Molecular Biology, Vol. 37, No. 5, 2003, pp. 707–715. Translated from Molekulyarnaya Biologiya, Vol. 37, No. 5, 2003, pp. 834–842.

BIOINFORMATICS

UDC 577.1

Search for Alternative RNA Secondary Structures Regulating Expression of Bacterial Genes

E. V. Lyubetskaya[1], L. A. Leont'ev[1], M. S. Gelfand[2], and V. A. Lyubetsky[1]

[1] *Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 101447 Russia; E-mail: lin@iitp.ru, lyubetsk@iitp.ru*

[2] *State Research Center GosNIIGenetika, Moscow, 113545 Russia*

Received February 18, 2003

Abstract—Expression of many bacterial genes is regulated by formation of alternative secondary RNA structure within the leader mRNA sequence. Our algorithm designed to search for these structures (basing on analysis of one nucleotide sequence) was applied to analyze operons of amino acid biosynthesis in alpha- and gamma-proteobacteria. The attenuators of these operons are predicted for genomes of some poorly known gamma-proteobacteria including *Shewanella putrefaciens*, attenuators of the tryptophan operon in some alpha-proteobacteria are also predicted.

Key words: attenuator structure, proteobacteria, expression regulation, search algorithm

PROBLEM STATEMENT AND SEARCH METHOD

Expression of many bacterial genes is regulated at the level of translation or by interaction of translation and transcription processes. In many cases alternative secondary RNA structures are formed in a certain RNA region, and these structures produce the main regulatory signal.

Examples are attenuators of the operons for amino acid synthesis [1] and regulatory structures of some operons for ribosomal proteins in *Escherichia coli* [2].

The standard approach to search for these regulatory signals is to generate secondary structures for new genomes using a set of 'patterns' (recognizing rules) based on the known similar structures from the well-studied genomes (see, for example, [1, 2]). However, it has a natural restriction: allowing one to extend the known regulation expression data on new genomes, it is of no use in studying new regulation systems. Among the numerous groups of bacteria, there are some without any well-known species. Moreover, it cannot be excluded that even the known genomes possess yet unknown regulatory systems.

For the above reasons, we have a problem: to search for regulatory RNA structures without using patterns of the known structures. In general, this problem probably cannot be solved, since the number of potential secondary structures is high and uncertain even for a short RNA fragment, and it remains unclear how to detect regulatory ones. However, we can consider the known fact that many regulatory interactions are based on formation of alternative structures. Using appropriate software, one can search for alternative RNA structures using the following procedure:

(1) Long enough segments upstream of the genes are scanned, and potential alternative secondary structures are detected.

(2) If a structure is 'good enough' (contains extended 'regular' helices) to exclude its random formation, it should be tested experimentally.

(3) If a predicted alternative structure is less reliable (i.e., the exact structures of alternative helices remain unclear, or the structure contains short or imperfect helices), then a possibility of forming *analogous* (in a special sense) alternative structures should be studied for the upstream segments of the same gene from the related genomes. Here it is assumed that the true secondary structures are conserved even if the nucleotide sequence is changed (this has proved true in many cases [1, 2]). The analysis is run first using the algorithm and then by hand, since we have no strict criterion for conservedness.

The biosynthesis of some amino acids and aminoacyl-tRNA synthetases in proteobacteria is regulated through a mechanism that alters the ratio of transcription and translation rates. It was shown experimentally for the operons *trp*, *his*, *ilvGMEDA*, *ilvBN*, *phe*, *thr*, *leu* from *E. coli* and *Salmonella typhimurium* [2]. This regulation is mediated by the ribosome.

Figure 1 presents the scheme of regulation. Two alternative conformations are shown for the secondary structure of the leader (noncoding) mRNA fragment located at the 5' end upstream of the genes coding for amino acid biosynthesis. Numbers show different

0026-8933/03/3705-0707$25.00

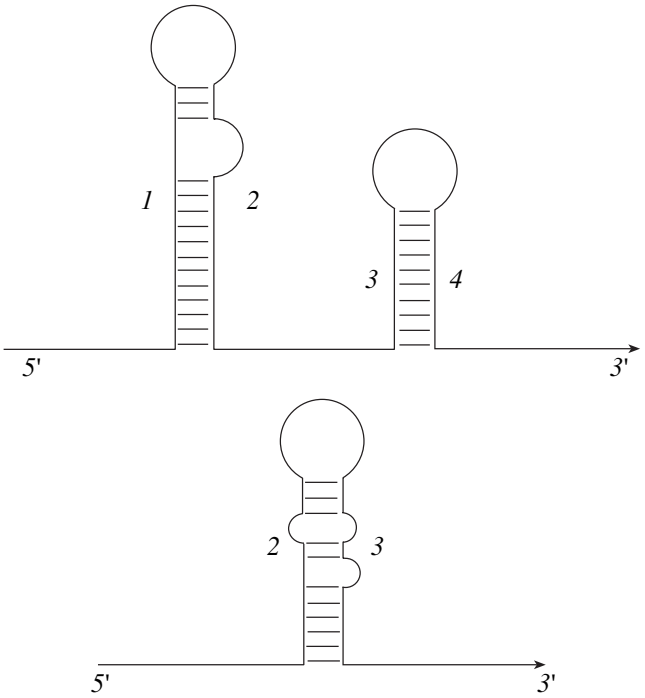


Fig. 1. Alternative secondary mRNA structure forming a typical attenuator of transcription.

