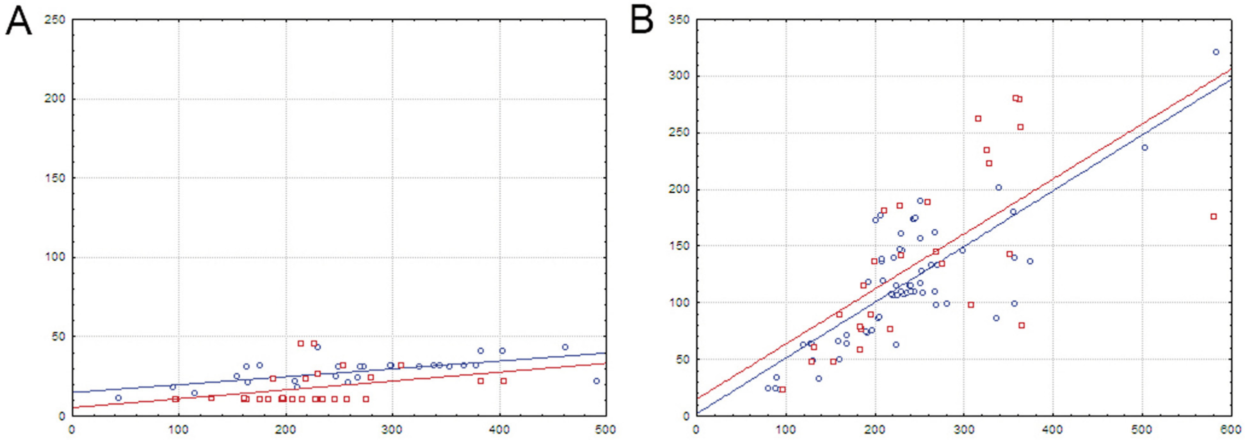

Fig 6. Two groups of divergons with double sites. A—the first group (constant inter-site distance); B—the second group (increasing inter-site distance). For details, see the text. The vertical axis is the inter-site distance. The horizontal axis is the intergenic distance. Notation as in Fig 4.

doi:10.1371/journal.pone.0132618.g006

Table 5. Interdependence of the intergenic distance and the distance to the proximal TF-binding site in divergons with common double sites.

	Linear regression coefficient (R^2)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,48 (0,38)	0,47 (0,40)	0,57 (0,37)
Operons with structural genes only	0,45 (0,38)	0,48 (0,42)	0,37 (0,24)
	Pearson correlation coefficient (p-value)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,62 ($p = 1{\cdot}10^{-7}$)	0,63 ($p = 2{\cdot}10^{-4}$)	0,61 ($p = 2{\cdot}10^{-4}$)
Operons with structural genes only	0,62 ($p = 1{\cdot}10^{-7}$)	0,64 ($p = 2{\cdot}10^{-4}$)	0,49 ($p = 4{\cdot}10^{-3}$)

doi:10.1371/journal.pone.0132618.t005

or both sides of the site). These additional boxes and their positions relative to the center of the binding motif were initially identified by visual analysis of logo diagrams of all aligned binding sites and their neighborhood for TFs forming orthologous groups.

To estimate the significance of this observation, additional boxes were compared to the boxes forming true sites and to the random sequences (pseudoboxes, as control) of the same length. The latter were taken from positions −20 and −21 nt from the binding site (Fig 9). Two pseudoboxes per each binding site were selected to allow for correct estimation of the statistical significance (see below).

The score for each half of a true palindromic site was calculated using the corresponding part of the PWM for this TF ($W_{\text{true left}}$ and $W_{\text{true right}}$, respectively). The same partial PWMs were used to calculate the scores of additional boxes ($W_{\text{near left}}$ and $W_{\text{near right}}$, respectively) and pseudoboxes. The score for each pseudobox was calculated twice using the left and right partial PWMs ($W_{\text{random left1,2}}$ and $W_{\text{random right1,2}}$). Additional boxes with the larger score from each pair ($W_{\text{near left}}$ or $W_{\text{near right}}$) were selected for further analysis. At that, each additional box was compared with the higher scoring one of two pseudoboxes in the same orientation (left or right).

