ISSN 0026-8933, Molecular Biology, 2006, Vol. 40, No. 3, pp. 440–453. © Pleiades Publishing, Inc., 2006.
Original Russian Text © V.A. Lyubetsky, L.I. Rubanov, A.V. Seliverstov, S.A. Pirogov, 2006, published in Molekulyarnaya Biologiya, 2006, Vol. 40, No. 3, pp. 497–511.

UDC 577.053

Model of Gene Expression Regulation in Bacteria via Formation of RNA Secondary Structures

V. A. Lyubetsky, L. I. Rubanov, A. V. Seliverstov, and S. A. Pirogov

*Institute of Information Transmission Problems, Russian Academy of Sciences,
Moscow, 127994 Russia; e-mail: slvstv@iitp.ru*

Received June 20, 2005; in final form, November 7, 2005

Abstract—A model was proposed for the classical attenuating mRNA regulation of gene expression via transcription termination. The model is based on the concept of secondary structure macrostates in the RNA regulatory region between the ribosome and RNA polymerase, utilizes resonant equations for estimating the deceleration of RNA polymerase by a set of hairpins located in this RNA region, and takes into account views on the initiation and elongation of transcription and translation. Special attention was paid to selecting the model parameters. To test the model, computations were performed to estimate, in particular, the probability of translation termination as dependent on the charged tRNA concentration and the amino acid concentration for several regulatory regions of the bacterial genome (as exemplified by *trp*E of *Streptomyces* spp., *Bradyrhizobium japonicum*, and *Escherichia coli*). Analysis was performed with different values of three parameters isolated as major ones. The resulting dependences agreed with the available experimental data, including those characterizing an enzymatic activity as dependent on the amino acid concentration in a culture (e.g., the anthranylate synthase activity as dependent on the tryptophan concentration in *S. venezuelae*). The following possible application was proposed for the model. Attenuating regulation is usually predicted on the basis of multiple sequence alignment, which requires several sequences. With the model, an individual sequence can be analyzed with proper parameters to generate a concentration–enzymatic activity curve. The curve characteristic of attenuation or its absence provides an additional argument for the presence or absence of attenuation.

DOI: 10.1134/S0026893306030113

Key words: attenuation, model of transcriptional regulation, mathematical model in genetics

INTRODUCTION

In recent times, it has been acknowledged that regulation of gene expression via mechanisms relying on the formation of RNA secondary structures plays a significant role. Such mechanisms affect the elongation of transcription or delay translation, utilizing various mediators such as the ribosome (in the case of classical attenuation), a regulatory protein, tRNA, or a cofactor (in the case of other attenuating regulations) [1–6]. The mechanisms have been studied mostly with γ-proteobacteria and bacilli [7–12]. New regulators include T-boxes [9, 15]; riboswitches, identified recently [4, 5, 13, 14]; and hypothetical regulatory LEU elements, found even more recently [16]. A large-scale search for new regulatory elements has been attempted [17, 18]. The results of such studies suggest a particular function for hypothetical genes and provide further insight into bacterial metabolic pathways [11, 14, 19, 20]. The history of studying classical attenuation has been described in brief in the introductions to certain works (see, e.g., [6, 8, 21]).

Studies on bioinformatics in the field include systematic experimental studies aimed at identifying well-known and new regulations by means of comparative genomics and a few works aimed at modeling a regulatory mechanism or its components [22–27].

Some of the above works have employed the Monte Carlo procedure in modeling the RNA secondary structure folding kinetics at the level of microstates and have posed the problem of modeling at the level of macrostates [22–24]. In some others, probabilistic simulation with the Monte Carlo procedure has been used to study the formation of pseudoknots in the RNA secondary structure [26, 27]. A model of the secondary structure folding kinetics has been proposed on the basis of earlier ideas [22–24] and an original technique to improve the operation of the Monte Carlo algorithm by excluding the repetition of Markov chain states. Our model utilizes another, rapid organization of the Monte Carlo procedure, also excluding repetitions. The probability of antitermination is computed by an explicit equation, as a sum of two components in [25]. One is the probability that the ribosome occurs at one of the regulatory codons and that an antiterminator is formed at the moment polymerase reaches a T-rich region. The other component is the product of the coefficient 0.5 and the probability that the ribosome leaves the stop codon until an antiterminator is formed. The coefficient 0.5 has been motivated by the idea that either a terminator or an antiterminator is formed with such probability in this situation. Further research has been based on computations according to this equation, which to us is rather unclear.

We propose a model for classical RNA attenuation of gene expression via transcription termination as described in [21, pp. 172–189]. The model is based on the idea of a secondary structure macrostate existing in the RNA regulatory region between the ribosome and RNA polymerase and on the resonance equations, which determine the rate of RNA polymerase deceleration by a set of hairpins located in the same region. In addition to defining the macrostate, we attempted to develop an equation that estimates the rate of RNA polymerase deceleration by the set of hairpins and, eventually, by the current macrostate. Another problem was to simulate the initiation and elongation of transcription for regulatory and nonregulatory codons.

Like earlier models [22–29], our model depends on many, rather arbitrary, decisions about how decompose the total process into components; which mathematical apparatus to choose for describing the components; which set of parameters to use and with which numerical values; and how to compare the modeling results with experimental data, which are scarce as yet. We think that these difficulties can be overcome by discussing and comparing the existing models with each other and with experimental data. It can be expected that such comparisons will stimulate experiments in the field.

DESCRIPTION OF THE MODEL

Definitions of Microstates and Macrostates. Transition Rate Constants

Let there be, here and from this point on, a fixed sequence with the four-letter alphabet {A, C, T, G}, be it a regulatory region of a bacterial genome or an arbitrary sequence. For instance, let it be a region starting from the promoter (if the promoter is known, which is a rare case) or the ribosome-binding site upstream of the leader peptide-coding sequence and ending at the end of the polyuracil tract.

The initial sequence is divided into sections of no less than 3 nt, which serve as arms of future helices: $\ldots a_i, \ldots, b_j, \ldots$. The pairing of any similarly sized sections a_i and b_j (which implies hydrogen bonding and stacking of the corresponding nucleotides throughout the lengths of sections a_i and b_j) yields helix γ_i. It is always implied that helix γ_i is nonextendable; i.e., the ends of sections a_i and b_j (the arms of helix γ_i) are flanked by noncomplementary nucleotides and that the region between the sections (the terminal loop of the helix) is no less than 3 nt in size. Generally speaking, the model allows any list of initial helices. The above definition concerns only one of all possible variants, taking as initial all nonextendable helices that meet the above requirements for the arm and loop sizes.

All these concepts have been described, for instance, in [21, pp. 172–189], along with the classical attenuating RNA regulation of gene expression as dependent on the concentration of an amino acid (or a charged tRNA, whose concentration depends on the concentrations of the corresponding amino acid and aminoacyl-tRNA synthase).

A hypohelix of helix γ_i is defined as any nonempty part $\bar{\gamma}_i$ of helix γ_i that consists of two connected arms of no less than 3 nt in size. The arms are hereafter defined as paired regions of a hypohelix or a helix; their ends are always designated as A, B, C, and D starting from the 5' end of the initial sequence. A terminal loop is defined as the RNA region located between the two arms of a hypohelix.

A microstate is defined as a (nonempty total) set of hypohelices that are nonextendable in the given set and lack pseudoknots. In this state, no two hypohelices are adjacent; i.e., A and D of one hypohelix are not the neighbor nucleotides of B and C of the other. In addition, empty set $\varnothing$ is a separate "initial" microstate. The term nonextendable in a set means that the arms of the hypohelices of the set cannot be extended. A pseudoknot is a pair of hypohelices such that exactly one arm of one hypohelix overlaps the loop of the other hypohelix (consequently, this arm is within the loop). The set of all hypohelices that belong to one helix and are involved in a particular microstate is termed a subhelix of the given helix in the given microstate.

For every microstate, each of its hypohelices and subhelices receives the same ordinal number as ascribed to the corresponding (nonextendable) helix; all helices of the initial sequence are numbered in a prefixed order.

