Mathematical Biology and Bioinformatics
2016. V. 11. № Suppl. P. t1-t14. doi: 10.17537/2016.11.t1

The translation of the original article
Markelova N.Y., Masulis I.S., Ozoline O.N. Matematicheskaya biologiya i bioinformatika. 2015. V. 10. № 1. P. 245–259. doi: 10.17537/2015.10.245

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

UDC: 579:252

REP-elements of the *Escherichia coli* genome and transcription signals: positional and functional analysis

Markelova N.Y., Masulis I.S., Ozoline O.N.*

Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Russia
Pushchino State Institute of Natural Sciences, Pushchino, Russia

Abstract. In the intergenic regions of the *Escherichia coli* genome there are 356 REP-elements, containing 1–12 repeated sequences with degenerated consensus. Their biological role is poorly understood, but multiplicity in the genome, preferential localization between convergent genes and ability to form hairpin structures have led to the assumption that REP-elements participate in the transcription termination and processes affecting stability of the corresponding RNAs. Though the direct experiments did not confirm the ability of the studied REP-sequence to stop RNA synthesis and some ambiguity regarding their primary function still exists. In this study, positional and functional analysis was undertaken for the entire set of annotated REP-sequences and the reduced efficiency of RNA synthesis behind the many REP-modules was observed. For all that some REP-modules did not affect the processivity of RNA synthesis, assuming their read-through transcription and further possibility to be involved in the regulatory events. We also observed REP-associated transcription activation and found overlapping promoters. The most unexpected was specific distribution of REP-sequences nearby the *promoter islands*, which assumes an insulator-like action of these sequences, maintaining transcriptional autonomy of the *islands*, and indicates functional significance of the *island*-born RNAs.

Keywords: *structural elements of bacterial genome, REP-sequences, promoters, promoter islands, mechanisms of transcription regulation.*

INTRODUCTION

Repetitive elements of 25–35 base pairs (bp) in length containing inverted sequences capable to fold in stable secondary structures were found in the genomes of eubacteria in early 1980s and named as REPs (Repetitive Extragenic Palindromic) [1]. A consensus sequence with the context: GC(g/t)GATGGCG(g/a)GC(g/t)…(g/a)CG(c/t)CTTATC(c/a)GGCCTAC has been proposed for these elements. The number of such repeats in the genome of *E. coli* according to different estimates and criteria range from 500 to 1000. The prevailing association of REP-elements with the ends of genes gave rise to multiple speculations concerning their physiological role, mechanisms of acquisition and evolution [2]. Thus, the ability of REPs to form hairpin structures in the RNA molecules allowed considering such motifs as potential transcriptional attenuators or terminators if they are located in the transcribed regions of the genome. Such a function has been verified experimentally and it was demonstrated that the sequence of REP-element by itself, being a part of template for

*Ozoline@rambler.ru

RNA synthesis, is not capable to block transcription elongation [3]. Later convincing evidences were obtained indicating the ability of REP motifs for promoting the Rho-dependent attenuation [4]. However, predestination of REP-elements to detain or stop transcription is still open.

Considerable efforts have been made to reveal the role of REP-elements in maintaining the mRNAs stability and in their protection from degradation by 3'–5'exonucleases [5–8]. A suggestion that the hairpins formed in RNAs may serve as a steric hindrance for migration of exonucleases was a rational basis to consider the stabilizing role of the structure-forming motifs. But these same hairpins can also be targeted by endonucleases that destroy the double-stranded RNAs or single-stranded loops within the hairpin (RNAse E) creating the 3'-ends for subsequent hydrolysis by exoribonucleases. Therefore, the assumption that REP-elements perform protective function against nucleases was initially controversial. Nevertheless, both *in vitro* and *in vivo* evidences were obtained indicating a decrease of mRNAs half-life upon deletion of REP-elements from the 3'-terminal untranslated regions [6]. Thus, the concept implying the predominant function of REP-sequence as mRNA protectors remained viable for over two decades.

The ability of self-complementary motifs in the structure of REP-elements to bind DNA gyrase, led to the assumption of their involvement in the modulation of DNA topological state [9]. DNA gyrase interacting with hairpin stem may cause relaxation of the DNA positive supercoils naturally accumulating near the 3'-ends of transcribed genes and suppressing the expression of the downstream genes [10]. REP-elements with high affinity to topoisomerases, if located between successive genes can reduce this tension and facilitate the downstream RNA synthesis [11].

With the accumulation of the whole genome sequences for a wide spectrum of microorganisms, an opportunity to search for similar elements in other bacteria and to carry out a comparative evolutionary analysis appeared. This analysis has been performed for the repeated sequences in the genomes of 66 bacterial species and allowed to classify REP motifs according to the context of conservative tetranucleotide blocks at the 5'-ends (GTAG and CGTC families) [12]. The nucleotide sequence of the hairpin-forming palindrome appeared to be much more variable. It has been found that the genomes of eubacteria are enriched by REPs of GTAG-type, though their number in pathogenic *E. coli* strains is almost twice lower than that in free living non-pathogenic bacteria of the same species [12]. This difference may indicate some interference of REP-elements with assimilation and / or expression of toxic genes. It is important to note that the way by which REP-elements appeared in the genomes remains the least understood. Thought the assumption of their expansion within the given genome due to gene conversion, occurring with a certain frequency as a consequence of recombination events, is quite realistic. However, it has not been documented experimentally and does not explain the specific localization of REP-elements in intergenic loci.

Thus the experimental data as well as theoretical considerations had not yet allowed formulating a holistic view on the biological role of REP-elements and the mechanisms of their origination, functioning and evolution. In the present study we have attempted to integrate new expression and genomic data with available information concerning the REP elements by focusing attention on their disposition relative to known genes and on their location in respect to promoters of different functional types.

METHODS AND ALGORITHMS

Genome of *Escherichia coli* and sets of REP-sequences

The nucleotide sequence of the *Escherichia coli* K12 MG1655 (*E. coli*) genome and the coordinates of REP-elements were taken from the NCBI GenBank (NC_000913.3). Positional analysis relative to the neighboring genes was performed for all 356 annotated REP elements

and for 224 samples most similar to the consensus REP-sequence GCCGGATGCGGCGTGAACGCCTTATCCGGCCTACGA ($2e^{-14} \le E \le 3e^{-4}$). The length of annotated REP-modules varies from 14 to 767 bp. The sequences in the second set ranged from 25 to 36 bp. Positional analysis was performed to estimate the distance between the nearest border of REP-element and the beginning or end of a neighboring gene.

