Mathematical Biology and Bioinformatics
2015. V. 10. № Suppl. P. t29–t38. doi: 10.17537/2015.10.t29

The translation of the original article
2014 Glazunova O.A., Kiselev S.S., Shavkunov K.S., Bykov A.A., Panyukov V.V., Ozoline O.N.
Matematicheskaya biologiya i bioinformatika. 2014. V. 9. № 2. P. 563-574. doi: 10.17537/2014.9.563

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

UDC: 579:252

Promoter islands in the genome of *E. coli*: comparative analysis against AT-rich sequences

Glazunova O.A.[1], Kiselev S.S.[1], Shavkunov K.S.[1,2], Bykov A.A.[1,3], Panyukov V.V.[4], Ozoline O.N.[1,2*]

[1]*Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Russia*
[2]*Pushchino State Institute for Natural Sciences, Pushchino, Russia*
[3]*Nizhny Novgorod State University, Biological Faculty, Nizhny Novgorod, Russia*
[4]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Russia*

Abstract. The functional properties of *E. coli* genome *promoter islands* (PIs), i.e. regions with abnormally high contents of transcription signals, were compared to those of genomic areas, abnormally enriched with A/T-pairs. It was found that two representative sets of these regions partially overlap, and their functional properties are similar in many parameters. At the same time, *promoter islands* are characterized by a higher potential for synthesis of short oligonucleotides, as compared to AT-rich sequences. Such RNAs may be the target products of these unusual sites or byproducts of their suppressed state. The *islands* are richer in inverted repeats than AT-rich regions, and much richer compared with regular promoters. Considering that such structural elements commonly serve as targets for interactions with dimers or tetramers of regulatory proteins, it can be assumed that transcription initiation from *island*-embedded promoters is under the control of cell regulatory networks. The resulting RNA products might, therefore, be required for normal cell functioning. This idea is also supported by experimentally confirmed high yield of oligonucleotide product from the *island* promoter inside the *yjgL* gene.

Key words: *promoter islands, AT-rich genomic regions, abortive synthesis, untranslated RNA, horizontal gene transfer.*

INTRODUCTION

We have previously demonstrated that the genome of *E. coli* bears regions with unusual functional properties [1–4]. We termed these areas as *promoter islands* because of an anomalously high density of transcription initiation points in their sequences [1], and found that the high capability of PIs to form transcription complexes is in no agreement with their low transcription activity [2, 3]. The heterochromatin-like state of the *islands* proved to be reasoned by their association with genes, appeared in the genome of *E. coli* as a result of horizontal transfer [2, 3]. Such exchange of genetic information is quite common in natural systems, providing evolutionary benefits to a microorganism that obtained an advantageous gene by chance [5]. However, most of the “alien” genes are not needed for the cells. Thus,

*Ozoline@rambler.ru

promoters fully capable of transcription initiation [4] are in most instances “silenced” by special xenogeneic inhibitors [6, 7]. Such inhibitory function in *E. coli* is fulfilled by a histone-like protein H-NS, which predominantly binds to A/T-pair-enriched DNA regions [8–10], including *promoter islands* [3]. Hence the unusual proportion of long and short RNA products originating from the PIs might well be due to the heterochromatin-like state of the *islands*. In this case a similar ratio is expected in other AT-rich regions, i.e. in other areas of predominant H-NS binding. For all that, a possibility for a specific biological role of short RNA products synthesized from the *islands* also cannot be excluded. If that is true, their synthesis ought to be under the control of special regulatory networks and therefore can be different in case of PIs, which were chosen for the high contents of transcription signals, and in genomic regions selected for A/T-pair richness.

In the present study we picked out the most AT-rich sequences equal in size to the shortest *island*, and compared the functional properties of these two sets.