The diagram of a microstate is a common parenthetical structure that reflects the arrangement of all hypohelices in the given microstate, each pair of parentheses (a chord, according to other terminology) having the same ordinal number as ascribed to the helix harboring the given hypohelix, which corresponds to the given parentheses (chord). The parenthesis structure reflects the arrangement of hypohelices according to common rules: several consequent hypohelices are shown with the same number of consequent parentheses ()()…(), and the location of hypohelix 1 in the loop of hypohelix 2 is shown with enclosed parentheses $((\ )_1)_2$, where the inner parentheses correspond to hypohelix 1 and the outer parentheses, to hypohelix 2. Helix numbers can be repeated many times in a diagram, because many hypohelices can be extracted from each helix. A diagram is easy to draw from a microstate, which is in fact a list of paired nucleotides. Yet a microstate cannot be recovered from its diagram: a diagram reflects only the geometry of hypohelices, indicating the helix wherefrom a hypohelix can be obtained for every pair of parentheses.

Every set consisting of helices $\gamma_1, \ldots, \gamma_k$ has a corresponding set of realizing microstates: this is any nonexpandable (in itself) set of subhelices $\bar{\gamma}_1 \subseteq \gamma_1, \ldots, \bar{\gamma}_k \subseteq \gamma_k$ (exactly one nonempty but not necessarily connected region $\bar{\gamma}_i$ is taken from each helix γ_i) without pseudoknots. As above, adjacent hypohelices (i.e., those with the pair of nucleotides A and D immediately following the pair of B and C) are combined.

A macrostate is defined as any nonempty diagram. A diagram is nonempty when it is realized in at least one microstate. For each microstate ω from macrostate Ω, the diagrams of ω and Ω coincide.

Bonding energy $E_{\bar{\gamma}_j}$ of hypohelix $\bar{\gamma}_i$ is obtained by summing the bonding energy of paired nucleotides over all consequent nucleotide pairs, using the stacking energy of bonds between adjacent pairs. Special provisions are made for the stacking of the first and last pairs of hypohelix $\bar{\gamma}_i$ and for the coaxial stacking of $\bar{\gamma}_i$, which depends on microstate ω wherefrom $\bar{\gamma}_i$ is taken. The energy is computed according to a published scheme with published numerical values [30–33].

For each hypohelix $\bar{\gamma}_i$ from microstate ω, the nucleotide number l_i is determined for the nucleotides that belong to its terminal loop and are beyond the loops and arms of other hypohelices of the given microstate. This parameter, l_i, depends on the microstate and is termed a length of the terminal loop of hypohelix $\bar{\gamma}_i$. At the same time, hypohelix $\bar{\gamma}_i$ is characterized by the total length of its terminal loop; i.e., the length of the loop is determined regardless of whether its nucleotides are paired within other hypohelices in the given microstate. The total length depends only on the hypohelix $\bar{\gamma}_i$ and is designated as l'_j.

By definition, microstate ω is characterized by two free energies: the energy of bonding in hypohelices and the energy of loops of hypohelices occurring in ω. Normalized free energies are considered herein; i.e., the energies are divided by the parameter $R \cdot T$, where T is 310 K. Consequently, all our equations yield dimensionless energy values. An example of such an equation is given below.

As in [30–33], the bonding energy in microstate ω is computed as

$$G_{\text{hel}}(\omega) = \frac{1}{RT}\sum_j E_{\bar{\gamma}_j}, \tag{1}$$

where j runs over all hypohelices of ω.

In the case of certain leader regions where attenuation has been predicted by their multiple sequence alignment and experimentally verified for some organisms involved in the alignment, Eq. (5), which is based on Eq. (1), yields an apparently erroneous result for the probability of a change in macrostate in our model: the probability of termination does not grow with increasing concentration (see Results of Model Computations and Discussion). In view of this, we propose a more general equation to be used in place of Eq. (1):

$$G_{\text{hel}}(\omega) = \frac{1}{RT}\sum_j \left(E_{\bar{\gamma}_j} - \alpha \frac{l'_j}{1 + \dfrac{l'_j}{l_{\max}}} \right). \tag{2}$$

Using Eqs. (5) and (2) at $\alpha > 0$, quite conceivable curves of termination probability were obtained with our model for the above leader regions. The additional summand (correction)

$$E(l') = -\alpha \frac{l'}{1 + \dfrac{l'_j}{l_{\max}}}$$

includes the parameters α and $l_{\max}$, where $l_{\max}$ is the length of loop l' such that $E(l')$ is equal to the half its asymptotic value. In our computations, the most typical estimates were $l_{\max} = 10$, $\alpha = 0$ (in many cases), and $\alpha = 5$–10 (in a few cases). The physical nature of the interaction responsible for this correction is unclear. It is possible to think about an additional energy of bonding between the RNA region realizing microstate ω and some stabilizing molecules or about an energy of the tertiary structure of this region, e.g., its pseudoknots and knots. The parameter α seems to be biologically significant. In particular, this is evident from our estimates of the cycle of changes in macrostate between two consequent transitions of the ribosome or polymerase. At $\alpha = 0$, the macrostate usually changes from tens to thousands of times in every such cycle, the number of changes reaching 50,000, or even more in some cases. The cycle grows appreciably shorter with increasing α. At $\alpha > 10$, the macrostate ceases to change at least once in every cycle, suggesting stabilization of the secondary structure between consecutive shifts of the ribosome or polymerase. It cannot be excluded that further refinement of the stacking and loop energies will obviate the need of this correction.

The loop energy of microstate ω is computed as

$$G_{\text{loop}}(\omega) = \sum_i (1.77\ln(l_i + 1) + B), \tag{3}$$

where i runs over all hypohelices of ω.

Numerical characteristics of potentially possible helices and states

Window size, nt	Mean helices with the arm length					Mean macrostates	Mean microstates
	3	4	5	6	>6		
10	0.09	–	–	–	–	0.09	0.09
20	1.93	0.55	0.17	0.05	–	2.91	2.92
30	5.48	1.73	0.60	0.38	0.27	17.35	18.31
40	11.02	3.73	1.54	0.68	0.69	105.2	116.4
50	17.89	6.16	2.92	1.00	1.27	501	578
60	24.91	8.42	4.35	1.27	1.80	1981	2325
70	32.15	10.82	5.69	1.59	2.16	8265	9887
80	39.56	13.00	7.04	1.92	2.44	33,713	40,801
90	53.55	17.55	9.18	3.09	3.73	219,097	284,627

This equation agrees well with the extended tables of the energies of all loops at all $l_i > 2$ [31, 33], provided that $B = 6.5$ for terminal loops, $B = 0$ for bilateral protrusions, and $B = 4$ for unilateral protrusions. The coefficient 1.77 (Flory parameter) has been justified in the theory of nonselfintersecting random migrations [34]. Cases where $l_i \le 2$ were considered separately, according to the tables available from [31, 33]. Namely, the following values were taken for the loop energy: –0.8 (at $l = 2$) for a bilateral protrusion and –6.2 (at $l =1$) or 4.5 (at $l = 2$) for a unilateral protrusion. Terminal loops of such lengths are excluded from our model.

Although Eq. (3) is part of a series similar to Edgeworth expansion, the available experimental data are insufficient for estimating its senior coefficients as yet.

Transitions between macrostates are classified as fast and slow. By definition, a fast transition occurs without changes in the corresponding macrostate. A slow transition is defined as a transition that changes the macrostate by exactly one pair of parentheses, that is, by ±1 chord. Generally speaking, any number of hypohelices can change upon any transition.

Exact probabilities (hereafter referred to as rates) of fast transitions from microstate ω to microstate ω' within one macrostate are unimportant for our model. It is assumed only that fast transitions within the total set of microstates ω in the given macrostate Ω lead to the Boltzmann–Gibbs steady-state probability distribution

$$p(\omega) = \frac{\exp(-(G_{\mathrm{loop}}(\omega) + G_{\mathrm{hel}}(\omega)))}{z(\Omega)}, \tag{4}$$

where $z(\Omega) = \sum_{w \in \Omega} \exp(-G_{\mathrm{loop}}(\omega) - G_{\mathrm{hel}}(\omega))$.

Slow transitions between microstates change the macrostate exactly by one chord. Their probabilities are described by Eqs. (5) and (6) [28, 29] in our model. The equations seem physically grounded to a certain extent: the degradation rate of a hypohelix is determined by its bonding energy, and the rate of hypohelix addition depends on how difficult it is to bring the arms of a future hypohelix close together.

Alternatively, slow transitions between microstates were described using Eq. (7). We cannot exclude that the equation for computing the slow transition probability should be chosen depending on the phylogenetic group (see Results of Model Computations and Discussion).

