Molecular Biology, Vol. 34, No. 4, 2000, pp. 461–467. Translated from Molekulyarnaya Biologiya, Vol. 34, No. 4, 2000, pp. 545–552.

UDC 577.1

Computer Analysis of Control Signals in Bacterial Genomes. Attenuators of Operons of Aromatic Amino Acids Metabolism

A. G. Vitreshchak[1] and M. S. Gel'fand[2]

[1] *Institute of Problems of Data Transmission, Russian Academy of Sciences, Moscow, 101447 Russia*
[2] *State Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 113545 Russia; E-mail: misha@imb.imb.ac.ru*

Received March 1, 2000

Abstract—Here we predict attenuators in operons of aromatic amino acids metabolism in the γ-subclass of proteobacteria *Salmonella typhi*, *Yersinia pestis*, *Vibrio cholerae*, *Haemophilus influenzae*, *Actinobacillus actinomycetemcomitans*, *Xanthomonas campestris*, and chlamydia *Chlamydia trachomatis*. Alternative secondary structures of mRNA governing transcription pathway selection as well as the related control leader peptide were constructed. Comparison with homologous *Escherichia coli* operons was used for the prediction. This control mechanism takes place in operons *trp* (tryptophan), *phe* (phenylalanine), and *pheST* coding for the large and small subunits of phenylalanine tRNA synthase in the considered γ-proteobacteria. Secondary structures of mRNA in the leader region of *tnaAB* operon possibly involved in ρ-dependent attenuation were predicted in certain enterobacteria.

Key words: computer analysis, functional signals, gene expression control, attenuation of transcription, aromatic amino acid metabolism, *tnaAB* operon

INTRODUCTION

Attenuation of transcription is an important control mechanism of gene expression widespread in various bacteria. Both nucleotide sequence and secondary mRNA structure are involved in the control process. Attenuation of transcription allows the cell to respond to changed concentration of the substances involved in vital metabolic cycles of the cell. For instance, in the case of operons of amino acid metabolism in γ-proteobacteria, attenuation provides for their response to changed amino acid concentration. The control structure in mRNA can force the transcribing RNA polymerase molecule either to pause during elongation, to prematurely terminate transcription, or to pass over potential terminator sequence. In addition, the RNA can include sites of binding to the control factors involved in the above events. Attenuation of transcription features existence of alternative secondary structures of mRNA and possible proceeding without the control protein.

MECHANISMS OF ATTENUATION

There are several mechanisms of activation of transcription in various bacteria: (1) a leader peptide provides for synchronization of transcription and translation. This mechanism works in operons of amino acid biosynthesis and tRNA synthase and is widespread in bacteria; (2) control protein binds secondary structure of mRNA. This mechanism is widespread in Gram-positive bacteria; (3) ρ-dependent attenuation (e.g., in the operon *tnaAB* of tryptophan utilization).

Let us consider attenuators of the first kind in detail at the example of *E. coli trp* operon [1]. As soon as mRNA region called pause hairpin is synthesized (1:2 at Fig. 1a), RNA polymerase pauses. The pause in mRNA synthesis is essential for motion synchronization of the RNA polymerase and the ribosome binding to the elongating mRNA. The ribosome binds to the RNA sequence of a short leader peptide containing several codons of a given amino acid (tryptophan codons in *trp* operon or phenylalanine codons in *phe* and *pheST* operons) and in the case of the amino acid deficit, ribosome makes a long pause on these codons. Spatial position of the ribosome obviates the pause hairpin 1:2 and, consequently, the secondary structure 2:3 called antiterminator is formed (Fig. 1a), which interferes formation of the terminator 3:4. In this case the RNA polymerase clearly passes the operon and synthesizes complete mRNA. In the case of sufficient concentration of the amino acid, the ribosome runs to the stop codon of the leader peptide and its spatial position obviates formation of the 2:3 antiterminator. This favors formation of the 3:4 terminator and transcription termination.

