Electronic version of an article published as JBCB, V.5, Issue 1, 2007, 155-180 (10.1142/S0219720007002576)

MODELING CLASSIC ATTENUATION REGULATION OF GENE EXPRESSION IN BACTERIA

VASSILY A. LYUBETSKY[1)]

SERGEY A. PIROGOV

LEV I. RUBANOV

ALEXANDER V. SELIVERSTOV[2)]

Institute for Information Transmission Problems RAS
Moscow, 127994, Russia
[1)]*lyubetsk@iitp.ru*, [2)]*slvstv@iitp.ru*

A model is proposed primarily for the classical RNA attenuation regulation of gene expression through premature transcription termination. The model is based on the concept of the RNA secondary structure *macrostate* within the regulatory region between the ribosome and RNA-polymerase, on hypothetical equation describing *deceleration* of RNA-polymerase by a macrostate and on views of transcription and translation initiation and elongation, under different values of the *four* basic model parameters which were varied. A special effort was made to select adequate model parameters. We first discuss kinetics of RNA folding and define the concept of the macrostate as a specific parentheses structure used to construct a conventional set of hairpins. The originally developed software that realizes the proposed model offers functionality to fully model RNA secondary folding kinetics. Its performance is compared to that of a public server described in [1]. We then describe the delay in RNA-polymerase *shifting* to the next base or its premature *termination* caused by an RNA secondary structure or, herefrom, a macrostate. In this description, essential concepts are the basic and excited states of the polymerase first introduced in [2]: the polymerase shifting to the next base can occur only in the basic state, and its detachment from DNA strand –only in excited state. As to the authors' knowledge, such a model incorporating the above mentioned attenuation characteristics is not published elsewhere. The model was implemented in an application with command line interface for running in batch mode in Windows and Linux environments, as well as a public web server [3]. The model was tested with a conventional Monte-Carlo procedure. In these simulations, the *estimate* of correlation between the premature transcription termination *probability p* and *concentration c* of charged amino acyl-tRNA was obtained as function $p(c)$ for many regulatory regions in many bacterial genomes, as well as for local mutations in these regions.

Keywords: Attenuation; transcription regulation model; mathematic modeling.

1. Introduction

The important role of RNA secondary structures in gene expression regulation is widely acknowledged. Its mechanism is based on affecting transcription elongation, translation delay, and involves various mediators, e.g. the ribosome in case of classic attenuation or regulatory proteins, tRNAs and co-factors in other attenuation types [4, 5, 6, 7, 8, 9]. Detailed evidence on attenuation is available for gamma-proteobacteria and Firmicutes [10, 11, 12, 13, 14, 15]. Regulations of novel type on the level of transcription and translation were proposed to involve T-boxes [16, 12], recently discovered riboswitches [17, 18, 7, 8], hypothetical regulatory LEU-element [19] as well as other elements, e.g. in chloroplasts [20]. Some alternative regulatory structures were sought using extensive

datamining [21, 22]. These studies advance functional annotation of hypothetical genes and contribute to filling the gaps in our understanding of bacterial metabolism [14, 17, 23, 24]. Attenuation regulation first gained attention in classic studies by C. Yanofsky and co-authors (ref. [25]). The historical record of attenuation research is briefly outlined in [4, 27, 9, 11].

Bioinformatic studies in this field include systematic efforts to find regulations of known and novel types using comparative genomic tools and a few attempts to model the regulatory mechanisms, the latter mainly focused on modeling RNA secondary structure kinetics. A dedicated web server was launched recently [1]. Our software that also implements kinetics modeling along with other functions is described in [27], and also is available as web server [3].

A host of approaches exists to model secondary folding kinetics, those are briefly reviewed, e.g., in introduction to [1]. Pioneering research on kinetics was published in [29-31] and formed the ideological basis for future work. Most studies employed Monte-Carlo technique to model kinetics of RNA secondary folding at the microstate level. Noteworthy, the adequate level to describe attenuation model is as yet not decided (each atom, each complementary pair, microstates or macrostates, i.e. clusters of microstates, and some others). In [32, 33] Monte-Carlo probabilistic modeling is applied to study pseudoknot formation in RNA secondary structure. A model of secondary structure folding kinetics was proposed based on original fast Monte-Carlo implementation developed to prevent the previously encountered states from being repeatedly sampled by the Markov chain. In [34] antitermination probability is estimated with explicit equation as a sum of two items: first, probability of the ribosome being at a regulatory codon when the antiterminator is formed as the polymerase reaches a *U*-rich region, and, second, probability of the ribosome leaving the stop codon when the antiterminator is not yet formed, multiplied by 0.5. The coefficient of 0.5 is introduced to account for mutually exclusive formation of either terminator or antiterminator.

Among other influential works in the field, in [35, 36] RNA folding kinetics is modeled with the approach proposed by A. Mironov; in [37] transformation rates are provided for certain minor specialized RNAs; in [38, 39] stochastic processes of RNA secondary structure formation are discussed with a conclusion that most effective approach to kinetics modeling is based on symmetric case of transition rate constant between secondary structures primarily introduced in [40] (here ref. Eq. (6)). This case is being called in [38, 39] Kawasaki rate.

The model proposed in this work differs in an attempt to describe secondary folding dynamics under conditions close to biological reality, where definitions of the primary sequence *region* involved in secondary structure folding, and the *time* of secondary structure dynamics are not imposed arbitrarily but are determined by shift events of the ribosome or polymerase. This has an important implication, as the secondary structure at the region between the ribosome and polymerase that determines the outcome (termination or antitermination) may exist in *non-equilibrium* state over a very short time of secondary structure dynamics within the current region. In other words, biologically plausible secondary structure is determined by the *formation dynamics* of the primary

sequence region between the ribosome and polymerase. When modeling only secondary folding kinetics without ribosome and polymerase shifts, in less than $5 \cdot 10^4$ secondary structure transitions our implementation of the model always computes an *equilibrium* structure quite close to one of those found by server [41] and program RNAstructure [44]. On average, it takes 18 sec for up to 120 bases-long sequences.

The model proposed in present study describes primarily classic RNA attenuation of gene expression regulation through premature transcription termination mechanism as described in [42], p. 172–189.

Our model, like the ones previously published, is susceptible to making arbitrary decisions on decomposing the entire process into elementary parts, on choosing mathematical tools to describe the parts, on defining parameters and their values, on means to juxtapose results of modeling with yet sparse experimental studies, etc. A perspective to cope with these uncertainties is seen in discussion and comparison of the models in the context of experimental evidence. There is hope that interpreting the models will also guide future experimental work. An earlier description of this model was published in [28] and mainly provided details of algorithmic realization and statistic properties of modeling results, such as average number of microstates in a macrostate, average helix length, helix number, etc.

2. Description of the Model

2.1. Definition of micro- and macrostates. Transition rate constant between mRNA secondary structure micro- and macrostates

Let the following be given: a sequence in four-letter alphabet $\{A, C, U, G\}$, a biological regulatory region in bacterial genome or a mutation of such region or random sequence. For instance, let it be a region from the promoter (when it is occasionally known) or from the ribosome-binding site before the leader peptide up to the end of *U*-run.

In the initial sequence, segments at least three bases-long are defined, the *stems* (otherwise, *shoulders*) of putative helices: … a_i, …, b_j, …. Pairing of any segments a_i and b_j with same length produces *helix* γ_s (formation of hydrogen bonds and stacking between the stems' bases is implied). The helix is always assumed to be *complete*, i.e. stems a_i and b_j are extended with complementary bases as far as possible, and the region spanning the segments is at least three bases-long (the helix *terminal loop*). Generally, the model allows for any set of primary helices. In the above example the set contains all helices that are both complete and imposed certain stem and loop constraints.

The details and concepts are described in e.g. [42], p. 172–189, including classic RNA attenuation triggered by charged tRNA, the latter being in turn dependent on concentrations of the amino acid and amino acyl-tRNA synthetase.

A *hypohelix* of helix γ_i is defined as any continuous region $\overline{\gamma}_i$ of γ_i consisting of two paired stems at least three bases-long. *Stems* are defined as paired segments of a hypohelix of the helix, and their termini are *designated A*, *B*, *C*, *D* starting from the 5'-end of primary sequence. A terminal loop is defined as an RNA region intercalating two stems of the hypohelix.

Microstate is a non-empty set of hypohelices, which are complete ***in the set***, lack pseudoknots, and are not contiguous (i.e., *A* and *D* of one hypohelix are not neighbors of *B* and *C* of the other hypohelix from same helix). An individual initial microstate is empty set $\varnothing$. *Completeness* ***in the set*** means that stems of constituent hypohelices can not be extended in this set. A *pseudoknot* is a pair of hypohelices with a stem of one hypohelix overlapping with the terminal loop of the other, thus being contained in this loop. All helices in the primary sequence are numbered in a fixed order, and in a microstate each hypohelix is tagged with the number of complete helix it is derived from. We also applied our model without the microstate completeness requirement and obtained somehow different results, which will be published elsewhere.

