Mathematical Biology and Bioinformatics
2023. V. 18. № 1. P. 1–14. doi: 10.17537/2023.18.1

========== **BIOINFORMATICS** ==========

Phylogeny and cross-regulation of the YjjM and LeuO transcription factors translated as multiple protein forms from one gene in *Escherichia coli*

Bessonova T.A.[1*], Rybina A.A.[2], Marakulina D.A.[1,3], Kaznadzey A.D.[4], Gelfand M.S.[2,4], Ozoline O.N.[1], Tutukina M.N.[1,2,4]

[1]*Institute of Cell Biophysics RAS* (*Federal Research Center "Pushchino Scientific Center for Biological Research of the RAS"*), *Pushchino, Russia*
[2]*Skolkovo Institute of Science and Technology, Moscow, Russia*
[3]*Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia*
[4]*Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia*

Abstract. Until recently, no examples of the in-frame translation of several proteins from one gene in bacteria were known. The first one was the VirF transcription factor controlling pathogenicity development in *Shigella flexneri* and CobB sirtuin in *Salmonella enterica.* Recently, we observed synthesis of shortened protein products for YjjM (LgoR) and LeuO functioning as transcription factors in *Escherichia coli.* To determine the evolutionary factors that could lead to the appearance of alternative start codons, we performed phylogenetic analysis and showed that each protein had a unique phylogenetic history, and additional starting methionines appeared only in Enterobacteria. Using the Western-blot analysis of proteins synthesized from the *Escherichia coli* K-12 MG1655 chromosome with the his-tagged *leuO* gene two shortened variants of LeuO, corresponding to translation starting from Met34 and Met48 were detected. Synthesis of all three LeuO forms was inhibited in the absence of the *yjjM* gene, suggesting interplay of these transcription factors. The YjjM recognition motif revealed from the ChIP-seq data appeared to be very similar to that of LeuO, known previously. Taking this into account, we compared ChIP and SELEX data for LeuO and YjjM and found six common targets. At least five of them were confirmed to be under control of these regulators by qRT-PCR. Interestingly, the effects were more prominent during anaerobic growth at 37°C simulating conditions inside a host organism. Two genes, coding for the enterobactin transporter FepA, and a repressor of genes responsible for flagellar biosynthesis and virulence YjjQ, were repressed, mainly by YjjM, only in these conditions, while *tsr* coding for the chemotaxis receptor protein was more repressed under lower temperature and higher aeration.

Key words: *Escherichia coli, transcription regulation, LeuO, YjjM (LgoR), regulon, pathogenicity control, phylogenetic analysis, evolution.*

INTRODUCTION

Transcription factors (TFs) play a crucial role in bacterial adaptation to environmental changes. The mechanisms of their participation in the regulation of transcription, originally proposed by the operon model of Jacob and Monod, turned out to be very diverse, and new aspects of TFs functioning are still being discovered. [1, 2]. Based on the number and scale of controlled targets, global and local transcriptional regulators are distinguished [3]. Here, we study two regulators – YjjM (LgoR), mainly involved in control of hexuronate metabolism, and LeuO, a more general regulator of cellular processes. A common feature of these TFs is

that both have several forms transcribed from a single gene, hence allowing for an additional mechanism to control their expression and functioning. Both YjjM and LeuO are involved in the regulation of the *Escherichia coli (E. coli)* virulence and motility [4, 5].

The first examples of multiple protein forms synthesized in-frame from one gene were the CobB sirtuin in *Salmonella enterica* and the VirF transcription factor controlling pathogenicity development in *Shigella flexneri*. In *Salmonella enterica,* the 236-amino-acid isoform CobB(S) and a longer 273-amino-acid isoform CobB(L) were observed, both of which were biologically active in the cell [6]. The N-terminal domain, present only in the CobB-L isoform, is required for the protein dimerization, while the short variant (CobB-S) has a deacetylase activity, targeting K73 of Alanyl-tRNA Synthetase [7]. The *cobB* gene is primarily expressed as two monocistronic *cobB* mRNAs from two promoters, one of which is mapped within the adjacent *ycfX* gene and the other located within *cobB*. The two isoforms are common among all Enterobacteriaceae except for *Erwinia sp*, where only CobB(L) is produced [6].

VirF is a key factor of host invasion by *Shigella spp*. Its synthesis is activated mainly in response to the temperature change when *Shigella* enters the host intestine. VirF belongs to the AraC family of TFs and controls expression of a cascade of genes involved in the intracellular proliferation of bacteria [8]. The *virF* gene encodes two independently translated polypeptides $VirF_{30}$ (30 kDa) and $VirF_{21}$ (21 kDa), lacking the N-terminal segment [8]. Same as for CobB, $VirF_{21}$ is translated from a mRNA transcribed from an internal start point which is located only one or two nucleotides upstream of the ATG84 start codon of $VirF_{21}$. The functions of the resulting protein products are largely different: $VirF_{30}$ is responsible for the activation of the virulence system, and $VirF_{21}$ negatively autoregulates the *virF* expression itself. Under lower temperatures outside a host organism, the *virF* expression is silenced by H-NS [8, 9].

In *E. coli*, there are several examples of overlapping genes coding for different proteins. Among them there are genes located on different DNA strands (*mbiA/yaaW*, *ytfP*/*yzfA*) and genes of toxin-antitoxin systems encoding proteins with in-frame translation (*hokA/mokA hokB/mokB*, *hokC/mokC*), but no examples of such genes encoding transcription factors have been found so far. However, in our pilot studies we detected synthesis of shorter forms for several TFs including YjjM, and LeuO.

YjjM is a GntR-family transcriptional regulator that belongs to the regulons of UxuR and ExuR, controlling hexuronate metabolism [10, 11]. It was initially predicted as an activator of the L-galactonate metabolism but later we have demonstrated that depending on the growth conditions YjjM regulates both hexuronate unitization and bacterial virulence in *E. coli*. Similarly to *virF*, the *yjjM* expression can be silenced by H-NS.

H-NS also controls expression of *leuO*, which encodes a dual global regulator. LeuO belongs to the LysR family [12], and was first described as an activator of leucine biosynthesis operon in *Salmonella Typhimurium* [13]. In further studies, LeuO was shown as a global regulator playing a role in pathogenicity in *Salmonella*, biofilm formation in *Vibrio cholerae* and, presumably, in *E. coli.* [4, 14, 15, 16]. SELEX experiments indicated that the LeuO targets overlapped with the targets repressed by H-NS [4, 5].

Here, using comparative genomics we study the conservation of alternative translation starts for these two TFs across bacterial species and among different *E. coli* strains, assess the synthesis of alternative LeuO protein forms in the wild type *E. coli* and its Δ*yjjM* derivative. We also analyze the intersection of the LeuO and YjjM regulons and test their participation in regulation of common targets experimentally.