Sets of promoter regions and positional analysis

Three sets of promoter regions were used for comparative positional analysis of annotated REP-elements and the transcription initiation sites. One of them was composed of 404 promoters with experimentally mapped transcription start points (TSPs). To make the size of this set comparable to the number of REP-elements, only those known promoters were selected that according to RegulonDB [13] are recognized by σ^{70}-RNA polymerase and do not have any other σ^{70}-dependent promoter nearby. REP-elements were searched within the range of ± 300 bp around these TSPs. The second compilation was comprised of 51 regulatory regions with 2–6 promoters located on both DNA strands and initiated RNA synthesis in opposite directions (a set of divergent promoters). Unfortunately, two genomic regions containing REP elements between divergent genes were excluded from this set since only one TSP was experimentally mapped in intergenic space. The third set included 434 *mixed promoter islands* selected earlier [14] and fitted the criteria that at least 8 potential TSPs are predicted in each sliding window of 100 bp throughout ≥ 360 bp. The algorithm PlatPromU [15] was used in this case to search for potential TSPs. It ignores the context of conserved elements recognized by the σ-subunits of RNA polymerase and, therefore, is able to detect promoters of all σ-factors. Only those potential TSPs were considered as significant signals if their scores exceeded the background level by at least 4 StD ($p < 0.00004$). The average size of these regions was 660 bp. The search of REP elements for the last two sets was carried out inside the promoter regions and within 300 bp area, flanking them on both sides. The mapping of potential promoters recognized by σ^{70}-RNA polymerase for genes *sapA*, *ymjA* and *puuP* was performed using PlatProm as it was described previously [16].

Functional analysis

To evaluate the ability of a transcription complex to overcome the potential barrier created by the hairpins of REP-elements we used the expression data obtained with high-density microarrays in [17]. The probes of these microarrays are 50 nucleotides in length and represent both strands of the entire genome of *E. coli* covering it with 25 bp overlap. The expression efficiency of any genomic locus can be estimated from the signal intensity of hybridization between the microarray probes and fluorescently labeled DNA copies of cellular RNAs (E_{exp}). Hybridization signals obtained from fragmented DNA were used as the control values for normalization (E_{gen}). A generally-accepted parameter $\log_2 R$, where $R = E_{exp}/E_{gen}$ was used as a measure of the transcription efficiency. The point-to point intensities of the hybridization signals over 1250 bp before the end of the REP-element and 250 bp thereafter, reflecting the number of RNAs transcribed in the whole genomic region, were used to evaluate the influence of REP-motifs on transcriptional read-through. To reveal the characteristic trend in the transcription efficiency over the 1500 bp region, the obtained profiles of hybridization signals were aligned to the far downstream boundaries of REP elements.

Searching for regions homologous to REP-elements

Multiplicity of the REP-elements in the genome suggests the possibility that REP-containing RNAs may form duplexes with complementary sequences in other RNAs or in bacterial chromosome. To search for such targets near functional promoters and *promoter islands* in the genome of *E. coli* K12 MG1655 (NCBI accession number NC_000913.3),

Microbial BLAST (NCBI) [18] was used. The complete nucleotide sequence of each annotated REP-element has been submitted as a query, and the search was carried out with the default settings. However, due to the specific structural organization of the target sequences, the analysis of low complexity regions was not turned off upon scanning. Homologous sequences longer than 14 bp (minimal length of annotated REP-element) found within the promoter regions or located in the vicinity of these areas were considered as significant. They were projected on the genome and the longest samples with at least 16 bp non-overlapping parts were selected.

RESULTS AND DISCUSSION

REP-element flanking the end of the *ymjA* is transcribed in the antisense direction

Previously, it was found that the REP-elements are often located in close proximity to an adjacent end of the gene. Typical REP-module, for instance, is located behind the end of *ymjA* gene (Fig. 1,**A**). It has a sequence GCCAGATGCGGCGTGAACGCTTTATCCGGACAAC that corresponds well to the consensus proposed by Stern *et al*. in 1984 (Fig. 1,**B**) and confers the capacity to form a stable secondary structure with a free energy of folding –10 kcal/mol (Fig. 1,**C**).

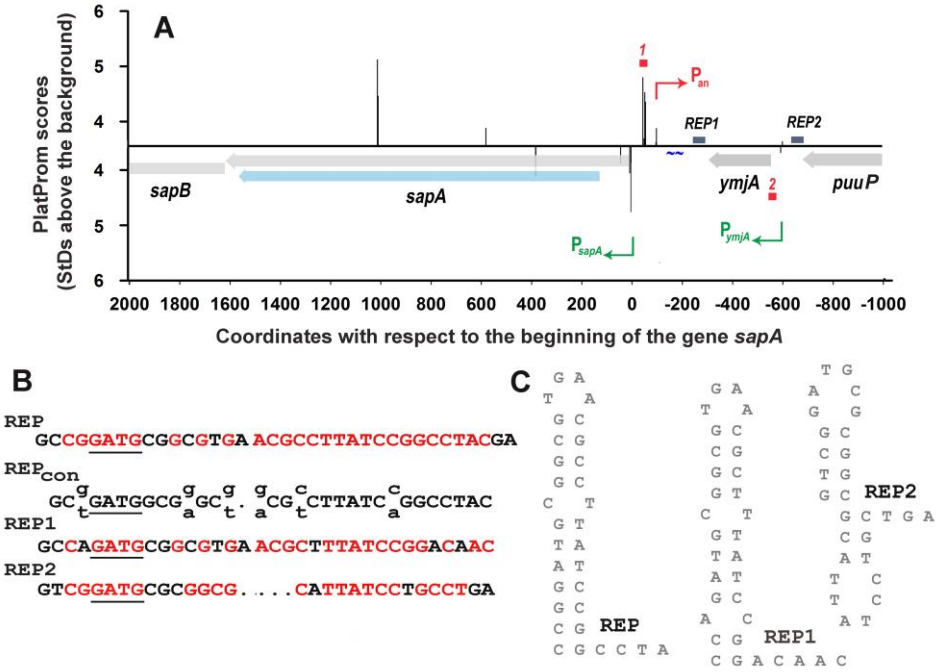

Fig. 1. **A.** The distribution of potential transcription start sites (bars) predicted by an algorithm PlatProm [16]. The height of the bars above and below the *X*-axis corresponds to the degree of promoter likelihood (score), calculated for the top and bottom strands of the genome, respectively. Broad arrows depicture the genes and the direction of their transcription. Blue arrow indicates an alternative open reading frame (ORF) for *sapA*. Dark boxes indicate the location of the REP1 and REP2. Red squares point out primers used for PCR amplification of the DNA fragment cloned in a pGEMAX vector; a wavy line denotes a probe for hybridization. **B.** The sequence of the most represented in the genome of *E. coli* K12 MG1655 REP-element (REP), consensus of REP-sequences proposed by Stern *et al.* [1] (REP_{con}) and the contexts of the REP1 and REP2 from the *ymgA-sapA* genomic locus. Nucleotides matching to REP_{con} are marked in red. The GATG motif is underlined. **C.** Hairpin structures predicted by RNA Structure software [20] for the REP, REP1, and REP2.