A microstate *diagram* is an ordinary parentheses structure describing the hypohelices localization in the microstate, with each closed pair of parentheses corresponding to a *hypohelix* and *tagged* with the number of its parental *helix*. The parentheses structure can be translated as follows: consecutive hypohelices correspond to consecutive closed pairs of parentheses, $(\,)_1(\,)_2...$; overlapping of first hypohelix with the terminal loop of second hypohelix is represented by embedded structure $((...)_1)_2$, where the inner parentheses denote the first hypohelix, and the outer ones – second hypohelix. Numbers of individual helices can represent multiple entries in the diagram, because several hypohelices can be derived from the same helix. A microstate, i.e. a set of all paired bases, uniquely defines its diagram. However, the diagram cannot be used to reconstruct the microstate, as it preserves only the "geometry" of hypohelices localization and information on which helix is allowed to provide a hypohelix for each pair of parentheses.

Any set of helices $\gamma_1,...,\gamma_k$ can produce a variety of its *realizing microstates*: any set of subhelices **complete *in the set*** $\overline{\gamma}_1 \subseteq \gamma_1,...,\overline{\gamma}_k \subseteq \gamma_k$ (each helix γ_i produces only one non-empty and not necessarily connected region $\overline{\gamma}_i$ with specific constraints on stems) without pseudoknots. As previously, *contiguous* hypohelices (i.e. having *A* and *D* bases neighboring with *B* и *C*) are merged.

Macrostate is any *non-empty* diagram ("non-empty" stands for it having at least one realizing microstate). For any microstate ω of macrostate Ω, diagrams ω and Ω are identical.

Bond energy $E_{\overline{\gamma}_j}$ of hypohelix $\overline{\gamma}_i$ is the sum of stacking bond energies of its adjacent base pairs. Special provisions are made to account for stacking of the first and the last pairs of hypohelix $\overline{\gamma}_i$ and coaxial stacking of $\overline{\gamma}_i$, which depends on microstate ω containing $\overline{\gamma}_i$. The energy is computed using the approach and numerical values published elsewhere [43-45, 41].

Each hypohelix $\overline{\gamma}_i$ from given microstate ω is assigned number l_i of nucleotides in its terminal loop that are *not contained in loops and stems of other hypohelices from* ω. This number is dependant on the microstate ω and defines *free length* l_i of the terminal loop of hypohelix $\overline{\gamma}_i$ in ω.

Microstate ω is by definition described with two free energies, *bond energy* and *loop energy* of ω. From here on, we considered only normalized energies obtained by dividing their values with $R \cdot T$, where $R = 0.001984\,(\text{kcal} \cdot \text{degree}^{-1} \cdot \text{mol}^{-1})$ and T equals e.g. $310K$. Therefore, in all our estimates the energy values are dimensionless.

Bond energy of microstate ω is estimated as follows:

$$G_{hel}(\omega) = \frac{1}{RT} \cdot \sum_i E_{\overline{\gamma}_i}, \quad (1)$$

where i varies over all hypohelices $\overline{\gamma}_i$ from ω, [43-45, 41].

Loop energy of microstate ω is estimated as follows:

$$G_{loop}(\omega) = \sum_i \left(1.77 \cdot \ln(l_i + 1) + B + \frac{C}{l_i} \right), \quad (2)$$

where i varies over all hypohelices $\overline{\gamma}_i$ from ω; Eq. (2) closely resembles that accepted in [29, 30]. Equation (2) is in good agreement with comprehensive tabulated data from [44, 41] for all loop energies under $l_i > 2$, assuming $B = 6.5$ for *terminal loops*, $B = 0$ for *double-strand bulges* (also known as internal loops) and $B = 4$ for *single-strand bulges*. Coefficient 1.77 (Flory parameter) is derived from the non self-intersecting random walk theory [46]. Results provided in Section 4 were obtained under $C = 5$; the case of $C = 0$ was also considered. Cases where $l_i \le 2$ are tackled separately according to the tables from [44, 41]: loop energies are set 0.8 (under $l = 2$) for a double-strand bulge and 6.2 (under $l = 1$) or 4.5 (under $l = 2$) for a single-strand bulge. Terminal loops of such lengths are excluded from our model. Although Eq. (2) is part of the Edgeworth expansion series, currently available experimental evidence does not seem to suffice for estimating its leading coefficients.

There exist "fast" and "slow" transitions between microstates. A *fast* transition by definition does not imply changes in corresponding microstate diagram. A *slow* transition by definition alters the microstate diagram by one pair of parentheses. Generally, in any transition the arbitrary set of hypohelices can be changed.

Absolute probabilities of fast transitions between microstates ω and ω' *from same macrostate* Ω are inessential in our model, instead, it operates with *rates*, see Eqs. (4-7). The critical assumption is that transitions in the set of all microstates ω from current macrostate Ω reproduce the Boltzmann-Gibbs stationary probability distribution:

$$p(\omega) = \frac{\exp\left(-(G_{loop}(\omega) + G_{hel}(\omega))\right)}{z(\Omega)}, \text{ where } z(\Omega) = \sum_{\omega \in \Omega} \exp\left(-G_{loop}(\omega) - G_{hel}(\omega)\right). \quad (3)$$

A slow transition from current microstate ω to any microstate ω' always occurs with altering macrostate ω by *one* pair of parentheses. There are two cases described with the equations below. In the model, the rate of any slow transition is described by Eqs. (4, 5) ("*asymmetric* case") [4, 27] under the assumption that hypohelix decomposition rate depends only on its bond energy and the hypohelix binding rate, the only spatial factor affecting the approach of the ends of its stems. Otherwise, an alternative Eq. (6) ("*symmetric* case") is used to describe the slow transition.

Note that equations, including (4-6), and tabulated values of all model parameters are defined symbolically in the program code, and thus can be easily changed in future implementations.

Thus, under slow transition in case of hypohelix decomposition current macrostate alters by one pair of parentheses, i.e. a transition occurs from microstate $\omega = \{\overline{\gamma}_{1i}, ..., \overline{\gamma}_{ki}\}$ (with all its hypohelices) to microstate $\omega' = \{\overline{\gamma}'_{1i}, ..., \overline{\gamma}'_{ki}\}$ (with all its hypohelices, where $\overline{\gamma}'_{li} = \varnothing$ under some chosen l, i, i.e. hypohelix $\overline{\gamma}'_{li}$ is virtually absent from ω'), hypohelices $\overline{\gamma}_{li}, \overline{\gamma}'_{li}$ are derived from same helix γ_l and correspond to one pair of parentheses, the *slow transition rate* is described with the following equation ("***asymmetric case***"):

$$K(\omega \to \omega') = \kappa \cdot \exp\left(G_{hel}(\omega) - G_{hel}(\omega')\right). \quad (4)$$

Under slow transition, binding of hypohelix, macrostate increases by one pair of parentheses and in the same notation is described as follows ("***asymmetric case***"):

$$K(\omega' \to \omega) = \kappa \cdot \exp\left(G_{loop}(\omega') - G_{loop}(\omega)\right). \quad (5)$$

Alternatively to Eqs. (4-5), any of the two slow transitions are described as the "***symmetric case***":

$$K(\omega \to \omega') = \kappa \cdot \exp\left\{\frac{1}{2}\left[\left(G_{loop}(\omega) + G_{hel}(\omega)\right) - \left(G_{loop}(\omega') + G_{hel}(\omega')\right)\right]\right\}. \quad (6)$$

In published studies $\kappa = 10^6$ s^{-1}; in [1] it is set to 10^7 by default. Here the value of κ was varied in the range from 31 s^{-1} to 10^6 s^{-1}. An equation similar to (6) was used in [39] to model RNA secondary structure folding. This equation was first introduced in [40] in relation to the kinetic Ising model. Eqs. (4-6) were derived to fulfill the detailed equilibrium requirement:

$$\frac{K(\omega \to \omega')}{K(\omega' \to \omega)} = \exp\left[E(\omega) - E(\omega')\right],$$

where $E(\omega)$ is energy of microstate ω under both approaches. Particularly, this accounts for coefficient $1/2$ in Eq. (6).

If macrostate dynamics is now to be described using dynamics of its realizing microstates, only two transitions will be possible: adding new hypohelix γ to current macrostate Ω and disappearance of initially present γ from Ω. Trivial averaging over all pairs of microstates $\omega \in \Omega, \omega' \in \Omega'$ produces the following equation for the *transition rate between macrostates* Ω and Ω' that applies to both the increase and decrease of a macrostate by one hypohelix:

$$K(\Omega \to \Omega') = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega'} p(\omega) \cdot K(\omega \to \omega'). \quad (7)$$

An effective original computer implementation of the described model is developed, i.e. an algorithm to compute the above sums without enumerating all microstate pairs, and is published in [47-48]. The program typically takes less than one minute to run in batch mode with default parameter settings to output all termination probabilities $p(c)$ under concentration values c ranging from 0 to 0.5 with step 0.05 and 100 independent

reiterations of the Monte-Carlo procedure for each c. It also logs down statistics and details of the run. Detailed representation of model trajectories can also be obtained to analyze secondary structure folding kinetics.

Original definitions of fast and slow transitions find their mathematical support from combinatorics in the below proposition, with its nontrivial proof published in [47-48].

Proposition 1. Let two microstates that realize the same macrostate be given. Then a chain of transitions from one microstate to the other is possible, such that each microstate of the chain belongs to the macrostate, through a number of steps, each realizing no more than two openings and two formations of base pairs. And otherwise: if two microstates originate from different macrostates, any such chain of transitions between them is not possible.