All equations, including Eqs. (5)–(7), and tabulated values are assessed by names and, consequently, are easy to change in our program. We were aimed, in particular, at designing a universal computer program so that the implemented model would follow the logic described here and any equations could be used for the regularities suggested. On request at lin@iitp.ru, we are ready to compute the termination probability $p(c)$ at any concentration c, using an original sequence, the list of equations numbered here, and the list of explicit parameter values mentioned here. If an equation or a parameter is not specified, its default value is used as described here.

When a slow transition consists in degradation of a hypohelix, that is, the macrostate is decreased by one chord upon transition from microstate $\omega = \{\bar{\gamma}_{1i}, \ldots, \bar{\gamma}_{ki}\}$ (where all hypohelices are indicated) to microstate $\omega' = \{\bar{\gamma}'_{1i}, \ldots, \bar{\gamma}'_{ki}\}$ (where all hypohelices are indicated, and $\bar{\gamma}'_{li} = \varnothing$ for particular l and i; i.e., hypohelix $\bar{\gamma}'_{li}$ is in fact absent from ω') with hypohelices $\bar{\gamma}_{li}$ and $\bar{\gamma}'_{li}$ belonging to one helix γ_l and correspond to one chord, the slow transition rate is defined as

$$K(\omega \longrightarrow \omega') = \kappa \exp(G_{\mathrm{hel}}(\omega) - G_{\mathrm{hel}}(\omega')). \tag{5}$$

When a slow transition consists in the addition of a hypohelix, that is, the macrostate is increased by one chord (the designations are as above), the reverse transition rate is

$$K(\omega' \longrightarrow \omega) = \kappa \exp(G_{\text{loop}}(\omega') - G_{\text{loop}}(\omega)). \quad (6)$$

In addition, we considered a variant other than Eqs. (5) and (6). The rate of either of the two slow transitions was obtained as

$$K(\omega \longrightarrow \omega') = \kappa \exp\left[\frac{1}{2}(G_{\text{loop}}(\omega) + G_{\text{hel}}(\omega)) - (G_{\text{loop}}(\omega') + G_{\text{hel}}(\omega'))\right]. \quad (7)$$

Compared with Eqs. (5) and (6), Eq. (7) is better suited to some operons of Gram-negative bacteria, including the tryptophan operons of *Escherichia coli* and some *Streptomyces* species (see Results of Model Computations and Discussion). Computations were performed in two variants for each leader region: one was based on Eqs. (5) and (6) and the other, on Eq. (7). The choice of one or several equations for describing the slow transition rates needs further systematic investigation by modeling and comparison with experimental findings.

We took $\kappa = 10^6$ s^{-1}, as recommended in [22–24]. However, interesting results can also be obtained with $\kappa = 10^4$ or 10^5 s^{-1}. It would be of interest and importance to experimentally estimate the correct value of κ, which is related to the viscosity of the cytoplasm.

Note that Eqs. (5)–(7) were selected so that the principle of detailed balance is obeyed:

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

where $E(\omega)$ is the energy ascribed to microstate ω in a particular variant. This explains the coefficient $\frac{1}{2}$ in Eq. (7).

When computations with Eqs. (5) and (6) were performed for *E. coli*, the amino acid concentration dependence obtained for the termination rate was apparently incorrect. At the same time, Eq. (7) allowed a conceivable result (Fig. 1). The results of the different modeling variants were comparable in other cases, although Eq. (7) produced better results as compared with Eqs. (5) and (6) in the case of *S. venezuelae.*

Considering the dynamics of a microstate on the basis of the dynamics of the microstates being realized, only two transitions are possible: a new hypohelix (chord) γ can be added to the current macrostate Ω or one of the existing hypohelices (chords) γ can disappear from state Ω. An apparent averaging over all microstate pairs $\omega \in \Omega$, $\omega' \in \Omega'$ yields the following equation for the transition rate from macrostate Ω to another macrostate Ω' regardless of whether the initial macrostate increases or decreases by one hypohelix:

$$K(\Omega \longrightarrow \Omega') = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega'} p(\omega) K(\omega \longrightarrow \omega').$$

We developed efficient algorithms for implementing individual components in our computer model. In particular, an algorithm was designed to compute the above sums without a search through all microstate pairs. The algorithms will be reported in forthcoming articles.

Our definitions of the fast and slow transition have a combinatorial support, which is provided by a statement proved elsewhere [28, 29]. It would be desirable to distinguish between fast and slow transitions in terms of constant rate, but such data are lacking as yet.

Statement 1. Let there be two microstates realizing one macrostate (which is equivalent to isomorphism of microstate trees: branches of a tree correspond to hypohelices, which are considered to be equivalent when belonging to one helix, and the order of direct descendants of each top is fixed and rendered isomorphic). Then, within one macrostate, a change from one microstate to another is possible via a chain of steps such that no more than two nucleotide pairs are disrupted or generated at each step. In contrast, when two microstates belong to different macrostates, such a chain is impossible between them.

Estimates of polymerase deceleration by a secondary structure formed in the mRNA region between the ribosome and RNA polymerase. A hairpin is defined as a chain of pairs of paired sections that are linearly arranged in each other's loops (i.e., the corresponding tree is linear), with small protrusions between neighbor pairs of sections and an arbitrary loop at the end of the chain of pairs. The first pair of sections is termed the stem of the hairpin. In a hairpin, every pair of paired sections (i.e., a hypohelix) has its loop, including all subsequent pairs of sections, protrusions, and loops. Note that a hairpin does not necessarily represent a microstate.

According to experimental data [34–36], the probability of termination as dependent on the terminator hairpin size is described by a curve known as a resonant curve in physics. Without discussing the physical process of interactions between a hairpin and polymerase, we used such a curve to describe the rate constant of polymerase jumps as dependent on the RNA secondary structure. The following naive comment can be made concerning Eq. (8). Polymerase has a positively charged region in a negatively charged surrounding, and this region is capable of Coulomb interactions with a negatively charged hairpin. This allows a picture to arise at a fixed distance r as described by

