PREDICTING sRNA GENES IN THE GENOME OF *E. COLI* BY THE PROMOTER-SEARCH ALGORITHM PlatProm

A.S. Brok-Volchanski[1], I.S. Masulis[1], K.S. Shavkunov[1], V.I. Lukyanov[1], Yu.A. Purtov[1], E.G. Kostyanicina[1], A.A. Deev[2], O.N. Ozoline[1*]

[1] *Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia; e-mail: ozoline@icb.psn.ru;* [2] *Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russia*

[*] *Corresponding author*

Abstract: The potentially transcribed regions in the genome of *E. coli* were searched for on a systematic basis using the novel pattern recognition software PlatProm. PlatProm takes into consideration both the sequence-specific and structure-specific features in the genetic environment of the promoter sites and identifies transcription start points with a very high accuracy. The whole genome scanning by PlatProm along with the expected promoters upstream of the annotated genes identified several hundred of very similar signals in other intergenic regions and in many coding sequences. Most of them are expected as start points for independent RNA transcripts, providing a unique opportunity to reveal genes encoding antisense and/or alternative RNAs. The potential PlatProm as a tool revealing sRNA genes is discussed.

Key words: promoters; genome regulatory regions; gene expression; promoter-search software; transcription; untranslated RNAs; genome annotation

1. INTRODUCTION

The current annotation of bacterial genomes relies on computational methods identifying coding sequences on the basis of certain specific features in base composition, codon usage, the database of *N*-terminal peptide sequences, matches for the ribosome binding site, transcription and translation terminators, homology with known genes in other bacterial species, expression efficiencies measured by microarrays, and other available information (Blattner et al., 1997; Lukashin and Borodovsky, 1998; Delcher et al., 1999; Besemer et al., 2001; Walker et al., 2002; Azad and Borodovsky, 2004). Although these methods are not perfectly precise, they allow identifying most protein-encoding genes, and the longest possible *N*-termini were generally selected in cases when multiple in-frame start codons were found. A high level of conservation and the specific features of the three-dimensional structures of rRNAs and tRNAs were employed to identify genes of these RNA species. However, the detection of sRNA genes (small untranslated RNAs) in bacterial genome is still problematic. These RNAs regulate diverse cellular functions, such as RNA processing, mRNA stability, translation, protein stability, and secretion. The first 13 were discovered fortuitously on the basis of high abundance or functions related to protein synthesis or activity. After genome-wide identification of these gene species became a focus of attention, 49 novel sRNA genes with yet unknown functions were found.

The largest contribution to the set of novel sRNAs was given by the approach primarily based on the sequence conservation within intergenic regions and exploiting microarrays as well as other techniques for experimental verification (Wassarman et al., 2001). Overall, 17 novel sRNAs and 6 ORFs were suggested by this combinatorial approach; however, sRNA genes that had evolutionary conserved secondary structures rather than nucleotide sequences might be overlooked as well as any sRNA genes overlapping the neighboring coding sequences. The former limitation was surmounted by Rivas et al. (2001), who distinguished conserved RNA secondary structures from a background of other conserved sequences using probabilistic models of expected mutational patterns in pairwise sequence alignments. Sequence-based structure comparison among genomes of closely related bacteria allowed detecting eight novel sRNAs. Carter et al. (2001) analyzed intergenic regions in terms of nucleotide and dinucleotide composition, occurrence frequency of sequence motifs typical of RNA structural elements, free energy of folding, and some other considerations, which led to the prediction of 562 sRNAs. Approaches relying mainly on the biological features of sRNAs were suggested by Vogel et al. (2003) and Zhang et al. (2003). In the former case, 50–400 nt RNA products were picked out from the total fraction of RNAs and their sequences were estimated. Most of them were derived from within the coding regions including the small fraction (~ 5 %) matching to the antisense strand. Fragments from intergenic regions comprise 18 % of all samples, and seven novel sRNA genes were identified among them. In the latter case (Zhang et al., 2003), a set of new sRNAs was chosen from the fraction of RNAs that co-immunoprecipitated with Hfq. Although only ~ 30 % of the known sRNAs interact with Hfq, this method allowed revealing five additional sRNAs.