2.2. Shifting of RNA-polymerase on DNA strand

A *hairpin* is defined here for our purposes as a chain of paired stems linearly nested in each other's terminal loops with *minor* bulges between contiguous stem pairs and an *arbitrary* terminal loop at the end of the chain; the first stem pair forms the *handle* of the hairpin. Each stem pair in the hairpin, i.e. a hypohelix, has a terminal loop comprising all such subsequent pairs, their terminal loops and bulges. A hairpin might not represent a microstate if the former does not meet the completeness requirement. Here the hairpin definition is used in a narrower sense than traditionally accepted.

Nucleotide z is *U-rich* if exists at least one word containing z at any position and exceeding a certain threshold in length (5 is default) and in relative frequent occurrence of character '*U*' (0.8 is default). Other letters are allowed at any position in the word including the position of z. For a set of all *U*-rich nucleotides, all intervals of maximal length are constructed; those define *U-rich regions*. Default parameter values are taken from experimental evidence [49].

Let us *denote* z a position in the given sequence, which coincides with the active center of the polymerase and where transcription takes place. If $z = n$, three scenarios are possible: shifting of the polymerase on $(n+1)$-nucleotide, premature termination at n-nucleotide or staying at n-nucleotide.

We will describe rate constants ν and μ for the first two scenarios. Given the assumption from [2] that the polymerase exists in two states, either basic or excited, at any nucleotide, let us denote probabilities of these states β and $(1-\beta)$, respectively. The polymerase can shift to the next nucleotide with rate constant λ_{pol} *only* in its basic state, and it can slip off DNA strand with rate constant λ_{ur} at a *U*-rich nucleotide *only* in its excited state (premature termination event). Only in excited state the polymerase binds with one of the hairpins. The default value $\lambda_{pol} = 40\ \mathrm{s}^{-1}$ is similar to that published in [50]; λ_{pol} describes the polymerase-DNA interaction. λ_{ur} is 0 at a non-*U*-rich nucleotide, and is $10\ \mathrm{s}^{-1}$ otherwise, the latter value is provided in [2]; λ_{ur} describes interactions facilitating release of RNA strand from the polymerase under certain conditions. The value of λ_{ur} is estimated in our model as being close to $10\ \mathrm{s}^{-1}$; corresponding calculations are provided below.

When transitions between the basic and excited states are fast, it was proved in [51] that transition rate constants between the states can be substituted by average values, probabilities β and $(1-\beta)$ of the states, to express transition rate constants of the shift and slippage of the polymerase in Eq. (8). Thus,

$$\nu = \beta \cdot \lambda_{pol} \quad \text{and} \quad \mu = (1-\beta) \cdot \lambda_{ur} . \tag{8}$$

In our assumption, β is determined *only* by the secondary structure (here, by macrostate Ω) of the region between the polymerase and ribosome, or, without ribosome, the region stretching from the upstream of polymerase up to the transcription start:

$$\nu(\Omega) = \beta(\Omega) \cdot \lambda_{pol} = \lambda_{pol} - F(\Omega) = \lambda_{pol} \cdot (1 - \frac{F(\Omega)}{\lambda_{pol}}) , \tag{9}$$

where $F(\Omega)$ is a "force" which corresponds to effective decrease of the rate constant of polymerase shifting on DNA strand in inverse seconds, s^{-1}. The second multiplier in (9) is dimensionless quantity, the factor of the reduction of normative polymerase rate by macrostate Ω. Eq. (9) introduces function F without specifying its form and thus, unlike Eq. (8), *does not* represent a *new* hypothesis, except for the statement that ν is determined *only* by the secondary structure. This statement is in strong agreement with the experiment, and there is no evidence to expect it otherwise.

Under no assumption of fast transitions between the basic and excited states, a Boolean variable is to be introduced in the model to specify the current state of polymerase. In our study, introducing this variable did not have considerable effect on results (ref. Sec. 4).

Let us now express μ from Eqs. (8-9) explicitly:

$$\mu = \frac{\lambda_{ur}}{\lambda_{pol}} \cdot F(\Omega) \tag{10}$$

A crucial step is to determine the form of function $F(\Omega)$. More precisely, $F(\Omega)$ is defined for any macrostate Ω as mathematic expectation over all its realizing microstates $p = p(c)$:

$$F(\Omega) = \sum_{\omega \in \Omega} p(\omega) \cdot F(\omega) \tag{11}$$

which reduces the problem to determining the dependence for $F(\omega)$. We will define $F(\omega)$ assuming that ω is a hairpin and that ω is an arbitrary microstate, thus rendering definitions Eqs. (8-11) *complete*.

Consider the first case. Let ω be a *hairpin consisting of a handle without bulges and a negligibly small loop*, with handle length h (the number of its constituent complementary base pairs). Four experimental points are known from [52]: $\langle 7; 0.11 \rangle$, $\langle 8; 0.4 \rangle, \langle 9; 0.54 \rangle, \langle 12; 0.2 \rangle$, where the first entry in the pairs is handle length h and the last one is *premature termination probability* within a *U*-run of length N. We will now derive an equation to describe *premature termination probability* in the form $1 - P(N)$,

where $P(N)$ is probability of the polymerase reaching the end of N bases-long U-run without a slippage event. Besides, three other experimental points were given in lectures by R. Landic (2004): $\langle 3;0.05\rangle, \langle 7;0.91\rangle, \langle 14;0.3\rangle$ with the same notation of the entries.

We will use this evidence to determine the **form of dependency** for $F(\omega)$ and then to estimate **parameters** in $F(\omega)$. First, $P(N)$ is to be explicitly expressed. Probability of the polymerase shifting from n- to (n+1)-nucleotide without termination is obviously $\nu/(\nu+\mu)$ and is not unity only within a U-run and only in presence of hairpins, i.e. when $\mu > 0$ or, identically, $F > 0$. Hence the probability of making N shifts within a U-run without slippage is approximately

$$P(N) = \left(\frac{\nu}{\nu+\mu}\right)^N = \left(\frac{1}{1+\frac{\lambda_{ur}}{\lambda_{pol}}\cdot\frac{F}{\lambda_{pol}-F}}\right)^N . \qquad (12)$$

Probability $1-P(N)$ is in direct proportion to the value of F.

Only now the abovementioned sets of experimental evidence can be interpreted: as length h of hairpin handle grows, "force" F of polymerase deceleration by hairpin ω increases, reaches its maximum and then decreases under fixed distance r. The deceleration is seemingly nonsymmetric over this maximum at certain value h_0 of the handle length. A naïve comment can be made here: a positively charged region exists in the negatively charged polymerase molecule that might be in Coulomb interaction with the negatively charged hairpin. In particular, it might explain nonsymmetric (over h_0) form of function $F(h)$. Therefore, we assume

$$F(h,r) = \frac{\delta}{L_1^2\cdot\left(\frac{1}{h}-\frac{1}{h_0}\right)^2+1}\cdot\exp\left(-\frac{r}{r_0}\right) \qquad (13)$$

where parameters δ, L_1, h_0, r_0 depend on polymerase characteristics. Thus, the first multiplier $F(h)$ in Eq. (13) indicates that the dependency of $F(h)$ is of resonant type. The second multiplier $F(r)$ in Eq. (13) implies exponential decrease of value $F(r)$ of the polymerase deceleration by a hairpin with the decrease of distance r between them, which seems natural.

Hence, a non-trivial task in hypothesis Eq. (13) is to determine dependency between $F(h,r)$ and length h, i.e. to select function

$$F(h) = \frac{\delta}{L_1^2\cdot(1/h-1/h_0)^2+1}$$

under fixed distance r. Under fixed r, $F(h,r)$ is denoted as $F(h)$. Function $F(h)$ is maximal at handle length $h = h_0$, converges to 0 at h converging to 0 and converges to constant $\delta\left(L_1^2/h_0^2+1\right)^{-1}$ at increasing h. Large values of h are not obtained in modeling. Dependency (13) was varied over a class of rational functions, therefore an

alternative form of dependency for $F(h)$, if exists, will have to be of completely different form. Hypothesis Eq. (13) accepted in this study is in agreement with the model predictions (ref. Sec. 4).

Unfortunately, Eq. (13) alone does not suffice for modeling. It is necessary to define the dependency between value F of the polymerase deceleration by a hairpin and *length l of terminal loop*, i.e. function $F(h,l)$. Besides, minor single- and double-stranded bulges occurring in the hairpin's handle are also to be accounted for. In modeling (ref. Sec. 3) bulges in the handle exceeding certain threshold in length (typically, 2) are precluded. Although hairpins with long loops in classic attenuation occur seldom in biological sequences, function $F(h,l)$ is to be formally defined in the model.

Mathematically rigor, albeit based on assumptions, generalization of $F(h)$ to $F(h,l)$ is provided in [47]. Let us now obtain $F(h,l)$ using heuristic approach. Dependency Eq. (13) can be rewritten as follows:

$$F(\omega) = \frac{\delta}{L_2^2 \cdot (p - p_0)^2 + 1} \cdot \exp\left(-\frac{r}{r_0}\right), \tag{14}$$

where $L_2 = 2L_1/\pi$ and $p = \pi/2h$, $p_0 = \pi/2h_0$. The generalization of $F(h)$ to $F(h,l)$ will be obtained by selecting function $p = p(h,l)$ to insert in Eq. (14). Dependency (14) will then be the unknown $F(h,l)$.