0026-8933/00/3404-0461$25.00 © 2000 MAIK "Nauka/Interperiodica"

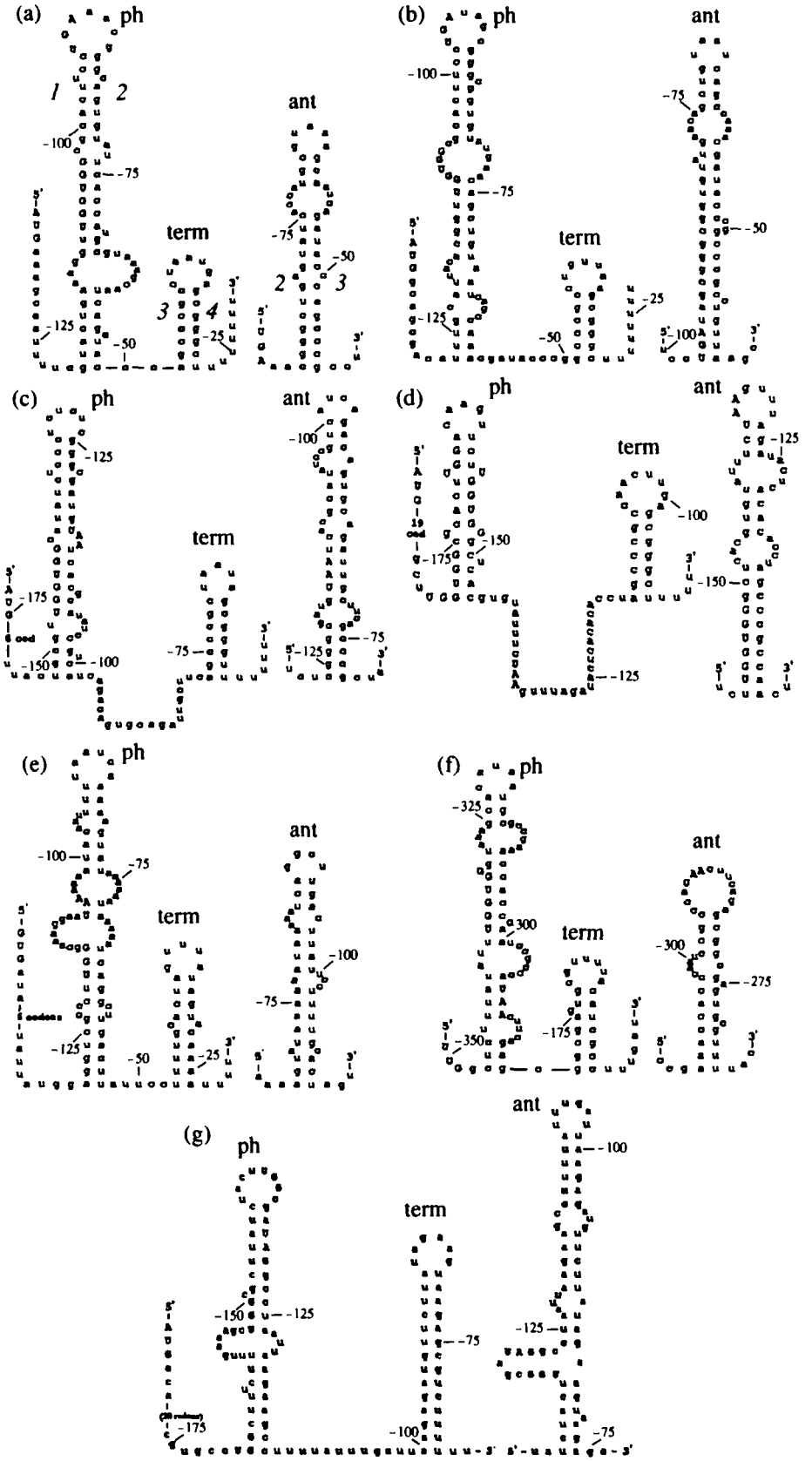

Fig. 1. Attenuators of *trp* operon; (a) *Escherichia coli*, (b) *Salmonella typhi*; (c) *Yersinia pestis*; (d) *Vibrio cholerae*; (e) *Haemophilus influenzae* (*trpEGDC*); (f) *Haemophilus influenzae* (*trpBA*); (g) *Chlamydia trachomatis*; ph, pause hairpin; term, terminator; ant, antiterminator; control tryptophan codons(UGG) as well as start and stop codons are given in uppercase.