The length and structure of binding motifs and thus PWMs among various orthologous groups of TFs differ, and hence the calculated scores could not be directly compared. To account for the diversity of motifs, all scores were normalized to the respective W_{true} values for

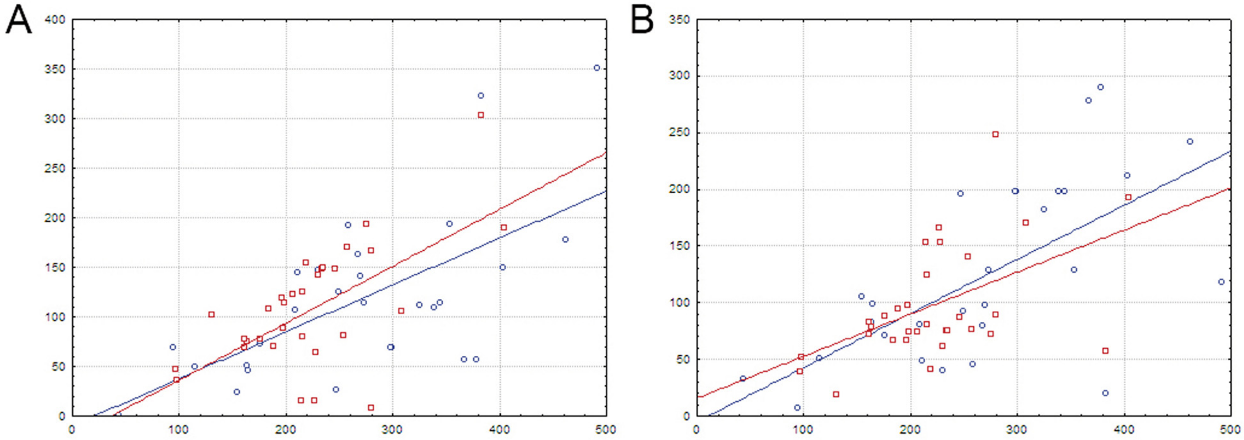

Fig 7. Distances between regulated genes and proximal TF-binding sites in divergons with common double sites. A—operons with a TF gene; B—operons with structural genes only. The vertical axis is the distance between the site center and the start codon. The horizontal axis is the intergenic distance. Notation as in Fig 4.

doi:10.1371/journal.pone.0132618.g007

Table 6. Interdependence of the intergenic distance and the distance to the proximal TF-binding site in divergons with separate double sites.

	Linear regression coefficient (R^2)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,24 (0,29)	0,24 (0,29)	0,25 (0,29)
Operons with structural genes only	0,26 (0,29)	0,27 (0,27)	0,27 (0,37)
Control divergons	0,31 (0,31)	0,33 (0,25)	0,31 (0,34)
	Pearson correlation coefficient (p-value)		
	Both subfamilies	FadR	HutC
Operons with a TF gene	0,54 ($p = 7{\cdot}10^{-9}$)	0,54 ($p = 2{\cdot}10^{-6}$)	0,54 ($p = 3{\cdot}10^{-3}$)
Operons with structural genes only	0,54 ($p = 1{\cdot}10^{-8}$)	0,52 ($p = 3{\cdot}10^{-6}$)	0,61 ($p = 6{\cdot}10^{-4}$)
Control divergons	0,56 ($p<1{\cdot}10^{-7}$)	0,50 ($p = 3{\cdot}10^{-7}$)	0,58 ($p = 1{\cdot}10^{-4}$)

doi:10.1371/journal.pone.0132618.t006

half-sites in the similar orientation as each given additional box:

$$S_{near} = \frac{W_{true} - W_{near}}{W_{true}} \tag{1}$$

$$S_{random} = \frac{W_{true} - W_{random}}{W_{true}} \tag{2}$$

where S_{near} and S_{random} denote normalized weights for additional boxes and pseudoboxes, respectively; W_{true}, W_{near}, W_{random} are respectively the weights of the true half-sites, additional boxes and pseudoboxes, calculated using PWM for the respective TF.