METHODS AND ALGORITHMS

The *E. coli* genome and sets of genomic sequences used in the work

In our study we used the genome of *E. coli* K12 MG1655 from NCBI GenBank (NC_000913.2) and the gene distribution map taken from RegulonDB [11]. Comparative analysis was carried out for genomic regions of five types. The first set consisted of 78 *promoter islands* collected previously [1, 2] as areas containing ≥ 8 potential transcription start points (TSPs) in each running window of 100 b.p., which span ≥ 300 b.p. The length of sequences taken into this set varied from 301 to 1102 b.p. Another set of regions taken for analysis included 95 AT-rich sequences 300 b.p. in length. It was formed using the ContentOf software (see below). The number of A/T-pairs in these sequences varied in the range from 216 to 238, giving a percent variation of 72÷79%, which is much higher that the content of A/T-pairs in the *E. coli* genome (49.2%). We have previously applied the three other types of genomic areas having multiple functional promoters, single promoters and control sequences, for a comparative analysis of the structural and functional characteristics of the PIs and usual promoters [2]. Each set included 78 sequences. The set of single promoters with minimal contents of transcription signals sufficient for appropriate gene expression included the sequences of standard length spanning from position –250 to +50 relative to TSPs. The set of multiple promoters included sequences with 3–9 experimentally verified transcription start points. Their size ranged from 318 to 847 b.p. and comprised the whole region delimited by the abovementioned borders of all single promoters. TSP coordinates for the latter group and for single promoters were taken from RegulonDB [11]. The control sequences had a standard length of 300 b.p. They were chosen from a vaster compilation used in [1] by the minimal promoter score, as calculated by the PlatProm algorithm.

Analysis of the *E. coli* genome with ContentOf software

The ContentOf software is a module of the software package aSHAPE [12]. This program plots the distribution histogram for regions with desired length (FragSz), ranging them by the amount of desired nucleotides (list of nucleotides Lst). Histograming is realized by reviewing every region in the genome with one base pair shift. The analysis of histograms allows the user to retrieve a part of the histogram and save it as a separate file for further study. Regions with increased A/T-content were picked out using the following parameters: FragSz = 300, Lst = {AT}.

Search for potential promoters

PlatProm algorithm [1] was used for searching of transcription initiation signals, which meet the criteria of structural and modular resemblance to promoters. The program scans both

DNA strands and for each position calculates the degree of its conformity (score) to a potential TSP. The estimated score was considered significant when it exceeded the background level by the value equaling to four standard deviations (StD), which was calculated using a previously prepared set of control sequences [1].

Search for direct and inverted repeats

We used Unipro UGENE (version 1.14.2) [13] to determine the repeated nucleotide sequence motives in the compared genomic regions. Direct and inverted repeats of ≥ 6 b.p. divided by a spacer of 5–10 b.p. were considered biologically significant in the context of the current tasks.

Analysis of contact sites with RNA polymerase and H-NS in the genomic DNA of *E. coli*

To assess the potential of RNA polymerase (RNAP) and H-NS to interact with selected areas of the genome, two data sets [14, 15] obtained by specific immunoprecipitation of DNA-protein complexes (ChIP-on-chip) were used. The method provides a means for elucidation of binding sites on the bacterial chromosome directly in the living cells for any target protein. This implies specific antibodies for retrieval of DNA complexes containing the target protein in isolated and fragmented chromatin (ChIP). The DNA extracted from the complexes is subsequently hybridized with high density microarrays (chip). The efficiency of RNAP binding to the genome was calculated as $\log_2$ from the ratio of microarray hybridization signals obtained with DNA after specific complex immunoprecipitation to signals from control DNA obtained without immunoprecipitation. A genomic region was considered as interacting with RNAP if the average $\log_2$ value for all microarray probes in the analyzed area was positive. The efficiency of interaction with H-NS was evaluated using the data already processed and published in [15]. In this case the genomic regions were deemed as interacting with H-NS when they overlapped with binding sites for at least 20 b.p.

Differential expression analysis

RNA products synthesized in *E. coli* cells were analyzed as described previously [2, 3]. The results of so far the only full-genome screening for 5'-end RNA sequences [16] were used as the input data for RNAMatcher software [2]. They consisted of 44 nt reads corresponding to 5′-ends of registered RNAs. The program searched for positions matching each of them in the genome and calculated the total number of identical reads in the dataset. If all 44 nucleotides coincided with the genomic sequence, such sample was assigned to a corresponding genomic position and was removed from the data set. The remaining set was searched for reads with lower coincidence on the 5'-end. Samples coinciding with the genome in 28–44 nucleotides were denoted as representing full-size RNA. Samples coinciding in only 9–20 nucleotides from the 5′-end (if the adaptor was found at the 3′-end) – were regarded as oligonucleotides. Samples with multiple matching to the genomic DNA were ascribed to all such positions in equal proportion.

Association of analyzed genomic regions with horizontally acquired genes

The relative disposition of the analyzed regions and genes, presumably obtained by *E. coli* via horizontal transfer, were determined using data obtained by five research groups [17–21]. The compared regions were considered as associated with foreign genes if they were embedded in the long *genomic islands* predicted by GIST [20] or IslandViewer [21] and if they overlapped with regulatory regions or coding sequences of foreign genes in ≥ 100 b.p. [17–19].