The predictive potential of transcription signals was exploited by Argaman et al. (2001) and Chen et al. (2001). The first team mostly relied on the accuracy of terminator prediction. Promoters were searched for upstream of terminators by approaches combining the consensus and weight matrix considerations. All the sequences 50–400 bp long located between the promoter and terminator within empty intergenic regions were compared to the genomes of other bacteria, and conserved sites were assayed experimentally. Expressions of 14 out of 24 predicted genes were detected. Chen et al. (2001) used RNAMotif and the thermodynamic scoring system to detect the set of terminators in the bacterial genome. Several hundreds were suggested as potential stop signals for genes yet to be discovered. Combining these data with a set of promoters predicted by profile-based software (Bucher et al., 1996) and requiring the presence of both signals on the same strand 45–350 bp away from each other, 227 candidates were selected. Expression of eight candidates was tested and confirmed in seven cases. Thus, both approaches employing transcription signals as a primary search criterion demonstrated a very good proportion between the predicted and the confirmed candidates.

What part of the overall population of sRNAs do the 62 verified genes comprise? More than 1000 other potential sites for sRNA synthesis were predicted in intergenic regions (Hersberg et al., 2003), and approximately the same number of transcripts generated from these regions was detected by the expression analysis (Tjaden et al., 2002). This essentially exceeds the estimates made for the total number of sRNAs in *E. coli* (50–200 genes) (Eddy, 1999; Wassarman et al., 2001). Many of the predicted genes may therefore be false positives. On the other hand, the discovered noncoding RNAs range from 45 to 370 nt in length. Considering that shorter RNAs (21–25 nt) in eukaryotic organisms are involved in RNA interference, processing, chemical modification, and stabilization, while gene silencing is controlled by longer RNAs, some untranslated RNA species may be overlooked due to commonly used size limitation. Moreover, most methods employed so far are focused on intergenic regions, excluding a possibility to find new transcripts generated from sequences encoding proteins, although experimental approaches demonstrate the possibility of both antisense and alternative (shortened) transcription from these sequences (Selinger et al., 2000; Vogel et al., 2003). This means that scanning techniques capable of predicting genes with parallel transcriptional output are required. Katz and Burge (2003) and Pedersen et al. (2004), who proposed methods revealing local RNA secondary structures within bacterial genes, made the first step in this direction. However, folding propensity may not be a general feature of intrinsic transcripts. Since the methods of comparative genomics cannot be used to detect them within ORF, approaches searching for transcription signals may have the highest potential here.

Promoter search algorithms proposed so far (Alexandrov and Mironov, 1990; Horton and Kanehisa, 1992; Mahadevan and Ghosh, 1994; Pedersen and Engelbrecht, 1995; Hertz and Stormo, 1996; Yada et al., 1999; Vanet et al., 1999; Leung et al., 2001; Gordon et al., 2003; Huerta and Collado-Vides, 2003) are usually based on the sequence preferences in the regions of specific contacts with RNA polymerase. Most of them can identify more than 80 % of promoters from testing compilations but even the best protocols at this level recognize a large portion (0.8–3.4 %) of non-promoter DNA as promoter-like signals (Horton and Kanehisa, 1992; Gordon et al., 2003). Within the genome size sequences, the background noise is, therefore, more than one order of magnitude greater than the required signal. That is why promoters are usually searched for as the most probable candidates of several promoter-like signals found within a limited region upstream of a particular gene (or terminator, as discussed above).

We tried to increase the performance of computational prediction by taking into account structural features in the genetic environment of the promoter sites, considering them as a generalized platform for transcription complex formation. The resultant algorithm PlatProm was used for promoter prediction within the entire genome of *E. coli*. The disposition of promoter-like signals according to the gene borders is discussed.