The sought dependency is of type

$$p(h,l,l'') = \frac{\pi}{2h + \theta_1 \cdot l + \theta_2 \cdot l''}, \tag{15}$$

where θ_1, θ_2 are certain parameters with negligible effect on modeling, hence we can assume $\theta_1 = \theta_2 = 1$, and l'' is *total length of all bulges* in the hairpin's handle. Assumption Eq. (15), i.e. dependency $p = \pi/(2h + l + l'')$, ***implies*** that loop and small bulges are considered as extensions of the handle, which seems natural. However, function $p(h,l,l'')$ in this form is not applicable, as under large l values in modeling the dependency from h becomes negligible. Therefore, under lack of bulges in the handle, i.e. under $l'' = 0$, we *define* function $p(h,l)$ with the equation commonly used in physics:

$$tg(p \cdot h) = \frac{2}{p \cdot (\beta \cdot l)},\ 0 < p \cdot h \le \frac{\pi}{2}. \tag{16}$$

Parameter β defines the effect of the hairpin's loop on its interaction with the polymerase. The meaning of β is not discussed here, and we assume $\beta = 1$. Heuristically, Eq. (16) is justified as follows. Let us express function $p(h,l)$ defined with Eq. (16) as a product of $\pi/2h$ and a power series of *reasonable* dimensionless parameter $l/2h$. It becomes obvious that the free term of the series under $l/2h = 0$ equals unity, i.e. then $p = \pi/2h$, which gives dependency Eq. (13), while to agree within a linear term under small $l/2h$, we obtain $p = \pi/(2h + l)$, which is what we expected to find under $l'' = 0$. This power series of $l/2h$ is easily expressed explicitly by rewriting (16) in the form of

$$\frac{l}{2h} = \frac{tg\left(\pi/2 - ph\right)}{ph},$$

expanding the second member of the equation into power series of $\pi/2 - ph$ and inverting this series. Comparison to $\pi/(2h+l)$ also requires expanding it into power series of $l/2h$.

Let us now **assume** presence of small bulges in the hairpin, i.e. revert to the **general hairpin definition** given in the Sec. 2.2. For hairpin ω with a set of paired segments separated by bulges and with a terminal loop let us define: if ω contains s segments of lengths $h_1,...,h_s$, and $s-1$ bulges of lengths $l_1,...,l_{s-1}$ and a loop of length l, then

$$p = \overline{p} \cdot \left(1 - \frac{l''}{2h+l}\right) \tag{17}$$

and $\overline{p}$ is found with Eq. (16), where $\overline{p}$ replaces p. In a similar heuristic justification of Eq. (17), the same values are obtained when $\overline{p}$ is approximated with $\overline{p} = \pi/(2h+l)$ in Eq. (17), and only linear terms are left in the expansion of $p = \pi/(2h+l+l'')$ into power series of *reasonable* dimensionless parameter $l''/(2h+l)$.

We also considered two cases similar to Eq. (17):

$$p = \overline{p} \cdot \left(1 - \frac{1}{2h + l \cdot \sin^2(\overline{p} \cdot h)} \cdot \sum_{i=1}^{s-1} l_i \cdot \sin^2\left(\overline{p} \cdot h(i)\right)\right) \tag{18}$$

and

$$p = \overline{p} \cdot \left(1 - \frac{1}{2h + l \cdot \cos^2(\overline{p} \cdot h)} \cdot \sum_{i=1}^{s-1} l_i \cdot \cos^2\left(\overline{p} \cdot h(i)\right)\right) \tag{19}$$

where *by definition* $h(i) = h_1 + ... + h_i$, $h = h(n) = h_1 + ... + h_n$. Because $0 < \overline{p} \cdot h < \pi/2$, factors in $\sin^2\left(\overline{p} \cdot h(i)\right)$ monotonously grow over all $h(i)$, while factors in $\cos^2\left(\overline{p} \cdot h(i)\right)$ decrease. Unlike Eq. (17), Eq. (18) accounts for a larger effect of a bulge nearby the loop under same total lengths $h(i)$, and its lower effect in Eq. (19).

In our modeling Eqs. (17-19) produce similar results due to absence in our data of any large bulges in the handle. It is assumed that $\delta < \overline{\lambda}_{pol}$, which gives $\nu(\Omega) > 0$ in (9).

Let us now consider the case of microstate ω. A diagram of microstate ω can be decomposed into a set of *elementary* diagrams characterized by presence of the *handle*, i.e. the outer pair of parentheses with assigned hypohelix. The initial diagram is uniquely defined as a linear succession of elementary diagrams from this set. In this set, the *i*-th elementary diagram corresponds to hypohelix γ_i with termini A_i and D_i assigned to the outer pair of parentheses. Hairpin ω'_i is *defined by* A_i, D_i and ω as follows: ω'_i starts with base pair $< A_i, D_i >$ and then expands into RNA region between A_i and D_i of the primary sequence with base pairings according to ω that preserves minor bulges until the first major bulge (default bulge size threshold is 2) or a fork is encountered in ω. Regions before the major bulge or fork form the *handle* of hairpin ω'_i, and those after – the *loop* of ω'_i.

Let us now define strength $F(\omega)$ of the effect of any microstate ω on RNA-polymerase by reducing it to the defined set of hairpins $\{\omega_i'\}$, the *root of microstate* ω. We considered two cases of defining $F(\omega)$:

$$F(\omega) = \sum_i F(\omega_i') \tag{20}$$

and

$$F(\omega) = \max_i \{F(\omega_i')\}\,. \tag{21}$$

In Eq. (20) values r_i are defined as above indicated; the sum contains an exponent rapidly decreasing with growth of distance r_i, which permits application of ordinary exponentially damped weight summing. In Eq. (21) r_i is another value called *straight distance* that is defined without accounting for all hairpins between the i-hairpin and polymerase, i.e. only two terminal nucleotides, A_j and D_j, are considered in all *j*-th hairpins occurring between the fixed *i*-th hairpin and polymerase.

Each leader region was analyzed using both Eq. (20) and Eq. (21). The results suggest that both approaches produce similar functions of premature termination probability in response to charged tRNA concentration. This similarity is accounted for by the fact that in all analyses of biological data one of the items in Eq. (20) is much larger and corresponds to the largest hairpin in set $\{\omega_i'\}$ of those closer to polymerase. But this item is difficult to determine formally.

2.3. Premature termination of RNA-polymerase on DNA strand and parameter values in function F(h)

Now consider the scenario of polymerase *premature termination* at a residue *within a U-rich region*. Polymerase with $z = n$ is found on *U*-rich nucleotide n with probability β in basic state and with probability $(1-\beta)$ in excited state. Termination of the polymerase with rate constant λ_{ur} is possible *only* in excited state. According to Eqs. (8-9), the probability is

$$\beta = 1 - \frac{F(\Omega)}{\lambda_{pol}} = 1 - \frac{\delta / \lambda_{pol}}{L_2^{\,2}(p - p_0)^2 + 1} \tag{22}$$

and the rate constant of premature termination is

$$\mu = \frac{\lambda_{ur} \cdot F(\Omega)}{\lambda_{pol}}. \tag{23}$$

Let us now obtain parameters δ, L_1 and h_0 in the expression for $F(h)$ in Eq. (13), i.e. for $F(h,r)$ under fixed r, using two different sets of experimental evidence. After inserting the four already mentioned data points in Eqs. (12, 13) under N=2 and $r = 0$ from [52], we obtain a set of four nonlinear equations with three unknowns, δ, L_1 and h_0. Solving this set under $\lambda_{ur} = 10\ \mathrm{s}^{-1}$ from [2] gives $\delta = 27, L_1 = 45, h_0 = 9.1$. Solving

similar set of equations using the data points from R. Landic under N=7 and $r = 0$ gives $\delta = 25, L_1 = 22.8, h_0 = 7$. The obtained values are of the same order of magnitude.

Importantly, using independent experimental data from [52] will provide very similar estimates of δ, L_1, h_0 in $F(h)$. In [52] four data points are given, $\langle 7; 0.11 \rangle, \langle 8; 0.32 \rangle$, $\langle 9; 0.35 \rangle, \langle 12; 0.09 \rangle$, where, in each pair, the first term is handle length h and second – probability q of premature termination, at the same fixed nucleotide and a fixed terminal loop. Then using

$$1 - q = \frac{\nu}{\nu + \mu} \quad \text{and} \quad \frac{\mu}{\nu} = \frac{\lambda_{ur}}{\lambda_{pol}} \cdot \frac{F(h)}{\lambda_{pol} - F(h)}, \; r = 0,$$

we obtain $\delta = 30, L_1 = 42.5, h_0 = 8.6$, $p_0 = 0.1826$. There is high agreement between the experimentally derived value of q and that predicted in our model using the latter parameter estimates. It is not trivial because exact approximation of four points by function $F(h, \delta, L_1, h_0)$ with only three parameters δ, L_1, h_0 is not always possible. Solving it with so high accuracy (ref. Table 1) substantiates our choice of function $F(h)$.

Table 1. Probability *q* of termination on a fixed nucleotide in correlation with the handle length in our model and the experiment [52].