mRNA sites involved in formation of the secondary structure. The first conformation (designated 1:2 and 3:4) is named *terminating*, and the second (designated 2:3) is named *antiterminating*.

If mRNA is folded into the antiterminating conformation, the RNA polymerase synthesizes the full-size transcript, while if mRNA is folded into terminating conformation, then transcription is stopped upstream of the structural operon genes. As seen from Fig. 1, the helix named *antiterminator* is formed by the "pausing" helix and *terminator*. The two secondary mRNA structures (terminating and antiterminating) are alternative to each other. Attenuators of operon transcription involved in biosynthesis of aromatic amino acids (*trp*, *pheA*, *pheST*) in some proteobacteria have been predicted using comparative analysis [1].

We developed the algorithm and the algorithm-implementing software [3], aiming to select for subsequent detailed manual analysis a set of the most similar secondary structures (and their helices) from pairwise comparison within the given set of regulatory regions from homologous operons in related genomes. Obviously, the criteria to decide about "good enough" (step 2) and "analogous" (step 3) helices are quite approximate, and therefore even the "best" and the "most conserved" alternative structure should be considered putative, even if optimally fitting the applied criteria. They should be verified either experimentally or using additional analysis, e.g., considering the presence of a sequence that encodes the signal peptide.

The essential peculiarity of the approach used in this work is that we tested a new algorithm developed by us to solve the problem formulated at step 1 basing on a *single* initial sequence. The set of related regulatory regions (see step 3) was used only for subsequent analysis and evaluation of the results; independent algorithms from [3] were used for this purpose.

This is the main feature distinguishing the suggested approach from earlier ones (see, e.g., [4–7]). These works either use comparison with a known pattern, or generate conserved secondary structure from analysis of compensatory substitutions in the aligned nucleotide sequences.

Standard definitions are the following. A secondary structure is composed of helices, and each helix is formed of two, left and right, ordered sets of the same number of nucleotide *stretches*. The neighboring stretches in each set are separated by spacers, either non-paired (named bulges), or paired with the second set (in this case the pair of spacers is named *internal loop*). The spacer between the two nucleotide stretches is named *external loop*. Each stretch *i* from the *start of the left set* should be complementary to the stretch *i* from the *end of the right set*. It is convenient to number these stretches in pairs, starting from the external loop. The reason for this numbering comes from the notion that the first (1–3 pairs of) stretches starting from the external loop are usually better detected by the algorithms than the subsequent "mix" of stretch pairs determined with lower precision.

Designate the ends of the left set as A and B, and the ends of the right set as C and D (nucleotides are always numbered in the direction from the beginning to the end of the initial sequence). The stretches from A to B and from C to D will be called *strands* of the helix (left strand and right strand, respectively). Consequently, a strand is a set of paired stretches with all bulges and internal loops (to be more exact, "halves" of the loops).

The algorithm [3], roughly quadratic to the size of initial data set, uses any nucleotide sequence as an input and produces a list of putative secondary structures within this sequence, alternative double-hairpin (terminating and antiterminating) or triple-hairpin (terminator-antiterminator-pausing). In the case of the double hairpin (with terminator A', B', C', D' and antiterminator A, B, C, D) the algorithm checks for some natural conditions, e.g., $B < A'$, $D < D'$ and certain minimal number of nucleotides for the stretches of the helices. Similarly, conditions of this type are checked in case of the triple-hairpin structure.

This work presents an example of analysis for regulation of expression for the genes involved in amino acid biosynthesis of operons *his*, *ilv*, *leu*, and *thr* in alpha- and gamma-proteobacteria. Our algorithm was applied to search for potential alternative mRNA structures.