MATERIALS AND METHODS

Genomic and protein data

Complete bacterial genomes of the latest version available at 8 March 2022, the respective translated coding DNA sequences (CDS), and annotation files in the GFF format were downloaded from the NCBI RefSeq FTP server [17]. The Pfam protein database, or Pfamseq version 35.0, was obtained from the EMBL-EBI FTP server [18]. Metadata for cross-referencing protein unique identifiers (ID) was downloaded from the FTP UniProt [19]. Taxonomy information was obtained from the FTP NCBI [20].

To find homologs of YjjM and LeuO, we first selected those translated CDS sequences that had been annotated as the respective proteins; records annotated as pseudogenes were discarded. The resulting datasets contained only close homologs mainly belonging to the *E. coli* strains and some Enterobacteriales species. To include more distant homologs, we expanded the dataset using several approaches.

To find YjjM homologs, BLAST 2.9.0+ was used [21]. BLASTP was run using the YjjM protein sequence from *E. coli* K-12 MG1655 (U00096.3) as a query against a local protein database constructed from translated CDS sequences (param eters: -outfmt 7 -evalue 1e-3 -num_descriptions 30000 -num_alignments 30000). Records with E-value < 0.001 and sequence identity > 30 % were retrieved from a local database using samtools v1.11 (samtools faidx command) [22].

LeuO belongs to a very diverse LysR family of TFs, and is very diverse by itself. As such, it was difficult to distinguish it from the other regulators of the same protein family. Thus, instead of using BLASTP we retrieved LeuO orthologs at the EggNog database version 5.0.0 (orthologous group ENOG501MX24) [23]. The resulting set of LeuO records included sequences from both the RefSeq and EggNog databases.

Phylogenetic tree reconstruction

Multiple alignment of protein sequences was obtained via MAFFT v7.475 with the default parameters [24]. Phylogenetic trees were built using FastTree v2.1.11 No SSE3 with the default options [25]. To reduce the number of nodes in each phylogenetic tree, protein sequences from organisms with the RefSeq category set to a representative genome were selected. Representative protein sequences were obtained based on linear clustering via MMseqs2 v13-45111 [26]. To prune phylogenetic trees, to align sequences and to visulaize the respective trees Biopython v1.76 [27] was used. Trees were displayed at the ITOL web-server [28]. Annotation files for ITOL were generated using *ad hoc* developed python scripts.

During construction of the YjjM phylogenetic tree for the *E. coli* strains, duplicated sequences were removed, and the YjjM protein sequence of *Kluyvera intermedia* str. N2-1 was used as an outgroup.The YjjM tree containing distant YjjM homologs was manually rooted by the cluster of YjjM homologs from Beta- and Alphaproteobacteria. Sequences of NhaR, OxyR, and LysR from the reference *E. coli* str. K-12 MG1655 were used as an outgroup for the LeuO phylogenetic tree.

Gene neighborhood analysis

To analyse the *yjjM* gene neighborhood, five open reading frames (ORFs) were taken downstream and upstream of the *yjjM* gene from each *E. coli* representative genome and used to construct orthologous groups. Local nucleotide database was built from these ORFs. Nucleotide BLAST search was run against a local database using ORFs of the reference *E. coli* str. K-12 MG1655 as query sequences. Sequences that were not included in the BLASTN output, were subjected to nucleotide all-against-all BLAST search. ORFs were plotted using DNA Features Viewer v3.0.3 [29].

Gene-target intersection analyses between LeuO and YjjM

To find common targets for LeuO and YjjM, data from the SELEX and ChIP-chip for LeuO protein obtained from [5, 15, 30] and the ChIP-seq data for YjjM (M.N. Tutukina, unpublished results) were used. The ChIP-seq data were analysed using MACS2 [31] and bedtools [32] were used to find the location of the ChIP-seq peaks. Motif logo for YjjM was built with the MEME suite [33]

Western-blot analysis

To study the synthesis of protein forms from the *leuO* gene, Western-blot analysis of *E. coli* K-12 MG1655 *leuO*-6xHis-tag and *E. coli* K-12 *ΔyjjM leuO*-6xHis-tag cells exponentially growing in LB media (OD_{600} = 0.4) was performed. Cells were grown either in conditions simulating free living bacteria (30°C aerobic) or growth inside a host organism (37°C anaerobic). Addition of nucleotides coding for 6 histidines to the 3'-end of the *leuO* gene was made using Gene doctoring [34]. Samples in a standard Laemmli loading buffer were run on a denaturing 10 % polyacrylamide gel at 20mA/gel [35], and then transferred onto a PVDF membrane using Trans-Blot Turbo System according to the manufacturer's protocol (Bio-Rad, USA). After protein transfer, the PVDF membrane was blocked for 1 hour in 5 % skimmed milk (Oxoid, UK) dissolved in the TBS buffer (50 mM Tris, 273 mM NaCl, pH 8.0). Then, the membrane was incubated 2 hours with the 6x-His tag rabbit polyclonal antibody (1:10000, Cat № PA-19838, Invitrogen, USA) in the TBS buffer with 1:1000 Tween-20 (TBS-T), following 1 hour incubation with secondary Anti-rabbit IgG HRP-linked antibody (1:10000, Cat № 7074S, Cell Signaling Technology, USA) in TBS-T. After three washes with TBS-T, the membrane was stained using Luminata Forte HRP substrate (Millipore, USA) and visualized on the ChemiDoc Imaging system (BioRad, USA) with 20 min exposure in the signal accumulation mode.

RNA isolation and qRT-PCR

For total RNA isolation, *E. coli* K-12 MG1655 and its *leuO* and *yjjM* deletion derivatives were grown on the M9 minimal medium with 5 % LB under conditions simulating free growth (30°C aerobic) and growth inside a host organism (37°C anaerobic). Cell growth was stopped at the exponential phase (OD_{600} = 0.2–0.4, after 4–4.5 hours) by centrifugation at 10 000 rpm 10 min at +4°C.

RNA was isolated using TriZol reagent (Invitrogen, USA) following the manufacturer's protocol. To remove any residual DNA, samples were treated with DNAse I (Promega, USA). Reverse transcription was performed using M-MulV RevertAid reverse transcriptase (Thermo Scientific, USA) as described in [36] using gene specific primers (Evrogen, Russia) listed in Table 1. Amplification was performed on the DT-lite PCR machine (DNA-technology, Russia). Amplification program for all primer pairs was as follows: 94 °C – 1 min; 94 °C – 20 sec, 56 °C – 30 sec, 72 °C – 25 sec, 72 °C – 10 sec (detection): 40 cycles. Antisense products of the *ysaA* gene were used as house-keeping [37].

qRT-PCR with the LeuO open reading frames was made with the RNA isolated from K-12 MG1655 *leuO*-6xHis-tag. *sapA* was used as the house-keeping gene.

RESULTS AND DISCUSSION

Conservation of alternative YjjM methionines in Proteobacteria

YjjM (LgoR) protein and its homologs fall into two clearly separated groups among Proteobacteria: highly conserved YjjM in Gammaproteobacteria and distant YjjM homologs in Alpha- and Betaproteobacteria (Fig. 1).