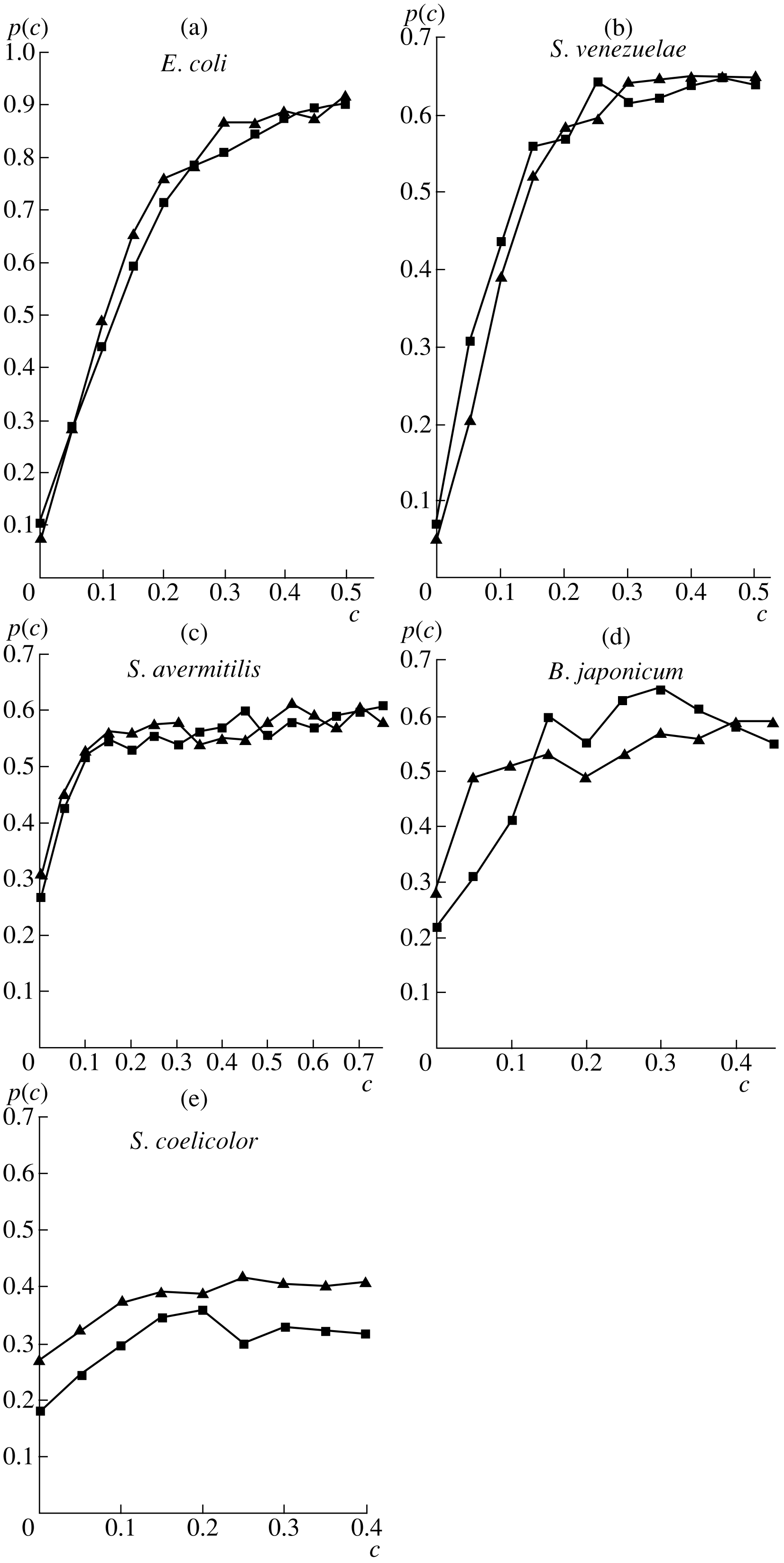


Fig. 1. Expression of *E. coli trp*E as dependent on tryptophan concentration in (a) *E. coli*, (b) *S. venezuelae*, (c) *S. avermitilis*, (d) *B. japonicum*, and (e) *S. coelicolor.* The estimates were obtained with Eq. (9) (squares) or Eq. (10) (triangles). $\alpha = 0$ (Figs. 1a–1d); $\alpha = 10$ (Fig. 1e). Abscissa, tryptophan concentration c; ordinate, termination probability $p(c)$.

Eq. (8) and its corollary (17): with increasing height h of the hairpin stem, polymerase deceleration force F generated by the hairpin increases to a certain maximum and then again decreases.

Thus, our model assumes that the force F of polymerase deceleration by hairpin ω has a meaning of an efficient decrease in the rate constant of polymerase movement along the DNA strand and is measured in s^{-1}, as determined by the following equation:

$$F(\omega) = \frac{\delta}{L_1^2(p-p_0)^2+1}\exp\left(-\frac{r}{r_0}\right), \qquad (8)$$

where r is the distance from the end D of hairpin ω to the start of polymerase. The parameters L_1, p_0, r_0, and δ depend only on the polymerase properties; their biological meaning is discussed below. The wave number p depends only on the hairpin. If the RNA region between the ribosome and polymerase includes several consecutive hairpins, which form hairpin set $\{\omega_i'\}$, the force $F(\omega)$ of the effect of this set on polymerase is computed as a sum of the forces generated by every particular hairpin ω_i'. In turn, the hairpin set $\{\omega_i'\}$ depends on the microstate ω, which is formed on the RNA region between the ribosome and polymerase according to the rule indicated below. Thus,

$$F(\omega) = \sum_i F(\omega_i'), \qquad (9)$$

where $F(\omega_i')$ and the corresponding r_i are computed from Eq. (8). Since Eqs. (8) and (9) include an exponent rapidly decreasing with distance, they are justified, to a certain extent, by summing with an exponentially decreasing weight, which is common in physics.

At the same time, special consideration is necessary for a telling argument suggesting that only one hairpin of the set $\{\omega_i'\}$ interacts with RNA polymerase. The principal difference is how to identify the exact hairpin involved. To consider this argument, we replaced Eq. (9) in our model and the program by the following equation, which includes the maximal value in place of the sum:

$$F(\omega) = \max_i\{F(\omega_i')\}. \qquad (10)$$

The summand corresponding to the ith hairpin is computed from Eq. (8), taking the distance r_i as established for the "stretched" sequence; i.e., all hairpins between hairpin i and polymerase are disregarded. On other words, to compute r_i, only two nucleotides, A_j and D_j, which correspond to the hairpin ends, are preserved from each hairpin j located between hairpin i and polymerase. Computations were performed in two variants for each leader region: according to Eqs. (9) and (10). As a comparison of the results showed, the final curve characterizing the termination probability as dependent on the amino acid concentration does not appreciably depend on whether variant (9) or (10) is chosen; yet such a dependence does occur in some cases. The results obtained with Eqs. (9) and (10) coincide because one of the summands is dramatically higher than any other summand in Eq. (9) for all biological sequences examined. The greatest summand corresponds to the most massive hairpin among all hairpins ω_i' sufficiently close to polymerase. Yet this summand cannot be determined formally, which explains the variant involving Eq. (10). The curves obtained with both variants are given in Results of Model Computations and Discussion.

We now determine the magnitude of the action of $F(\omega)$ from an arbitrary microstate ω, on RNA polymerase, reducing it to the determination of a certain hairpin set $\{\omega_i'\}$—a root of microstate ω. The diagram of microstate ω can be expanded in a single way into a chain of nonexpandable diagrams, each being identifiable by the presence of a stem, that is, the outermost pair of parentheses with the corresponding helix number. In this chain, hypohelix γ_i having the outer ends A_i and D_i and ascribed to the outmost parentheses, corresponds to the ith nonexpandable diagram (nonexpandable microstate).

By definition, hairpin ω_i' starts with hypohelix γ_i (i.e., with nucleotide pair $\langle A_i, D_i\rangle$) and continues along the RNA region located between A_i and D_i in the initial sequence according to nucleotide pairings in hairpin ω until a large protrusion (as defined by a certain threshold, which is strictly higher than 2 by default) or a branching appears in ω, while small protrusions existing in ω are left unchanged. The regions preceding this point are taken as the arms of hairpin ω_i', and the internal region is taken as the loop of hairpin ω_i'. Then, Eq. (8) is applied to hairpin ω_i'.

$F(\Omega)$ is determined as a mathematical expectation over all microstates ω realizing the given macrostate Ω:

$$F(\Omega) = \sum_{\omega\in\Omega} p(\omega)F(\omega). \qquad (11)$$

The rate constant of polymerase passing from a nucleotide to the subsequent one is determined from the following equation:

$$\nu(\Omega) = \bar{\lambda}_{\rm pol} - F(\Omega). \qquad (12)$$

It is assumed that $\delta < \bar{\lambda}_{\rm pol}$. In the case of Eq. (10), this means that $\nu(\Omega) > 0$. The same is true for all calculations in the case of Eq. (9).

A principal equation for computing the parameter p was derived earlier [28, 29], and only the result is given here. When a hairpin consists of a stem and a loop, p is obtained from the equation

$$\tan(ph) = \frac{2}{pl}, \quad 0 < ph < \frac{\pi}{2}, \tag{13}$$

where h is the stem height, that is, the number of paired nucleotides in the stem. When a hairpin consists of several paired sections with small protrusions between them and an arbitrary loop at the end, it is possible to assume the following. Let a hairpin contain s sections with heights $h_1, \ldots, h_s$; $s - 1$ protrusions between them with lengths $l_1, \ldots, l_{s-1}$; and a loop of length l. Then

$$p = \bar{p}\left(1 - \frac{1}{2h + l\sin^2(\bar{p}h)}\sum_{i=1}^{s-1} l_i \sin^2(\bar{p}h(i))\right), \tag{14}$$

where $h(i) = h_1 + \ldots + h_i$ by definition and, consequently, $h = h(n) = h_1 + \ldots + h_n$, and $\bar{p}$ can be obtained from an equation similar to Eq. (13):

$$\tan(\bar{p}h) = \frac{2}{\bar{p}l}, \quad 0 < \bar{p}h < \frac{\pi}{2}. \tag{15}$$

Since $0 < \bar{p}h < \frac{\pi}{2}$, the multiplier $\sin^2(\bar{p}h(i))$ increases monotonically with $h(i)$.