The distributions of S_{near} and S_{random} (Fig 10) were significantly different for all three subfamilies, FadR, HutC, and YtrA (the Wilcoxon rank-sum test, $p<0.001$). Moreover, the average W_{near} value approximately equals half of the average W_{true} value, while average W_{random} value is close to zero, confirming that the used control is correct. These boxes may play a role in the regulation, though their exact function should be a subject of further experimental study.

Conclusions

In this work we identify regulated genes and binding sites for 1252 GNTR-family TFs from the 64 orthologous groups and three subfamilies, FadR, HutC, and YtrA. Using these data, we predict most favorable DNA-protein contacts by analysis of the correlations between amino acids of the TFs and nucleotides of the corresponding binding motifs. Correlation analysis shows that, despite significant differences in the structure of TFs from different subfamilies, main

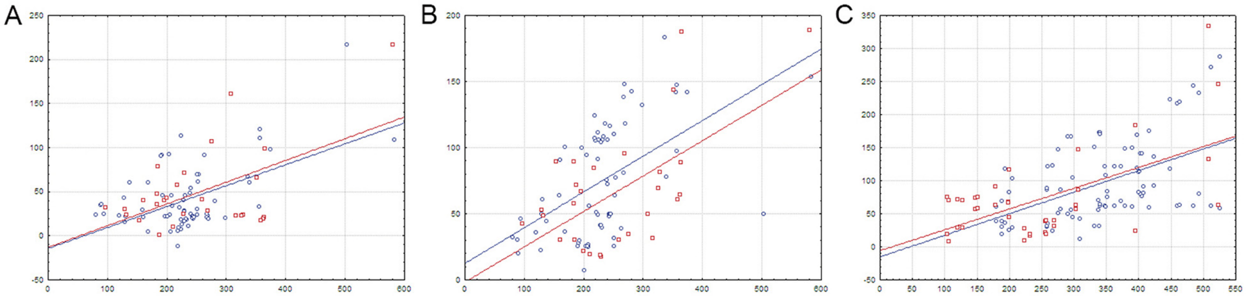

Fig 8. Distances between regulated genes and proximal TF-binding sites in divergons with separate double sites. A—operons with a TF gene; B—operons with structural genes only; C—the control group. The vertical axis is the distance between the site center and the start codon. The horizontal axis is the intergenic distance. Notation as in Fig 4.

doi:10.1371/journal.pone.0132618.g008

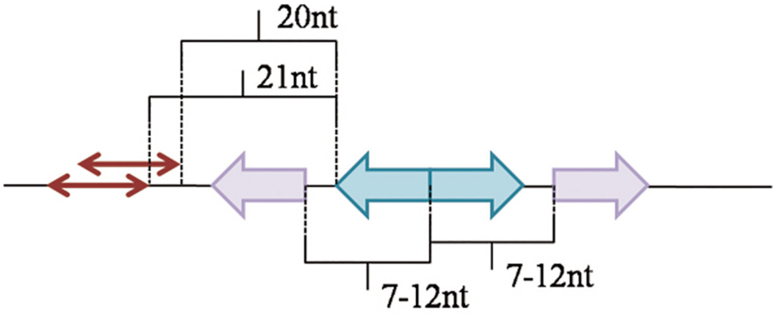

Fig 9. Positioning of additional boxes and control pseudoboxes. True binding half-sites are shown in blue; additional boxes, in violet; pseudoboxes, in red arrows. For details, see the text.

doi:10.1371/journal.pone.0132618.g009

predicted contacts (Arg-G, Asn-A, Asp-C *etc*) are quite similar and conform well to the DNA-protein contacts known for FadR from *E. coli* and AraR from *B. subtilis*, as well as to the interaction trends described in literature.