There are two additional conserved methionines, located 19 and 37 amino acids downstream (Met19 and Met37) of the main annotated start, from which translation of the

YjjM shortened forms can start in the *E. coli* strains and some related species. Structure of the YjjM protein tree and respective multiple sequences alignment indicate that both additional methionines might have emerged within the Enterobacteriales order (Fig. 1).

Table 1. Gene-specific primers for qRT-PCR

Primer	Nucleotide sequence
fes_PCR	5′– GGAATGGCAGCGTCTGAATG –3′
fes_RT	5′– TCCACTGCCAGACGTTAGTG –3′
sdiA_PCR	5′– GATGCTGTTGCGTTTTCAGA –3′
sdiA_RT	5′– GCCTCAGGGTAATTGGTGTA –3′
tsr_PCR	5′– CGGAGCTGATCCAACTGTTA –3′
tsr_RT	5′– GGAGGCATTGTTATCGCTGA –3′
ybeQ_PCR	5′– ACGTCAAGTTGCTGCGATAA – 3′
ybeQ_RT	5′– GTGCTTCACAATGACCTTGC – 3′
acrE_PCR	5′– CATGCCAGGTTTTTCCTCCT – 3′
acrE_RT	5′– GCGAACTTCGGCTATACGAT – 3′
fepA_PCR	5′– TTCATTCCCTGGCCTTGTTG – 3′
fepA_RT	5′– GCGCCTGTAAGTTCTGCTC – 3′
ysaA _RT	5′– ACGACCTCCATCGCACCAAACG – 3′
ysaA _PCR	5′– GCAGTAGCCTGTCATCAGTGTGAAGAT – 3′
leuO_antiRT_-41 (1)	5′ – CAGGATTATTTCTCTGCATTC – 3′
leuO_RT1 (2)	5′ – ACTCCGCCGTCTCTGGATG – 3′
leuO_antiRT_132 (3)	5′– TTCTGGGAATGTCGCAACC –3′
leuO_antiRT_270 (4)	5′ – CAGGCATTGCAACTAGTACA – 3′
leuO_RT3 (5)	5′ – CTGCGAGGTCAGAATGCTGT – 3′
sapA_PCR	5′ – ACCTTTAACCCATCCAAAGC – 3′
sapA_RT	5′ – CGGCATTCATTTTACGAGTG – 3′

From Figure 1 we can observe that Met19 is present in some Enterobacteriaceae (e.g. *Klebsiella* sp., *Citrobacter* sp., *Pectobacterium* sp., *Escherichia coli*), whereas Met37 is found not only in Enterobacteriaceae but also in genomes from other families including Yersiniaceae (e.g. *Chania multitudinisentens* RB-25, *Serratia* sp., *Yersinia hibernica* str. CFS1934), Erwiniaceae (*Erwinia billingiae* Eb661, *Pantoea dispersa* str. Lsch), Budviciaceae (e.g. *Leminorella richardii* str. NCTC12151), and Morganellaceae (e.g. *Providencia* sp.). This indicates that Met37 and hence the shortest protein variant, has presumably appeared earlier than Met19. Some species may have later lost one of these methionines due to secondary mutations.

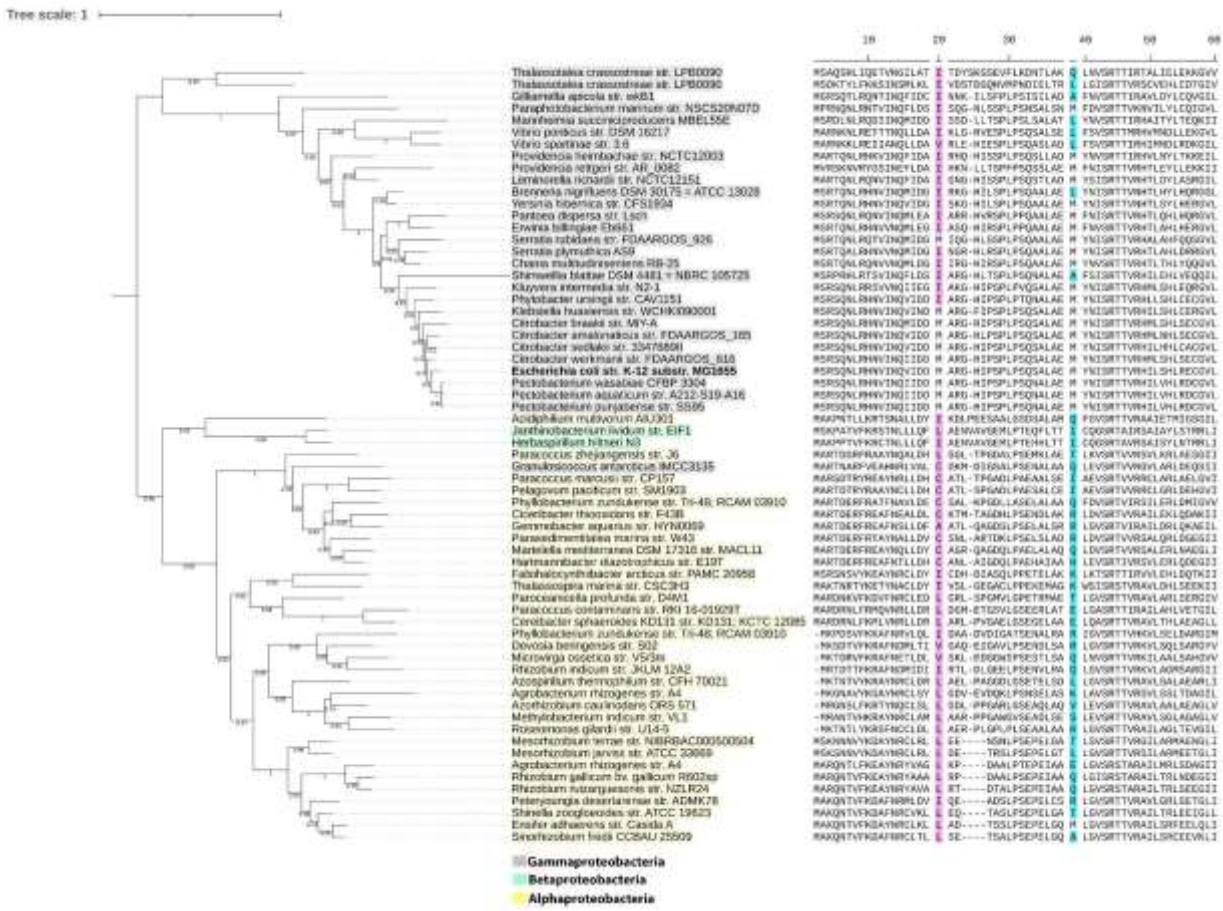

Fig. 1. Phylogenetic representation of YjjM and its homologs among Gamma-, Beta-, and Alphaproteobacteria. YjjM protein phylogenetic tree was inferred using maximum-likelihood algorithm. The tree was manually rooted by the branch consisting of YjjM homologs from Beta- and Alphaproteobacteria. First sixty positions of respective multiple sequence alignment are displayed. Non-methionine amino-acid residues aligned to Met19 and Met37 of the reference *E. coli* str. K-12 MG1655 are marked pink and blue, respectively. Full size picture