Targeted registration of short RNAs synthesized from *island* promoters

Cells of *E. coli* K12 MG1655 were grown in LB medium until early stationary phase. The intracellular contents of short RNA molecules was determined for cells cultured at 37°C in 100 ml of medium in a 0.5 liter flask with splitters (optimal growth conditions) or in microaeration conditions (two 50 ml flasks filled for 80%). The agitation rate was equal for both cultures. Following inoculation with overnight culture the medium was incubated for 8 h and cells were harvested by centrifugation. Short RNA fraction was derived with the use of mirVana™ kit (Ambion, USA). The 3'- and 5'-ends of isolated RNAs were modified with T4-RNA-ligase by attaching the adaptors with certain sequences (NEBNext^R Multiplex Small RNA Library Prep Set, New England Biolabs). Detection of target transcripts was done with two primer pairs. One primer in each pair corresponded to the 5'-end of the 5'-end or 3'-end adapter, while the other one matched the junction point between the 3'-end of the adapter and the target RNA. PCR was conducted on a programmable thermocycler DT-322 (DNA Technology, Russia). Amplicons were fractionated in 2% agarose gel containing ethidium bromide and photographed in UV-light.

RESULTS AND DISCUSSION

Promoter islands only partly overlap with the most AT-rich genomic regions

The most AT-rich regions of the *E. coli* genome were selected among the 300 b.p. long fragments. The number of A/T-pairs in such fragments varied in the range from 71 (23.7%) to 238 (79.3%) with a mean value of 147.6. A set of AT-rich fragments comparable in size to those of the previously composed sets (each of 78 fragments) was made using the threshold A/T-pair content of ≥ 216.

Table 1. Coordinates of left borders of AT-rich regions in the genome of *E. coli* K12 MG1655 and A/T-pair number

83984_224	953799_219	1752597_222	2783004_216	3453432_221	4258373_216
156970_223	986363_216	1811105_217	2784026_220	3467822_216	4266437_216
237009_216	996808_227	1984520_216	2882221_226	3579989_218	4280712_217
330784_216	1102507_223	2031798_224	2901625_228	3581077_218	4324863_225
384000_228	1196754_220	2054647_216	2903460_219	3582489_218	4435575_221
389066_221*	1210337_220	2101896_218	2986249_216	3632404_216	4474835_229
400358_220	1211312_226	2104051_218	2988958_227	3648953_217	4539632_231
522097_224	1211712_217	2190227_230	2989846_217	3651672_216	4540652_217
527819_218	1214738_216	2342291_219	2993025_226	3767666_238	4554376_225
567808_219	1463065_230	2453650_216	2993691_216	3795016_225	4570120_216
576058_218	1528094_223	2467454_230	3117099_219	3797231_232	* blue color indicates A/T-rich regions with no less than 50% overlap with coding sequences
582518_233	1542767_218	2468158_232	3134360_217	3798815_218	
583280_216	1544015_217	2589013_220	3183150_229	3802238_224	
584855_217	1570098_219	2627556_218	3267089_223	3802871_230	
651121_217	1580563_218	2758162_223	3285166_225	3834657_216	
735388_216	1596164_218	2772310_221	3358908_227	4233532_217	
752055_222	1622551_218	2772814_222	3383267_216	4249396_220	

A total of 95 sequences distributed along the whole genome were selected, which appeared to be located not only in intergenic spaces, but also inside genes, similarly to *promoter islands* (Table 1). Fifty-eight of them (61%) are situated inside of or adjacent to *promoter islands* (Fig. 1), and in two cases AT-rich fragments fell into the same *island*. However, 37 (38.9%) sequences enriched with A/T-pairs had no overlap with *islands*, while 22 *islands* (28.2%) did not overlap with fragments from the AT-rich set. Hence, the similarity of *island* and promoter sequences is dictated not solely by their richness in A/T-pairs.