Operons for amino acid biosynthesis in gamma-proteobacteria and predicted attenuators

Species	Operon, gene							
Escherichia coli	**trpEDCBA	**pheST	**pheA	**hisGDCBHAFI	**thrABC	**leuABCD	**ilvGMEDA	**ilvBN
Salmonella typhi	**trpED #trpCBA	**pheST	**pheA	*hisGDCBHAFI	*thrABC	*leuABCDxx	*ilvGMEDA	*ilvBN
Yersinia pestis	**trpEGDCBA	*pheST	**pheA1 **pheA2	*hisGDCBAHAFI	*thrABC	*leuABCD	*ilvGMEDA	*ilvBN
Vibrio cholerae	**trpEGDCBA	#pheST	**pheA	*hisGDCBHAFI	*thrABC	*leuABCDx	*ilvGMEDA	No
Haemophilus influenzae	#trpEGDC #trpBA	#pheST	#pheA	*hisGDCBHAFI	*thrABC	*leuABCD	No	#ilvBN
Shewanella putrefaciens	*trpEGDCBA	#pheST	*pheA	*hisGDCBHAFI	*thrABC	*leuABCD	*ilvGDA	No
Actinobacillus actinomycet-emcomitans	#trpEG #trpD #trpFC #trpxBA	#pheST	#pheA	#hisGDC #hisBH #hisF	*thrAB #thrCx	No	#ilvGE	#ilvBN
Pasteurella multocida	#trpEG #trpDC #trpBA	#pheSTxx	#pheA	*hisG #hisDCBxHAFxA	*thrABCxxxx	*leuABCD	*ilvGMxDA	No
Klebsiella pneumoniae	#trpEDC #trpB	*pheS pheT	*pheA	#hisD #hisBHA #hisFI	#thrABC	#leuA #leuC	*ilvGxEDA	*ilvBN
Pseudomonas aeruginosa	#trpE	#pheST	#pheA	#hisGDC #BH #hisAF	#thrA #thrC	#leuA #leuB #leuC #leuD	No	#ilvBN
Xanthomonas axonopodis	#trpExGxDCx #trpB #trpA	#pheST	#pheA	#xhisGDCBHAFI	*thrAB #thrC	No	*ilvCGMxleuA	No
Erwinia carotovora	*trpEGDCBA	*pheST	*pheA	*hisGDCBHA	*thrABC	*leuABCD	*ilvGED	#ilvBN
Xylella fastidiosa	#trpEGDC #trpBAx	#pheST	#pheA	#xhisGDCBHAFI	#thrABC	No	*xilvGAxleuA	No

* Putative new candidates with leader peptides detected as regulatory structures using our algorithm.

** Known structures with leader peptide detected by our algorithm and reported earlier [6, 7].

No prediction obtained with our algorithm for one of the following reasons: absence of the leader peptide for a good alternative structure, absence of a good alternative structure, absence of a putative terminator (polyT), too short region upstream of the first gene in the operon.