Close YjjM homologs from Alteromonadales (*Thalassotalea* sp), Orbales (e.g. *Gilliamella apicola* str. wkB1), Pasteurellales (e.g. *Mannheimia succiniciproducens* MBEL55E) and Vibrionales (*Vibrio* sp.) lack candidate starts for shorter protein forms as well as distant YjjM homologs from various Alpha- and Betaproteobacteria, the latter forming a separate clade on the tree. Rare exceptions are YjjM from *Ensifer adhaerens* str. Casida A (Alphaproteobacteria) and *Paraphotobacterium marinum* str. NSCS20N07D (Vibrionales) that possess Met37; it is not clear whether these starts are functional.

Interestingly, non-methionine amino acid residues (leucine, isoleucine, valine) that are aligned to Met19 from YjjM of the reference *E. coli* in Gamma- and some Alphaproteobacteria have physical properties similar to methionine, all being non-polar aliphatic residues. Conversely, Met37 shares the position with polar residues, neutral glutamine (Q) and threonine (T); positive lysine (K) and histidine (H); and negative glutamic acid (E). This may indicate selection towards maintaining tertiary structure of the DNA-binding domain of YjjM.

In some pathogenic *Shigella* species - such as *S. sonei* Ss046, *S. boydii* Sb227, *S. flexneri* 2a and 5, and others, Met37 was annotated as the YjjM translation start. The same possible start was annotated for the pathogenic *E. coli* O157:H7 EDL933. According to our search for ORFs in the *yjjM* genes from these genomes, this is likely a result of miss-annotation. However, taking into account that YjjM belongs to the GntR family of regulators with the N-

terminal DNA-binding domain, this heterogeneity may reflect differences in its interaction with the target promoters. We thus checked conservation of YjjM among the *E. coli* strains.

YjjM has two sequence variants with different genomic context in the *E. coli* strains

YjjM is present in almost all tested *E. coli* strains differing mainly in the N-terminal DNA-binding domain and in the very end of the C-terminal ligand-binding domain (Suppl. Fig. 1). To check whether this variation is related to the genomic context of the *yjjM* gene, we studied gene co-localisation pattern for all the strains on the YjjM tree. In all cases, the upstream gene neighborhood was conserved. Respective gene functions were related either to DNA replication (*dnaT*, *dnaC*, and possibly *yjjA*) or carbohydrate metabolism (*opgD* and *lgoD*) (Fig. 2,B).

The downstream context was more variable, demonstrating five different gene combinations located between *yjjM* and *tsr* that encodes the methyl-accepting chemotaxis protein (Fig. 2,B). Separation of the protein tree into two main branches correlated with the presence of the *lgoT* gene coding for galactonate/H^+ symporter right next to *yjjM*: in a large cluster *yjjM* is co-localised with *lgoT* and *tsr* while a smaller branch lacks *lgoT*. Instead, there are genes encoding other transporters: Na^+/H^+ antiporter, C4-dicarboxylate TRAP transporter substrate-binding protein, TRAP transporter small permease, and TRAP transporter large permease. Considering the presence of *tsr* downstream of the *yjjM* gene in both groups of *E. coli*, we suggest that replacement of the *lgoT* was probably a result of a genomic rearrangement. Based on sequence variations affecting the key functional domains and the difference in the genomic context we might assume that the functions of YjjM from two branches of the tree are slightly different. This observation can be especially important because the small branch includes mostly pathogenic strains, e.g. biofilm-forming uropathogenic UPEC132 [38].

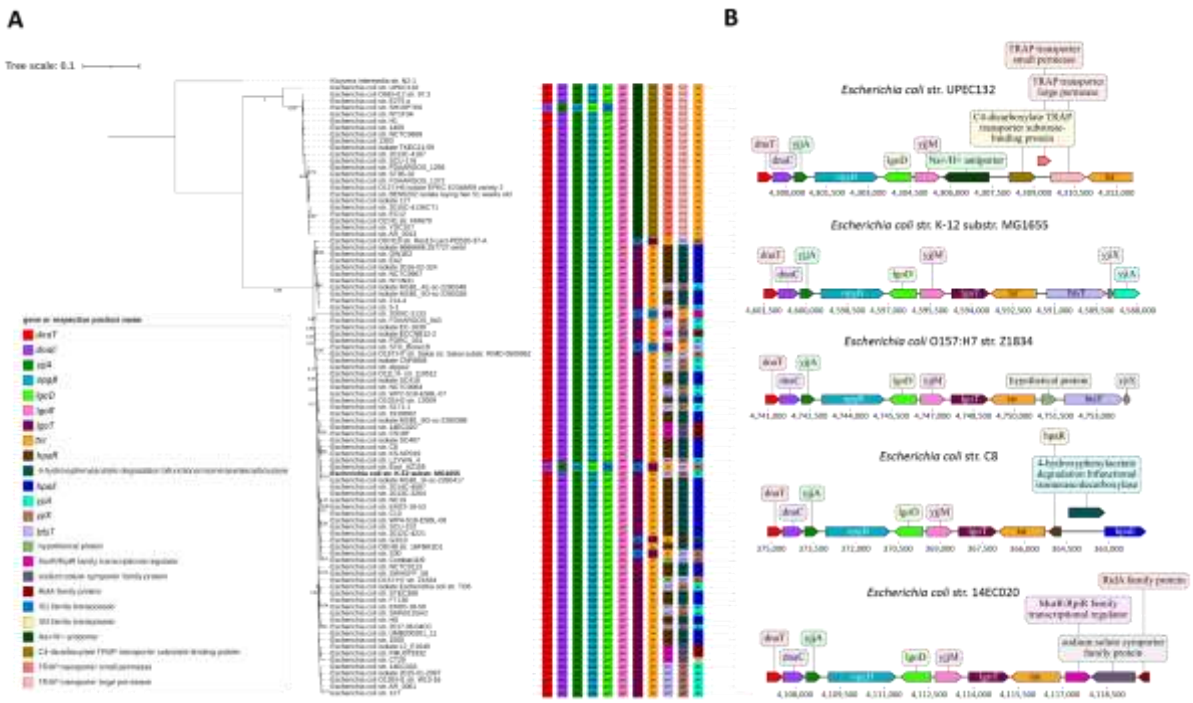

Fig. 2. YjjM protein phylogenetic tree based on YjjM protein sequences of *Escherichia coli* strains and the genomic neighbourhood of the *yjjM* gene in respective genomes. Each gene and respective orthologous group is marked the same colour according to the colour scheme specified to the left of the tree (**A**) YjjM protein phylogenetic tree combined with the genomic context of the *yjjM* gene for each *E. coli* strain presented in the tree. (**B**) Five main variants of the *yjjM* gene neighbourhood from (**A**) visualised for several *E. coli* strains. Gene colouring is the same as in (**A**). Full size picture

LeuO phylogenetic tree reconstruction: three alternative methionines can initiate translation of shortened protein forms

Same as for YjjM, LeuO forms two main branches on the phylogenetic tree: one clade contains close homologs from species belonging to Enterobacteriales and Vibrionales orders, while another includes more distant homologs from Gamma-, Alpha-, and Betaproteobacteria (Fig. 3).