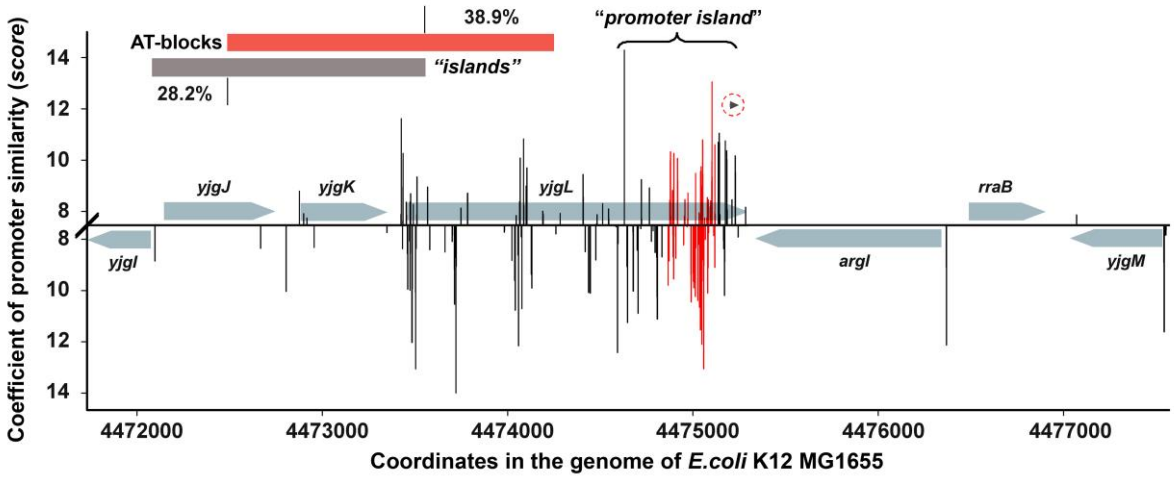

Fig. 1. Distribution of potential transcription initiation points (bars) predicted by PlatProm algorithm. The height of bars above and below the X-axis reflects the *score*, calculated for the upper and lower strands of the genome, correspondingly. The position of genes and the direction of their transcription are marked by grey arrows. The *promoter island* is indicated by a curly bracket, while the AT-rich region is shown in red color. The arrow in dashed circle denotes the area of synthesis and the size of the only RNA product registered from the *island* in [16]. Insertion on the left shows the scheme of overlap for the sets of *promoter islands* and AT-rich sequences.

Frequency of inverted repeats in *islands* is higher compared to that in other sets

Promoter search using PlatProm algorithm is based on the presence of a number of structural and functional modules in them. Along with motives having high or minor conservation, which are recognized by the exchangeable σ-subunits of RNAP or regulatory proteins of transcription machinery, these modules also include various A/T-tracks promoting anisotropic bends, flexible loops and thermodynamically unstable sites in the DNA double helix. That is why single and multiple promoters have a higher content of A/T-pairs than sequences from gene coding regions (Fig. 2,A). In this aspect *promoter islands* look rather like *superpromoters* and are not very much different from AT-rich sequences (Fig. 2,A).

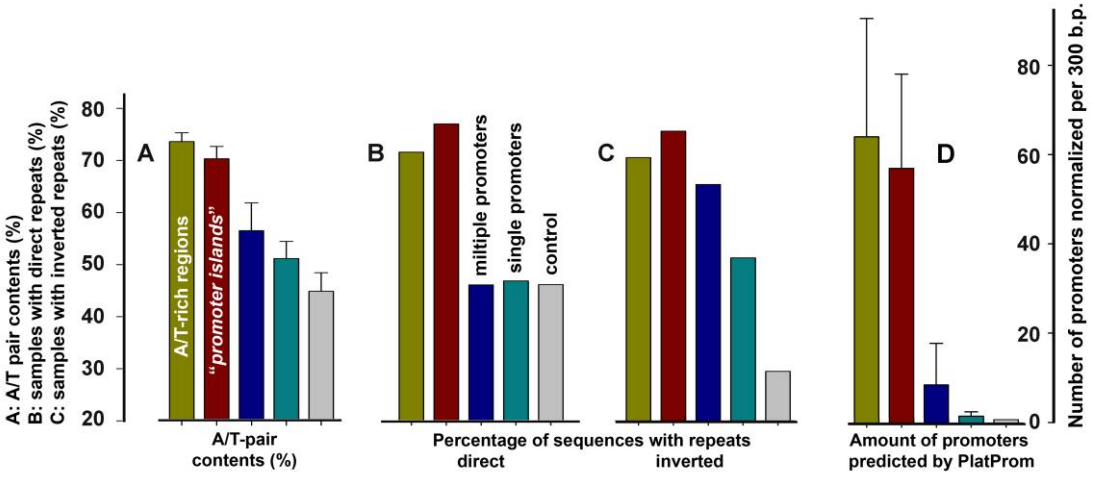

Fig. 2. Content of A/T-pairs (**A**), direct (**B**) and inverted (**C**) repeats, as well as potential promoters (**D**) in the compared sets of genomic sequences. The rate of A/T-pairs and promoters predicted by PlatProm with significance of $p < 0.00004$ was calculated for each sequence, averaged in each of the five sets and plotted with ± StD. The search for repeated motives was done using Unipro UGENE software (version 1.14.2) [13]. Repeats of ≥ 6 b.p. were considered significant.