A specific feature of genetic environment of *ymjA* is the presence of the second REP-element, located in the upstream intergenic region (REP2). Both REP-elements belong to the GATG-family. There are only 23 genes in the genome of *E. coli*, surrounded by REP-

elements on both sides. Although REP2 may be functionally linked to the *puuP* gene, its effect on the expression of *ymjA* is also quite likely. The REP2 is slightly shorter than the REP1, but the degree of similarity with the consensus sequence is almost the same (Fig. 1,**B**), though the structure with the minimum free energy of folding (–6 kcal/mol) is markedly different from that of two others (Fig. 1,**C**). The conformational properties of the known REP-elements may, therefore, be substantially different.

A particular interest to the study of this genomic area was stimulated by the finding of a potential promoter P_{an}, capable to initiate synthesis of a new RNA from the intergenic region *ymjA–sapA*, which potentially can be terminated on the REP1, or the synthesis of longer RNA, antisense to mRNA *ymjA*. In the latter case, REP1 appeared to be included into the RNA product. Since both possibilities are of fundamental importance, the fragment of the genome, containing the intergenic region *sapA/ymjA* and the entire coding region of *ymjA* (PCR-amplified with primers 1 and 2, Fig. 1,**A**), was cloned into the expression vector pGEMAX [19]. Total RNA was isolated from the cells of *E. coli* K12 MG1655 transformed by the vector containing this insert, and RNAs hybridizing to radioactively labeled probe complementary to antisense product (the wavy line in Figure 1) were detected in these samples. The presence among them of the RNA product longer than 400 bp implies that transcription can pass-through the entire regulatory region including REP1.

Positioning of REP-elements relative to the borders of the genes is not random

For the nucleotide sequence of 380 bp, assumingly incorporated into the potentially regulatory RNA transcribed from P_{an}, the search for homologous regions in the genome was carried out using Microbial Nucleotide BLAST [18]. Only the REP1 sequence of this query extended by 2 bp at the 5'-end showed multiple occurrence in the genome (Fig. 1,**B**). This extended REP1 has more than 200 homologous segments in the bacterial chromosome and the most similar to REP1 sequence REP (Fig. 1,**B**) with an even greater number of homologues was selected as a context model for further analysis. REP is a part of three annotated REP-elements, and the sequence of its first 33 base pairs is precisely repeated 44 times in the genome. There are some differences in the first half of the REP as compared to REP_{con} proposed by Stern *et al*. [1]. If REP sequence (Fig. 1,**B**) was used as a query in Microbial Nucleotide BLAST, 224 homologous motifs were found and the distance of their 5'- or 3'-ends from the borders of the neighboring genes was estimated (Fig. 2).

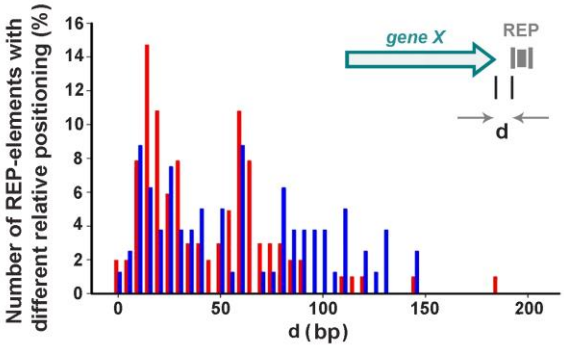

Fig. 2. The distribution of distances (*d*) from the 3'-ends of annotated genes for sequence motifs homologous to REP (Fig. 1,**B**). Red bars represent repeats located between convergent genes, blue – between collinear genes. For tandem repeats the shortest distance to the 3'-end of the coding sequence was taken into account, in the case of convergent orientation of flanking genes only minimal value of the two possible ones was accounted. The orientation of motifs relative to the direction of transcription was ignored.

In accordance with the published data [21], a significant portion of these motifs lay at a distance of 10–20 bp from the nearest end of the gene (Fig. 2). This fact was exploited in a model suggesting participation of REP-elements in the regulation of mRNA translation by

provoking stalling of elongation complex followed by recruitment of RNAse R that destroys mRNAs from their 3'-ends [21]. Though, such a mechanism of action was experimentally approved for repeats remote from the genes termini at a distance less than 16 nucleotides [21], a considerable part of 224 scrutinized motifs are more distant from the stop codon generating a second peak at a distance of 60–65 nucleotides (Fig. 2). In the case of collinear genes 80 % of REP-motifs (blue bars in Figure 2) are located at a distance longer than 15 bp. This analysis was further repeated for another set containing all 356 REP-elements annotated in the genome of *E. coli* (Fig. 3,**B**). For converged genes in this case we estimated the distances to both neighboring genes, and the position of REP-element was marked by the genomic coordinates of the first and the last pair (according to annotation) rather than by borders of the conserved textual motif. The distances from 1 to 20 bp (52.5 %) were obviously dominant, including 45 % of cases where distances were 15 bp or less. However, the more remote location of the REP-elements with the distance larger than 70 bp was also observed. Thus, although many REP-elements may be involved in the attenuation of translation and mRNA degradation mediated by RNase R as proposed in [21], the biological significance of almost half of these modules still requires a different conceptual understanding.