Apart from identifying usual palindromic binding sites of GntR-family TFs, we also demonstrate that these motifs may sometimes be extended by additional boxes. They may possibly be involved in alternative TF dimerization, or participate in recruiting additional subunits of TFs, their oligomerization and co-operative regulation, thus allowing for the more flexible and precise transcription control.

Analysis of the divergon structure in the FadR and HutC subfamilies revealed some tendencies in the site localization. A single site in a divergon is usually positioned approximately in the middle of the intergenic area and thus may regulate both operons. It is also interesting to note that for divergons with a TF gene, distance between the single binding site and the structural operon increases slower than the distance to the operon comprising a TF gene, as the intergenic distance grows. This might reflect the fact that TF auto-regulation is slightly weaker than regulation of the corresponding structural genes. Double sites are presumably either involved in the cooperative regulation of both operons and are localized in the center of the intergenic area, or each site in the pair independently regulates its own operon and tends to be near it. Thus we classify dual binding sites in divergons into co-operative and operon-specific

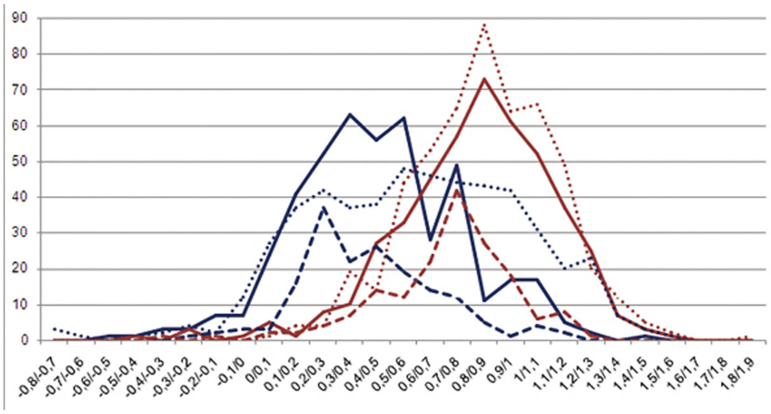

Fig 10. Distribution of S_{near} and S_{random} in the FadR, HutC and YtrA subfamilies. The vertical axis—the number of S values falling in the given interval. The horizontal axis—intervals of S values. Blue color denotes S_{near} values; red color—S_{random} values. FadR subfamily data is shown in continuous lines; HutC subfamily, in dotted lines; YtrA subfamily, in dashed lines.

doi:10.1371/journal.pone.0132618.g010

ones. Unfortunately, we do not find any functional differences between these two types of divergons, since there is no evident distinction in their gene content.

Supporting Information

S1 Table. Genome abbreviations.
(DOC)

S1 File. GntR-family TFs, corresponding binding sites and regulated operons. Datasets for the correlation analysis.
(XLS)

S2 File. Contingency tables of amino acid-nucleotide pairs. Pairs which occur significantly more often than expected are colored red (strongly preferred) and yellow, those occurring less often than expected are colored blue (strongly avoided).
(XLS)