There are three methionines positioned at +34 (Met34), +48 (Met48) and +63 (Met63) from the annotated translation start in the *E. coli* which could be alternative starting points for the translation of shortened LeuO forms. All three of them are present mainly in Enterobacteriales and Vibrionales. Occasionally, alternative starts can be found in other Gammaproteobacteria species including *Pseudomonas* spp. (Met34 and Met48) and *Acinetobacter* spp. (Met48), with few exceptions likely being secondary mutations (Fig. 3).

It is worth noting that there is no specificity to alternative substitutions to the second (Met34) and third (Met48) start methionines. In the same position we observe non-polar aliphatic (valine, isoleucine, leucine), non-polar aromatic (phenylalanine), and polar basic (arginine) residues. We assume that LeuO could initially have had alternative inner methionines responsible for short protein forms, and while they were mostly found among Gammaproteobacteria species, in other species they were later lost due to secondary mutations.

Compared to Met48, Met63 is present in a smaller group of bacteria, all belonging to the Enterobacteriales and Vibrionales orders. Most organisms from this branch have both Met48 and Met63. There are minor exceptions such as *Salmonella sp.* SJTUF14178 and *Enterobacter cloacae* complex str. AR_0002 lacking Met48. In *Cedecea lapagei* str. NCTC11466 and *Mangrovibacter* sp. MFB070 Met63 is replaced by Threonine, which is most likely a random mutation.

Alternative start methionines were presumably originally present in the ancestor LeuO protein, and retained in some Gammaproteobacteria such as Vibrionales, Enterobacteriales (Met63, Met48, and Met34), Cellvibrionales, Oceanospirillales, several Pseudomonadales and Moraxellales species (Met34 and Met48), whereas in other bacteria they were replaced by different amino-acid residues.

Another candidate translation start methionine located at +21 can be found in the *E. coli* strains alignment (Suppl. Fig. 2).

To check the possibility of synthesis of the LeuO shortened forms in bacterial cells we made western-blot using anti-his antibodies and *E. coli* K-12 MG1655 strain where 6 histidines had been added to the 3'-end of *leuO* on the chromosome.

LeuO can be translated as three shortened protein forms in *E. coli* K-12 MG1655

Western blot was performed for the *E. coli* K-12 MG1655 *leuO*-6xHis-tag and *E. coli* K-12 Δ*yjjM leuO*-6xHis-tag cells growing either at 30°C and aeration or at 37 °C and anaerobic conditions. In addition to the main LeuO form (36.5 kDa), two shortened products were detected (Fig. 4,A). Their weight of 31.2 and 29kDa suggests that at least two alternative methionines, Met34 and Met48 can act as the translation starting points. These methionines are located between the alpha-helices of the HTH-DNA binding domain of LeuO (Fig. 4,B) indicating that DNA-binding properties of shortened protein variants can differ from that of the full-length regulator.

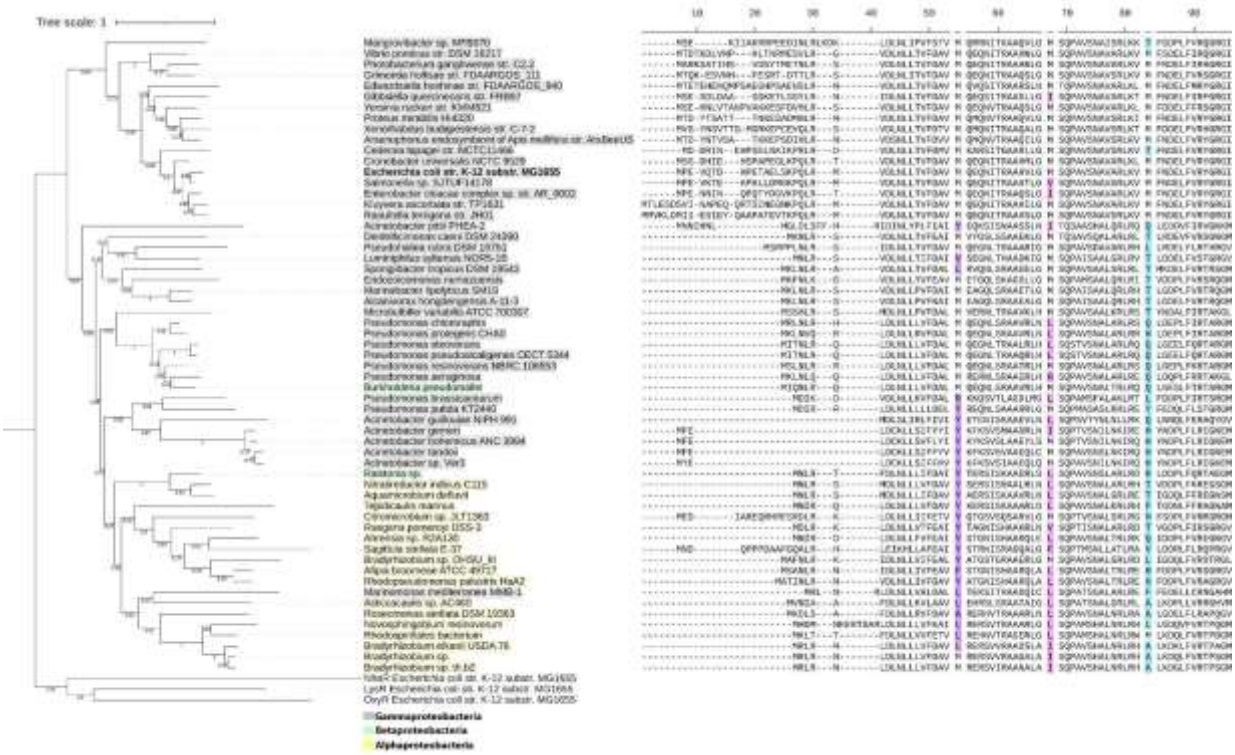

Fig. 3. Phylogenetic representation of LeuO and its homologs among Gamma-, Beta-, and Alphaproteobacteria. LeuO protein phylogenetic tree was inferred using maximum-likelihood algorithm. The tree was manually rooted by the branch leading to NhaR, OxyR, and LysR of the reference *E. coli* K-12 MG1655. First ninety eight nine positions of respective multiple sequence alignment are displayed. Non-methionine amino-acid residues aligned to Met48 and Met63 of the reference *E. coli* str. K-12 substr. MG1655 are marked pink and blue, respectively. Full size picture