Let us consider attenuation of the third kind at the example of *tnaAB* operon [1]. The leader peptide sequence is directly followed by the site of ρ-protein binding. In the case of excessive tryptophan, the ribosome runs to the stop codon of the leader peptide and prevents binding of ρ-protein to the mRNA, which allows RNA polymerase to synthesize complete mRNA. In the case of tryptophan deficit, the ribosome pauses at the tryptophan codons, ρ-protein binds the mRNA and induces premature termination of transcription. By contrast to the first kind attenuation, attenuation of the third kind is feedback-controlled (amino acid utilization). There are two hairpins [2] with possible control role in the leader peptide region of *E. coli tnaAB* operon, while ρ-protein has low affinity to mRNA regions with secondary structures.

COMPARATIVE APPROACH TO ATTENUATOR PREDICTION

First of the above mechanisms of transcription attenuation in *E. coli* was experimentally demonstrated for the operons of aromatic amino acid metabolism *trp* [1, 3], *phe* [3], and *pheST* [4], as well as for other ones: *leu*, *thr*, *ilvGMEDA*, and *ilvBN* [1]. However, analogous mechanisms in other bacteria including those with completely sequenced genome remain unknown. Here we try to predict transcription attenuators in other γ-proteobacteria *Salmonella typhi*, *Yersinia pestis*, *Vibrio cholerae*, *Haemophilus influenzae*, *Actinobacillus actinomycetemcomitans*, and *Xanthomonas campestris*, as well as in representative of another group *Chlamydia trachomatis*.

This work relies on comparative or phylogenetic approach to predicting secondary RNA structure. In the course of RNA evolution, the functionally significant elements of secondary RNA structure remain more conserved than the nucleotide sequence and the secondary RNA structure can be determined by aligning two or more sequences and the search for common conserved regions and helices, accounting mutual location of the helices, certain homologous bases, etc. This approach was initially applied by Levitt in 1969 to predict tRNA spatial structure [5] and was successfully applied to other molecules since then [6, 7]. However, comparative approach was applied to structural RNA molecules, while we extended it to control regions of mRNA. Previously we considered control of translation initiation of ribosomal proteins [8] while here we consider the signals responsible for attenuation of transcription. Note that the results obtained by comparative analysis often outperform those obtained by free energy minimization, since the latter problem is unstable in terms of calculation—even minor change in parameters (e.g., energy of hydrogen bond formation or stacking of neighboring base pairs) significantly rearranges the optimal secondary structure [9].

The analysis was carried out in the following way. The chains of genes orthologous to the genes of biosynthetic operons *trp* and *phe* as well as operon *pheST* were identified in all studied bacteria. The following order of genes was revealed. Operon *trp* in *E. coli*: *trpE*, *trpG/D*, *trpF/C*, *trpB*, and *trpA*. The order of the genes was the same in *S. typhi*, *Y. pestis*, and *V. cholerae*; apparently, these are the desired operons. In addition, attenuators of transcription were found in the leader region of these operons (see below). The case of *H. influenzae* and *A. actinomycetemcomitans* seems more interesting. The tryptophan operon was divided into two parts (*trpEGDC* and *trpBA*) and in the second one the genes of tryptophan synthesis are preceded by an apparent gene of alcohol dehydrogenase [10]. Note that the both formed operons have potential attenuators (see below). The order of genes in operons *phe* and *pheST* is the same in all bacteria.

Spatial mRNA structures responsible for transcription attenuation were predicted by comparison with the *E. coli* structures. Energy of certain hairpins at 37°C was estimated using parameters from [11].

As a result, positive signal structures for operons *trp*, *pheA*, *pheST*, and *tnaAB* were constructed.

RESULTS

Operon *trp* contains the genes of tryptophan synthesis. Transcription control of the operon includes attenuation mechanism. Excessive tryptophan induces premature termination of transcription (before the first functional codon), while the whole operon is expressed at deficient tryptophan. Alternative structures involved in attenuation as well as the leader peptide sequence with codons significant for the control were found experimentally in *E. coli*. We predict attenuation of transcription going by a similar pathway in *S. typhi*, *Y. pestis*, *V. cholerae*, and *H. influenzae* as well as in *C. trachomatis* (Figs. 1a–1g). The two latter cases are the most interesting. In *H. influenzae* genome *trp* operon is divided into two parts: *trpEGDC* and *trpBA* (Figs. 1e, 1f). Presumably, attenuation of transcription is specific for the both operons.