Direct and inverted repeats are also important features of normal promoters [1, 22, 23]. Most of them interact with regulatory proteins. Thus, even not having the information about the contexts of all nucleotide sequences recognized by transcription factors, PlatProm uses discriminatory potential of repeated motifs. The context of these modules is not necessarily AT-rich, hence their presence or absence provides an independent criterion for comparison of the investigated sets.

The search for repeated sequences was carried out using Unipro UGENE software (version 1.14.2) [13]. The results are demonstrated in Fig. 2,B and C. Surprisingly, the sets of single and multiple promoters did not differ by the presence of direct repeats from the control sequences taken from within the coding parts of genes. This might be due to the preset potential for coding regular protein structures. Thus, even short tracks of identical amino acids can be encoded by triplets, which make direct repeats in the DNA. An even more intriguing finding was the equality of single and multiple promoters containing tandem repeats. This observation can presumably be explained by their relatively infrequent use for interaction with regulatory proteins compared to inverted repeats (Fig. 2,C) The percentage of *islands* with direct repeats turned out to be higher than that for the group of AT-rich sequences; still, the abovementioned parity of the three other sets does not allow considering it as a significant fact.

The type of distribution of inverted repeats in the analyzed groups proved to be completely different (Fig. 2,C). Only 29% samples from the control group had such modules. The percentage of sequences from the three promoter groups, which carry inverted repeats, grew consistently with an increase of the amount of potential promoters. This is exactly in line with the expectation, considering that inverted repeats are targets for interaction with enantiomorphic dimers of regulatory proteins. The number of such sites in multiple promoters is expected to be even higher because of the certain contribution from conservative elements of divergent promoters. The fact that the *islands* retained this tendency implies their evolutionary selection in view of their promoter function. In AT-rich sequences this tendency is broken: having higher density of potential promoters (Fig. 2,D and example in Fig. 1), the content of sequences with inverted repeats appeared to be lower than that in *islands*. This can be possibly accounted as evidence speaking for the different kinds of evolutional pressure on these two types of sequences and, therefore, for possible difference in their functions.

Promoter islands and AT-rich sequences have different profiles of synthesized RNA products

We have previously determined that 75 out of 78 *islands* (96.2%) are located near genes acquired by *E. coli* via horizontal transfer [2, 3]. This is significantly higher than the percent of single and multiple promoters with similar localization (50.0% and 53.9%, respectively). The rate of such regions in AT-rich sequences was virtually the same compared to that in *islands* (92.6%), being in good correspondence with the assumption that foreign genes have specific genomic surroundings enriched with A/T-pairs [20]. Perhaps this very peculiarity lays the basis for their preferential binding with histone-like protein of the bacterial nucleoid – H-NS [6–9, 24, 25]. Thus it is no wonder that we determined a similar high efficiency of interaction with this protein for the set of A/T-rich sequences (Fig. 3,A). Since these loci partly overlap with *islands* (insert in Fig. 1), often reside in gene regulatory regions (Table 1) and have high density of promoter-like sites (Fig. 2,D), it is no surprise that they interact with RNAP with equal efficiently compared to the *islands*, and better than sequences from the other sets (Fig. 3,B). The question was: do they vary from the *islands* in their relative ability to initiate the synthesis of short and long RNA products?

To answer this question we carried out a differential analysis of data published in [16]. These data contain information about 5’-end sequences up to 44 nucleotides long, of the RNAs synthesized in bacterial cells. The analysis was carried out as described previously

[2, 3], but all reads corresponding to the genome in at least 28 nucleotides were considered as full-sized products, while previously it had been done for only 44-nucleotide sequences. Reads coinciding with the genome in 9–20 5'-end nucleotides were referred to short RNA products. The data used [16] clearly confirmed initiation of RNA ≥ 28 nucleotides in only 15 *islands* (19%) and 15 AT-rich sequences (16%). This is far less that the rate of single and multiple promoters, successfully driving the synthesis of full-sized RNA (83% and 87%, correspondingly). Upon normalization of the number of products from the transcribed regions to the size of the analyzed regions, the difference between these sets became sharper (Fig. 3,C). Thus, based on a larger dataset (reads of 28–44 nucleotides) we confirmed the low capacity of the *promoter islands* to drive the synthesis of long RNAs [2, 3] and testified the same property for the AT-rich sequences.