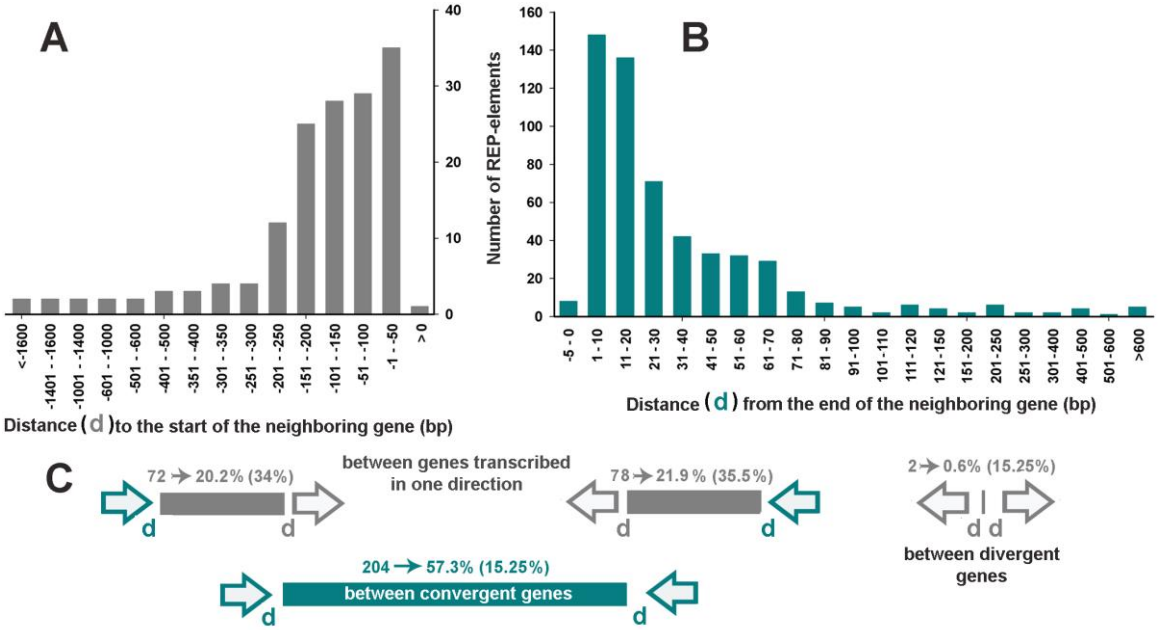

Fig. 3. The distribution of distances (*d*) from the annotated REP-elements relative to the starts (**A**) and ends (**B**) of neighboring genes. In all cases, the distance to the nearest border of the REP-element was estimated. Each REP-element annotated in the genome was considered as a single module, even if it consisted of a few recurring motifs. Both distances to the neighboring genes in the case of convergent and divergent genes were accounted. Motif orientation relative to the direction of transcription was ignored. Genes flanking REP-elements in all possible mutual orientations are indicated by gray and cyan arrows in the panel **C**. Intergenic regions are indicated by rectangles. Their length is proportional to the number of gene pairs of each type (collinear, convergent and divergent) in the whole set of intergenic regions containing REP-elements. The number of gene pairs and their percentage in the set are indicated above schemes. The percentage of gene pairs with the same orientation in the genome is indicated in parentheses.

According to the latest *E. coli* genome annotation, the REP-containing spacers separating collinear genes vary in length from 47 to 6176 bp. On average they are longer (375 bp) than the average length of the intergenic region in the genome (241 bp). At least in part this is due to the lack of REP-elements in operons, where the intergenic distances are usually short and sometimes the stop codon TGA of one gene overlaps with the start codon ATG of the next one. In many intergenic regions with REP-elements the space is sufficient to allow the

presence of new genes or promoters for neighboring genes. We, therefore, investigated the distribution of REP-elements relative to the initiation codons of nearest genes (the case of collinear or divergent genes) and found that the proximal boundary of REP-elements is usually located within 250 bp or less from ATG (Fig. 3,**A**). The starting points of transcription initiation for ~ 90 % of known promoters are concentrated exactly in this area [22] with the maximum in the first 40 bp. This means that the positioning of REP-elements relative to the starts of genes as well as to their ends is not occasional and may be dependent on functional destination of these genomic modules.

REP-elements can suppress the read-through transcription or let it to pass

The majority of REP-elements (57.3 %, Fig. 3,**C**) are located between the convergent genes, and the percentage of such intergenic regions in the set of REP-containing spacers is 3.8-fold higher than the percentage of corresponding gene pairs in the genome. This obvious positional preference provoked the assumption that REP-elements may be involved in the transcription termination and/or subsequent RNA processing. The bias is strengthened by the fact that REP-elements were found only between the two pairs of divergent genes (Fig. 3,**C**) that is 25-fold lower than expected. Nevertheless, in the regions separating the collinear genes, where transcription attenuation or termination are of the same demand as for convergent genes, REP-elements are also underrepresented (1.65-fold lower than expected) (Fig. 3,**C**). Therefore, an importance of REP-elements for transcription termination remains ambiguous.

Thus, in the next step, we used the genome-wide expression data obtained from high-density microarrays [17] in order to investigate the profiles of transcription for genes located upstream of REP-elements, along REP-elements themselves, and for short genomic segments flanking REP-modules. Efficiency of transcription was estimated as the intensity of hybridization signals between the tiled microarray probes and fluorescently labeled DNA copies of cellular RNAs (E_{exp}). Registered fluorescent signals were normalized to the hybridization signals obtained for the same probes with samples of fragmented genomic DNA labeled by another fluorescent dye (E_{gen}). The value of $E = \log_2 R$, where $R = E_{exp}/E_{gen}$, was used as a measure of transcription efficiency and $E = 0$ indicates the presence of one copy of the corresponding RNA per cell.

Totally 558 genomic regions (408 convergent and 150 collinear genes) flanked by REP-elements were examined. Transcriptional level of many genes and their flanking regions was very low ($\sum E < 0$), which masked the changes mediated by REP-elements. Such profiles have been combined in a separate group and were not subjected to further analysis. The rest profiles were divided into three groups. The first set was composed of transcribed genes, for which the levels of hybridization signals dropped down on the borders of REP-modules (Fig. 4,**A**). There were 80 regions of this type and 21 samples with REP-associated depression were found between collinear genes. The length of REP-elements in this group ranged from 15 to 281 bp. Only in one case inhibition of RNA synthesis have been registered on both ends of the REP-module covered by overlapping transcription from both DNA strands. The size of this unique REP is 17 bp, and it is located between the convergent genes *metE* and *ysgA* separated by 41 bp intergenic region. It has some homology with REP_{con} (Fig. 1), is surrounded by A(T)-tracks on both ends and forms a perfect hairpin with free energy of folding –18.6 kcal/mol: *aaaa*tccaaACCGGGTGGTAATACCAcccggt*ctttt* (uppercase letters show 17 bp REP-sequence, complementary sites are underlined). This REP-element, therefore, represents a part of the classical ρ-independent transcription terminator tethered to both convergent genes. Thus, a large number of profiles demonstrated REP-elements-mediated repression of RNA synthesis (Fig. 4,**A**), which mechanism is well understood.

The second group was composed of 34 profiles indifferent to the presence of REP-element (Fig. 4,**B**). Twenty of them lay between convergent genes and in one case stable transcription

was registered in both directions (between *ycbL* and *aspC*). REP-sequences of this group may be incorporated into the RNA products. The length of such REP-modules ranged from 21 to 388 bp, and the size of corresponding intergenic regions varied from 42 to 1238 bp, that in many cases is sufficient for a new product to be encoded.