Progress of RNA Polymerase along the DNA Strand

Consider the situation where polymerase jumps from a nucleotide belonging to a T-rich region. When the 3' end of polymerase, hereafter designated as z, is on the nth nucleotide, polymerase can jump to the $(n + 1)$th nucleotide or leave the nucleotide sequence. A T-rich region is defined as follows. Nucleotide z is termed T-rich (within the region) if there is at least one word containing z in any position that exceeds a certain threshold (6 by default) in length and a certain threshold (0.8 by default) in T density. The word may contain exclusions (i.e., non-T) in any position including the termini, and z is not necessarily T. Let us form all intervals of maximal sizes in a set of T-rich nucleotides. Such intervals are termed T-rich regions and have no overlap.

From the position $z = n$ (ground position), polymerase can proceed to position $z = n + 1$ with the rate constant $\bar{\lambda}_{\text{pol}}$ or can undergo a transition into the excited state n^* with a certain rate constant. From the excited state, polymerase can fall off the sequence with the rate constant λ_{ur} or can undergo a reverse change to the ground state $z = n$ [35]. If the transitions between n and n^* are fast, this scheme of polymerase movements can be replaced by its average variant: let the transition from n to $n + 1$ take place with the constant $\beta \cdot \bar{\lambda}_{\text{pol}}$ and falling-off take place with the constant $(1 - \beta) \cdot \lambda_{\text{ur}}$, where β is the probability for polymerase to occur in the ground state and $(1 - \beta)$ is the probability for polymerase to occur in the excited state. Taking $\beta \cdot \bar{\lambda}_{\text{pol}} = \nu(\Omega) = \bar{\lambda}_{\text{pol}} - F(\Omega)$, we obtain $(1 - \beta) \cdot \bar{\lambda}_{\text{pol}} = F$; i.e., $(1 - \beta) = \frac{F}{\bar{\lambda}_{\text{pol}}}$. Eventually, the falling-off rate of polymerase is determined as

$$\mu = \mu_{\text{out}} = \frac{\lambda_{\text{ur}}F}{\bar{\lambda}_{\text{pol}}}. \tag{16}$$

Assuming that the above transitions are not necessarily fast, it is possible to derive a more sophisticated equation, which is still similar to Eq. (16). The ratio $\frac{\bar{\lambda}_{\text{pol}}}{\lambda_{\text{ur}}}$ is a natural parameter and is equal to 4 according to other authors [35].

This scheme was applied to numerically estimate the parameters L_1, p_0, and δ. When a hairpin consists of one stem with a negligibly small loop, $p = \frac{\pi}{2h}$ and

$$F(h) = \frac{\delta}{L_2^2\left(\frac{1}{h} - \frac{1}{h_0}\right)^2 + 1}\exp\left(-\frac{r}{r_0}\right), \tag{17}$$

where $L_2 = \frac{\pi}{2} \cdot L_1$ and $h_0 = \frac{\pi}{2p_0}$. Obviously, the probability that polymerase passes from nucleotide n to nucleotide $(n + 1)$ and does not fall off is $\frac{\nu}{\nu + \mu}$. Thus, the probability differs from unity only in a T-rich region and only in the presence of hairpins (i.e., when $\mu > 0$, or, equivalently, $F > 0$). The probability for polymerase to perform N movements along a T-rich region without falling off it is apparently

$$\left(\frac{\nu}{\nu + \mu}\right)^N. \tag{18}$$

These equations can be applied to experimental data on the termination rate as dependent on the stem height (i.e., the number of nucleotide pairs h in the case of an 8-nt uracil tract, $N = 7$) in *E. coli* [35–37]: $\langle 3; 0.2\rangle$, $\langle 7; 0.8\rangle$, and $\langle 14; 0.2\rangle$. In these pairs, the first value is the stem height and the second one is the termination rate. In the approximation where F is a function of h and $r = 0$, F depends on three parameters: h_0, L_1, and δ. Hence, we have a system of three nonlinear equations in three unknowns. Its solution yields

$$h_0 = 7, \quad L_1 = 14.5, \quad \text{and} \quad \delta = 25,$$

whence it follows that $p_0 = \frac{\pi}{14}$.

These numerical values are only tentative and need refinement; in particular, the phylogenetic group of the organism should be taken into account.

The r_0 value is chosen to be similar to the RNA polymerase dimension from the RNA exit site to the transcription site because statistical evaluation of r_0 requires additional data.

Progress of the Ribosome along the mRNA Strand

On nonregulatory codons, the rate constant $\bar{\lambda}_{\text{rib}}$ of ribosome shifting by one codon is taken as $\lambda_{\text{rib}} = \bar{\lambda}_{\text{rib}} = 15\ \text{s}^{-1}$. On regulatory codons, the rate constant depends on the concentration c of the corresponding amino acid according to the Michaelis–Menten equation:

$$\lambda_{\text{rib}}(c) = \frac{\bar{\lambda}_{\text{rib}}c}{c_0 + c}, \qquad (19)$$

where c is the charged tRNA concentration, c_0 is the charged tRNA concentration such that the ribosome moves on regulatory codons at the half-maximal rate (the maximal rate is $\bar{\lambda}_{\text{rib}} = 15\ \text{s}^{-1}$), and $\bar{\lambda}_{\text{rib}}$ is the value of this function at concentration c so high that the ribosome moves on regulatory and nonregulatory codons at the same rate.

Since the model should allow comparison with experimental data and it is unclear how to experimentally estimate c_0, we used the following approach. The amino acid concentration dependence of the charged tRNA concentration is similarly described by the Michaelis–Menten equation. Substituting this equation in Eq. (19) yields a similar equation, where c is the amino acid concentration in the cell. Yet the concentration is measured in a culture, rather than within a cell, in experiments. Hence, another substitution is used to obtain the same Eq. (19), where c is the amino acid concentration in the culture. The corresponding c_0 is the Michaelis–Menten parameter reflecting these two processes: the effect of the amino acid concentration in the culture on the amino acid concentration in the cell and the effect of the amino acid concentration in the cell on the charged tRNA concentration, which, in turn, affects the ribosome movement rate on regulatory codons. Thus, the constant c_0 lacks a direct biological interpretation. The above reasoning implies that aminoacyl-tRNA synthases and tRNAs occur in sufficient amounts. In these three cases, $\bar{\lambda}_{\text{rib}}$ is the same and coincides with the rate of translation of nonregulatory codons.

Ribosome Landing on the Shine–Dalgarno (SD) Sequence

As soon as the SD sequence, the start codon (atg or gtg) of the leader peptide sequence, and additional $s_0 + s_1$ nucleotides (the distance between the ribosome P site and the transcription point) are transcribed, it is possible for the ribosome to bind mRNA. Let us assume (keeping in mind that this reasoning only helps to construct the equation) that in the complex, the ribosome and charged tRNA act as two arms to bind, respectively, the SD sequence and the start codon, and that the mRNA region including the SD sequence and the stop codon occurs in macrostate Ω. Hence, transitions are possible between states $\langle\Omega, \textit{freerib}\rangle$ and $\langle\Omega, \textit{boundrib}\rangle$. The rate constants are designated as K_{in} for the direct transition (left to right) and K_{out} for the reverse transition. According to the common rule,

$$K_{\text{in}} = \sum_{\omega \in \Omega} \sum_{\omega' \in \langle\Omega,\, \textit{boundrib}\rangle} p(\omega)K(\omega \longrightarrow \omega'), \qquad (20)$$

where $K(\omega \longrightarrow \omega') = \kappa_{SD}\exp(G_{\text{loop}}(\omega) - G_{\text{loop}}(\omega'))$ and $\kappa_{SD} = 10\ \text{s}^{-1}$. It seems possible to assume that $G_{\text{loop}}(\omega) = G_{\text{loop}}(\omega')$; then, $K(\omega \longrightarrow \omega') = \kappa_{SD}$ if at least three consecutive nucleotides of the SD sequence are open and the start codon is completely open in microstate ω. Otherwise, $K(\omega \longrightarrow \omega') = 0$.

The reverse transition is characterized by the rate constant $K_{\text{out}} = \sum_{\omega' \in \langle\Omega,\, \textit{boundrib}\rangle} \sum_{\omega \in \Omega} p(\omega')\, K(\omega' \longrightarrow \omega)$, where $K(\omega' \longrightarrow \omega) = \kappa\exp(G_{\text{hel}}(\omega') - G_{\text{hel}}(\omega))$ and $\kappa = 10^6\ \text{s}^{-1}$ is the standard closing rate constant. State $\langle\Omega, \textit{boundrib}\rangle$ allows a transition to state $\langle\Omega, \textit{freerib}\rangle$ or a shift of the ribosome to the first (after the start) codon. As a result of this shift, the ribosome can no longer be released (until it reaches the stop codon of the leader peptide sequence and momentarily falls off), initiation is completed, and transitions are considered that are characteristic of stable movement, when the ribosome and polymerase are both associated with an mRNA strand.