Chlamydia trachomatis is evolutionarily distant from proteobacteria. It only has *trpBA* operon and separate *trpC* gene. These genes appear in *C. trachomatis* as a result of a horizontal transfer from some enterobacterium [12]. The leader region of *trpBA* contains a potential attenuator (Fig. 1g); however, we failed to construct potential attenuators in *trpC*. Even in the very close *E. coli* and *S. typhi*, the attenuators have slightly different form and, apparently, rigid conservatism of the secondary structure is not required—alternative secondary structures of the RNA and position of control and stop codons of the leader peptide sequence relative to them has the key significance for the control.

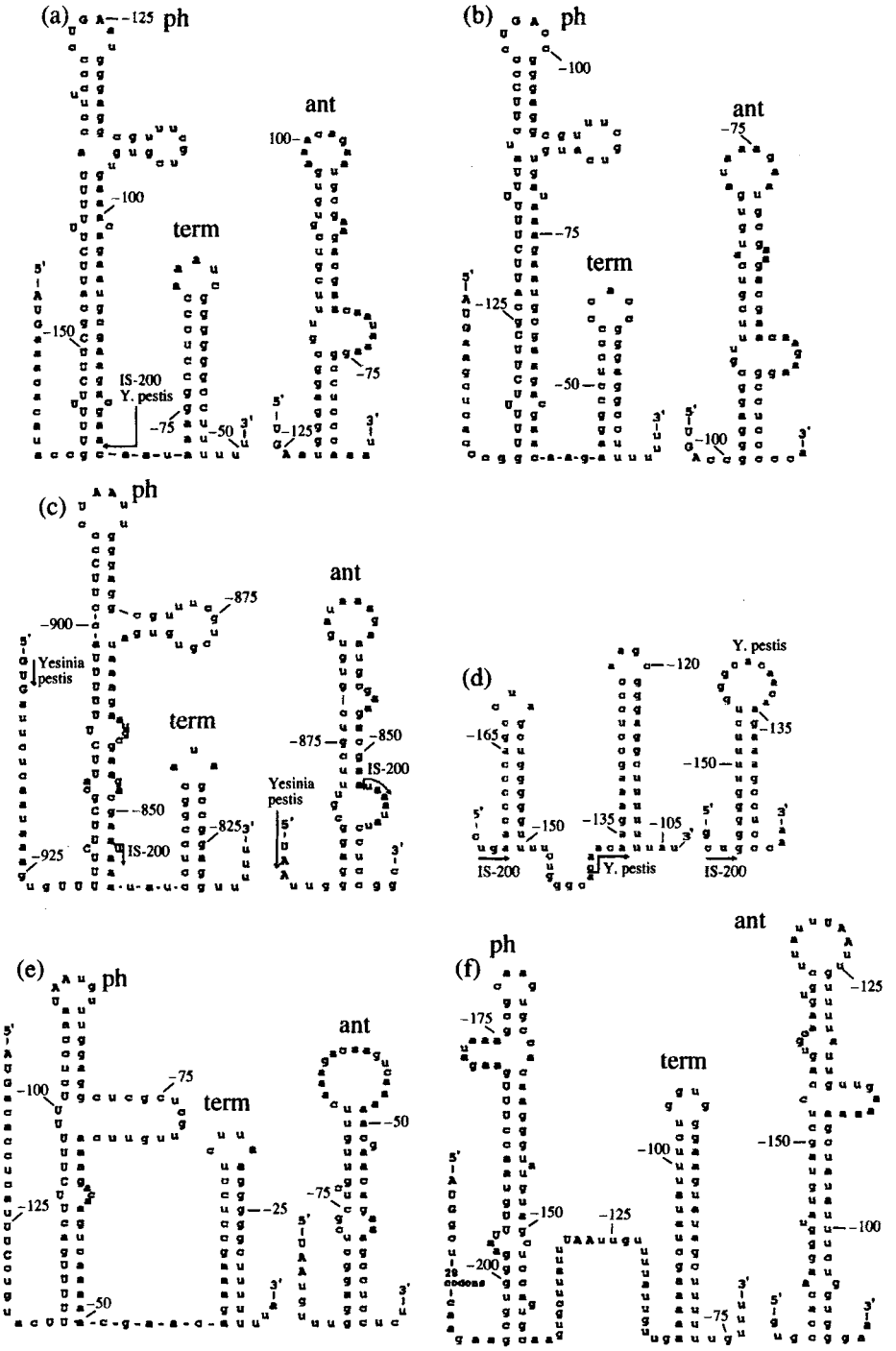

Fig. 2. Attenuators of *pheA* operon; (a) *Escherichia coli*, (b) *Salmonella typhi*; (c) *Yersinia pestis* (1st att); (d) *Yersinia pestis* (2nd att); (e) *Vibrio cholerae*; (f) *Haemophilus influenzae*; (g) *Actinobacillus actinomycetemcomitans*; (h) *Xanthomonas campestris*; ph, pause hairpin; term, terminator; ant, antiterminator; control phenylalanine codons (UUU and UUC) as well as start and stop codons of the leader peptide sequence are given in uppercase.

The hairpin energies are presented in the table. Difference of the hairpin free energy between organisms are exemplified by the *trp* operon. Apparently, the hairpin energy value is not significant, since most of them are stable enough and the principle of attenuation is satisfied. Most likely, the predicted