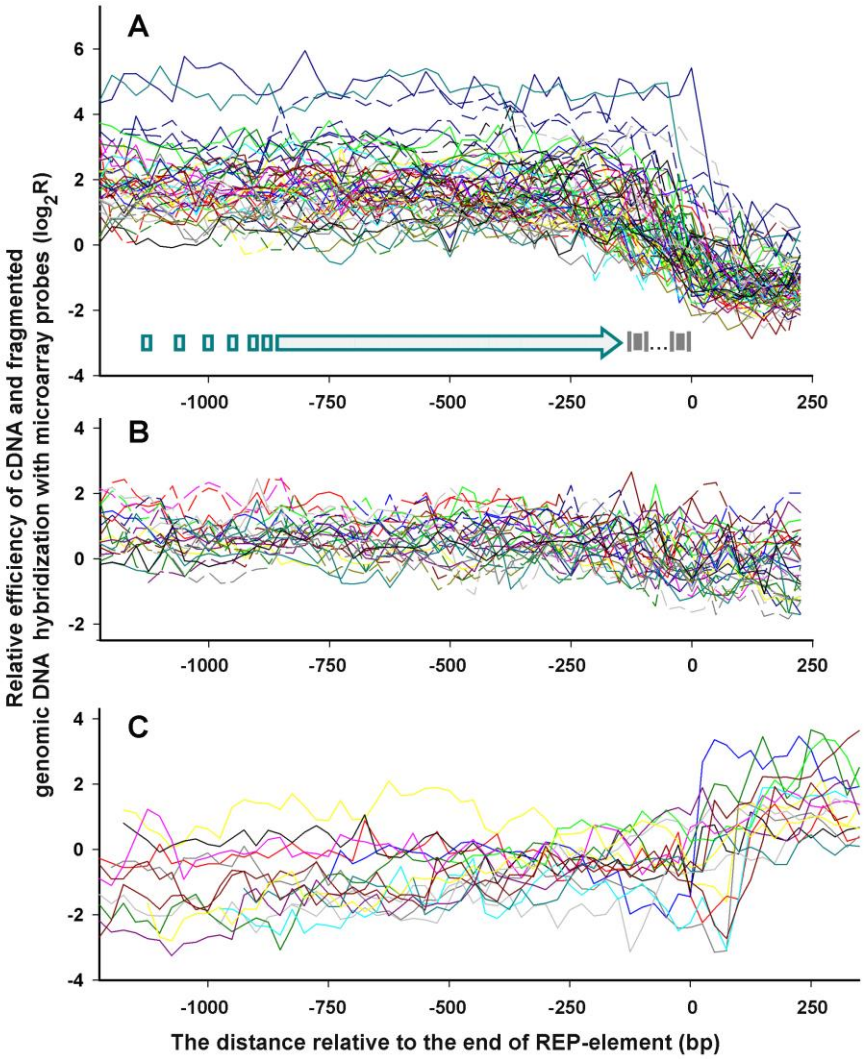

Fig. 4. The superposition of the transcription profiles for genes located upstream of REP-elements and adjacent regions immediately downstream to REPs (see scheme on the panel **A**) according to the data of the experiment «A» in [17]. All trajectories are aligned in respect to the ends of REP-elements (position 0) and are combined on the panels **A**, **B** and **C** depending on the profile type. All hybridization signals from microarray probes corresponding to the whole upstream gene, REP-element itself and 250 bp flanking its downstream boundary were plotted for genes shorter than 1250 bp. For longer genes the profiles cover only their last 1250 base pairs.

The most striking was the third group (Fig. 4,**C**). It included 16 tracks showing an increase in hybridization signal upon passage across 15–388 bp REP modules. Corresponding intergenic regions ranged in size from 56 to 1024 bp. Such a transcriptional behavior can be explained by the presence of functional promoters overlapping with REP-elements. For seven of nine genomic regions of this type located between the collinear genes the transcription start sites were really mapped experimentally (genes *fhuA*, *uof*, *sucA*, *poxB*, *phoP*, *tyrB* and *uxuR*) [13]. For two remaining genes (*yegP* and *yhaH*) promoters were predicted by PlatProm [16]. In 8 cases, REP-elements lay upstream of the transcription start points at a distance of 60–101 bp, and only in one case this module falls into the early transcribed region (from the position +25).

For convergent genes of this group promoters within intergenic area were not searched experimentally, though in three cases TSPs were predicted *in silico*. The observed increase in the value of hybridization signals (Fig. 4,**C**) almost in all cases, therefore, can be explained by the presence of functional or predicted promoters. Thus, the possible impact of REP-elements on the promoter activity, which has not been investigated so far, was decided to examine.

Known promoters and early transcribed regions are depleted in REP-sequences

Three sets of promoter sequences were used for the analysis (Fig. 5 and 6). One of them was composed of 404 known promoters activated by σ^{70}-RNA polymerase. To simplify further interpretation, only "single" promoters similar to P_{dapB} were selected for this set (Fig. 6,**A**). It should, however, be taken into account that some new promoters may be later found in these areas. The second set included 51 intergenic regions separating divergent genes with promoter orientation similar to that exemplified in Figure 1,**A** (P_{sapA} and P_{an}). These genomic areas included from 2 to 6 promoters with approved activity in both directions.

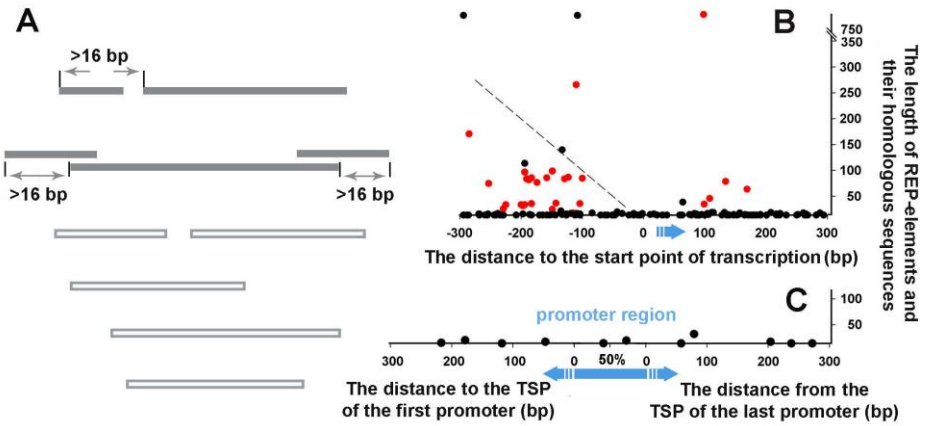

Fig. 5. The disposition of known REP-elements and their homologous sequences (red and black symbols, respectively) relative to the transcription start points of 404 single (**B**) and 51 multiple divergent promoters (**C**). The positions of 5'-ends of REP-elements or their homologues relative to the TSPs versus the length of corresponding repeated motifs are plotted. The blue arrow on the panel **B** shows the direction of transcription. **C:** The area occupied by divergent promoters (the average size 127 bp) is shown by blue double-headed arrow. Since these areas vary in length, positions of REPs and REP-similar sequences are indicated as a percentage of the distance between the left boundary of intergenic region and the left end of REP- or RER-similar sequence from the whole size of intergenic space. The dashed line on the panel **B** separates (on the left side) the REP-motifs non-overlapping with the promoter start points (position 0) from few REP-sequences covering promoter or located in the early transcribed region. The strategy used to select sequences homologous to REP-elements is schematically shown in the panel **A**. Grey boxes correspond to the sequences that are considered as independent homologues. Light boxes, substantially overlapping with the query (long gray box), were discarded.