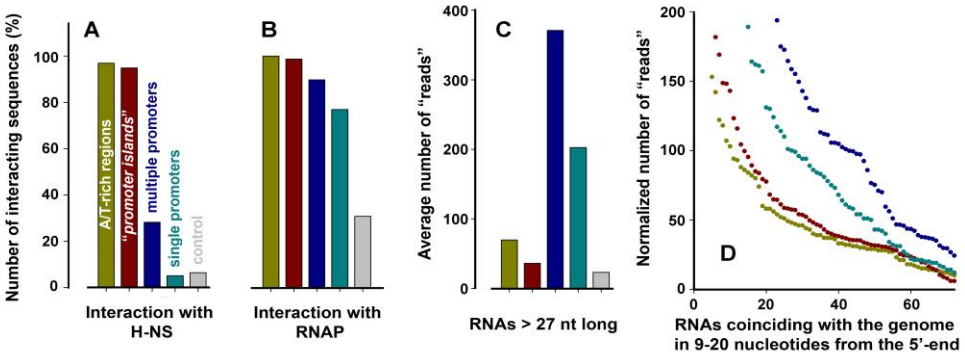

Figure 3. Capability of the compared sets of genomic sequences to interact with H-NS (**A**) and RNAP (**B**), and to drive the synthesis of long (**C**) and short (**D**) RNAs. The potential of the studied sets for interaction with H-NS and RNAP was assessed by the number of sequences forming appropriate complexes (data from [14, 15]). Evaluation of transcription activity was done using the results of RNA-seq [16]. Reads of different length ranges shown in panels **C** and **D** were summarized for each genomic region. The sum was normalized on the length of 300 b.p. for the *islands* and multiple promoters with irregular size. Panel **C** demonstrates the average number of long reads for the transcribed sequences. Ranged graphs in panel **D** plot the amount of short reads registered for each analyzed region and normalized on the fragment lengths. Their color legends correspond to those on the other panels.

Reads of 9–20 nucleotide length were registered for virtually each of the analyzed sequences. The number of potentially short RNA products with starts falling inside the *islands* exceeded the amount of similar products from AT-rich regions. This means that despite the similar localization in the genome (often close to foreign genes) and high likeness of structural and functional properties, the *promoter islands* differ from AT-rich sequences in a somewhat higher ability to synthesize short RNAs.

Promoters of the *yjgL*-associated *island* promote synthesis of short RNAs

Differential analysis allowed evaluating the ratio of long and short RNA products in different parts of the genome. Still, the obtained results are insufficient to justify the synthesis of those particular RNAs registered among the published reads. First of all, many of them can be products of degradation. Secondly, the 5'-end adapter used in [16] was attached to the RNAs with T4 RNA ligase, while the 3'-end adapter included 9 random nucleotides at its 3'-end. It was included into the DNA copies in the reverse transcription reaction, and the random sequences ensured obtaining the copies of all RNA species. Yet this strategy gave no possibility of exact determination of the 3'-end of the product, since the 9-nucleotide random sequence might have come across a complementary target also in the internal regions of RNAs. The applied differential analysis reduced the error rate, because all regions with longer reads were excluded to select producers of potentially short RNAs. Besides, only samples carrying the adapter sequence at the 3'-end were taken for this selection. However, all this does not abolish the uncertainty, thus complete genome sequencing of short RNAs following

their preliminary ligation with adapters at both ends is required. We accomplished such ligation for the fraction of short RNAs from cells grown in different conditions.