signals for *S. typhi*, *Y. pestis*, *V. cholerae*, and *C. trachomatis* are attenuators. However, in the case of *H. influenzae* the hairpins of attenuators *trpEDGC* and *trpBA* are not stable enough, while the principle of attenuation is satisfied. The attenuator is active for a short time period during RNA polymerase movement along a short DNA fragment, so that the structure of attenuator does not necessarily have maximum energy (by

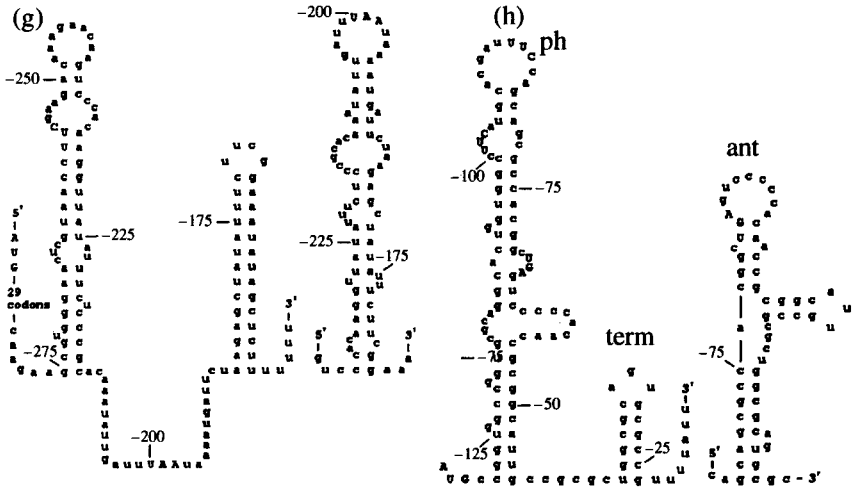

Fig. 2. (Contd.)

contrast to structural RNA such as tRNA). Relatively unstable structures can exist for a short time period and have regular significance; hence, it remains unclear if the signals revealed in *H. influenzae* operons *trpEDGC-* and *trpBA* are attenuators or artifacts. This problem can be resolved by comparative analysis when the genomes of other bacteria of the Pasteurellaceae family where *H. influenzae* belongs become available or by experimental verification.

Operon *phe*. Regulation of this operon also includes the attenuation mechanism. Excessive phenylalanine induces termination of transcription while the whole operon is expressed at deficiency of this amino acid. Operon *phe* is composed of a single gene *pheA* coding for chorismate mutase P and prephenate dehydrogenase. We predict a similar attenuation mechanism in other γ-proteobacteria: *S. typhi*, *Erwinia herbicola* (data not shown), *Y. pestis* (particular case), *V. cholerae*, *H. influenzae*, *A. actinomycetemcomitans*, and *X. campestris* (Fig. 2).