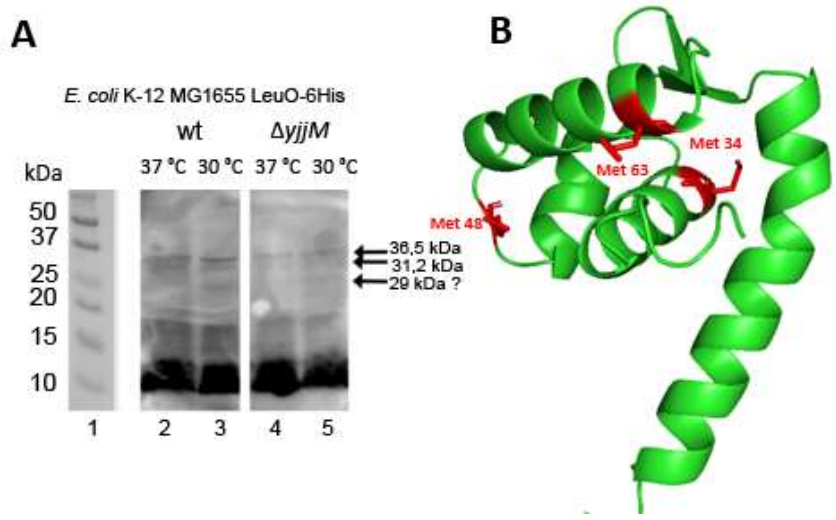

Fig. 4. [**A**] LeuO-6xHis synthesis detection from the bacterial chromosome in *E. coli* K-12 MG1655 *leuO*-6xHis and its derivative *E. coli* K-12 MG1655 Δ*yjjM leuO*-6xHis. Above the lanes the culture growth conditions indicate: 37 °C, anaerobically and 30°C, aerobically. [**B**] Structure of the LeuO DNA binding domain. Positions of alternative methionines are indicated by arrows.

In the *yjjM* deletion mutant (Fig. 4,A, lines 4–5) bands corresponding to at least two longest LeuO protein forms were lower, assuming regulatory interference of LeuO and YjjM. Very little changes were observed when shifting cells from free-living conditions to similar to intestine conditions (compare lines 2 and 3, 4 and 5 in Fig. 4,A).

We then tested the levels of the *leuO*-mRNAs from which each protein form can be translated in the wild type K-12 MG1655 and its *yjjM* deletion mutant using qRT-PCR with different primer pairs (primer positioning relative to the ATG codons is shown in Fig. 5,A). From Figure 5,B, we see that the levels of mRNAs that might correspond to the protein products translated from the main ATG (Met1) and Met34 (primers 1–2 and 3–5) are also slightly lowered being in line with the western blot data.

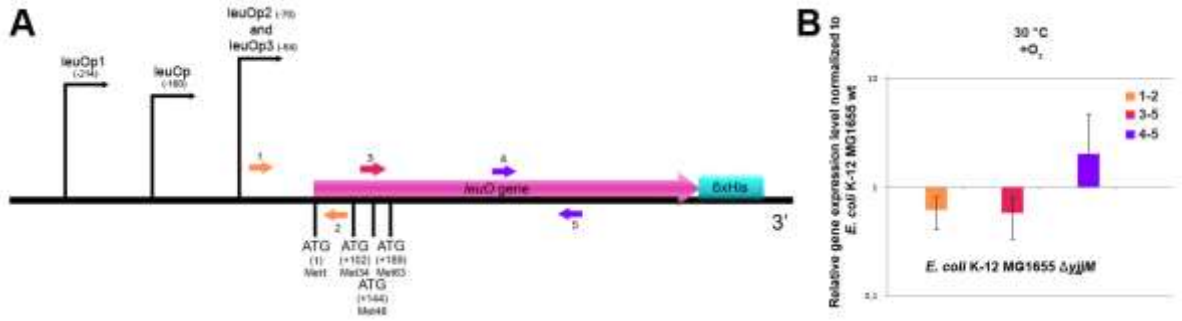

Fig. 5 [**A**] Scheme of the *leuO* gene and its upstream region. Promoters and their positions relative to ATG are shown by black arrows (promoter positioning is taken from [39]). Primers used in qRT-PCR are shown by coloured arrows; colour corresponds to the respective bars in Fig. 5,B. [**B**] Expression dynamics of three *leuO* gene regions, which contain additional start ATG codons in *E. coli* K-12 MG1655 leuO-6xHis-tag. Cells were grown aerobically at 30 °C. Calculations were made by the $2^{-\Delta\Delta Ct}$ method. Data were normalized to the wild type *E. coli* and presented as logarithmic histogram. Error bars represent the standard error of mean (SEM) from three biological and three technical replicates.

LeuO and YjjM share their regulons and have 6 common targets related to bacterial virulence

We then analyzed the available ChIP-seq data for YjjM in *Escherichia coli* K-12 MG1655 exponentially growing on D-glucose (Tutukina M. *et al* unpublished data), selected 60 most representative targets, and using MEME suite revealed the motif for YjjM binding in these conditions (Fig. 6).

Fig. 6. DNA-binding motif for YjjM protein identified by MEME [33] based on our ChIP-seq data. $n = 60$.

Keeping in mind that YjjM belongs to the GntR family of transcription regulators; the A/T-rich motif shown in Figure 6 was rather unexpected. However, we noticed that it was very similar to the A/T-rich motif of LeuO revealed in [15]. Therefore, we compared the sets of LeuO and YjjM targets based on the available ChIP-seq and SELEX data [5, 15, 30] and ChIP-seq data for YjjM (Tutukina M. *et al* unpublished data). As a result, we revealed 6 promoter regions containing binding sits for both TFs (Table 2), including *sdiA* – the common target for these regulators in *E. coli* and *Salmonella enterica*. Being located between divergently transcribed genes/operons these 6 promoter regions can control expression of at least 24 genes, of which 5 encode transcription factors and one small regulatory RNA (sRNA) SdsN. Many genes of the LeuO and YjjM common regulon are related to bacterial virulence and pathogenicity in *E. coli* (Table 2).