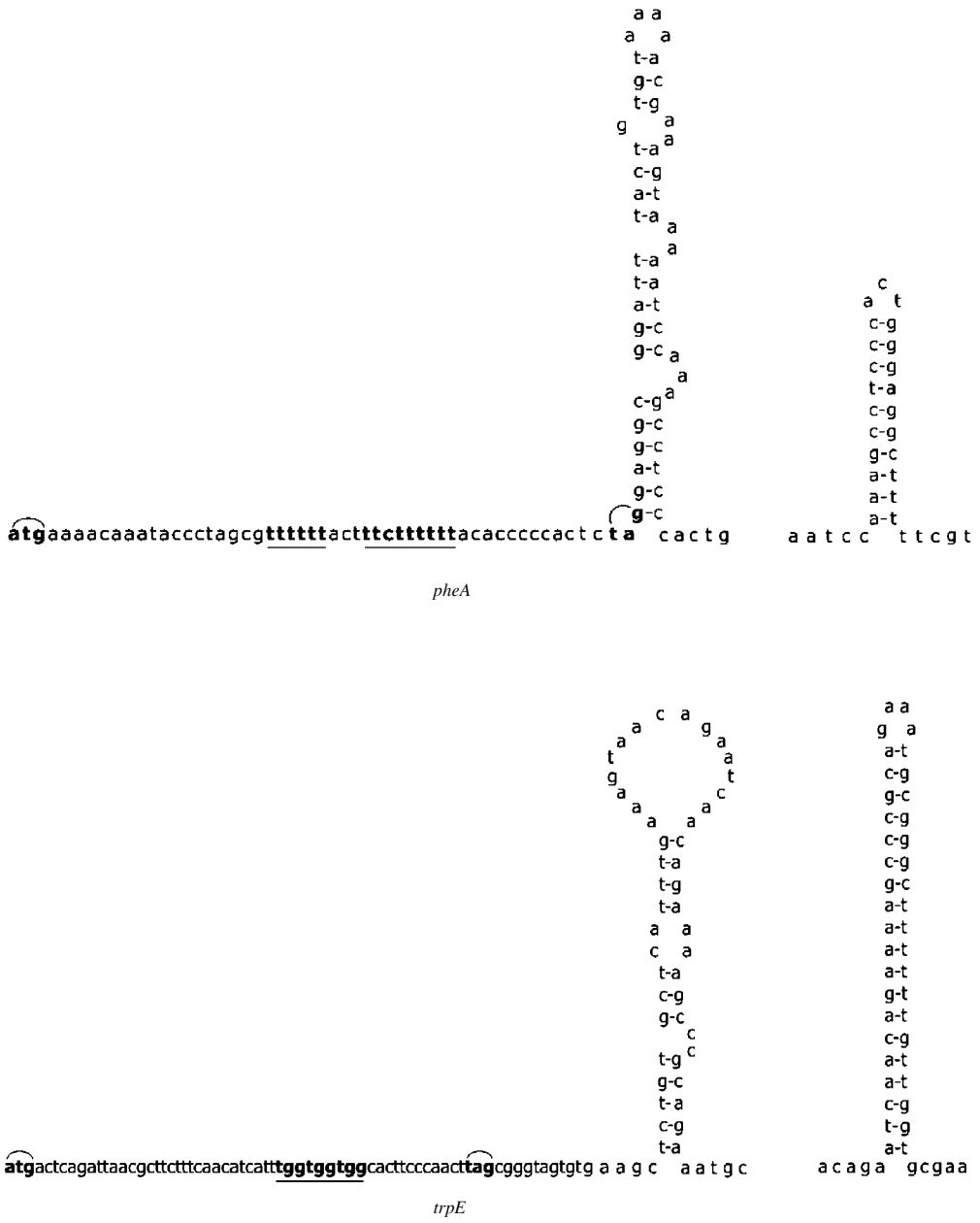
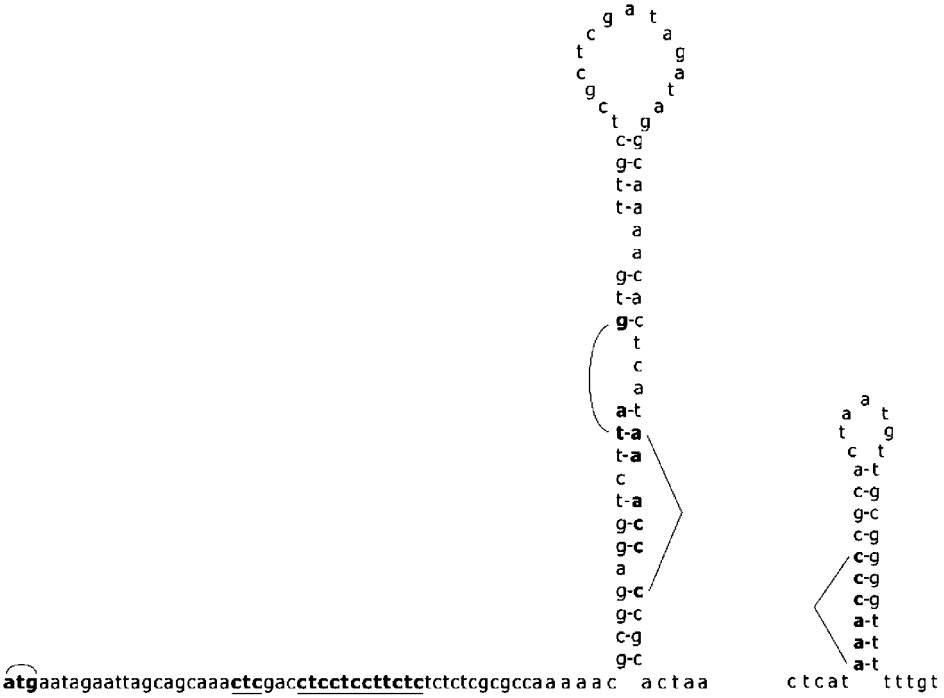


Fig. 2. Attenuators of transcription predicted for *Shewanella putrefaciens*.

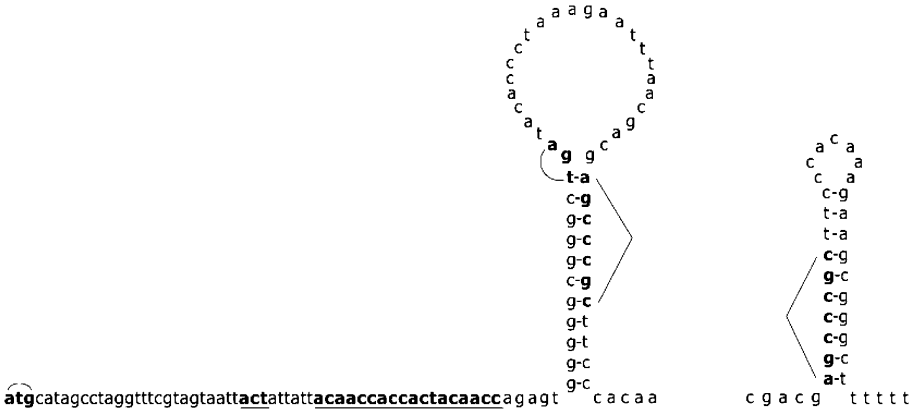
It should be noted, however, that the algorithm can be used to study some other types of secondary structure (not shown in this work). The algorithm uses some approaches suggested earlier [8]. Its detailed description is given elsewhere [9], and the executive program file is available at http://www.iitp.ru/lyubetsky; no description is given here. Importantly, 'parallel' analysis is possible, i.e., the algorithm can be applied using multiple processors with joined memory; this is essential for genome-wide studies. For this purpose, the algorithm is designed to scan the input sequence with a fixed window and to run calculations for all windows in parallel.