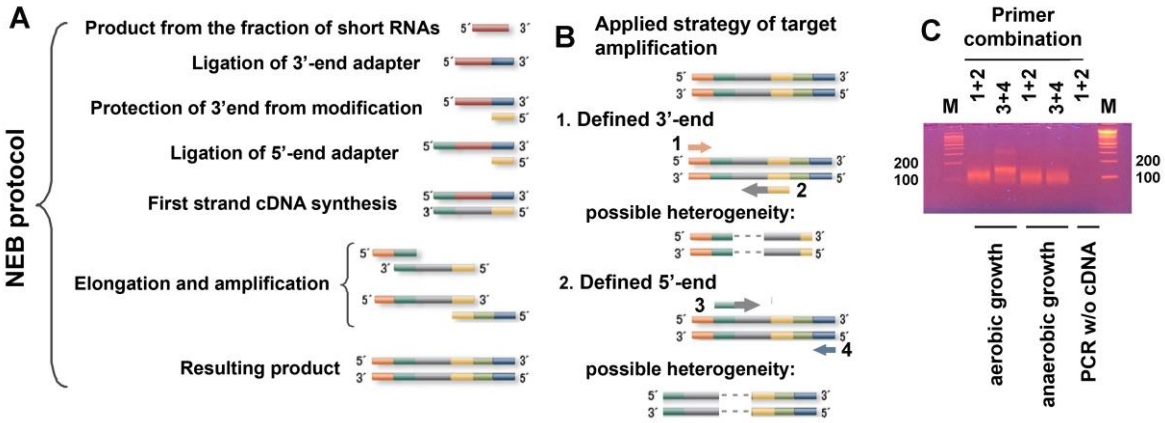

Fig. 4. Detection of the 26 nt long RNA synthesized from the *island*-associated promoter of the *yjgL* gene (see Fig. 1), which was recovered during analysis of RNA-seq results [16]. The protocol of sample preparation proposed by NEB (**A**) and our strategy for subsequent target amplification (**B**) are shown schematically. The sought RNA sequence: AATCCACCGGAAAATTTACGTATAGC. Products of PCR were obtained using primers: 5'- AATGATACGGCGACCACC-3' (1), 5'-CTCTTCCGATCTTGCTATACGTAAA-3' (2), 5'-TCCGACGATCAATCCACC-3' (3) and 5'-CAAGCAGAAGACGGCATAC-3' (4). Underlined are the regions corresponding to the sequence of the target RNA. Amplicons were visualized in 2% agarose. Sample without (w/o) cDNA was used as a negative control. The gel was calibrated with DNA markers from Evrogen (left lane) and NEB (right lane).

Fig. 4 schematically illustrates the sample preparation strategy. Adaptor ligation was done using protocols and reagents from NEB (Fig. 4,A). Detection of RNA expected to be synthesized from the *island* promoter was realized via PCR with special primer pairs, and in each of them one primer corresponded to the junction point between the 3'-end of an adapter and the target oligonucleotide (Fig. 4,B). Fig. 4,C shows an example of PCR product obtained for RNA from the *yjgL*-associated *island* (marked with a dashed circle in Fig. 1). Providing the exact correspondence to RNA-seq data the length of the fragment is expected to be 95 b.p. when using primers 1 and 2 and 101 b.p. with primers 3 and 4, but in both cases the products can vary in length (Fig. 4,B). It appeared that under normal growth conditions (Fig. 4,C), the bacterial cells contain not only the sought type of RNA (26 nucleotides), but also a longer product (~250 nucleotides). Since it was amplified with primer pair 3 + 4, we suggest that its synthesis is initiated from the promoter of the short RNA. In any case, this experiment confirms the synthesis of the 26-nucleotide RNA initiating from an *island* promoter, and not from one of the strong promoters from overlapping with AT-rich region (indicated by red in Fig. 1).

CONCLUSION

The transition of genomics from studying individual genes to analysis of complete genomes enabled elucidation of a number of novel structural and functional elements, including *promoter islands*. Their emerging near horizontally acquired genes might require significant corrections to the existing views on the mechanisms of accelerated evolution. Another peculiar feature of the *islands* is their unexpectedly low ability to initiate the synthesis of full-sized RNAs. According to the available data [1–3] this might be due to heterochromatization of the *islands* by the histone-like protein H-NS, which interacts with A/T-pair-rich sequences as primary targets. In this instance the structural and functional properties of *islands* are expected to resemble those of the AT-rich sequences. The present study was carried out to verify this assumption, which was largely confirmed. Still, it turned out that the *islands* contain inverted repeats more often, compared to AT-rich regions, and far

more often than regular promoters. Such structural modules often act as targets for interaction with dimeric and tetrameric forms of regulatory proteins. Thus it can be anticipated that transcription initiation from *island* promoters is to a greater extent controlled by the regulatory networks of the cells, than from AT-rich sequences.

Previously it has been shown that the *islands* can synthesize short RNA products, but it is still unclear, whether they are target products of these regions or by-products of their suppressed state. A greater capacity of the *islands* to form such products, compared to AT-rich regions (Fig. 3,C), higher contents of potential targets for interaction with regulatory proteins (Fig. 2,C) and confirmed synthesis of a short RNA from within the *yjgL* gene (Fig. 4,C) argument for their biological significance. For instance, *promoter islands* in microbial genome may serve as peculiar factories for the production of special short RNA. Though the biological role of such RNAs requires conceptual comprehension, the results of comparative analysis testify different patterns of evolutionary selection in respect of the *promoter islands*, which are picked out by the high density and even distribution of potential transcription signals, and AT-rich sequences, selected based on the high contents of A/T-pairs in a genomic region of a given length.