Author Contributions

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

References

1. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics. 2007 Jul 1; 23(13):i347–53. PMID: 17646316
2. Rodionov DA. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem Rev. 2007 Aug; 107(8):3467–97. PMID: 17636889
3. Pérez-Rueda E, Collado-Vides J. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. Nucleic Acids Res. 2000 Apr 15; 28(8):1838–47. PMID: 10734204
4. Lustig B, Jernigan RL. Consistencies of individual DNA base amino acid interactions in structures and sequences. Nucleic Acids Res. 1995 Nov 25; 23(22):4707–11. PMID: 8524664
5. Morozov AV, Siggia ED. Connecting protein structure with predictions of regulatory sites. Proc Natl Acad Sci U S A. 2007 Apr 24; 104(17):7068–73. PMID: 17438293
6. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Proc Natl Acad Sci U S A. 1976 Mar; 73(3):804–8. PMID: 1062791
7. Marabotti A, Spyrakis F, Facchiano A, Cozzini P, Alberti S, Kellogg GE, et al. Energy-based prediction of amino acid-nucleotide base recognition. J Comput Chem. 2008 Sep; 29(12):1955–69. doi: 10.1002/jcc.20954 PMID: 18366021
8. Mirny LA, Gelfand MS. Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res. 2002 Apr 1; 30(7):1704–11. PMID: 11917033
9. Gromiha MM, Fukui K. Scoring function based approach for locating binding sites and understanding recognition mechanism of protein-DNA complexes. J Chem Inf Model. 2011 Mar 28; 51(3):721–9. doi: 10.1021/ci1003703 PMID: 21361378
10. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. Annu Rev Biochem. 2010; 79:233–69. doi: 10.1146/annurev-biochem-060408-091030 PMID: 20334529
11. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein—DNA binding specificity predictions with structural models. Nucleic Acids Res. 2005 Oct 24; 33(18):5781–98. PMID: 16246914
12. Mahony S, Auron PE, Benos PV. Inferring protein—DNA dependencies using motif alignments and mutual information. Bioinformatics. 2007 Jul 1; 23(13):i297–304. PMID: 17646310
13. Huang N, De Ingeniis J, Galeazzi L, Mancini C, Korostelev YD, Rakhmaninova AB, et al. Structure and function of an ADP-ribose-dependent transcriptional regulator of NAD metabolism. Structure. 2009 Jul 15; 17(7):939–51. doi: 10.1016/j.str.2009.05.012 PMID: 19604474

14. Ravcheev DA, Li X, Latif H, Zengler K, Leyn SA, Korostelev YD, et al. Transcriptional regulation of central carbon and energy metabolism in bacteria by redox-responsive repressor Rex. J Bacteriol. 2012 Mar; 194(5):1145–57. doi: 10.1128/JB.06412-11 PMID: 22210771

15. Desai TA, Rodionov DA, Gelfand MS, Alm EJ, Rao CV. Engineering transcription factors with novel DNA-binding specificity using comparative genomics. Nucleic Acids Res. 2009 May; 37(8):2493–503. doi: 10.1093/nar/gkp079 PMID: 19264798

16. Camas FM, Alm EJ, Poyatos JF. Local gene regulation details a recognition code within the LacI transcriptional factor family. PLoS Comput Biol. 2010 Nov 11; 6(11):e1000989. doi: 10.1371/journal.pcbi.1000989 PMID: 21085639

17. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 2002 Jul 26; 320(5):991–1009. PMID: 12126620

18. Zheng M, Cooper DR, Grossoehme NE, Yu M, Hung LW, Cieslik M, et al. Structure of Thermotoga maritima TM0439: implications for the mechanism of bacterial GntR transcription regulators with Zn2+-binding FCD domains. Acta Crystallogr D Biol Crystallogr. 2009 Apr; 65(Pt 4):356–65. doi: 10.1107/S0907444909004727 PMID: 19307717

19. Aravind L, Anantharaman V. HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. FEMS Microbiol Lett. 2003 May 16; 222(1):17–23. PMID: 12757941

20. Rigali S, Derouaux A, Giannotta F, Dusart J. Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. J Biol Chem. 2002 Apr 12; 277(15):12507–15. PMID: 11756427

21. Rigali S, Schlicht M, Hoskisson P, Nothaft H, Merzbacher M, Joris B, et al. Extending the classification of bacterial transcription factors beyond the helix—turn—helix motif as an alternative approach to discover new cis/trans relationships. Nucleic Acids Res. 2004 Jun 24; 32(11):3418–26. PMID: 15247334

22. König B, Müller JJ, Lanka E, Heinemann U. Crystal structure of KorA bound to operator DNA: insight into repressor cooperation in RP4 gene regulation. Nucleic Acids Res. 2009 Apr; 37(6):1915–24. doi: 10.1093/nar/gkp044 PMID: 19190096