The search for REP-elements with the left end laying in the ± 300 bp area around the transcription start point of single promoters gave 28 annotated REP-sequences (red symbols in Figure 5,**B**). Assuming that all REP-elements are located in intergenic regions and 204 of 356 are disposed between the convergent genes (Fig. 3,**C**), only 152 REP-sequences may be close to the promoter regions. The latest version of *E.coli* MG1655 genome has 4318 genes that are expressed as 2628 operons. Six hundred fifty-eight units of these 2628 operons are transcribed in opposite directions from overlapping regulatory regions and contain only two of known REPs. This implies that the remaining 150 REP-modules are distributed between 1970 (1970 = 2628 – 658) promoter regions (one REP-element per 13 promoter regions on average). Some of these intergenic regions have single promoter, while others – several transcription signals. Limiting the length of the analyzed area around TSPs by ± 300 bp and using samples with only a single promoter (Fig. 5,**B**) did not significantly alter the expected

proportion (one REP-element per 14 promoters). The promoter sampling is, therefore, rather representative.

As expected, based on the data shown in Figure 4,**C**, the majority of known REP-elements found nearby single promoters lay on the left from their TSPs (red symbols in Figure 5,**B**). Only one of them, starting at position –110 and spanning for 256 bp overlaps with the promoter and the transcription start point. Only 5 of the annotated REP-elements fall into the transcribed region. Searching for additional REP-like sites in the given set of 404 promoters revealed several long homologous sequences including two insertion elements of 767 bp in length (black symbols in Figure 5,**B**), but they have not changed the tendency of REP-modules to avoid location within the transcribed regions. At the same time the 14–15 bp REP-homologues were found on both sides around TSPs in equal proportion. Similar depletion in REP-like sequences has also been observed for 51 genomic loci containing divergent promoters flanked by transcribed regions on both sides (Fig. 5,**C**). It is, therefore, likely that REP-modules are specifically eliminated not only from the genes coding sequences but also from their promoter regions.

REP-elements associated with *promoter islands* chose another strategy of disposition

Promoter islands (exemplified in Figure 6,**A**) may be considered as an antipodes of single promoters (P_{dapB} in Figure 6,**A**) and in some extent as analogs of regions with multiple promoters capable for bidirectional transcription (*sapA*/*ymjA* regulatory region in Figure 1,**A**). They contain a huge number of potential TSPs for RNA synthesis, but the transcription in most cases is limited to the production of only short oligonucleotides [14, 16, 23].

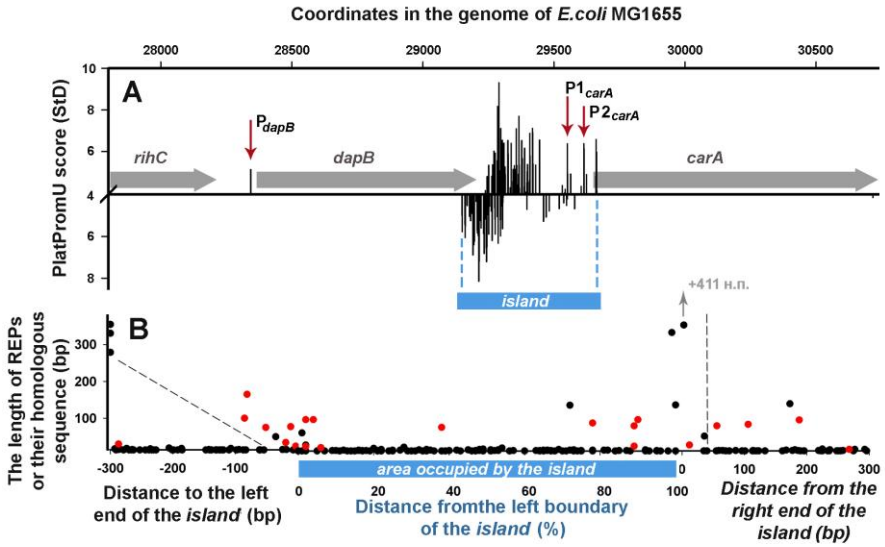

Fig. 6. A. An example of *mixed promoter island* identified by algorithm PlatPromU [14]. Red arrows point out the transcription starts for the single (P_{dapB}) or tandem promoters ($P1_{carA}$ and $P2_{carA}$). **B**. The distribution of known REP-elements and their homologues sequences relative to the borders of 434 *mixed promoter islands*. The positions of 5'-ends of REP-elements or REP-like sequences relative to the borders of *promoter islands* are shown in red and black symbols, respectively. Two dashed lines on the panel **B** separate the REP-motifs without any overlap with promoter regions (on the left from the left line and on the right from the right vertical line). Blue rectangle schematically represents genomic region occupied by *islands*. Their average size is 660 bp and the borders correspond to the last TSPs located on either strand. Promoter regions, therefore, were assumed to occupy 50 bp upstream of the left border and downstream from the right boundary of the *island*. Since *islands* differ in length, the positions of REPs and REP-like sequences are indicated as a percentage of the distance between the left end of REP (or homologous sequence) and the left boundary of the *island* from the whole size of the *island*. The vertical arrow denotes the position of the large insertion element (767 bp in length), similar to other three homologous insertion sequences shown in Figure 5.

Abnormally low transcriptional output of the *islands* presupposes the existence of special factors suppressing the ability of their promoters to initiate transcription. Some of them, like the inherent propensity to maintain a low level of negative DNA supercoiling [14, 23] or the ability to bind the histone-like protein of bacterial nucleoid H-NS [24] have already been discussed. However, the involvement of REP-elements in the transcription termination (Fig. 4), and the asymmetry in their association with known promoters (Fig. 5) prompted us for a more detailed study of the mutual disposition of these modules and *promoter islands* (Fig. 6). As a result, we found more than 20 REP-elements or their extended homologues in the immediate vicinity of the *islands*, but in contrast to the single promoters (Fig. 5,**B**) almost all of them lay in a zone that overlaps with promoters of the *islands* (bounded by dashed lines in Figure 6). Theoretically, they would have the same negative effect on the transcription processivity as shown in Figure 4,**A** and contribute independently to the suppression of the RNA synthesis initiated from the multiple promoters of the *islands*.