It is interesting to compare the organisms relatively distant from *E. coli* but close to each other: *H. influenzae* (Fig. 2f) and *A. actinomycetemcomitans* (Fig. 2g). In *E. coli* and *H. influenzae* the linker between the pause hairpin and terminator is several and 28 nucleotides long, respectively. Comparison of *H. influenzae* and *A. actinomycetemcomitans* demonstrated that it is conserved in the latter case as well.

A special situation is found for *Y. pestis*. An IS element (IS-200) in reverse orientation has integrated in the leader region of this operon. In particular, it has

Free energy of *E. coli* attenuator hairpins and predicted attenuators (kcal/mol)

Bacterium	Operon *trp*			Operon *phe*			Operon *pheST*		
	ph	term	ant	ph	term	ant	ph	term	ant
E. coli	−18.6	−11.4	−19.5	−25.5	−16.1	−21.2	−32.5	−10.0	−16.4
S. typhi	−10.0	−9.1	−15.9	−25.5	−16.4	−19.6	−23.8	−12.7	−19.6
Y. pestis	−15.2	−8.8	−13.9	−12.7	−10.7	−11.7	−26.8	−10.7	−12.5
				−13.3	−14.8	−12.8			
V. cholerae	−10.7	−11.4	−19.4	−14.8	−12.1	−11.7	–	–	–
H. influenzae	−4.0	−7.3	−4.8	−10.1	−12.1	−14.4	−22.2	−5.2	−6.0
	−5.8	−5.8	−7.5						
A. actinomycetemcomitans	–	–	–	−17.8	−13.1	−5.0	–	–	–
X. campestris	–	–	–	−12.6	−5.0	−16.0	–	–	–
C. trachomatis	−10.7	−12.0	−11.6	–	–	–	–	–	–

Note: *H. influenzae* operon *trp*: hairpin energies of attenuators *trpEGDC* and *trpBA* are given in the first and second lines, respectively; *Y. pestis* operon *phe*: hairpin energies of attenuators remote and close to the translation start are given in the first and second lines, respectively.

Designations: ph, pause hairpin; term, terminator; ant, antiterminator.