2. METHODS AND ALGORITHMS

Compilations. The training set contained 400 σ^{70}-promoters with single start point. The testing set contained 290 known promoters with single or multiple transcription start points. Overlapping and homologous promoters were removed from both sets. All sequences were 411 bp long (–255/+155 according to the start point, nominated as 0). The control set contained 400 sequences taken from the coding regions of convergently transcribed genes.

Weight matrices. Three types of weight matrices were used to formalize structural organization of the promoter sites. The matrices of the first type reflect distribution of nucleotides in the conservative elements (–35 and –10) and dinucleotides near the start point (position –1) and in the 5′-flanking region of element –10 (‘extended –10 element’). These matrices are designed exactly as described by Hertz and Stormo (1996) and contain 6 × 4 (or 1 × 16) scores equal to natural logarithms of normalized frequencies of the appearance of each nucleotide (dinucleotide) at each position of the preliminary aligned promoters. Occurrence frequency of particular nucleotides (dinucleotides) in genome was used for normalization. Allowed variations of the spacer and the distance between the start point and the element –10 were 14–21 and 2–9 bp, respectively. This means that 64 alignments were tested for each sequence to find the maximal score. Any deviations from the optimal spacer (17 bp) and optimal distance (6 bp) were penalized based on their frequencies. Optimal matrices were generated by the procedure of expectation–maximization.

Table -1. Weight matrix for $(T)_n$ in the region –83/–76

n	Scores	FS
< 4 (penalty)	–0.02	–0.01
4	0.27	0.14
5	0.74	0.38
≥6	1.32	0.69

Table -2. Weight matrix for TA in the region –49/–47

Order	Position	FS
1 check	–48	0.56
2 check	–47	0.52
3 check	–49	0.39
Penalty	No	–0.13

The $(T)_n$ or $(A)_n$ ($n \geq 4$) tracts interacting with RNA polymerase α-subunits (Wada et al., 2000) or stabilizing the transcription complex by a properly induced bend (Hivzer et al., 2001) were accounted by the set of 26 simplified matrices scoring the presence of these elements in 26 regions from –20 to +34 (distributed with a periodicity of ~ 1 helix turn). An example of such matrices, accounting the presence of $(T)_n$ in the region –83/–76 is shown in Table 1. Positive scores represent log probabilities to find 5′-end of $(T)_n$ in the region. The penalty is estimated as a probability of $(T)_n$ absence. Final scores (FS) were reduced by the coefficient estimated as a ratio of the average information content in the region to the information content at the sixth position of aligned element – 35 (the least significant). This reduction was aimed to balance the endowments of structure-specific elements with the contributions of conservative base pairs.

Flexible dinucleotides TA, supporting adaptive isomerization of the DNA on the protein surface (Ozoline et al., 1999a; Masulis et al., 2002), were accounted by 20 cascade matrices exemplified in Table 2. The presence of TA is first checked at the position where probability to find it is maximal and than at the adjacent points. An absence of TA is penalized. The whole set of such matrices reflects a regular distribution of TA in the region of – 98 to + 24.

A large contribution to the sensitivity of PlatProm is made by mixed A/T-tracts, putatively involved in the polymerase sliding along the DNA (Ozoline et al., 1999b). These elements are distributed with periodicities of

~ 1 and ~ 1.5 helix turns. To take into account this regularity, we scored the presence of paired rather than single A/T-tracts: www(*n*)www (w = A = T; where *n* is 7, 8, 13, or 14 random base pairs) in the region of – 139 to + 11. Overall 15 cascade matrices (similar to that shown in Table 2) were used for this purpose.

Ideal direct and inverted repeats (5–11 bp long separated by 5 or 6 bp) were considered as putative targets for interaction with transcription regulators. Their presence was scored as natural logarithms of the lengths. Contributions of all elements were summarized, giving the total score. An average score for non-promoter sequences amounted to – 3.8. The value of Std was equal to 3.0. An average score of promoters from the testing set was 7.29.