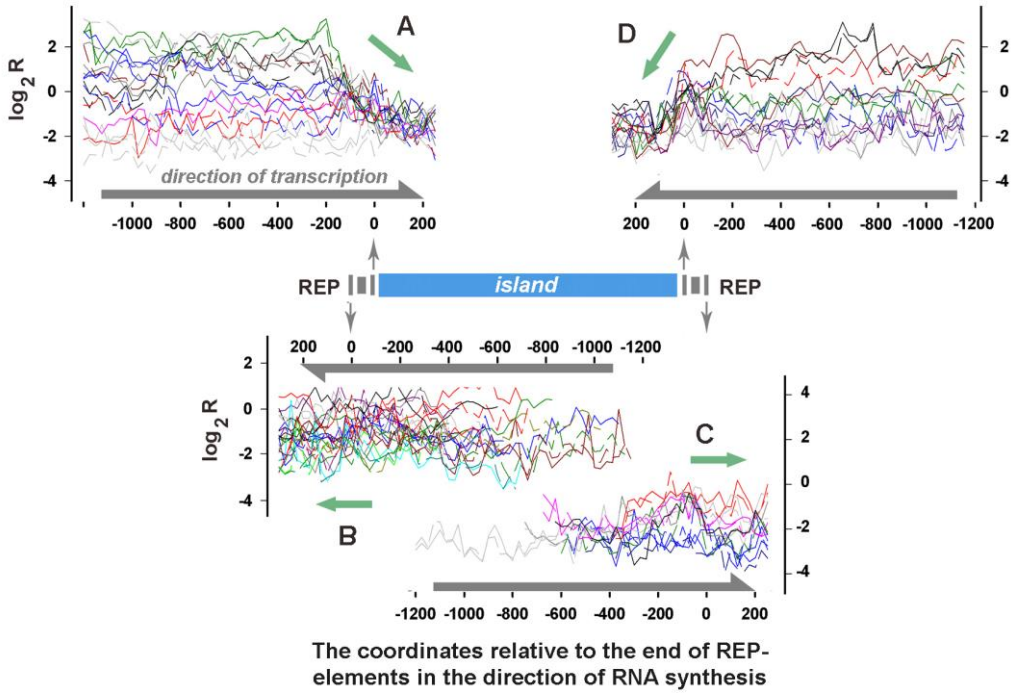

Fig. 7. Superimposed transcriptional profiles for both DNA strands along the *promoter islands* (**B** and **C**) and flanking genomic regions (**A** and **D**) (experiments «A» and «B» in [17]). All trajectories were aligned to the distal borders of REP-elements in the direction of RNA-synthesis (position 0). Profiles of hybridization signals on panels **A** and **D** are the same in length since they cover standard in size regions flanking the REP-elements on the left (**A**) or the right (**D**) side. Hybridization profiles on panels **B** and **C** vary in length since they correspond to the different in size regions between the REP-elements and the right (**B**) or the left (**C**) border of *promoter islands*. Green arrows accentuate the presence (**A** and **D**) or absence (**B** and **C**) of changes in the level of transcription upon passing REP-elements.

However, REP-elements do not affect outward transcription initiated inside the *islands* (Fig.7,**B** and 7,**C**), whereas inward transcription was usually decreased exactly at the borders of REP-elements (Fig.7,**A** and 7,**D**). Thus it is likely that REP-modules associated with *promoter islands* contribute to their isolation from the external incoming transcription.

CONCLUSION

Despite of more than thirty-year history of REP-elements study, almost all aspects of their origination, genomic expansion and possible functions yet remain obscure. The analysis of their positional preferences and involvement in transcription processivity reported in present work (Fig. 4,**A**) in general confirmed the implication of these genomic modules in

transcriptional termination. Nevertheless the fact that some REPs are ignored by transcriptional machinery, and some have even been associated with increased rather than decreased levels of RNA production, assumes that transcriptional termination may represent only a special case in their functional repertoire. The potency to hamper transcription due to the formation of hairpins by the inverted repeats is *a priori* expected, but this function does not necessarily represents all the responsibilities consigned to REP-modules in the genome. The typical location of REP-sequences in the terminator regions of genes (Fig. 2,**B**, 3,**B** and 3,**C**) infers that many cellular RNAs contain REP-like modules at their 3'-ends. Thus it cannot be excluded that these RNAs are specifically subjected to REP-mediated targeted degradation.

Transcriptional passage across REP-elements documented here (Fig. 4,**B**) means also that REP-sequences can appear inside the stable RNAs and, therefore, can be engaged in regulatory duplex formation with many other RNAs, containing similar motifs. SapZ RNA synthesized in the intergenic region *ymjA*/*sapA* in an antisense direction relative to *ymjA* (Fig.1,**A**) may belong to this class of regulatory REP-containing RNAs. It is clear that the scale of such regulatory actions depends on the number of potential targets. Thus, REP-elements found in the vicinity of promoters in the non-transcribed regions (Fig. 3 and Fig. 5), obviously, cannot participate in this type of regulatory events. However, the REP-elements functioning as constituents of special non-coding RNAs can trigger a cascade of multiple interactions, combining into a single network several RNAs or even regulatory proteins.

Three small untranslated RNAs (SroC, C0362 and C0664), containing REP-elements, have already been discovered. The most studied of them is SroC [25, 26] that was shown to interact with another small RNA, GcvB, known as a global post-transcriptional regulator of many mRNAs including those of ABC transporters (Opp, Dpp) and transcription factors (Lrp, PhoP and CsgD). It has been established that SroC binds GcvB via REP-sequence and this particular interaction mediates targeted GcvB degradation with the participation of RNAase E. Thus, the interaction of different transcripts with key regulatory RNAs mediated by REP-motifs may be the basis of their active participation in the network of intermolecular communications.

The most striking result of this work is the discovery of an unexpected REP-associated changes in the transcription profile nearby *promoter islands*. Taking into account the low transcriptional activity of the *islands*, the presence of elements with the potential terminator function at their borders has been quite predictable. However, in contrast to the expected suppression of outward transcription, we found REP-associated decrease in RNA synthesis directed towards the *islands* from neighboring regions (Fig. 7,**A** and 7,**D**). Hairpin structure works as transcription terminator depending on the presence and position of the homonucleotide track and operates as a bidirectional terminator only if such tracks flank both sides of the hairpin. Therefore unidirectional suppression of RNA synthesis in the vicinity of the REP-containing *islands* by itself is also quite natural. Surprising was the fact that the REP-elements flanking promoter *islands* are oriented in such a way as being aimed to isolate this atypical genomic regions from the external transcription. We consider this phenomenon as a sign of evolutionary fixed necessity to maintain an autonomous status of the *islands* protecting them from the transcriptional interference.

The study was supported by the grant of the Russian Scientific Foundation (project №14-14-00985).