One more interesting feature of the algorithm should be noted. The search for optimal (in a given sense) secondary structures usually implies optimizing a numeric argument-dependent function, where the argument is a variable corresponding to the whole set of the secondary structures or at least their helices. It is known that for complex functions or broad argument variation this sort of optimization is not easy to

hisG



leuA



thrA

Fig. 2. (Contd.)

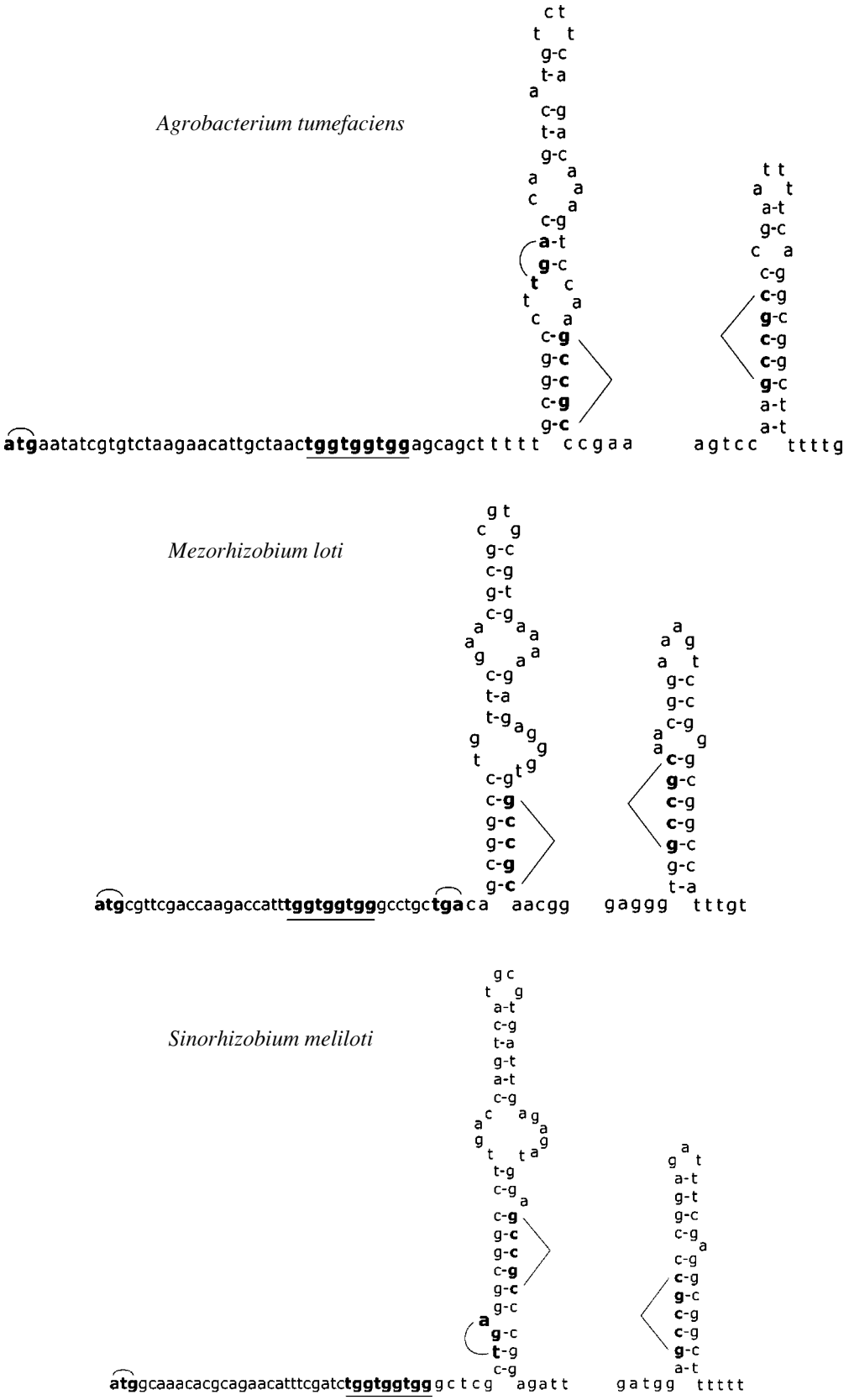


Fig. 3. Attenuators of the operons *trpEG* in alpha-proteobacteria.