The distribution of promoters predicted by PlatProm within the genome of *E. coli*. Only highly reliable signals scoring ≥ 8.2 ($p < 0.00005$) were used. The genome map of *E. coli* K12 (NCBI, GenBank entry U00096) was basically used for this purpose, but ~ 300 genes additionally annotated in the Colibri DataBase were added. The allowed distances between the transcription start points and the coding sequences were deduced based on positional coordinates typical of the known promoters (Figure 1). Although ~ 90 % of them are < 250 bp far from coding sequences, some genes have leaders as long as 600 bp and more. That is why we considered 750 bp as the distance at which the predicted promoter may be ascribed to the downstream gene. Other promoter-like points were sorted according to their location relative to gene borders. The map of known and predicted transcription start points is available by request (ozoline@icb.psn.ru).

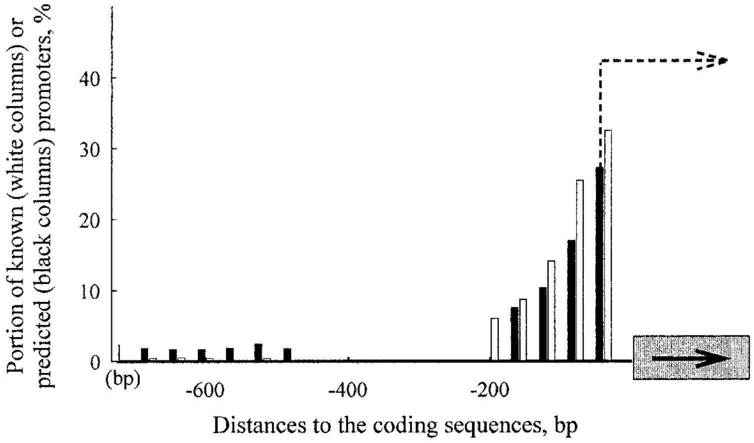

Figure -1. Distributions of known and predicted promoters relative to coding sequences. Arrows indicate orientation of the gene (rectangle) and the direction of transcription.

3. RESULTS AND DISCUSSION

The current version of PlatProm identifies 84.8 % promoters of testing compilation at the level when zero false positives were found in the control set (scores > 3.0). This means that the combination of sensitivity (((true positives (TP))/(TP + false negatives (FN))) × 100 = 84.8 %) and specificity (((TN/(TN + false positives (FP))) × 100 = 100.0 %) of our approach is better than in the case of algorithms based on neural networks (80.6 % and 99.14 %, respectively; Horton and Kanehisa, 1992); logic grammar formalism (68.7 % and 82.23 %; Leung et al., 2001); and sequence alignment kernel (82 % and 84 %; Gordon et al., 2003). Another three parameters characterizing performance of promoter search algorithms are AE (average error) = (((FN + FP)/(N + P)) × 100), CC (correlation coefficient) = (TP × TN – FP × FN)/((TP + FP) × (TN + FN) × (TP + FN) × TN + FP))$^{1/2}$, and accuracy of the transcription start point prediction among others promoter-like signals detected in the same region. AE and CC for PlatProm are 6.3 and 0.87, respectively, which is better than reported originally (16.5 and 0.67) (Gordon et al., 2003). Positions with maximal scores coincided with experimentally estimated start points or were located at the nearest two positions in ~ 80 % cases, while this value is usually lower than 50 % (Hertz and Stormo, 1996; Huerta and Collado-Vides, 2003).

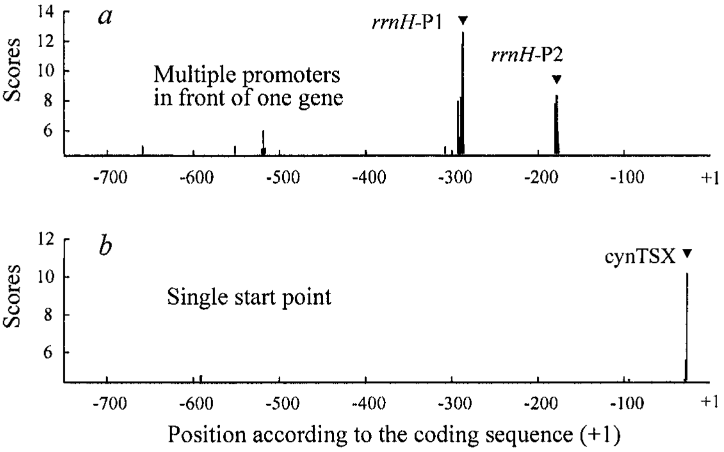

Figure -2. Distributions of promoter-like signals (columns) within 750-bp regions upstream of (*a*) rrnH and (*b*) cynT genes. The positions of known promoters are indicated by triangles.