REFERENCES

1. Stern M.J., Ames G.F.L., Smith N.H., Robinson E.C. and Higgins C.F. Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell.* 1984. V. 37. P. 1015–1026. doi: 10.1016/0092-8674(84)90436-7.

2. Higgins C.F., McLaren R.S., Newbury S.F. Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? *J. Gene.* 1988. V. 72. P. 3–14. doi: 10.1016/0378-1119(88)90122-9.
3. Stern M.J., Prossnitz E., Ames G.F.L. Role of the intercistronic region in post-transcriptional control of gene expression in the histidine transport operon of *Salmonella typhimurium*: involvement of REP sequences. *Mol. Microbiol.* 1988. V. 2. P. 141–152. doi: 10.1111/j.1365-2958.1988.tb00015.x.
4. Espeli O., Moulin L., Boccard F. Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *Mol. Biol.* 2001. V. 314. P. 375–386. doi: 10.1006%2Fjmbi.2001.5150.
5. Merino E., Becerril B., Valle F., Bolivar F. Deletion of a repetitive extragenic palindromic (REP) sequence downstream from the structural gene of *Escherihia coli* glutamate dehydrogenase affects the stability of its mRNA. *Gene*. 1987. V. 58. P. 305–309. doi: doi:10.1016/0378-1119(87)90386-6.
6. Newbury S.F., Smith N.H., Robinson E.C., HiIes I.D., Higgins C.F. Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell.* 1987. V. 48. P. 297–310. doi: 10.1016/0092-8674(87)90433-8.
7. Bachellier S., Clement J.M., Hofnung M. Short palindromic repetitive DNA elements in enterobacteria: a survey. *J. Res. Microbiol*. 1999. V. 150. P. 627–639. doi: 10.1016/S0923-2508(99)00128-X.
8. Khemici V., Carpousis A.J. The RNA degradosome and poly(A) polymerase of *Escherichia coli* are required *in vivo* for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol Microbiol*. 2004. V. 51. P. 777–790. doi: 10.1046/j.1365-2958.2003.03862.x.
9. Ton-Hoang B., Siguier P., Quentin Y., Onillon S., Marty B., Fichant G., Chandler M. Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. *J. Nucleic Acids Res.* 2012. V. 40. P. 3596–3609. doi: 10.1093/nar/gkr1198.
10. Moulin L., Rahmouni A.R., Boccard F. Topological insulators inhibit diffusion of transcription-induced positive supercoils in the chromosome of *Escherichia coli*. *Mol Microbiol.* 2005. V. 55 P. 601–610. doi: 10.1111/j.1365-2958.2004.04411.x.
11. Messing S.A, Ton-Hoang B., Hickman A.B., McCubbin A.J., Peaslee G.F., Ghirlando R., Chandler M., Dyda F. The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. *J. Nucleic Acids Res.* 2012. V. 40. P. 9964–9979. doi: 10.1093/nar/gks741.
12. Di Nocera P.P., De Gregorio E., Rocco F. GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes. *J. BMC Genomics.* 2013. V. 14. P. 522. doi: 10.1186/1471-2164-14-522.
13. Salgado H., Peralta-Gil M., Gama-Castro S., Santos-Zavaleta A., Muñiz-Rascado L., García-Sotelo J.S., Weiss V., Solano-Lira H., Martínez-Flores I., Medina-Rivera A. et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013. V. 41. P. D203. doi: 10.1093/nar/gks1201.
14. Panyukov V.V., Kiselev S.S., Shavkunov K.S., Masulis I.S., Ozoline O.N. Mixed *promoter islands* as genomic regions with specific structural and functional properties. *Mathematical Biology and Bioinformatics*. 2013. V. 8. P. t12–t26. doi: 10.17537/2013.8.t12.
15. Kiselev S.S., Ozoline O.N. Structure-specific modules as indicators of promoter DNA in bacterial genomes. *Mathematical Biology and Bioinformatics*. 2011. V. 6. P. t1–t13. doi: 10.17537/2011.6.t1.
16. Shavkunov K.S., Masulis I.S., Tutukina M.N., Deev A.A., Ozoline O.N. Gains and unexpected lessons from genome-scale promoter mapping. *Nucl. Acids Res.* 2009. V. 37. P. 4919–4931. doi: 10.1093/nar/gkp490.

17. Reppas N.B., Wade J.T., Church G.M., Struhl K. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol. Cell.* 2006. V. 24. P. 747–757. doi: 10.1016/j.molcel.2006.10.030.
18. NCBI Microbial Nucleotide BLAST. URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes (accessed: 07.06.2015).
19. lgarashi K., lshihama A. Bipartite functional map of the *E. coli* RNA polymerase α subunit: involvement of the C-Terminal region in transcription activation by CAMP-CRP. *Cell.* 1991. V. 65 P. 1015–1022. doi: 10.1016/0092-8674(91)90553-B.
20. Mathews D.H. RNA secondary structure analysis using RNAstructure. *Current Protocols in Bioinformatics*. 2014. V. 46. P. 12.6.1–12.6.25. doi: 10.1002/0471250953.bi1206s46.
21. Liang W., Rudd K.E., Deutscher M.P. A Role for REP sequences in regulating translation. *Mol. Cell*. 2015. V. 58. № 3. P. 431–9. doi: 10.1016/j.molcel.2015.03.019.
22. Brok-Volchanski A.S., Masulis I.S., Shavkunov K.S., Lukyanov V.I., Purtov Yu.A., Kostyanicina E.G., Deev A.A., Ozoline O.N. Predicting sRNA genes in the genome of *E. coli* by the promoter-search algorithm PlatProm In: *Bioinformatics of Genome Regulation and Structure II.* Eds. Kolchanov N., Hofestaedt R. Springer, 2005. P. 11–20. doi: 10.1007/0-387-29455-4_2.
23. Panyukov V.V., Ozoline O.N. Promoters of *Escherichia coli* versus promoter islands: function and structure comparison. *PLoS ONE*. 2013. V. 8. P. e62601. doi: 10.1371/journal.pone.0062601.
24. Purtov Yu.A., Glazunova O.A., Antipov S.S., Pokusaeva V.O., Fesenko E.E., Preobrazhenskaya V.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. *J. Bioinformatics and Computational Biology*. 2014. V. 12. № 2. P. 322–331. doi: 10.1142/S0219720014410066.
25. Vogel J., Bartels V., Tang T.H, Churakov G., Slagter-Jäger J.G., Hüttenhofer A., Wagner E.G. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.* 2003. V. 31 P. 6435–6443. doi: 10.1093/nar/gkg867.
26. Miyakoshi M., Chao Y., Vogel J. Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA. *EMBO J.* 2015. V. 34. P. 1478–1492. doi: 10.15252/embj.201490546.

Received March 22, 2016.
Published March 28, 2016.