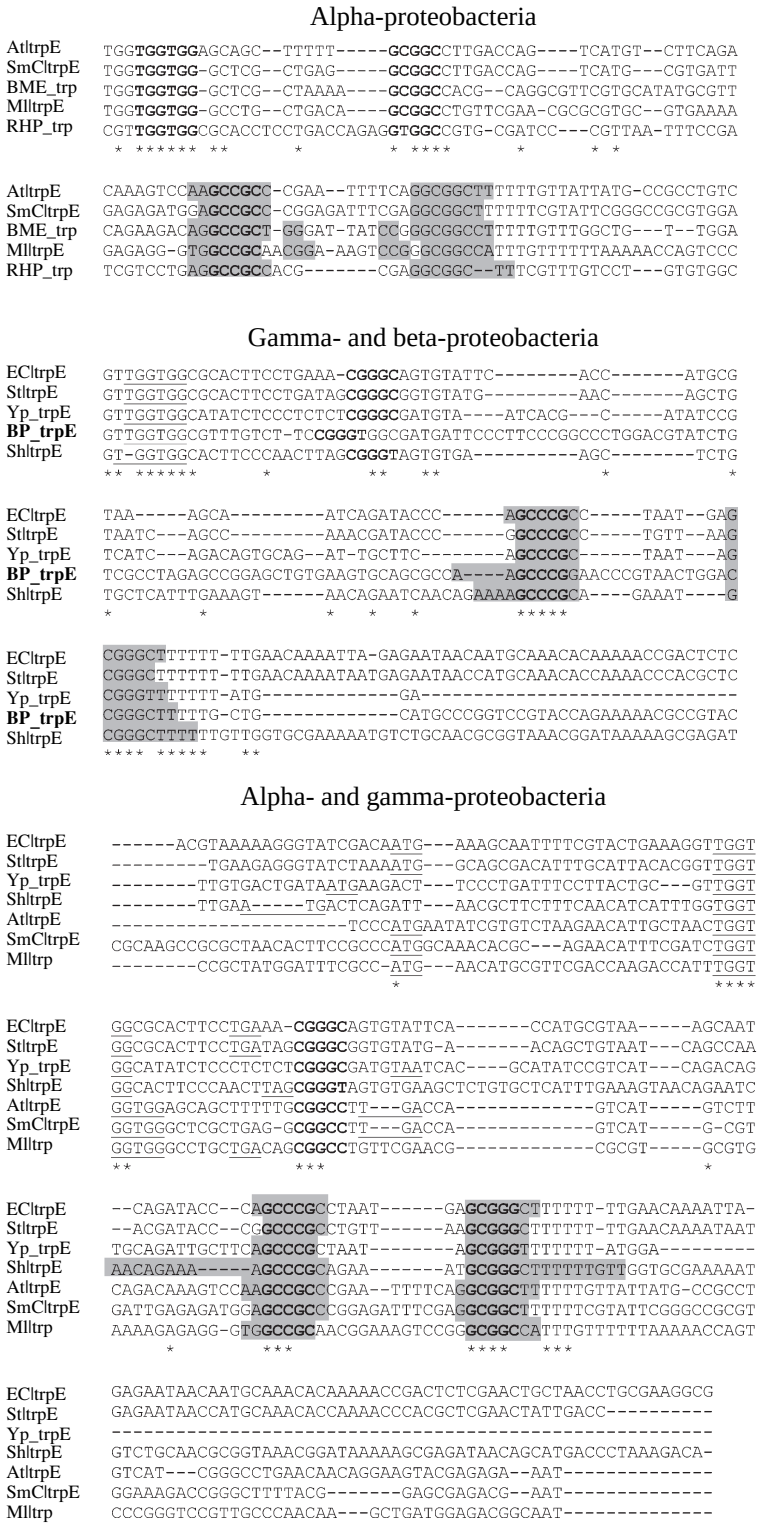


Fig. 4. Alignment of the regions upstream of *trpE* (the first gene of the tryptophan operon) in alpha-, beta-, and gamma-proteobacteria. Shown are regulatory loops, leader peptides with tryptophan codons, conserved sites within regulatory loops. The terminators are in gray boxes; regulatory, start and stop codons are underlined; conserved repeat included in both terminator and antiterminator is in bold. The following parameters were used for alignment: deletion opening 10.0, deletion elongation 0.05, transition weight 0.5. Abbreviations: At, *Agrobacterium tumefaciens*; SmC, *Sinorhizobium meliloti*; BME, *Brucella melitensis*; Ml, *Mesorhizobium loti*; RHP, *Rhodopseudomonas palustris*; EC, *Escherichia coli*; St, *Salmonella typhi*; Yp, *Yersinia pestis*; BP, *Bordetella pertussis*; Sh, *Shewanella putrefaciens*.

implement algorithmically. Essentially, the idea of our algorithm implies not optimization of the whole function, but rather a nonlinear ordering in a set of specially selected values for secondary structure patterns and elements. The search for the optimal ratio of the secondary structure parameters is easier and more reasonable than optimization of the numeric function.