Table 2. Common gene/operon targets for YjjM and LeuO related to bacterial pathogenicity. Genes, which expression in the *leuO* and *yjjM* deletion mutatnts was further checked by qRT-PCR, are in bold

Genes/operons	Functions
ybeQ ⇔ *ybeR*	Oxidative stress response ⇔ cell wall precursor metabolism
sdiA ⇔ *sdsN*	TF involved in regulation of cell division ⇔ sRNA involved in the metabolism of oxidized nitrogen compounds
*fepE-entF- ybdZ-**fes*** ⇔ ***fepA**-entD*	Enterobactin gene system: biosynthesis proteins (YbdZ, Fes, EntD) and outer membrane transporter of enterobactin or colicines B and D (FepA)
btsT ⇔ ***tsr***	pyruvate:H+ symporter ⇔ chemotaxis signaling protein I
envR ⇔ ***acrE**-acrF*	TF (AcrEF repressor) ⇔ Membrane pump protein of the AcrEF drug-efflux system
yjjPB ⇔ ***yjjQ**-bglJ*	Putative succinate exporters ⇔ TF of flagellar biosynthesis genes (YjjQ) and TF controlling expression of many operons (BglJ)

Transcription efficiency in these six genomic loci in the wild type K-12 MG1655 and its *yjjM* and *leuO* deletion derivatives was analyzed with qRT-PCR. The results are presented in Figure 7. The most affected gene was *tsr* transcribed from the gene adjacent to *yjjM*. The *tsr* expression was inhibited by both YjjM and LeuO, with the *yjjM* deletion effect being more pronounced under free living conditions. YjjM also repressed *yjjQ*, which encodes the *flhDC* repressor forming a part of the virulence feedback network. This effect was detected only under conditions simulating a host organism's intestine. The same was observed for *fes* and *fepA*, encoding proteins involved in enterobactin metabolism. In general, effects of *yjjM* deletion on all tested targets were higher under conditions close to that in a host organism. However, the variability of the obtained results, reflected in the error bars, is of concern. At least partially, it can be explained by the heterogeneity of the cell population: due to insufficient mixing some cells remained floating while others started forming biofilms. On the contrary to YjjM, LeuO seems to act mainly at 30°C, except for *tsr* (Fig. 7).

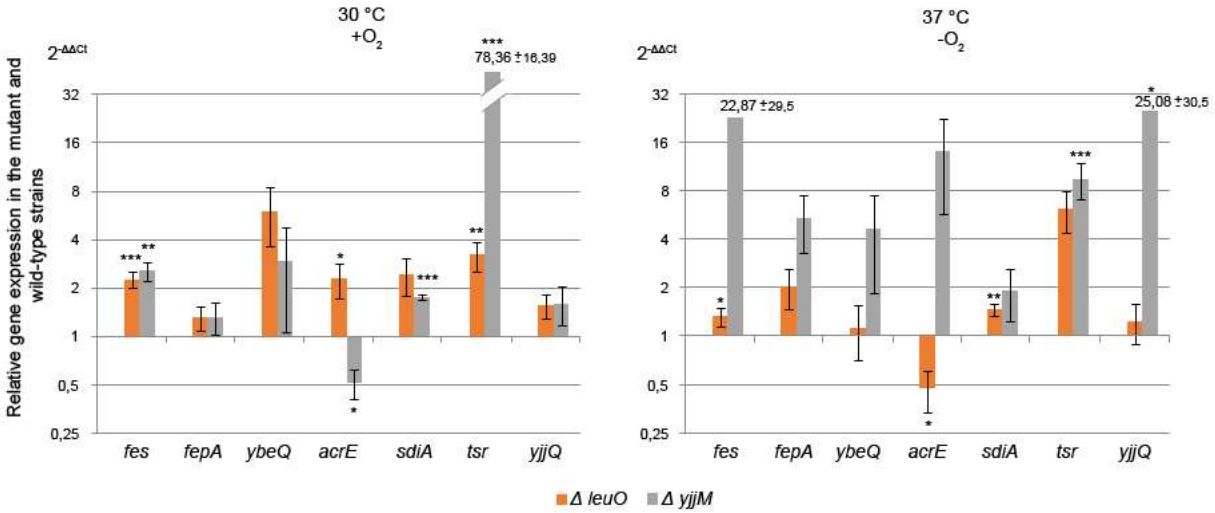

Fig. 7. Expression dynamics of six common target genes/operons of LeuO and in the *E. coli* K-12 MG1655 cells aerobically growing at 30°C, and anaerobically at 37 °C. Calculations were made by the $2^{-\Delta\Delta Ct}$ method. Values obtained are plotted relative to the data for the wild type K-12 MG1655 (= 1) as a logarithmic histogram. Error bars represent the standard error of mean (SEM) from three biological and 4–8 technical replicates. Significant differences were defined by $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) compared to (K-12 MG1655).

Conclusion

We determined several putative starts for translation of shortened protein forms for the YjjM and LeuO transcriptional regulators in *E. coli*. These candidates, namely Met19 and Met37 in YjjM, and Met34, Met48 and Met63 in LeuO, are conserved among different Enterobacterial species, which might indicate their functional significance associated with the synthesis of short protein isoforms. Phylogenetic tree analysis suggests that alternative start codons of YjjM and LeuO have most likely emerged from a common ancestor of Enterobacteriales. In this work, synthesis of two predicted short variants for the LeuO protein from *E. coli* was confirmed *in vivo*.

To date, little is known about cases with several protein forms being produced in-frame from one bacterial gene and thus differing just in their N-terminal domains. Both studied transcription factors possess DNA-binding domains in their NTDs suggesting that shortened forms might have altered ability to bind DNA targets. We showed that YjjM and LeuO had similar A/T-rich binding motifs and six common targets, differentially regulated under various temperature conditions. The statistically significant dependence of at least three TF genes (*sdiA*, *acrE*, and *yjjQ*) on deletion of YjjM or LeuO increases their regulatory potential, and it is likely that YjjM plays a special role in the transition of the bacterial population from free-living to invasive conditions.

Acknowledgements: This work was supported by RFBR 18-34-01006 (TAB).

Part of this project was made during the Summer School for Molecular and Theoretical Biology-2018 in Poznan, Poland.

REFERENCES

1. Jacob F., Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*. 1961. № 3. P. 318–356. doi: 10.1016/s0022-2836(61)80072-7
2. Jacob F., Perrin D., Sanchez C., Monod J. Operon: a group of genes with the expression coordinated by an operator. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*. 1960. № 250. P. 1727–1729.
3. Busby S.J.W. Transcription activation in bacteria: ancient and modern. *Microbiology*. 2019. V. 165. № 4. P. 386–395. doi: 10.1099/mic.0.000783
4. Shimada T., Bridier A., Briandet R., Ishihama A. Novel roles of LeuO in transcription regulation of *E. coli* genome: antagonistic interplay with the universal silencer H-NS. *Molecular Microbiology*. 2011. V. 82. № 2. P. 378–397. doi: 10.1111/j.1365-2958.2011.07818.x
5. Ishihama A., Shimada T., Yamazaki Y. Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Research*. 2016. V. 44. № 5. P. 2058–2074. doi: 10.1093/nar/gkw051
6. Tucker A.C., Escalante-Semerena J.C. Biologically active isoforms of CobB sirtuin deacetylase in *Salmonella enterica* and *Erwinia amylovora*. *Journal of Bacteriology*. V. 192. № 23. P. 6200–6208. doi: 10.1128/JB.00874-10
7. Umehara T., Kosono S., Söll D., Tamura K. Lysine Acetylation Regulates Alanyl-tRNA Synthetase Activity in *Escherichia coli*. *Genes* (*Basel*). 2018. V. 9. № 10. P. 473. doi: 10.3390/genes9100473
8. Di Martino M.L., Romilly C., Wagner E.G.H., Colonna B., Prosseda G. One Gene and Two Proteins : a Leaderless mRNA Supports the Translation of a Shorter Form of the *Shigella* VirF Regulator. *mBio*. 2016. V. 7. № 6. P. 1–10. doi: 10.1128/mBio.01860-16
9. Prosseda G., Falconi M., Giangrossi M., Gualerzi C. O., Micheli G., Colonna B. The *virF* promoter in *Shigella*: More than just a curved DNA stretch. *Molecular Microbiology*. 2004. V. 51. № 2. P. 523–537. https://doi.org/10.1046/j.1365-2958.2003.03848.x