Experimental part of the study was supported by the Russian Science Foundation (grant № 14-14-00985), while bioinformatics analysis by the Russian Foundation of Basic Research (grant № 13-04-0997).

REFERENCES

1. 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.
2. Panyukov V.V., Ozoline O.N. Promoters of *Escherichia coli* versus *promoter islands*: function and structure comparison. *PLoS ONE*. 2013. V. 8. Article No. e62601.
3. 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
4. Shavkunov K.S., Tutukina M.N., Masulis I.S., Ozoline O.N. Promoter islands: the novel elements in bacterial genomes. *J. Biomol. Struct. Dyn.* 2011. V. 28. P. 1128–1129.
5. Syvanen M. Evolutionary implications of horizontal gene transfer. *Annu. Rev. Genet*. 2012. V. 46. P. 341–358.
6. Lucchini S., Rowley G., Goldberg M.D., Hurd D., Harrison M., Hinton J.C. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathog.* 2006. V. 2. Article No. e81.
7. Navarre W.W., Porwollik S., Wang Y., McClelland M., Rosen H., Libby S.J., Fang F.C. Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science*. 2006. V. 313. P. 236–238.
8. Dorman C.J. Horizontally acquired homologues of the nucleoid-associated protein H-NS: implications for gene regulation. *Mol. Microbiol.* 2010. V. 75. P. 264–267.
9. Dorman C.J. Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nat. Rev. Microbiol.* 2013. V. 11. P. 349–355.
10. Dorman C.J. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. *Plasmid.* 2014. V. 75. P. 1–11.
11. 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., Salgado-Osorio G., Alquicira-Hernández S., Alquicira-Hernández K., López-Fuentes A., Porrón-Sotelo L., Huerta A. M., Bonavides-Martínez C., Balderas-Martínez Y. I., Pannier L., Olvera M., Labastida A., Jiménez-Jacinto V., Vega-Alvarado L., Del Moral-

Chávez V., Hernández-Alvarez A., Morett E., Collado-Vides J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucl. Acids Res.* 2013. V. 41. P. D203–D213.

12. Panyukov V.V., Nazipova N.N., Ozoline O.N. Software Package aSHAPE as a Tool to Study the Spatial Conformation of the Bacterial Genome Sites. *Mathematical biology and bioinformatics*. 2011. V. 6. P. 211–227. doi: 10.17537/2011.6.211
13. Okonechnikov K., Golosova O., Fursov M., the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics.* 2012. V. 28. P. 1166–1167.
14. 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.
15. Kahramanoglou C., Seshasayee A.S.N., Prieto A.I., Ibberson D., Schmidt S., Zimmermann J., Benes V., Fraser G.M., Luscombe N.M. Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucl. Acids Res.* 2011. V. 39. P. 2073–2091.
16. Dornenburg J.E., DeVita A.M., Palumbo M.J., Wade J.T. Widespread antisense transcription in *Escherichia coli*. *mBio.* 2010. V. 1. Article No. e00024-10.
17. Lawrence J.G., Ochman H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA.* 1998. V. 95. P. 9413–9417.
18. Nakamura Y., Itoh T., Matsuda H., Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* 2004. V. 36. P. 760–766.
19. Price M.N., Dehal P.S., Arkin A.P. Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biol.* 2008. V. 9. Article No. R4.
20. Huang Q., Cheng X., Cheung M.K., Kiselev S.S., Ozoline O.N., Kwan H.S. High-density transcriptional initiation signals underline genomic islands in bacteria. *PLoS ONE*. 2012. V. 7. Article No. e33759.
21. Langille M.G., Brinkman F.S. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics.* 2009. V. 25. P. 664–665.
22. Ozoline O.N., Deev A.A. Predicting antisense RNAs in the genomes of *Escherichia coli* and *Salmonella typhimurium* using promoter-search algorithm PlatProm. *J. Bioinf. Comput. Biol.* 2006. V. 4. P. 443–454.
23. 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
24. Oshima T., Ishikawa S., Kurokawa K., Aiba H., Ogasawara N. *Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Res.* 2006. V. 13. P. 141–153.
25. Dorman C.J. H-NS, the genome sentinel. *Nat. Rev. Microbiol.* 2007. V. 5. P. 157–161.

Received April 08, 2015.
Published April 15, 2015.