Handle length *h*	*q* in our model	*q* in the experiment
7	0.11	0.11
8	0.33	0.32
9	0.39	0.35
12	0.08	0.09

Let us now define parameter λ_{ur}. From [2], $\lambda_{ur} = 10\ \text{s}^{-1}$; we will estimate λ_{ur} in our model using the given $\delta = 30, L_1 = 42.5, h_0 = 8.6$, $p_0 = 0.1826$ and the experimentally obtained ratio of the polymerase delay time at the eighth nucleotide in slippage event to the same time in shifting event being 4 in macrostate Ω of a hairpin with the handle and loop lengths 11 and 6, respectively [2]. By estimating force $F_0 = F(\Omega)$ we obtain $F_0 = 15.5$. Average time $1/\mu$ before slippage divided by average time $1/\nu$ before shifting is

$$4 = \frac{\nu}{\mu} = \frac{\beta}{1-\beta} \cdot \frac{\lambda_{pol}}{\lambda_{ur}}, \text{ then } 4 \cdot \frac{\lambda_{ur}}{\lambda_{pol}} = \frac{\lambda_{pol}}{F_0} - 1$$

and $\lambda_{ur} = 15.8$, which gives values of the same order of magnitude.

2.4. Ribosome sliding on mRNA strand

At *non-regulatory* codons rate constant λ_{rib} of ribosome elongation on one codon is assumed to be standard $\overline{\lambda}_{rib} = 15\ \text{s}^{-1}$, i.e. $\overline{\lambda}_{rib} = 45\ \text{s}^{-1}$ per nucleotide. On regulatory codons we assume it to depend on concentration c of charged amino acyl-tRNA, according to the Michaelis-Menten law:

$$\lambda_{rib}(c) = \frac{\overline{\lambda}_{rib} \cdot c}{c_0 + c}. \qquad (24)$$

Here c_0 is concentration of amino acyl-tRNA, at which the ribosome slides on regulatory codons at a rate two times less than maximal $\overline{\lambda}_{rib} = 45\ s^{-1}$; $\overline{\lambda}_{rib}$ is its value at concentration c that is high enough to provide for the ribosome sliding rate on regulatory codons being as high as on non-regulatory codons. From here on, concentration is expressed in c/c_0 units, hence we avoid the need to define c_0. In other words, it is *assumed* that c_0=1.

2.5. Ribosome binding to the Shine-Dalgarno sequence

The process is initiated with the polymerase starting from the promoter, where protein-DNA regulation can occur. The ribosome then binds to the Shine-Dalgarno (SD) sequence and to the start codon of the leader peptide gene. After the SD sequence, the start codon of the leader peptide gene and next $s_0 + s_1$ nucleotides ($s_0 + s_1$ is the distance between P-site of the ribosome and transcription center z of the polymerase) are transcribed, the ribosome can bind to mRNA. The ribosome binding may be silenced by secondary structures shielding the SD sequence and start codon and, probably, by protein-DNA interaction in these domains. To model these effects, the *ribosome binding rate constant* K_0 was incorporated in the model in its simplest form:

$$K_0 = \lambda_0 \cdot \frac{d_{open}}{d_{\max}}, \qquad (25)$$

where d_{open} is maximal number of consecutive open nucleotides in the SD sequence (*provided* that the start codon is open), d_{max} is the length of SD sequence (normally, d_{max}=6) and λ_0 is translation initiation rate constant, the reciprocal to the time of ribosome binding under absence of silencing factors. When the ribosome binds, the secondary *downstream* structure is inherited in the model.

Because λ_0 is likely unknown, we modeled a case when the ribosome already covers the SD sequence and, thus, the leader peptide gene start codon. Here the required is not the λ_0 value but the initial instant position of the polymerase. Let *sp* be the distance from the 5'-edge of the leader peptide start codon to 5'-edge of the polymerase at the instance the ribosome has bound with mRNA. For a class of biological sequences, this binding occurs unimpeded immediately as the SD sequence, the leader peptide start codon and the next $s_0 = 12$ nucleotides (here the ribosome size being $s_0 = 12$ nucleotides) are transcribed. This is likely the case of bacterial operons considered below. For these, it was assumed that $sp = 13$ in modeling described in Tables 2-6, 8, 9. Widely varying parameter *sp* has a surprisingly low effect on the result of modeling (ref. Table 7).

3. Modeling of Premature Transcription Termination and its Parameter Values

With classic attenuation regulation, the purpose of modeling was estimation of $p = p(c)$, the correlation between termination probability p and concentration c of charged tRNA.

In equilibrium, $c = c_{aa} \cdot c_{aaS} \cdot d$, where c_{aa} is amino acid concentration, c_{aaS} is concentration of amino acyl-tRNA synthetase and d is a coefficient. One of the factors was varied, c_{aa} for amino acid biosynthesis operon or c_{aaS} for amino acyl-tRNA synthetase, while keeping the other one constant. To account for amino acid concentration in substrate, which was studied in [53], one can set $c_{aa} = c_{aaOut} \cdot d'$, where c_{aaOut} is amino acid concentration in substrate and d' is corresponding coefficient. **Each of the cases** produces values of c_0, d and d' affecting the c-axis *only*, thus not discriminating between the cases in simulation studies. The assumption of nonlinear dependencies greatly complicates calculations and is beyond the scope of this work.

We determined function $p = p(c)$ for many leader regions of various bacterial amino acid operons and amino acyl-tRNA synthetases. Function $p(c)$ was built with repeating the modeled process certain number of times (usually 10^3 -10^4) under some increment of c. Each run gives one of two possible outcomes: premature termination of the polymerase on a *U*-run within primary sequence or its successful passing of the *U*-run, therefore $p(c)$ was calculated as a *fraction* of times the termination occurred.

Apart from the secondary structure kinetic parameters that were fixed in Sec. 2.1 with their typical values, others were varied: r_0 within the range 0.1–5, $\delta = 30, L_2 = 27.1$, $p_0 = 0.1826$, $\kappa = 31 - 2000\ \mathrm{s}^{-1}$, $sp = 13 - 50$. The "size" of the ribosome from its P-site to its 3'-end is $s_0 = 12$ nt, and that of the polymerase, from its beginning y on RNA strand to the transcription center, $s_1 = 5$ nt. All modeling on biological and artificial sequences was done **under fixed parameter values**, unless explicitly indicated otherwise. Thus, in Tables 2, 4, 9 containing numerical values of $p(c)$, all the below mentioned sequences were modeled under $r_0 = 1$, $\kappa = 10^3\ \mathrm{s}^{-1}$, $sp = 13$. Concentration c of loaded tRNAs was varied from 0 to 0.5 with increment 0.05. In point 0.5 dependency $p(c)$ stabilized in all cases. The concentration is expressed in conventional units c / c_0; therefore interval $(c_{\min}, c_{\max})$ of concentrations where attenuation really occurs cannot be derived from our data. It might be surmised from the form of function $p(c)$: regulation occurs within the region of its monotonous increase. In this work we do not consider an important issue of transition from nondimensional concentration to its physical units and from nondimensional probability (frequency) $p(c)$ to physical units of enzyme activity, except in Table 3. The latter data can be used to tackle this issue but it is unfortunately very scarce.

Given a fixed primary sequence, the modeling process contains: the *window* between 3'-end x of the ribosome and beginning y of the polymerase. The transcription center is designated z, i.e. $z = y + s_1$. In the window, transition occurs from macrostate Ω to Ω', with macrostates allowed to contain only hypohelices overlapping with the window by at least three nucleotides, thus defining *macrostates in current window*; and there is macrostate Ω, i.e. a non-empty diagram. Macrostates describe the secondary RNA structure in the window. The window does not exist before binding of the polymerase (an empty macrostate), and before binding of the ribosome the window opens at the first nucleotide of the primary sequence and closes at the 5'-edge y of the polymerase.

The primary steps of the modeling are: after transcribing the start codon and subsequent $s_0 + s_1$ nucleotides, the ribosome attempts to bind with the SD sequence and

the start of leader peptide gene according to the described rule. As it occurs, the ribosome occupies the start of the leader peptide and the following characteristics become fixed: left boundary x of the window at the position "start of the leader peptide gene plus s_0", right window boundary y at the position determined by the beginning of the polymerase, and the secondary structure inherited in the window from the past.

We also modeled binding of the polymerase to the promoter, but with our sequences it did not affect the values of $p(c)$. In data below, the ribosome started immediately from the SD domain, and, at the same time, the 5'-edge of the polymerase was located at *sp* nucleotides downstream the 5'-edge of the start codon.

Standard Monte-Carlo technique is used for modeling. *Current state* is described with parameters $< x, y, z, T, \Omega >$. During initiations, description of the state includes condition ς of whether or not binding of the polymerase or ribosome occurred; this condition is not considered on further steps. The *neighborhood* of given state Ω, *centered* in Ω, is a set of all states with non-zero probability of transition from Ω. If given neighborhood contains *n* states, and corresponding transition rate constants are $k_1, \ldots, k_n$ with $k = \sum k_i$, the transition state (*next* state in the trajectory) is determined by choice of i with probability k_i / k.

Physical time in seconds between each two successive transitions in the model is calculated giving the overall transcription time of regulatory region in the model, which can approximate this time in biological reality. Physical time t is taken to have exponential distribution $1 - \exp(-kt)$. Notably, the values of $\lambda_{sd}, \lambda_{rib}, \lambda_{pol}, K(\Omega \to \Omega')$ are of considerably different orders of magnitude, thus causing certain inconveniencies.