RESULTS AND DISCUSSION

We analyzed genomes of the following gamma-proteobacteria: *E. coli*, *S. typhi*, *Klebsiella pneumoniae*, *Yersinia pestis*, *Vibrio cholerae*, *Haemophilus influenzae*, *Actinobacillus actinomycetemcomitans*, *Pasteurella multocida*, *S. putrefaciens*, *Pseudomonas aeruginosa* and of the following alpha-proteobecteria: *Agrobacterium tumefaciens*, *Sinorhizobium meliloti*, *Mezorhizobium loti*, *Rhodopseudomonas palustris*, *Caulobacter crescentus*, *Brucella melitensis*. Beta-proteobacterium *Bordetella pertussis* was also studied; we analyzed operons orthologous to the operons for amino acid synthesis in *E. coli trpEDCBA*, *pheA*, *pheST*, *hisGDCBHAFI*, *ilvBN*, *ilvGMEDA*, *leuABCD*, *thrABC*.

Search for orthologs was run using GenomeExplorer software [7]. After finding secondary structures with our algorithm, we aligned nucleotide regulatory regions using the Clustal program [10] in order to verify the results.

We found 121 operons in genomes of the studied gamma-proteobacteria. These results are shown in the table: the algorithm has found putative attenuator structures in 59 cases out of 121. We also can suppose that these structures really do not exist in part of the 62 cases when the algorithm found no triple-hairpin regulatory structure in the regulatory area; in this work these results are classified as negative.

The algorithm allowed us to find all known published attenuator structures in operons of the species listed in the table. Beside rather simple predictions for the genomes taxonomically similar to the known ones, we obtained some interesting new results, e.g., for the first time we predicted the existence of attenuators in the genome of *S. putrefaciens* for operons *ilvGMEDA*, *ilv*B, *trp*, *phe*A, *his* (Fig. 2). It should be noted that these regulatory regions cannot be aligned.

We also found potential attenuator regions for operons *trp*EG in alpha-proteobacteria *A. tumefaciens*, *S. meliloti*, *M. loti*, *R. palustris*, *B. melitensis*, *B. pertussis* (part of these results are shown in Fig. 3). These are first examples of attenuators found in genomes of alpha-proteobacteria (no earlier data found in published works). These regulatory regions can be aligned neither with each other, nor with similar regions from gamma-proteobacteria (Fig. 4). At the same time, the GC-rich complementary regions of terminator and antiterminator are conserved. Even the triplets of the respective words are similar (we designate them with numbers 2, 3, 4 from left to right): CGGGc–GCCCG–CGGGc in gamma-proteobacteria and GCGGC–GCCGC–GCGGC in alpha-proteobacteria (the words "2–3–4" are located at the end of the right strand in the pausing hairpin, and at the beginning and the end of the left terminator strand; they define pairing of the respective strands 2 and 3 or 3 and 4 in the attenuator structure). This points to the homology of attenuators, i.e., their common ancestor origin. Increased conservation in this region can be explained by the necessity of three (not two as in an ordinary helix) complementary substitutions to preserve both terminator and antiterminator helices.

Finally, we found an attenuator for operon *hisSxG* from *C. crescentus*. In all other cases our algorithm has shown motivated absence of attenuator regulation in proteobacteria.

In conclusion, testing has demonstrated the applicability of our algorithm in preliminary recognition of alternative RNA structures involved in regulation of gene expression. In future we plan to systematically analyze complete bacterial genomes, including those of poorly known taxonomic groups; to improve the algorithm, adding automatic analysis of leader peptides and other biologically essential structures; to consider the structures more complicated than the alternative helices; and to study evolution of attenuator structures.

ACKNOWLEDGMENTS

The authors are grateful to A.A. Mironov and A.G. Vitreschak for advice and helpful discussion of the results; to M.A. Shirshin for visualization of the results and for the help in programming and calculation; and to the anonymous reviewer who drew our attention to a recent work [11].

This work was supported by the Howard Hughes Medical Institute (grant 55000309).