23. Sharadamma N, Khan K, Kumar S, Patil KN, Hasnain SE, Muniyappa K. Synergy between the N-terminal and C-terminal domains of Mycobacterium tuberculosis HupB is essential for high-affinity binding, DNA supercoiling and inhibition of RecA-promoted strand exchange. FEBS J. 2011 Sep; 278(18):3447–62. doi: 10.1111/j.1742-4658.2011.08267.x PMID: 21787377

24. Brinkman AB, Ettema TJ, de Vos WM, van der Oost J. The Lrp family of transcriptional regulators. Mol Microbiol. 2003 Apr; 48(2):287–94. PMID: 12675791

25. Wiethaus J, Schubert B, Pfänder Y, Narberhaus F, Masepohl B. The GntR-Like Regulator TauR Activates Expression of Taurine Utilization Genes in Rhodobacter capsulatus. J Bacteriol. 2008 Jan; 190(2):487–93. PMID: 17981966

26. Lee MH, Scherer M, Rigali S, Golden JW. PlmA, a new member of the GntR family, has plasmid maintenance functions in Anabaena sp. strain PCC 7120. J Bacteriol. 2003 Aug; 185(15):4315–25. PMID: 12867439

27. Zhang L, Leyn SA, Gu Y, Jiang W, Rodionov DA, Yang C. Ribulokinase and transcriptional regulation of arabinose metabolism in Clostridium acetobutylicum. J Bacteriol. 2012 Mar; 194(5):1055–64. doi: 10.1128/JB.06241-11 PMID: 22194461

28. Franco IS, Mota LJ, Soares CM, de Sá-Nogueira I. Functional domains of the Bacillus subtilis transcription factor AraR and identification of amino acids important for nucleoprotein complex assembly and effector binding. J Bacteriol. 2006 Apr; 188(8):3024–36. PMID: 16585763

29. Franco IS, Mota LJ, Soares CM, de Sá-Nogueira I. Probing key DNA contacts in AraR-mediated transcriptional repression of the Bacillus subtilis arabinose regulon. Nucleic Acids Res. 2007; 35(14):4755–66. PMID: 17617643

30. Yang C, Rodionov DA, Li X, Laikova ON, Gelfand MS, Zagnitko OP, et al. Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of Shewanella oneidensis. J Biol Chem. 2006 Oct 6; 281(40):29872–85. PMID: 16857666

31. Belitsky BR. Bacillus subtilis GabR, a protein with DNA-binding and aminotransferase domains, is a PLP-dependent transcriptional regulator. J Mol Biol. 2004 Jul 16; 340(4):655–64. PMID: 15223311

32. Bramucci E, Milano T, Pascarella S. Genomic distribution and heterogeneity of MocR-like transcriptional factors containing a domain belonging to the superfamily of the pyridoxal-5'-phosphate dependent enzymes of fold type I. Biochem Biophys Res Commun. 2011 Nov 11; 415(1):88–93. doi: 10.1016/j.bbrc.2011.10.017 PMID: 22020104

33. Magarvey N, He J, Aidoo KA, Vining LC. The pdx genetic marker adjacent to the chloramphenicol biosynthesis gene cluster in Streptomyces venezuelae ISP5230: functional characterization. Microbiology. 2001 Aug; 147(Pt 8):2103–12. PMID: 11495988

34. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009 Jun 26; 324(5935):1720–3. doi: 10.1126/science.1162327 PMID: 19443739

35. Quail MA, Dempsey CE, Guest JR. Identification of a fatty acyl responsive regulator (FarR) in Escherichia coli. FEBS Lett. 1994 Dec 19; 356(2–3):183–7. PMID: 7805834

36. Rodionov DA, Gelfand MS. Computational identification of BioR, a transcriptional regulator of biotin metabolism in Alphaproteobacteria, and of its binding signal. FEMS Microbiol Lett. 2006 Feb; 255 (1):102–7. PMID: 16436068