4. Simulation Results and Discussion

4.1. The case of proteobacteria

(1) Results of modeling are shown partially, mainly for the leader regions of *trpE*-containing operons of alpha- and gamma-proteobacteria and for corresponding mutants. Experimentally known mutations are taken from [25]: in the leader region of tryptophan operon in *E. coli* G was replaced by A at the 75th and 132nd positions from the transcription start. These mutants are designated trpL75 and trpL132, respectively. The trpL75 mutation destabilizes the antiterminator secondary structure and leads to **inevitable termination**. The trpL132 mutation results in destabilization of the terminator structure and thus **reduces the termination frequency**, Table 2. Similar results of modeling are published in [54, 28] for other bacteria, while any experimental evidence on their mutations is lacking.

Thus, Table 2 shows the values of function $p = p(c)$ for *trpE*-gene containing operons of *Rhodopseudomonas palustris*, *Rhizobium leguminosarum*, *Sinorhizobium meliloti*, *Escherichia coli*, *Vibrio cholerae*. Presence of classic attenuation in these bacteria is corroborated partly by experiment [55, 25] and partly by analyses of multiple alignments of native leader regions. Alignment data is available for actinobacteria [19] and proteobacteria [11].

(2) In silico mutations were introduced at a very limited number of positions *only* in the left box of antiterminators, which had a great impact on modeling result (ref. Fig. 1

and Table 4). More precisely, with biological sequences probability $p(c)$ under high concentration c was at least twice as high as under low c, while it almost did not depend on concentration with mutants. In other words, *termination effectiveness* $p(0.5)/p(0)$ was about 2 at the lowest for native leader regions, and about 1 – for the mutants. Somewhat an exception is the *trpE* leader region in *S. meliloti*, where termination effectiveness was about 2 in the mutant but about 7 in wild type.

Table 2. Modeled termination probability $p(c)$ (%) vs. concentration c of charged trp-tRNA in *E. coli* and its mutants trpL75 and trpL132, as well as in other proteobacteria.

c	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
E. coli (wild type)	22	36	53	60	65	69	67	70	68	72	70
E. coli (trpL75)	75	70	72	72	71	73	73	71	71	71	70
E. coli (trpL132)	2	6	11	14	14	18	19	17	20	19	19
V. cholerae	11	22	41	56	61	66	71	72	74	74	72
R. leguminosarum	12	22	34	42	48	52	52	60	58	60	59
R. palustris	1	11	25	31	33	35	38	38	37	40	37
S. meliloti	10	12	19	26	32	36	37	40	39	41	39

(3) In [25] expression of the *trpE* operon in *E. coli* was studied in vitro. Anthranilate concentrations were obtained over 30 min under high and low concentrations of tryptophanyl-tRNA, and ratios shown in the second column of the Table 3. The third column contains antitermination probability ratios estimated in the model under fixed parameter values. The sequences used in modeling are shown in Fig 1.

Table 3. Comparison of expression effectiveness of gene *trpE* in *E. coli* under different tryptophanyl-tRNA concentrations in experiment [25] and the model.

Genome	Anthranilic acid produced without/with translation of leader peptide	$\frac{1-p(0.05)}{1-p(0.5)}$
E. coli (wild type)	2.1	2.1
E. coli (trpL75)	1.2	1.0
E. coli (trpL132)	1.7	1.2

(4) Total estimated time between each pair of successive transitions between states along the same modeling trajectory gives physical time of the whole attenuation process, typically 2-3 sec. These estimates are of the same order of magnitude as would be expected in biological reality from the average polymerase rate and length of primary sequence. Notably, in all our simulations computing one trajectory is considerably (4 to 1000 times) less than this physical timing, i.e. the model is *faster* than the biological process.

(5) In addition, our program and server [3] implement a special mode to ignore the behavior of the ribosome and polymerase to model only secondary structure kinetics in a long run. Modeling a fixed nucleotide sequence over sufficiently long time produces the following results.

(a) For sequences up to 170 bases in length, the program and server [3] always compute the minimal energy secondary structure with accuracy either ±12% or ±2.8 kcal/mol over $5 \cdot 10^4$ transitions. It was proved for sequences up to 120 bases-long under parameter ρ value 70 (*maximal allowed loop length*), and for up to 170 bases-long sequences under parameter ρ value no less than the maximal loop length in structures computed by the RNAstructure (Mfold) program v.4.2 [41, 44] (e.g., 87 for the *hisL* operon of *Escherichia coli* K12, 125 for the *trpE* operon of *Rhodopseudomonas palustris*, 136 for the *trpE* operon of *Vibrio cholerae*). Exceptional was the *trpBEGDC* operon of *Corynebacterium diphtheriae*, where this pattern was observed for 96% model trajectories, and with slightly less accuracy – for the rest 4%. For sequences with length 120-170 bases under $\rho = 70$ the lowest accuracy of the server predictions is ±21%, since the RNAstructure program only computes structures with loop length above 70 and so our server should not find such structures *a priory*. However, even that accuracy level seems acceptable. The RNAstructure program [44] implements a slightly different approach to compute energies and different rules to find secondary structures. E.g., server [41] outputs structures that do not meet the completeness requirement and contain hypohelices with lengths 1 and 2 not allowed in our model and program. Thus, certain inconsistency between the energy and secondary structure predictions between servers [3] and [41] is unavoidable. Parameter ρ can be explicitly specified in our program and server [3], with $\rho = 70$ set by default.

(b) For all sequences shorter than 120 bases, among $\tau = 3$ of computed minimal structures, 100% modeled trajectories contained a minimal energy structure found by the RNAstructure program (sometimes, accurate within one hypohelix). This structure is further referred to as an *equilibrium* structure, and its corresponding energy – as *minimal* energy. Exceptional appeared the same sequence from *Corynebacterium diphtheriae*, for which this pattern was observed in 93% trajectories. Number of minimal energy structures τ is specified by the user.

(c) Average time of finding structures from items (a), (b) under $N_1 = 5 \cdot 10^4$ transitions and $\rho = 70$ for sequences up to 100 bases-long is less than a second, maximal time – 6 sec.; for sequences up to 120 bases-long those are 18 and 220 sec, respectively; for sequences up to 170-bases long – 4 and 89 min. Under maximal loop lengths obtained from RNAstructure program (typical lengths shown in item (a)) timing is 5 and 115 min, respectively. All above figures are for 3GHz PC.

A heuristic approach was used in this mode to speed up computations of equilibrium structures and minimal energies based on coercive decrease (e.g., following logarithmic law) of the computed slow transition rate under increase of multiple hits into the resulting macrostate. After computing transition rates from current macrostate Ω into each neighboring macrostate Ω', value $K(\Omega \to \Omega')$ estimated with Eq. (7) was substituted by

$$\tilde{K}(\Omega \to \Omega') = K(\Omega \to \Omega') \cdot \frac{Q}{Q + \ln N(\Omega')},$$

where $N(\Omega')$ is final number of hits in macrostate Ω', and $Q > 0$ is a parameter that allows to keep the base of the logarithm when customizing the dependency. With this correction, cycles occur less frequently along the trajectory thus decreasing CPU time. Particularly, good results are often obtained with parameter pairs $< N_1 = 2 \cdot 10^4, Q = 1 >$ and $< N_1 = 10^4, Q = 5 >$. Typical characteristics of performance to find equilibrium structures and minimal energies are available at server [3].

The authors interpret the results of items (1)–(5) in support to the model logic. The choice of parameter settings and form of dependencies needs further discussion and is greatly dependant on additional experimental evidence. Methodology of developing such models and analyzing relevant experimental data are proposed above.

Therefore, among primary results of this work one can consider developing a *versatile program* implementing the described logic of the model with a possibility to modify any equation and/or variable upon need. The program is *available from the authors*, with its current online implementation available online as a web server [3]. Requests for computations with custom nucleotide sequences and/or parameter settings/dependencies (otherwise, default settings will be used) are welcome at lin@iitp.ru.

The model output is not limited to $p(c)$ values under given concentrations. A default program output of each run contains final positions of the ribosome and polymerase including codon type, number of transitions, physical time to termination or antitermination events, etc., with statistic characteristics of these variables over a series of runs in the final output, thus facilitating analysis of model behavior under given c. Special effort was made to facilitate studies of secondary structure kinetics. Conventional 2-D representation of the structure is a poor description of transition between the states. Instead, we used natural 1-D representation in the form of initial sequence with color-coding of hypohelices in current state. The model trajectory is thus a chronological succession of such representations with cumulative modeled time to the current transition over previous transitions in seconds. For better visualization, any recurring cycles of any length and depth are excluded, and local stability of color schemes for same hypohelices is maintained. The authors are unaware of a similar representation of secondary structure dynamics. Representation of the model trajectory allowed the program to find secondary structures forming between the events of the ribosome or polymerase move. With wide windows in the leader region of gene *trpE* in *E. coli* these structures are always far from equilibrium, which seems natural.