If we neglect multiple events of the ribosome landing on and falling off the mRNA strand, it suffices to compute the mathematical expectation of the time between transcription of the start codon and the transition of the ribosome on the first codon following the start codon (bound state). This time is

$$T = \frac{K_{\text{in}} + K_{\text{out}} + \bar{\lambda}_{\text{rib}}}{K_{\text{in}}\bar{\lambda}_{\text{rib}}}.$$

Hence, it is possible to state that, after transcription of the start codon (and additional $3 + s_0 + s_1$ nucleotides), a step is possible that consists in the binding of the ribosome P site with the first codon following the start codon. The rate constant of this process is

$$K_{\text{in}}^{\text{eff}} = \frac{K_{\text{in}}\bar{\lambda}_{\text{rib}}}{K_{\text{in}} + K_{\text{out}} + \bar{\lambda}_{\text{rib}}}.$$

After this, the ribosome progresses along the mRNA strand until the stop codon. Modeling of the ribosome landing is possible only when the transcription start site is known, because the secondary structure of the region including the SD sequence is assessable only in such cases.

DESCRIPTION OF THE MODELING SCHEME

In the case of classical attenuating regulation, the objective of modeling was to numerically evaluate the probability of termination $p = p(c)$ as dependent on the concentration c of charged tRNA (or the amino acid in the cell) for various biological regulatory regions; several other dependences related to $p(c)$ were also analyzed. To construct the $p = p(c)$ dependence, the process considered in our model was iterated a specified number of times (e.g., 10^3–10^4 times, which yielded much the same results) for every c taken from a grid with a particular pitch between knots, and $p = p(c)$ was computed as the portion of cases where termination takes place. The parameter c_0 in Eq. (19) was taken as unity; i.e., c_0 was a unit of measure on the axis c. The parameter r_0 was selected from an interval of 2–8.

There are some experimental characteristics that can be compared with the modeling results. For instance, the ratio has been estimated for the probabilities $p = p(c)$ observed at rather high and rather low concentrations. Plots have been reported for the activity of an enzyme (e.g., anthranyltransferase) as dependent on the amino acid concentration (e.g., the tryptophan concentration in the culture [37]). A transition from relative units, which are commonly accepted on the axes of $p(c)$ plots, to physical units of measuring the activity and concentration requires separate investigation. Such plots were compared qualitatively.

For the initial fixed RNA sequence, the current state of the model is characterized as follows.

(1) There is a window between the ribosome 3' end x and the polymerase start y. The dimension of the ribosome from the P site to the 3' end is designated s_0 (10–12 nt; 12 nt by default); the dimension of polymerase from y, where the RNA strand is released, to the transcription point is designated s_1 (2–7 nt; 5 nt by default). The transcription point is designated z, and it is always true that $z = y + s_1$. Within the window, the secondary structure changes from one macrostate Ω to another macrostate Ω'. The macrostates include only helices that overlap the window by both arms or at least by three nucleotides; i.e., the point in question is the macrostate in the window (for a current window).

(2) There is list T of (potential) helices overlapping the window by both arms (at least by the minimal hypohelix length, that is, 3 nt). This is a trivial component of the state in the sense that it is easy to compute anew for each state, having the initial set of helixes.

(3) There is macrostate Ω, which is also known as a nonempty diagram or a secondary structure in the windows.

There is no window until polymerase lands (empty macrostate). When polymerase has landed and the ribosome still has not, the window starts at the first nucleotide of the initial sequence (point 0) and ends at the current position of the start of polymerase. A nonempty macrostate Ω consisting of one chord can arise in the window for the first time. Then, another chord can be added to the first one or the macrostate can reverse to the initial empty state, and so on.

One of two possible outcomes is tracked in modeling: (1) polymerase falls off at one of the nucleotides belonging to the polyuracil tract of the initial sequence or (2) polymerase proceeds throughout the polyuracil tract.

Initiation of the RNA regulation process (from the landing of polymerase to the landing of the ribosome).

(1) Polymerase binds to the promoter and makes several steps to achieve the start of the leader peptide sequence according to the common rules.

(2) As soon as polymerase has transcribed the start of the leader peptide sequence and additional $s_0 + s_1$ nucleotides, the ribosome attempts to bind to the SD sequence with the rate constant reflecting the dependence on the quality of this sequence and the secondary structure covering it. Immediately after the binding, the ribosome is positioned at the start of the leader peptide sequence. Two parameters are fixed at this moment: the left end x of the window at the start of the leader peptide sequence + s_0 and the right end y of the window at the position currently occupied by the start of polymerase.

Transitions during RNA Regulation after the Formation of Window [x, y]

(1) Polymerase shifts rightward by 1 nt, which increases the window by 1 nt and may extend the helix list T. Alternatively, polymerase falls off at a T-rich region.

(2) The ribosome shifts rightwards by one codon. The window decreases by 3 nt; generally speaking, the helix list T is reduced; and macrostate Ω changes. The leftmost parenthesis (with the appropriate right parenthesis) is excluded from the diagram of Ω if the corresponding helix is beyond the new list T. The resulting new macrostate Ω, which can be empty, is fixed in the current window.

(3) The secondary structure is rearranged; i.e., the macrostate in the window changes. The window and the helix list T remain the same.

Completion of Modeling

If polymerase falls off at the polyuracil tract, the modeling is terminated. Otherwise, modeling ends when polymerase proceeds throughout the polyuracil tract. If the ribosome fails to shift during a particular transition, it is possible to fix the period of time before transition. Such periods are summed throughout the time until the first shift of the ribosome. The distribution of the periods contains useful information.

Organization of Transitions during the Modeling

The modeling utilizes the Monte Carlo procedure in the standard mode. The state is described by the set $\langle x, y, z, T, \Omega\rangle$. The parameter ς which reflects whether the ribosome has landed or not, is included in the description for the period of initiation and may be omitted afterwards. The surrounding of the given state Ω (with the center in Ω) is defined as a set of all states the transitions from Ω to which are possible with a nonzero probability. If the surrounding includes n states and the corresponding transition rate constants are, respectively, $k_1, \ldots, k_n$ (let $k = \sum k_i$), the state resulting from a transition (such a state is considered to be the next state on the given trajectory) is defined as a realization of the random parameter $i \longrightarrow \frac{k_i}{k}$. At certain steps in modeling, the time before transition is determined as the time it takes to realize the random parameter $t \longrightarrow ke^{-kt}$. It should be noted that the values of λ_{SD}, λ_{rib}, λ_{pol}, and $K(\Omega \longrightarrow \Omega')$ often differ considerably in the order of magnitude.

Representation of data in the program implementing our model was a rather nontrivial challenge. It is necessary for the efficient realization of the above scheme of transitions that the program keep not only the set $\langle x, y, z, T, \Omega\rangle$, but also the total set of microstates possible for the current window (or at least for the surrounding of macrostate Ω) together with the corresponding sets of all possible microstates. As computations showed, 80% or an even greater portion of all helices have an arm size of 3–4 nt in any sequence. Hence, the vast majority of potential macrostates each contain one microstate, while cases of several microstates occurring in one macrostate are rare. As an example, the qualitative characteristics of typical helices and potential secondary structures are given in the table, which summarizes the results of analyzing the leader region upstream of the *S. avermitilis* MA-4680 tryptophan operon with the use of a sliding window of fixed width with subsequent averaging of the tabulated parameters over all different positions of the window.

It is clear that the sizes of the macrostate and microstate sets grow rapidly with the increasing width of the window and approaches 1,000,000 at a width of 80–100 nt, while at least 70% of macrostates each include only one microstate. In view of this, the description of data in our program is based on the set of all microstates possible in the current window, and the microstates are then grouped in macrostates by similarity of diagrams. We used a four-level combined linear-list data structure (Fig. 2).