The percentage of recognized promoters increases up to 90 % if ± 5-bp variations in the positioning of the start point were allowed. For the training set, this value is 94 %, while the percentage of known promoters recognized in the genomic DNA is 91.4 %. Figure 2 exemplifies distribution of promoter-like signals within the regulatory regions of two genes. In the case of gene encoding 16S RNA, both known promoters (*rrnH*-P1 and *rrnH*-P2) are surrounded by other promoter-like signals forming two clusters (Figure 2*a*). Huerta and Collado-Vides (2003) already mentioned this phenomenon and assumed that additional sites might be used for polymerase trapping.

Putative promoters were found upstream of 2229 genes. Most of them form only one cluster or appear as single promoter-like point, like the known promoter *cynTSX* (Figure 2*b*). The rest form several clusters, providing a possibility to predict multiple promoters with potentially different regulation. Similarly to known promoters, the predicted sites are usually located less than 250 bp away from coding sequences (Figure 1, black columns). The information given by this set of predicted promoters facilitates their experimental identification and sometimes gives a chance to find functional promoters in the places where they otherwise may be overlooked.

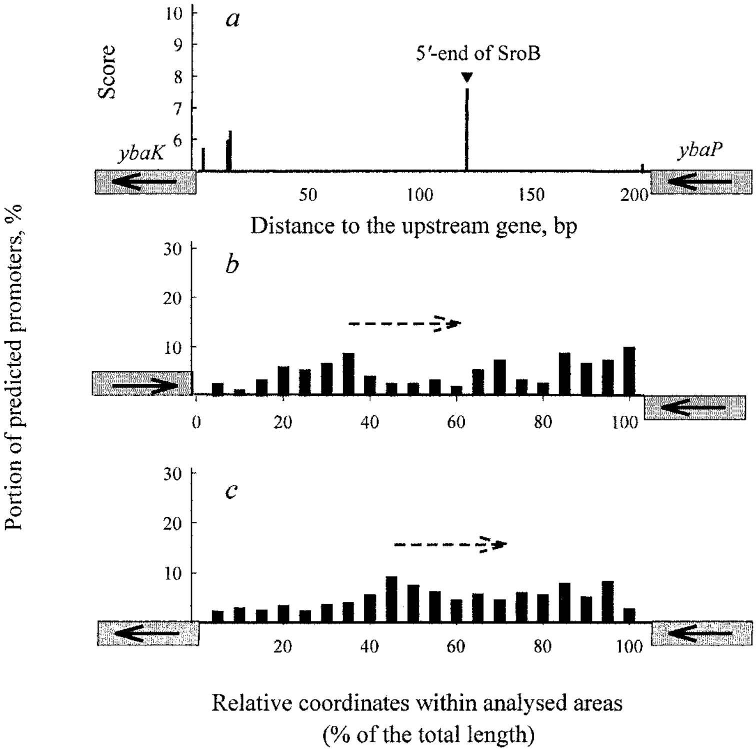

Figure -3. Distribution of promoter-like points between *yba*K and *yba*P genes. The 5'-end of the SroB sRNA is indicated (*a*); relative positioning of the predicted promoters relative to adjacent genes (rectangles) (*b* and *c*). Solid arrows show directions of transcription; dashed arrows, orientation of the predicted promoters.