(6) As to the authors' knowledge, there is only one alternative public resource that implements secondary structure kinetics modeling [1]. Comparing performance of this server and our server reveals certain differences in modeling results. Thus, for the region between the leader peptide stop codon and the end of the terminator *U*-run in the leader region of operon *trpBEGDC* in *Corynebacterium diphtheriae* the compared server always computes a structure with energy value significantly different from the minimal energy. The output structure does not belong to the set of equilibrium structures, all, unlike the output structure, containing a conservative terminator. A similar situation is observed for the tryptophan operon of

Sinorhizobium meliloti in 90% cases. For other sequences, e.g. the above discussed leader region of operon *trpEGDCBA* in *Corynebacterium glutamicum* the server fails to compute the solution after an hour of CPU time. This behavior is observed under maximum allowed computing time of the latest version 5 of the server with its default parameter settings. Computing time of our server is orders of magnitude less for the same input sequences; in both cases $\kappa = 10^7\ s^{-1}$ was used.

(7) Model parameters were varied to estimate robustness of the result. Tables 5-7 show the values of probability $p(0)$ under varied parameters r_0, κ, sp in turn, with other parameters fixed. As mentioned earlier, data in Table 7 reveals only minor effect of varying parameter *sp* in a wide range of values (initial distance between the ribosome and polymerase) on modeling results. Hence, translation initiation and delay timing have negligible impact in modeling. Similar data (not shown) on probability $p(0)$ vs. parameter δ, which was varied between 15 and 40, suggests that our assumed $\delta = 30\ \mathrm{s}^{-1}$ is optimal in the same sense.

Table 4. Termination probability $p(c)$ vs. charged tRNA concentration for mutant sequences shown in Fig. 1. In all cases, this varying completely inhibited regulation, except for *S. meliloti*, where it was partially inhibited leading to growth of $p(0)$.

c	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
E. coli	72	72	71	73	72	71	72	72	72	71	70
V. cholerae	71	71	71	76	73	73	74	74	73	75	75
R. leguminosarum	63	60	56	55	55	55	53	55	55	56	56
R. palustris	38	41	41	41	38	40	41	39	42	42	41
S. meliloti	28	25	29	34	37	36	38	40	41	41	41

Table 5. Premature termination frequency under null amino acid concentration $p(0)$ vs. different r_0 values, i.e. distance between 3'-end of the hairpin and the beginning of polymerase, for *trpE* leader regions shown in Fig 1.

r_0	0.1	0.5	1	2	3	4	5
E. coli	18	19	22	29	33	34	38
V. cholerae	10	11	11	12	16	21	27
R. leguminosarum	11	12	12	18	23	29	35
R. palustris	0.2	0.4	1	2	5	7	9
S. meliloti	8	9	10	11	17	19	20

Table 6. Premature termination frequency under null amino acid concentration $p(0)$ vs. different κ values, i.e. “cytoplasm viscosity”, for *trpE* leader regions shown in Fig 1.

κ	31	62	125	250	500	1000	2000
E. coli	53	51	44	43	30	22	23
V. cholerae	29	17	12	11	12	11	13
R. leguminosarum	34	35	25	18	12	12	16
R. palustris	7	5	2	0.5	0.5	1	1.4
S. meliloti	20	19	18	19	14	10	8

Table 7. Premature termination frequency under null amino acid concentration $p(0)$ vs. different *sp* values, i.e. initial distance between the ribosome P-site (transcribed codon) and the 5'-edge of polymerase, for *trpE* leader regions shown in Fig 1.

sp	13	20	30	40	50
E. coli	22	23	21	21	21
V. cholerae	11	14	12	13	11
R. leguminosarum	12	13	14	13	13
R. palustris	1	1	1	0.7	0.7
S. meliloti	10	9	10	9	9

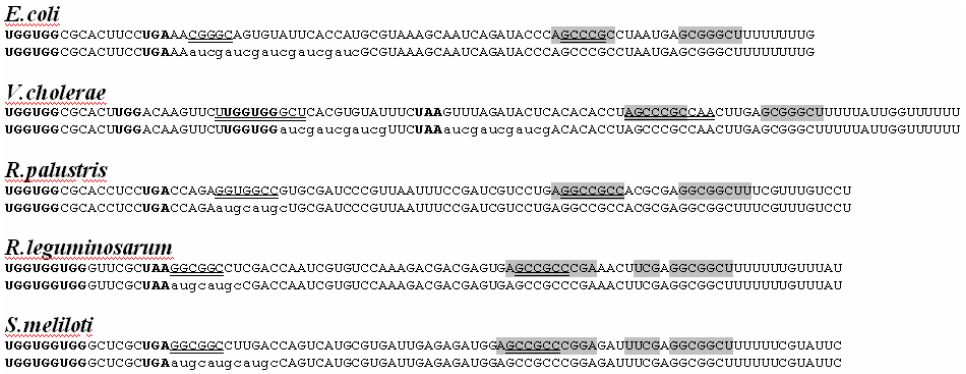

Fig. 1. Biological leader sequences are shown in the first line under corresponding species names, antiterminator underlined, terminator shaded, regulatory and stop codons set in bold face. Mutants are shown in the second line, with mutated bases set in lowercase.

4.2. The case of histidine operon

Symmetric case Eq. (6) and $\kappa = 10^3\ \mathrm{s}^{-1}$ were used for the tryptophan operon discussed above. For the histidine operon concentration-independent function $p(c)$ is obtained under various κ values. In contrast, asymmetric case Eqs. (4-5) under $\kappa = 10^6$ produces a reasonable solution, Table 8. We assume that the locus of the leader region in interest might be essential in choosing between the two cases. Notably, regulatory regions of the histidine operon are considerably longer than those of tryptophan operon, with different consensus sequences, etc., which might suggest presence of a factor responsible for the difference. In the model we use different, albeit logically similar, definitions of transition constants between RNA secondary structures, the symmetric and asymmetric cases.

Table 8. Termination probability $p(c)$ vs. concentration c of charged his-tRNA.

c	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
E. coli, *his*	18	32	56	59	59	62	70	70	68	63	64

4.3. The case of Streptomyces

For some leader regions with attenuation regulation predicted from the structure of their multiple alignments and/or proved experimentally, Eq. (1) of hypohelix energy provides misleading estimates: termination probability $p(c)$ does not grow under growing concentration c. This equation was used in this study in its current form known to be preliminary. Here we propose the following modification to it:

$$G_{hel}(\omega) = \frac{\sum_j \left(E_{\overline{\gamma}_j} - \alpha \cdot \frac{l'_j}{\left(1 + l'_j / l_{\max}\right)} \right)}{RT}. \tag{26}$$

The difference from Eq. (1) is the additional term («correction»)

$$E(l') = -\alpha \cdot \frac{l'}{\left(1 + l'/l_{\max}\right)}, \tag{27}$$

which contains parameters α and $l_{\max}$, where $l_{\max}$ is loop length l', under which $E(l')$ equals half of its asymptotic value. In our modeling, typical settings were $l_{\max} = 10$ and α=0 or 10. For the phylogenetic lineage of proteobacteria $\alpha = 0$ (then $l_{\max}$ is eliminated), and for streptomycetes $\alpha = 10$. An interaction that corresponds to this correction might be related to additional energy of binding the RNA region having a realized macrostate with stabilizer molecules, to the tertiary structure energy of this region, e.g., knots and pseudoknots.

Results for two *Streptomyces* species are shown in Table 9; similar results were obtained for other *Streptomyces* spp.

Table 9. Termination probability $p(c)$ vs. concentration c of charged trp-tRNA in case of certain *Streptomyces*. Parameter settings are $\alpha = 10$, $l_{\max} = 10$ and, here only, $r_0 = 5$.

c	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
S. avermitilis, trpE	16	34	43	41	45	48	50	55	55	55	59
S. coelicolor, trpE	16	27	35	40	46	45	45	47	47	48	52

To conclude, let us again outline the scope of potential applications for this model. It can be used to obtain additional evidence in predicting attenuation regulation on the basis of multiple alignments by estimating function $p(c)$ of termination probability vs. concentration of charged tRNA. The evidence is a pronounced and relatively smooth monotonous growth of the function. Another field is research on the effect of point mutations in the leader region on evolutionary stability and expression effectiveness of the gene under given regulation.

5. Acknowledgments

The authors are grateful to professor M. Gelfand for very valuable discussion and professor A. Mironov, who confided this problem to one of the authors once during

summer 2004 absorbingly enough to provoke fruitful research on this for over two years, and who generously helped by answering ample questions. Also, we wish to thank K. Gorbunov and L. Rusin for much of advice and continuous help with simulations. The work was partially supported by grant ISTC 2766.