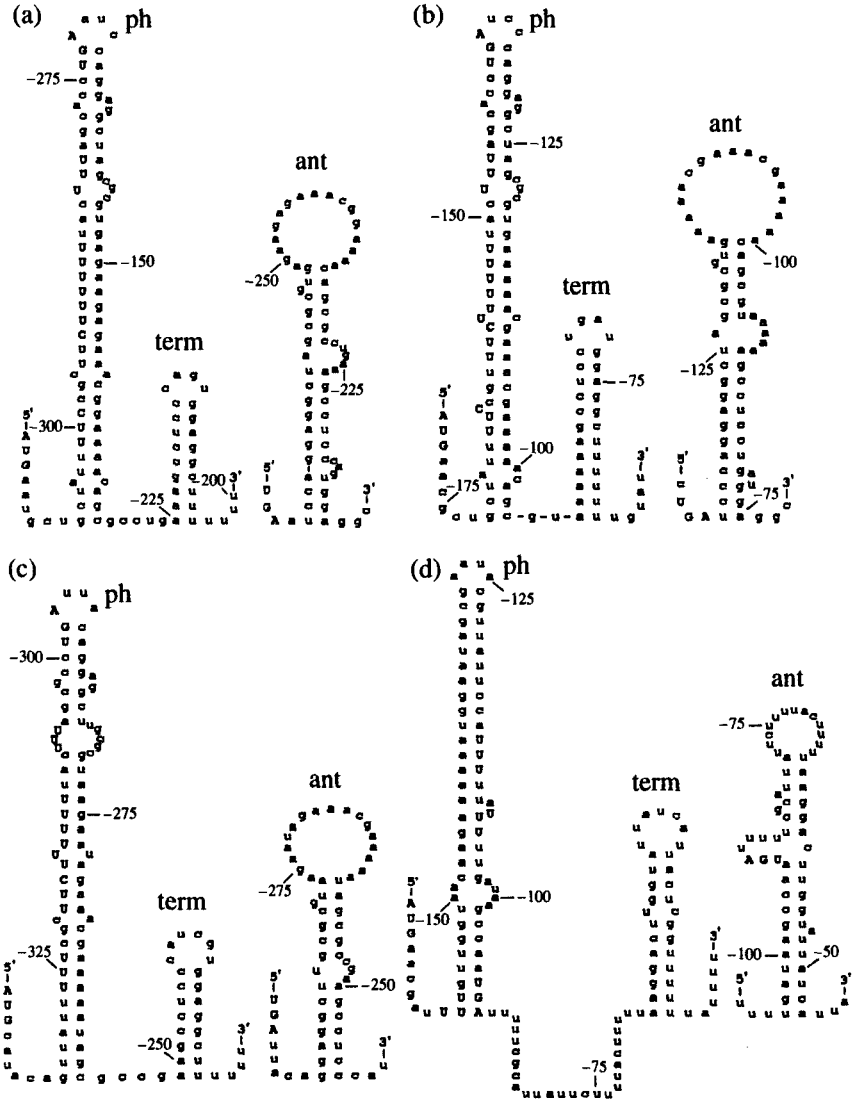

Fig. 3. Attenuators of *pheST* operon; (a) *Escherichia coli*, (b) *Salmonella typhi*; (c) *Yersinia pestis*; (d) *Haemophilus influenzae*; ph, pause hairpin; term, terminator; ant, antiterminator; control phenylalanine codons (UUU and UUC) as well as start and stop codons of the leader peptide sequence are given in uppercase.

intervened into the attenuator, splitting it so that one part contains the pause hairpin while the other one contains the terminator. IS-200 has terminal hairpins [13], and one of them is attached to the broken attenuator part with the pause hairpin as a terminator. Thus, an antiterminator is formed and the attenuator is preserved (Fig. 2c). However, the length of IS-200 is over 700 bp and the region of this attenuator is spaced by over 800 bp from the first functional gene; it remains unknown if the genes and attenuator are functional in this case. One can also compare the new attenuator and *Y. pestis* attenuator before IS-200 integration, which is very similar (just a few nucleotides substitutions) to *pheA* attenuator from *E. coli* (Fig. 2a). An arrow indicates the site of IS integration in *Y. pestis* attenuator. Another part of the split attenuator containing the terminator and another terminal hairpin of IS-200 fit together and form an antiterminator. Thus the secondary structures of attenuator are formed although without the leader peptide (Fig. 2d). It is unclear if it can be considered as an "emerging attenuator."

In addition, finding of potential attenuator in an organism relatively remote from *E. coli*—*X. campestris*—is of interest (Fig. 2h).

Operon *pheST*. Regulation pattern of this operon is similar to that of *pheA* operon—it responds to concentration of phenylalanine tRNA. Attenuators of transcription were constructed for the following organisms: *S. typhi*, *Y. pestis*, and *H. influenzae* (Fig. 3).

Operon *tnaAB*. Both secondary structures possibly controlling ρ-dependent attenuation in *E. coli*