REFERENCES

1. Vitreschak A.G., Gelfand M.S. 2000. Computer-assisted analysis of regulatory signals in bacterial genomes. Attenuators of the operons involved metabolism of aromatic amino acids. *Mol. Biol.*, **34**, 545–552.
2. Landick R., Yanofsky C. 1996. Transcription Attenuation. In *Escherichia coli and Salmonella. Cellular and Molecular Biology.* Ed. Neidhardt F.C. Washington DC: ASM Press, 1, Ch. **81**, 1263–1286.
3. Gorbunov K.Yu, Lyubetskaya E.V., Lyubetsky V.A. 2001. On two algorithms to search for alternative secondary RNA structure. *Informatsionnye Protsessy.* **1**, 178–187. (http://www.jip.ru/).

4. Gorodkin J., Stricklin S.L., Stormo G.D. 2001. Discovering common stem-loop motifs in unaligned RNA sequence. *Nucleic Acids Res*. **29**, 2035–2044.
5. Eddy S.R. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*. **3**, 1–16.
6. Waterman M.S. 1989. In *Mathematical Methods for DNA Sequences*. Ed. Waterman M.S. Boca Raton, FL: CRC Press, 185–224.
7. Mironov A.A., Vinokurova N.P., Gelfand M.S. 2000. Software for analysis of bacterial genomes. *Mol. Biol.* **34**, 253–262.
8. Vereshchagin N.K., Lyubetsky V.A. 2000. An algorithm to analyze secondary structure of RNA. *Trudy Nauch.-Issled. Semin. Logich. Tsentra IF RAN*. Moscow: Izd. RAN, **14**, 99–109.
9. Leont'ev L.A., Lyubetskaya E.V., Lyubetsky V.A. 2002. A modified algorithm to search for alternative RNA secondary structures and results of calculations. *Informatsionnye Protsessy.* **2**, 100–105 (http://www.jip.ru/).
10. Jeanmougin F., Thompson J.D., Gouy M., Higgins D.G., Gibson T.J. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci*. **23**, 403–405.
11. Lathe W., Suyama M., Bork P. 2002. Identification of attenuation and antitermination regulation in prokaryotes. *Genome Biology.* 3, preprint 0003. (http://genomebiology.com/2002/3/6/preprint/0003).
12. Vereshchagin N.K., Lyubetsky V.A. 2000. An algorithm to analyze secondary structure of RNA. *Trudy Nauch.-Issled. Semin. Logich. Tsentra IF RAN*. Moscow: Izd. RAN, **14**, 99–109.
13. Pesole G., Liuni S., D'Souza M. 2000. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*. **16**, 439–450.
14. Vitreschak A.G., Rodionov D.A., Mironov A.A, Gelfand M.S. 2002. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* **30**, 3141–3151.
15. Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. 2002. Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms. *J. Biol. Chem.* **277**, 48949–48959.
16. Vitreschak A.G., Bansal A.K., Titov I.I., Gelfand M.S. 1999. Computer-assisted analysis of regulatory signals in complete bacterial genomes. Initiation of translation in operons of the ribosomal proteins. *Biofizika.* **44**, 601–610.
17. Switzer R.L., Turner R.J., Lu Y. 1999. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic operon by transcriptional attenuation: control of gene expression by an mRNA-binding protein. *Progress Nucleic Acid Res. Mol. Biol*. **62**, 329–367.
18. Liu J., Turnbough C.L. Jr. 1994. Effects of transcriptional start site sequence and position on nucleotide-sensitive selection of alternative start sites at the pyrC promoter in *Escherichia coli*. *J. Bacteriol*. **176**, 2938–2945.
19. Rutberg B. 1997. Antitermination of transcription of catabolic operons. *Mol. Microbiol*. **23**, 413–421.
20. Perkins J.B., Pero J.G. 2001. In *Bacillus subtilis and its relatives: from genes to cells*. Eds. Sonenshein A.L., Hoch J.A., Losick R. Washington, DC: American Soc. Microbiol., 279–293.
21. Miranda-Rios J., Navarro M., Soberon M. 2001. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA*. **98**, 9736–9741.
22. Stasinopoulos S.J., Farr G.A., Bechhofer D.H. 1998. Bacillus subtilis tetA(L) gene expression: evidence for regulation by translational reinitiation. *Mol. Microbiol*. **30**, 923–932.
23. Keener J., Nomura M. 1996. Regulation of ribosome synthesis. In *Escherichia coli and Salmonella. Cellular and Molecular Biology.* Ed. Neidhardt F.C. Washington DC: ASM Press, 1, Ch. **90**, 1417–1431.
24. Decatur A., McMurry M.T., Kunkel B.N., Losick R. 1997. Translation of the mRNA for the sporulation gene *spoIIID* of *Bacillus subtilis* is dependent upon translation of a small upstream open reading frame. *J. Bacteriol*. **179**, 1324–1328.