3.1 Putative promoters in intergenic regions

More than 600 promoters were predicted between the genes transcribed convergently or from the opposite DNA strand. Their positioning does not show any preference (Figures 3*b* and 3*c*). The average distance between non-overlapping genes in the genome of *E. coli* is 148 bp. The average length of intergenic regions containing potential promoters is larger (440 and 298 for the sets in Figures 3*b* and 3*c*, respectively), thus suggesting the existence of additional genes. Several hundreds of sRNA genes are predicted in such regions (see above). Thus, the presence of functional promoters for their expression is also expected. Figure 3*a* shows the potential promoter for the predicted SroB sRNA (Vogel et al., 2003).

3.2 Potential promoters for antisense transcription

The 709 promoter-like signals detected on the opposite strand of protein-encoding sequences are perhaps of the highest significance (Figure 4*b*). the antisense RNAs produced from such promoters may block translation by base pairing with mRNAs or regulate their processing and stability. Such RNAs control expression of many plasmid and transposon genes, but they were not considered so far as typical of the bacterial genes (Wagner et al., 2002). All the genes bearing promoter-like signals were tested previously for the synthesis of antisense RNAs (Selinger et al., 2000), and in all but one cases such products were found. Although the data from the expression analysis cannot be considered as strong evidence, at least some of the found signals may be true positives. Assuming antisense transcription, Vogel et al. (2003), who analyzed short RNA products and detected 21 RNAs generated from the antisense strand, obtained other data. Eight of the RNAs may be produced as run-through transcripts from the neighboring genes, while appearance of the remaining 13 requires a different explanation, and antisense transcription may be the most evident. At least five of them may be produced from promoters predicted by PlatProm.

3.3 Potential promoters for alternative transcription

A genome-wide promoter screening unexpectedly detected 379 genes having promoters with a propensity to produce shortened RNA products from the sense strand (Figure 4*a*). At least 46 of the 275 previously detected short RNAs (Vogel et al., 2003) may be initiated at such promoters, thus supporting the suggestion that bacterial genes may have a parallel transcriptional output. The role of such RNAs remains vague. Some of them may encode alternative proteins, but the appearance of functional promoters in coding sequences may also have other reasons. Thus, additional promoters may be involved in polymerase trapping or may intensify the transcription of properly oriented downstream genes even if they are > 750 bp away from these promoters. On the other hand, the preferred location of promoter-like sites at the beginning of genes allows a speculation that coordinates of some genes require correction, here, our data may be used to pick them out.

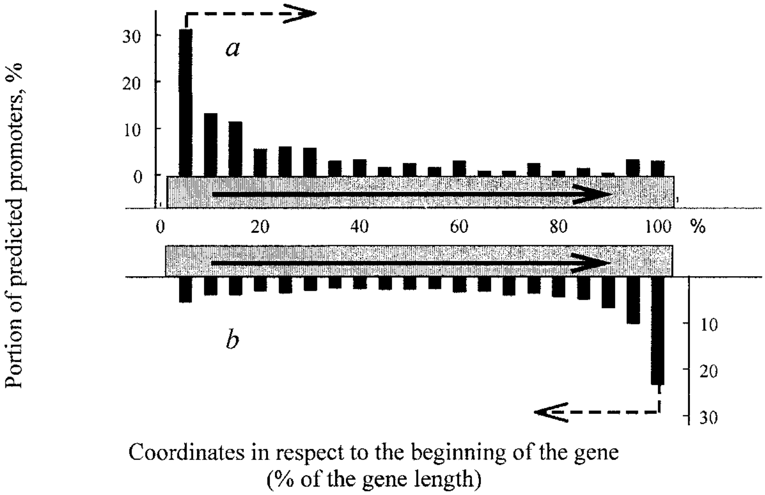

Figure -4. Relative positioning of the predicted promoters within the coding sequences: (*a*) potential promoters for alternative transcription and (*b*) promoters for antisense transcription.

In any case, PlatProm provides a unique opportunity of predicting independently transcribed regions within coding sequences.

ACKNOWLEDGMENTS

The work was supported by RFBR grant No. 03-04-48339 and RFBR–Ministry of Industry and Science (Moscow oblast) grant No. 04-04-97280.