6. References

1. L.V. Danilova, D.D. Pervouchine, A.V. Favorov, A.A. Mironov, "RNAKINETICS: A web server that models secondary structure kinetics of an elongating RNA", *Journal of Bioinformatics and Computational Biology* **4**(2), 589–596 (2006).
2. H. Yin, I. Artsimovitch, R. Landick, J. Gelles, "Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules", *PNAS* **96**(23), 13124–13129 (1999).
3. World Wide Web server at http://lab6.iitp.ru/rnamodel
4. T.M. Henkin, C. Yanofsky, "Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions", *Bioessays* **24**(8), 700–707 (2002).
5. F.J. Grundy, T.M. Henkin, "The T box and S box transcription termination control systems", *Front Biosci*. **8**, d20–31 (2003).
6. F.J. Grundy, T.M. Henkin, "Regulation of gene expression by effectors that bind to RNA", *Curr. Opin. Microbiol.* **7**(2), 126–131 (2004).
7. M. Mandal, R.R. Breaker, "Gene regulation by riboswitches", *Nat. Rev. Mol. Cell. Biol.* **5**(6), 451–463 (2004).
8. A.G. Vitreschak, D.A. Rodionov, A.A. Mironov, M.S. Gelfand, "Riboswitches: the oldest mechanism for the regulation of gene expression?" *Trends in Genetics* **20**(1), 44–50 (2004).
9. C. Yanofsky, "The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis*", *Trends in Genetics* **20**(8), 367–374 (2004).
10. E.M. Panina, A.G. Vitreschak, A.A. Mironov, M.S. Gelfand, "Regulation of aromatic amino acid biosynthesis in gamma–proteobacteria", *Journal of Molecular Microbiology and Biotechnology* **3**(4), 529–543 (2001).
11. A.G. Vitreschak, E.V. Lyubetskaya, M.A. Shirshin, M.S. Gelfand, V.A. Lyubetsky, "Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis", *FEMS Microbiology Letters* **234**, 357–370 (2004).
12. F.J. Grundy, T.M. Henkin, "Conservation of a transcription antitermination mechanism in amino acyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria", *J. Mol. Biol*. **235**(2), 798–804 (1994).
13. F.J. Grundy, T.M. Henkin, "The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria", *Mol. Microbiol*. **30**(4), 737–749 (1998).
14. B.A. Murphy, F.J. Grundy, T.M. Henkin, "Prediction of gene function in methylthioadenosine recycling from regulatory signals", *J. Bacteriol.* **184**(8), 2314–2318 (2002).
15. E.M. Panina, A.G. Vitreschak, A.A. Mironov, M.S. Gelfand, "Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria", *FEMS Microbiol. Lett.* **222**, 211–220 (2003).
16. T.M. Henkin, B.L. Glass, F.J. Grundy, "Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes", *J Bacteriol*. **174**(4), 1299–1306 (1992).
17. N. Sudarsan, J.E. Barrick, R.R. Breaker, "Metabolite-binding RNA domains are present in the genes of eukaryotes", *RNA* **9**(6), 644–647 (2003).

18. D.A. Rodionov, A.A. Vitreschak, A.A. Mironov, M.S. Gelfand, "Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes", *J. Biol. Chem*. **277**(50), 48949–48959 (2003).
19. A.V. Seliverstov, H. Putzer, M.S. Gelfand, V.A. Lyubetsky, "Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria", *BMC Microbiology* **5**(54) (2005).
20. A.V. Seliverstov, V.A. Lyubetsky, "Translation regulation of intron containing genes in chloroplasts", *Journal of Bioinformatics and Computational Biology* (in print) (2006).
21. J.E. Barrick, K.A. Corbino, W.C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J.K. Wickiser, R.R. Breaker, "New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control", *Proc. Natl. Acad. Sci. USA* **101**(17), 6421–6426 (2004).
22. C. Abreu-Goodger, N. Ontiveros-Palacios, R. Ciria, E. Merino, "Conserved regulatory motifs in bacteria: riboswitches and beyond", *Trends Genet*. **20**(10), 475–479 (2004).
23. A.A. Vitreschak, D.A. Rodionov, A.A. Mironov, M.S. Gelfand, "Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation", *Nucleic Acids Research* **30**(14), 3141–3151 (2002).
24. A.G. Vitreschak, D.A. Rodionov, A.A. Mironov, M.S. Gelfand, "Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element", *RNA* **9**(9), 1084–1097 (2003).
25. A. Das, I.P. Crawford, C. Yanofsky, "Regulation of Tryptophan Operon Expression by Attenuation in Cell free Extracts of *E. coli*", *Journal of Biological Chemistry* **257**(15), 8795–8798 (1982).
26. Nudler E., Mironov A.S., "The riboswitch control of bacterial metabolism", *Trends Biochem. Sci.* **29**, 11–17, 2004.
27. R. Landick, C.L. Turnbough, C. Yanofsky, "Transcription attenuation", in Neidhardt F.C., Curtiss R., Linn E.C. (eds.), *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd ed., American Society for Microbiology: Washington, DC, pp. 1263–1286, 1996.
28. V.A. Lyubetsky, L.I. Rubanov, A.V.Seliverstov, S.A. Pirogov, "Model of gene expression regulation in bacteria via formation of RNA secondary structures", *Molecular Biology* **40**(3), 440–453 (2006).
29. A.A. Mironov, A.E. Kister, "Theoretical analysis of RNA secondary structure formation kinetics during transcription and translation. Accounting for imperfect helices", *Molecular Biology (Mosk)* **19**, 1350–1357 (1985), in Russian.
30. A.A. Mironov, A.E. Kister, "A theoretical analysis of structural restructuring during formation of secondary RNA structures", *Molecular Biology (Mosk)* **23**, 61–71 (1989), in Russian.
31. A.A. Mironov, V.F. Lebedev, "A kinetic model of RNA folding", *BioSystems* **30**, 49–56 (1993).
32. A. Xayaphoummine, T. Bucher, F. Thalmann, H. Isambert, "Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations", *Proc. Natl. Acad. Sci. USA* **100**, 15310–15315 (2003).
33. A. Xayaphoummine, T. Bucher, H. Isambert, "Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots", *Nucleic Acids Res*. **33** (Web Server issue), W605–10 (2005).
34. J. Elf, M. Ehrenberg, "What Makes Ribosome-Mediated Transcriptional Attenuation Sensitive to Amino Acid Limitation?" *PLoS Computational Biology* **1**(1), e2 (2005).
35. F. Liu, Z.C. Ou-Yang, "Monte Carlo simulation for single RNA unfolding by force", *Biophys. J.* **88,** 76–84 (2005).
36. A.P. Gultyaev, F.H. van Batenburg, C.W. Pleij, "The computer simulation of RNA folding pathways using a genetic algorithm", *J. Mol. Biol.* **250,** 37–51 (1995).

37. B. Onoa, I. Tinoco, Jr., "RNA folding and unfolding", *Curr. Opin. Struct. Biol.* **14**, 374–379 (2004).
38. R.M. Dirks, M. Lin, E. Winfree, N.A. Pierce, "Paradigms for computational nucleic acid design", *Nucleic Acids Res.* **32,** 1392–1403 (2004).
39. C. Flamm, W. Fontana, I.L. Hofacker, P. Schuster, "RNA folding at elementary step resolution", *RNA* **6,** 325–338 (2000).
40. K. Kawasaki, "Diffusion constants near the critical point for timedependent Ising models", *Phys. Rev*. **145**, 224–230 (1966).
41. M. Zuker "Mfold web server for nucleic acid folding and hybridization prediction", *Nucleic Acids Res.* 31(13), 3406–15 (2003).
42. P. Berg, M. Singer, *Genes and genomes*, Mill Valley, University Science Books, 1991.
43. D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure", *J. Mol. Biol.* **288**, 911–940 (1999).
44. D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, D.H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure", *PNAS* **101**(19), 7287–7292 (2004).
45. I. Dima, C. Hyeon, D. Thirumalai, "Extracting Stacking Interaction Parameters for RNA from the Data Set of Native Structures", *Journal of Molecular Biology,* **347**(1), 53–69 (2005).
46. G.F. Lawler, L.N. Coyle, *Lectures on Contemporary Probability*, AMS, 1999.
47. S.A. Pirogov, K.Y. Gorbunov, V.A. Lyubetsky, "Macro- and microstates in a model of attenuator regulation of gene expression in bacteria", *Proceedings of the RAS conference "Issues of control and modeling in complex systems, VII International conference",* 27 June – 1 July 2005, RAS (Samara), 210–215 (in Russian).
48. V.A. Lyubetsky, S.A. Pirogov, "A model of attenuator regulation in bacteria", *Proceedings of the RAS conference "Issues of control and modeling in complex systems, VII International conference",* 27 June – 1 July 2005, RAS (Samara), 205–210 (in Russian).
49. S. Lynn, L. Kasper, J. Gardner, "Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the *thr* operon attenuator", *J. Biol. Chem.* **263,** 472–479 (1988).
50. S.L. Gotta, O.L. Miller jr., S.L. French, "rRNA Transcription Rate in Escherichia coli", *Journal of Bacteriology*, **173**(20), 6647–6649 (1991).
51. A.D. Ventzel, M.I. Freydlin, *Fluctuations in dynamic systems under minor random perturbances*. Moscow, Nauka, 1979 (in Russian).
52. K. Wilson, P. von Hippel, "Transcription termination at intrinsic terminators: the role of the RNA hairpin", *Proc. Natl. Acad. Sci. USA* **92,** 8793–8797 (1995).
53. C. Lin, A.S. Paradkar, L.C. Vining, "Regulation of an anthranilate synthase gene in *Streptomyces venezuelae* by a *trp* attenuator", *Microbiology*, **144**, 1971–1980 (1998).
54. Lyubetsky V.A., Seliverstov A.V., "Estimating effectiveness of tryptophan biosynthesis regulation in bacteria with a classic attenuation model", *Information processes* **6**(1), 55–57 (2006) (in Russian).
55. D.M. Heery, L.K. Dunican, "Cloning of the *trp* gene cluster from a tryptophan-hyperproducing strain of *Corynebacterium glutamicum*: Identification of a mutation in the *trp* leader sequence", *Applied and Environmental Microbiology*, **59**, 791–799 (1993).