10. Pérez-Rueda E., Collado-Vides J., Segovia L. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Computational biology and chemistry*. 2004. V. 28. № 5–6. P. 341–350. doi: 10.1016/j.compbiolchem.2004.09.004
11. Suvorova I.A., Tutukina M.N., Ravcheev D.A., Rodionov D.A., Ozoline O.N., Gelfand M.S. Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria. *Journal of Bacteriology*. 2011. V. 193. № 15. P. 3956–3963. doi: 10.1128/JB.00277-11
12. Henikoff S., Haughn G.W., Calvo J.M., Wallace J.C. A large family of bacterial activator proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 1988. V. 85. № 18. P. 6602–6606. doi: 10.1073/pnas.85.18.6602
13. Hertzberg K.M., Gemmill R., Jones J., Calvo J.M. Cloning of an EcoRI-generated fragment of the leucine operon of *Salmonella typhimurium*. *Gene*. 1980. V. 8. № 2. P. 135–152. doi: 10.1016/0378-1119(80)90033-5
14. Hernández-Lucas I., Calva E. The coming of age of the LeuO regulator. *Molecular Microbiology*. 2012. V. 85. № 6. P. 1026–1028. doi: 10.1111/j.1365-2958.2012.08175.x
15. Dillon S.C., Espinosa E., Hokamp K., Ussery D.W., Casadesús J., Dorman C.J. LeuO is a global regulator of gene expression in *Salmonella enterica* serovar Typhimurium. *Molecular Microbiology*. 2012. V. 85. P. 1072–1089. doi: 10.1111/j.1365-2958.2012.08162.x
16. Hernández-Lucas I., Gallego-Hernández A.L., Encarnación S., Fernández-Mora M., Martínez-Batallar A.G., Salgado H., Oropeza R., Calva E. The LysR-type transcriptional regulator LeuO controls expression of several genes in *Salmonella enterica* serovar Typhi. *Journal of Bacteriology*. 2008. V. 190. № 5. P. 1658–1670. doi: 10.1128/JB.01649-07
17. *NCBI RefSeq FTP*. URL: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/ (accessed 20.01.2023)
18. *FTP EMBL-EBI.* URL: https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam35.0/ (accessed 20.01.2023)
19. *FTP UniProt.* URL: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/ (accessed 20.01.2023)
20. *FTP NCBI*. URL: https://ftp.ncbi.nih.gov/pub/taxonomy/ (accessed 20.01.2023)
21. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *Journal of Molecular Biology*. 1990. V. 215. № 3. P. 403–410. doi: 10.1016/S0022-2836(05)80360-2
22. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. V. 25. № 16. P. 2078–2079. doi: 10.1093/bioinformatics/btp352
23. Huerta-Cepas J., Szklarczyk D., Heller D., Hernández-Plaza A., Forslund S.K., Cook H., Mende D.R., Letunic I., Rattei T., Jensen L.J., von Mering C., Bork P. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*. 2019. V. 47. № D1. P. D309–D314. doi: 10.1093/nar/gky1085
24. Katoh K., Misawa K., Kuma K., Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. 2002. V. 30. № 14. P. 3059–3066. doi: 10.1093/nar/gkf436
25. Price M.N., Dehal P.S., Arkin A.P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*. 2009. V. 26. № 7. P. 1641–1650. doi: 10.1093/molbev/msp077

26. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. 2017. № 35. P. 1026–1028. doi: 10.1038/nbt.3988
27. Cock P.J., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., de Hoon M.J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009. V. 25. № 11. P. 1422–1423. doi: 10.1093/bioinformatics/btp163
28. Letunic I., Bork P. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*. 2021. V. 49. № W1. P. W293–W296. doi: 10.1093/nar/gkab301
29. Zulkower V., Rosser S. DNA Features Viewer: A sequence annotation formatting and plotting library for Python. *Bioinformatics*. 2020. V. 36. № 15. P. 4350–4352. doi: 10.1093/bioinformatics/btaa213
30. *TEC database*. URL: https://shigen.nig.ac.jp/ecoli/tec/top/ (accessed 20.01.2023).
31. Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. 2008. V. 9. Article No. R137. doi: 10.1186/gb-2008-9-9-r137
32. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010. V. 26. № 6. P. 841–842. doi: 10.1093/bioinformatics/btq033
33. Bailey T.L., Johnson J., Grant C.E., Noble W.S. The MEME Suite. *Nucleic Acids Research*. 2015. V. 43. № W1. P.W39–W49.
34. Lee D.J., Bingle L.E.H., Heurlier K., Pallen M.J., Penn C.W., Busby S.J.W., Hobman J.L. Gene doctoring: a method for recombineering in laboratory and pathogenic *Escherichia coli* strains. *BMC Microbiology*. 2009. V. 9. № 252. P. 1–14 doi: 10.1186/1471-2180-9-252
35. Laemmli U.K. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature*. 1970. № 227. P. 680–685. doi: 10.1038/227680a0
36. Igarashi K., Ishihama A. Bipartite functional map of the *E. coli* RNA polymerase alpha subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. *Cell*. 1991. V. 65. № 6. P. 1015–1022. doi: 10.1016/0092-8674(91)90553-b
37. Purtov Y.A., Glazunova O.A., Antipov S.S., Pokusaeva V.O., Fesenko E.E., Preobrazhenskaya E.V., Shavkunov K.S., Tutukina M.N., Lukyanov V.I., Ozoline O.N. Promoter islands as a platform for interaction with nucleoid proteins and transcription factors. *Journal of Bioinformatics and Computational Biology*. 2014. V. 12. № 2. P. 1441006. doi: 10.1142/S0219720014410066
38. Yang D., Dong J., Su X., Zhang W., Zhang L., Li L., Lv L., Guo L. Functional analysis of the uropathogenic *Escherichia coli* R049 gene. *Microbiological Research*. 2015. V. 171. P. 39–44. doi: 10.1016/j.micres.2015.01.002
39. *Regulon DB Database*. URL: http://regulondb.ccg.unam.mx/gene?term=ECK120000524&organism=ECK12&format=jsp&type=gene (accessed 20.01.2023).

Received 11.01.2023.
Revised 26.01.2023.
Published 31.01.2023.