The structure was constructed anew for each change of the window in the early versions of the program, which was quite acceptable for windows up to 40–50 nt in width but dramatically increased the computation time with wider windows. In view of this, in the final version, the current sets of macrostates and microstates are not constructed anew after the right or the left margin of the window shifts. Rather, starting from a certain threshold width of the window, the sets are reconstructed from the existing sets. Although more sophisticated algorithmically, this procedure considerably improves the efficiency of the program. As computations showed, the Monte Carlo procedure with 1000 iterations performed for every concentration c yields a curve that does not change with increasing number of iterations.

RESULTS OF MODEL COMPUTATIONS AND DISCUSSION

An initial sequence started from the SD sequence of the leader peptide region and ended with the end of the polyuracil tract of the transcription terminator.

As an example, we describe here the results of computations performed for the anthranylate synthase genes of three *Streptomyces* species (*S. venezuelae* ISP5230, *S. avermitilis* MA-4680, and *S. coelicolor* A3(2)). An alignment of their 5'-untranslated regions is shown in Fig. 3 [16]. In addition, the results of computations are described for the *trp*E anthranylate synthase gene of α-proteobacteirum *Bradyrhizobium japonicum* and γ-proteobacterium *E. coli.* Mass computations will be considered in a forthcoming article.

Analysis of the model and numerical computations showed that L_1, r_0, and α are the most critical parameters of the model. Computations were performed varying these parameters within the ranges specified above. For F, we used $\delta = 25$. In the cases illustrated in Figs. 1a and 1b, we used $L_1 = 14.5$, $p_0 = 0.167$, $r_0 = 2$, the variant with Eq. (7), and $\alpha = 0$. In the cases shown in Figs. 1c–1e, we took $L_1 = 10$, $p_0 = 0.12$, $r_0 = 5$, and the variant with Eqs. (5) and (6); the α values are given in the figure legends. Both variants of computing the force F—with Eq. (9) and with Eq. (10)—are shown in Fig. 1.

The sizes of the ribosome and polymerase did exert an effect, but this effect was appreciably weaker and the same for all organisms and genes considered. This finding made it possible to select common $s_0 = 12$ and $s_1 = 5$ for all these genes. As for p_0, its value should be

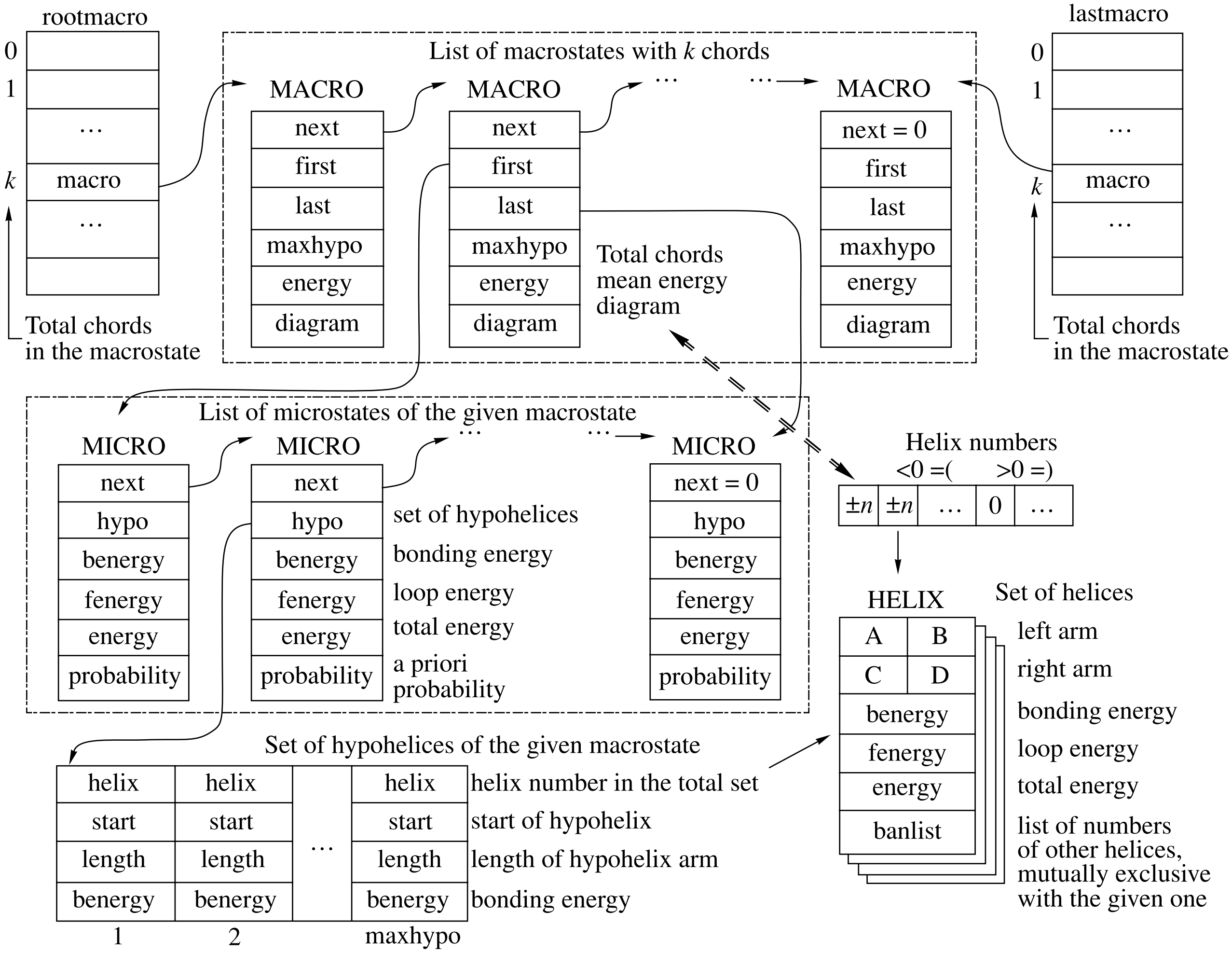
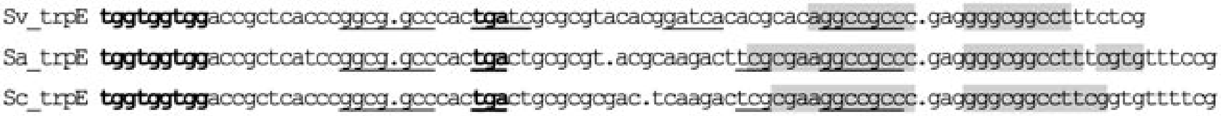


Fig. 2. Structure of data characterizing one state in our model.



```
Sv_trpE tggtggtggaccgctcacccggcg.gcccactgatcgcgcgtacacggatcacacgcacaggccgccc.gaggggcggccttctcg
Sa_trpE tggtggtggaccgctcatccggcg.gcccactgactgcgcgt.acgcaagacttcgcgaaggccgccc.gaggggcggcctttcgtgtttccg
Sc_trpE tggtggtggaccgctcacccggcg.gcccactgactgcgcgcgac.tcaagactcgcgaaggccgccc.gaggggcggccttcggtgttttcg
```

Fig. 3. Alignment of *trp*E 5'-untranslated regions of *Streptomyces* species *S. venezuelae* ISP5230, *S. avermitilis* MA-4680, and *S. coelicolor* A3(2). The regulatory and stop codons are in bold; the antiterminators are underlined; the terminators are shaded.

important in principle, but computations showed that p_0 only slightly affects the character of the dependence and mostly shifts the region of termination probabilities.

The results of computations with Eqs. (9) and (10) and the above parameters are plotted for each gene in Fig. 1, with the ordinate showing $p(c)$. In the cases of *S. venezuelae, S. avermitilis, B. japonicum,* and *E. coli*, we used $\alpha = 0$. The case of *S. coelicolor* provides a different example: computations with $\alpha = 0$ yielded a strongly descending plot of the termination probability $p(c)$, and we were forced to use the correction to Eq. (2) with $\alpha = 10$. It was difficult to expect that the model would not work in this case while working with the two other *Streptomyces* species, because the multiple sequence alignment (Fig. 3) showed that attenuating regulation is equally possible in all three cases. Moreover, such regulation has been proved experimentally in the case of *S. venezuelae.*

In all five cases the following was observed: at low concentrations of tryptophan, the rate of termination increased as the tryptophan concentration increased, while at large concentrations, there occurred saturation and a flattening of the curve. The difference between the results obtained with Eqs. (9) and (10) was insignificant for *E. coli, S. venezuelae*, and *S. avermitilis* (Fig. 1). In the case of *B. japonicum*, Eq. (9) seemed more preferable than Eq. (10). In the case of *S. coelicolor*, Eq. (9) was preferable at low concentrations. At higher concentrations, the transcription termination rate started to decrease, while a near monotonic growth of $p(c)$ was obtained with Eq. (10).