37. Feng Y, Zhang H, Cronan JE. Profligate biotin synthesis in α-proteobacteria—a developing or degenerating regulatory system? Mol Microbiol. 2013 Apr; 88(1):77–92. doi: 10.1111/mmi.12170 PMID: 23387333

38. Condemine G, Berrier C, Plumbridge J, Ghazi A. Function and expression of an N-acetylneuraminic acid-inducible outer membrane channel in Escherichia coli. J Bacteriol. 2005 Mar; 187(6):1959–65. PMID: 15743943

39. Jochmann N, Götker S, Tauch A. Positive transcriptional control of the pyridoxal phosphate biosynthesis genes pdxST by the MocR-type regulator PdxR of Corynebacterium glutamicum ATCC 13032. Microbiology. 2011 Jan; 157(Pt 1):77–88. doi: 10.1099/mic.0.044818-0 PMID: 20847010

40. Xu Y, Heath RJ, Li Z, Rock CO, White SW. The FadR-DNA Complex. Transcriptional control of fatty acid metabolism in Escherichia coli. J Biol Chem. 2001 May 18; 276(20):17373–9. PMID: 11279025

41. van Aalten DM, DiRusso CC, Knudsen J. The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR. EMBO J. 2001 Apr 17; 20(8):2041–50. PMID: 11296236

42. Jain D, Nair DT. Spacing between core recognition motifs determines relative orientation of AraR monomers on bipartite operators. Nucleic Acids Res. 2013 Jan 7; 41(1):639–47. doi: 10.1093/nar/gks962 PMID: 23109551

43. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, et al. GenBank. Nucleic Acids Res. 1999 Jan 1; 27(1):12–7. PMID: 9847132

44. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1; 25 (17):3389–402. PMID: 9254694

45. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug 19; 5:113. PMID: 15318951

46. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 1996; 266:418–27. PMID: 8743697

47. Gelfand MS, Koonin EV, Mironov AA. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucleic Acids Res. 2000 Feb 1; 28(3):695–705. PMID: 10637320

48. Mironov AA, Vinokurova NP, Gel'fand MS. Software for analyzing bacterial genomes. Mol Biol (Mosk). 2000 Mar-Apr; 34(2):253–62.

49. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, et al. RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. Nucleic Acids Res. 2010 Jul; 38(Web Server issue):W299–307. doi: 10.1093/nar/gkq531 PMID: 20542910

50. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004 Jun; 14(6):1188–90. PMID: 15173120

51. Hill T, Lewicki P (2007) STATISTICS: Methods and Applications. StatSoft, Tulsa, OK.

52. Martínez-Antonio A, Janga SC, Thieffry D. Functional organisation of Escherichia coli transcriptional regulatory network. J Mol Biol. 2008 Aug 1; 381(1):238–47. doi: 10.1016/j.jmb.2008.05.054 PMID: 18599074

53. Tan K, McCue LA, Stormo GD. Making connections between novel transcription factors and their DNA motifs. Genome Res. 2005 Feb; 15(2):312–20. PMID: 15653829

54. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays. 1998 May; 20 (5):433–40. PMID: 9670816

55. Madar D, Dekel E, Bren A, Alon U. Negative auto-regulation increases the input dynamic-range of the arabinose system of Escherichia coli. BMC Syst Biol. 2011 Jul 12; 5:111. doi: 10.1186/1752-0509-5-111 PMID: 21749723

56. Suvorova IA, Ravcheev DA, Gelfand MS. Regulation and evolution of malonate and propionate catabolism in proteobacteria. J Bacteriol. 2012 Jun; 194(12):3234–40. doi: 10.1128/JB.00163-12 PMID: 22505679

57. Mota LJ, Sarmento LM, de Sá-Nogueira I. Control of the arabinose regulon in Bacillus subtilis by AraR in vivo: crucial roles of operators, cooperativity, and DNA looping. J Bacteriol. 2001 Jul; 183(14):4190–201. PMID: 11418559