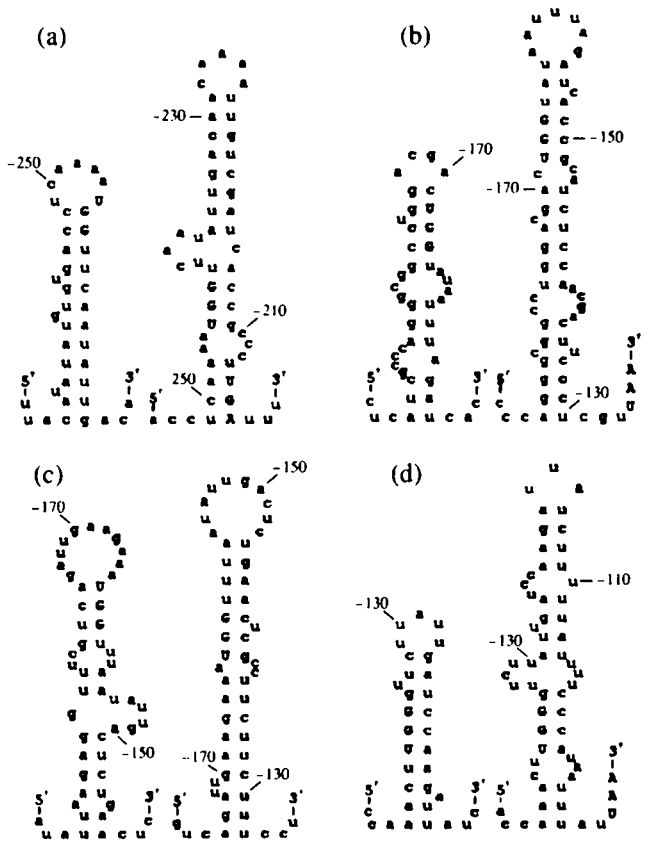

Fig. 4. Secondary structure in the control region of *tna* operon RNA; (a) *Escherichia coli*, (b) *Enterobacter aerogenes*; (c) *Proteus vulgaris*; (d) *Haemophilus influenzae*.

were also found in the following bacteria: *Enterobacter aerogenes*, *Proteus vulgaris*, and *H. influenzae* (Fig. 4).

CONCLUSION

On the basis of the obtained data we conclude that comparative analysis can be used to predict unknown control signals of genes from various bacteria with sequenced genome. In addition, note several facts following from the above considerations: due to low conservation of the secondary structures in *tnaAB* operon, a more accurate conclusion on their control significance can be drawn only provided a greater number of sequenced similar genomes. At the same time, the predicted control signals of *trp*, *phe* and *pheST* operons are more conserved and have all properties of attenuator. This allows us to propose control significance of the found signals. Note that after integration of IS element in the attenuator (*phe* operon in *Y. pestis*) the terminal hairpins of the element fitted the attenuator hairpins and, thus, new potential attenuators were formed. It is common knowledge that IS elements exhibit symmetry of various kinds during their integration, and their "jumps" can play a role in evolution of control signals ("hairpin translocation").

ACKNOWLEDGMENTS

Our thanks are due to V.A. Lyubetskii, A.A. Mironov, and E. Panina for valuable criticism. This work was supported in part by the program Human Genome (project no. 65-99), Russian Foundation for Basic Research (project no. 99-04-48247), Merck Genome Research Institute (project no. 244), and INTAS (project no. 99-1476).

REFERENCES

1. Landick, R., Yanofsky, C., *et al.*, *Escherichia coli and Salmonella. Cellular and Molecular Biology*, Neidhardt, F.C., Ed., Washington DC: ASM, 1996, vol. 1, ch. 81, pp. 1263–1286.
2. Gish, K. and Yanofsky, C., *J. Bacteriol.*, 1995, vol. 177, pp. 7245–7254.
3. Keller, E.B. and Calvo, J., *Proc. Natl. Acad. Sci. USA*, 1979, vol. 76, pp. 6186–6190.
4. Grunberg-Manago, M., *Escherichia coli and Salmonella. Cellular and Molecular Biology*, Neidhardt, F.C., Ed., Washington DC: ASM, 1996, vol. 1, ch. 91, pp. 1432–1457.
5. Levitt, M., *Nature*, 1969, vol. 224, pp. 759–763.
6. Woese *et al.*, *Nucleic Acids Res.*, 1980, vol. 8, pp. 2275–2293.
7. Moazed, D., Stern, S., and Noller, H., *J. Mol. Biol.*, 1986, vol. 187, pp. 399–416.
8. Vitreshchak, A. and Gelfand, M.S., *Biofizika*, 1999, vol. 44, no. 4, pp. 601–610.
9. Waterman, M.S., *Matematicheskie metody dlya analiza posledovatel'nostei DNK* (Mathematical Methods for DNA Sequence Analysis), Moscow: Mir, 1999, chapters 7, 8.
10. Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S., *Nucleic Acids Res.*, 1999, vol. 27, pp. 2981–2989.
11. Zuker, M. and Stienger, P., *Nucleic Acids Res.*, 1981, vol. 9, pp. 133–148.
12. Stephens, R.S. *et al.*, *Science*, 1998, vol. 282, pp. 754–759.
13. Mahillon, J. and Chandler, M., *Microbiology and Molecular Biology Reviews*, 1998, vol. 62, pp. 725–774.