To summarize, we proposed a model of attenuating regulation and implemented it in a computer program. The model is based on explicit concepts (as concerns the description of dependences and the selection of parameter values) that are accessible for analysis and strongly formulated. Computations performed for biological examples with our model yielded results that qualitatively agreed with experimental findings. A method was proposed for specifying the parameters on the basis of the initial data. The computer program allows a wide variation of the dependences and parameters used in the model. Based on computations, we observed that the model is more sensitive to some parameters and relatively resistant to certain others, and we obtained numerical estimates both for biologically informative parameters and for parameters that are artifactual and are related to the Monte Carlo procedure. We determined several numerical characteristics that are internal but significant for any model in the field: the typical arm size, the ratio between the numbers of microstates and macrostates, the cycles between two consecutive transitions of the ribosome and polymerase, etc.

Sequestration (that is, blocking by an RNA hairpin) of the ribosome-binding site is reflected in our model. However, this issue will be discussed separately, as well as allowances for the effect of protein–DNA transcription regulation, i.e., the binding of a repressor or activator protein to DNA close to the promoter. With our model, we intend to predict the effect of point mutations altering the regulatory regions on attenuation, including prediction of the evolutionary stability of organisms. Then the model will be included into a wider model of the regulation of gene expression and metabolism in bacteria. In addition, the model has another application. Attenuating regulation is usually predicted on the basis of multiple sequence alignment, which requires several sequences. With our model, an individual sequence can be analyzed with proper parameters to generate a concentration–enzymatic activity curve. The curve characteristic of attenuation or its absence provides an additional argument for the presence or absence of attenuation.

ACKNOWLEDGMENTS

We are grateful to A.A. Mironov for many explanations and the formulation of the problem, M.S. Gelfand for the opportunity to report this work at his workshop, and K.Yu. Gorbunov for advice and help in computations.

This work was supported by the ISTC foundation (grant no. 2766).

REFERENCES

1. Henkin T.M., Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: How RNA provides instructions for transcription termination/antitermination decisions. *Bioessays*. **24**, 700–707.
2. Grundy F.J., Henkin T.M. 2003. The T box and S box transcription termination control systems. *Front. Biosci*. **8**, 20–31.
3. Grundy F.J., Henkin T.M. 2004. Regulation of gene expression by effectors that bind to RNA. *Curr. Opin. Microbiol*. **7** (2), 126–131.
4. Mandal M., Breaker R.R. 2004. Gene regulation by riboswitches. *Nature Rev. Mol. Cell. Biol*. **5**, 451–463.
5. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. 2004. Riboswitches: The oldest mechanism for the regulation of gene expression? *Trends Genet*. **20**, 44–50.
6. Yanofsky C. 2004. The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis. Trends Genetics*. **20** (8), 367–374.
7. Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. 2001. Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. *J. Mol. Microbiol. Biotechnol*. **3**, 529–543.
8. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. 2004. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: Comparative genomics analysis. *FEMS Microbiol. Lett*. **234**, 357–370.
9. Grundy F.J., Henkin T.M. 1994. Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in Gram-positive bacteria. *J. Mol. Biol*. **235**, 798–804.
10. Grundy F.J., Henkin T.M. 1998. The S box regulon: A new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Mol. Microbiol*. **30**, 737–749.
11. Murphy B.A., Grundy F.J., Henkin T.M. 2002. Prediction of gene function in methylthioadenosine recycling from regulatory signals. *J. Bacteriol*. **184**, 2314–2318.
12. Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. 2003. Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria. *FEMS Microbiol. Letts*. **222**, 211–220.
13. Sudarsan N., Barrick J.E., Breaker R.R. 2003. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*. **9**, 644–647.
14. Rodionov D.A., Vitreschak A.A., Mironov A.A., Gelfand M.S. 2003. Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes. *J. Biol. Chem*. **277**, 48949–48959.
15. Henkin T.M., Glass B.L., Grundy F.J. 1992. Analysis of the *Bacillus subtilis tyrS* gene: Conservation of a regula-

tory sequence in multiple tRNA synthetase genes. *J. Bacteriol.* **174**, 1299–1306.

16. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. 2005. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology.* **5**, 54.
17. Barrick J.E., Corbino K.A., Winkler W.C., Nahvi A., Mandal M., Collins J., Lee M., Roth A., Sudarsan N., Jona I., Wickiser J.K., Breaker R.R. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA.* **101**, 6421–6426.
18. Abreu-Goodger C., Ontiveros-Palacios N., Ciria R., Merino E. 2004. Conserved regulatory motifs in bacteria: Riboswitches and beyond. *Trends Genet.* **20** (10), 475–479.
19. Vitreschak A.A., Rodionov D.A., Mironov A.A., Gelfand M.S. 2002. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* **30**, 3141–3151.
20. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. 2003. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA.* **9**, 1084–1097.
21. Singer M., Berg P. 1998. *Genes and Genomes*. Moscow: Mir.
22. Mironov A.A., Kister A.E. 1985. Theoretical analysis of secondary RNA structure formation kinetics in the course of transcription and translation: Account of defective helices. *Mol. Biol.* **19**, 1350–1357.
23. Mironov A.A., Kister A.E. 1989. Theoretical analysis of structural rearrangements in the course of secondary RNA structure formation. *Mol. Biol.* **23**, 61–71.
24. Mironov A.A., Lebedev V.F. 1993. A kinetic model of RNA folding. *BioSystems.* **30**, 49–56.
25. Elf J., Ehrenberg M. 2005. What makes ribosome-mediated transcriptional attenuation sensitive to amino acid limitation? *PLoS Comput. Biology.* **1** (1), e2.
26. Xayaphoummine A., Bucher T., Thalmann F., Isambert H. 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA.* **100**, 15310–15315.
27. Xayaphoummine A., Bucher T., Isambert H. 2005. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.* **33** (Web Server issue), W605-10.
28. Pirogov S.A., Gorbunov K.Yu., Lyubetsky V.A. 2005. Macro- and microstates in the attentuation model of gene expression regulation in bacteria. *Trudy 7 Mezhd. Konf. "Problemy upravleniya i modelirovaniya v slozhnykh sistemakh"* (Proc. 7th Int. Conf. "Problems of Control and Modeling in Complex Systems"), Samara: Ross. Akad. Nauk, 210–216.
29. Lyubetsky V.A., Pirogov S.A. 2005. The model of attenuation regulation in bacteria. *Trudy 7 Mezhd. Konf. "Problemy upravleniya i modelirovaniya v slozhnykh sistemakh"* (Proc. 7th Int. Conf. "Problems of Control and Modeling in Complex Systems"), Samara: Ross. Akad. Nauk, 205–210.
30. Mathews D.H., Sabina J., Zuker M., Turner D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
31. Mathews D.H., Disney M.D., Childs J.L., Schroeder S.J., Zuker M., Turner D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA.* **101**, 7287–7292.
32. Dima I., Hyeon C., Thirumalai D. 2005. Extracting stacking interaction parameters for RNA from the data set of native structures. *J. Mol. Biol.* **347**, 53–69.
33. RNA Structure, Turner Lab, http://rna.chem.rochester.edu.
34. Lawler G.F., Coyle L.N. 1999. *Lectures on Contemporary Probability*, AMS.
35. Yin H., Artsimovitch I., Landick R., Gelles J. 1999. Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules. *Proc. Natl. Acad. Sci. USA.* **96**, 13124–13129.
36. Wilson K., von Hippel P. 1995. Transcription termination at intrinsic terminators: The role of the RNA hairpin. *Proc. Natl. Acad. Sci. USA.* **92**, 8793–8797.
37. Lynn S., Kasper L., Gardner J. 1988. 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.
38. Lin Cong, Paradkar A.S., Vining L.C. 1998. Regulation of an anthranilate synthase gene in *Streptomyces venezuelae* by a *trp* attenuator. *Microbiology.* **144**, 1971–1980.