ISSN 0032-9460, Problems of Information Transmission, 2008, Vol. 44, No. 4, pp. 333–351. © Pleiades Publishing, Inc., 2008.
Original Russian Text © V.A. Lyubetsky, E.A. Zhizhina, L.I. Rubanov, 2008, published in Problemy Peredachi Informatsii, 2008, Vol. 44, No. 4, pp. 52–71.

LARGE SYSTEMS

Gibbs Field Approach for Evolutionary Analysis of Regulatory Signal of Gene Expression

V. A. Lyubetsky[1], E. A. Zhizhina[2], and L. I. Rubanov[1]

Kharkevich Institute for Information Transmission Problems, RAS, Moscow
lyubetsk@iitp.ru zhizhina@iitp.ru rubanov@iitp.ru

Received February 12, 2008; in final form, April 22, 2008

Abstract—We propose a new approach to modeling a nucleotide sequence evolution subject to constraints on the secondary structure. The approach is based on the problem of optimizing a functional that involves both standard evolution of the primary structure and a condition of secondary structure conservation. We discuss simulation results in the example of evolution in the case of classical attenuation regulation.

DOI: 10.1134/S0032946008040066

1. INTRODUCTION AND STATEMENT OF THE PROBLEM

The problem of reconstructing the evolution of a set of related species or genes (proteins) based on the current members of that set is well known and has long been studied (see, e.g., [1–3]). The evolution is described by a *phylogenetic tree*, which defines related states of the evolutionary process and represents series of evolutionary events leading from an ancestral sequence at the tree root to given extant sequences at the tree leaves. The problem of reconstructing the evolution is commonly posed as one of two statements: either a phylogenetic tree itself is constructed along with ancestral sequences at all inner nodes, or the tree is assumed to be known and only the ancestral sequences at the inner nodes are sought for. Both ancestral and current sequences are composed of letters in a certain alphabet. We here consider sequences in the four-letter alphabet {A, C, T, G} and call these letters nucleotides.

According to modern conceptions, genes play a main role in the species genome, and an equal role is played by particular regions of the genome usually located upstream from the genes. Such a region can enable and sustain a sufficiently high level of functioning of the corresponding gene ("expressing" the gene), or it can disable the gene operation or, more exactly, reduce the level of functioning of the gene ("non-expressing" the gene). This region is called a regulation site or regulatory domain. Expressing or non-expressing a gene are two alternate states of a regulation site. The former is called antitermination (gene expression is present), and the latter is called termination (gene expression is absent). These states are realized by special complex mechanisms involved separately or in combination. We consider one of these mechanisms, called classical attenuation regulation. This mechanism is described in [4] at the biological level and in [5] in rigorous mathematical terms. Classical attenuation regulation was first predicted in [6], and a fundamentally important step toward modeling it was taken in [7].

Although this paper does not cover the attenuation regulation mechanism itself, we recall it in the biological context using classical attenuation regulation as an example. It is the type of regulation

[1] Supported in part by the International Science and Technology Center, project no. 3807.
[2] Supported in part by the Russian Foundation for Basic Research, project nos. 07-01-92216-NCNIL-a and 08-01-00105-a, and the U.S. Civilian Research and Development Foundation, grant no. RUM1-2693-MO-05.

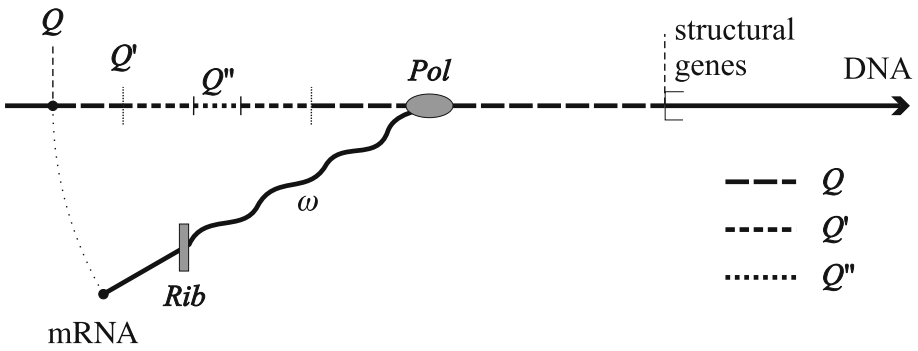

Fig. 1. Mechanism of classical attenuation regulation. The RNA polymerase *Pol* transcribes the sequence Q upstream from the structural genes if possible. The ribosome *Rib* translates the leader peptide gene Q'. The *Rib* motion at the regulatory codons Q'' is controlled by the concentration of the regulated amino acid. The mRNA secondary structure ω between *Rib* and *Pol* decelerates *Pol* and occasionally tears it off the sequence Q. If *Pol* reaches the structural genes, then they are expressed, i.e., transcribed in turn and then translated. We use the same symbols Q, Q', and Q'' on both DNA and mRNA for the respective whole sequence upstream from the structural genes, leader peptide gene, and regulatory codons.

in Examples 1–3 below. According to the basic doctrine of molecular biology, DNA information is read in two stages: first, the template RNA (mRNA, a single-strand counterpart of DNA, a nucleotide chain carrying the information) is synthesized; second, the protein is synthesized on the mRNA. The mRNA synthesis is called transcription and is realized by a molecular mechanism called RNA polymerase. Protein synthesis on mRNA is called translation and is realized by a molecular mechanism called the ribosome. During protein synthesis, the ribosome reads three nucleotides (a codon) at a time and appends an amino acid to the growing protein chain in accordance with that codon (see Fig. 1). A codon is a nucleotide triplet corresponding to the amino acid under a universal law; several codons often encode the same amino acid. Regulation of the gene expression (protein synthesis output) depends on external conditions and occurs at several levels. Here we consider only one type of regulation, attenuation regulation.

Attenuation regulation is based on the possibility of forming alternative secondary structures such that one structure allows protein synthesis and the other prohibits it. The structure is folded as follows. Nucleotides of the mRNA, even if widely separated on the chain, can be coupled by adjacent twins of complementary pairs from the fixed list of possible pairs. Thus, the nucleotides are coupled by quadruples. The list of complementary pairs usually comprises G-C and T-A pairs; the G-T pair is included with some constraints. Each letter of the sequence can participate in only one pair at a time. The pairing is a kind of hydrophobic bond called stacking; a hydrogen bond of the paired nucleotides is also sometimes considered (all these notions are formulated in detail, e.g., in [4, 5]). One method for calculating the energy of a helix formed by stacking was proposed in [5].

A group of consecutive complementary pairs makes a helix. The two segments of the site that make a helix are called helix shoulders. The region between two shoulders is called a loop. The helix shoulders are not necessarily contiguous regions and may include a few unpaired nucleotides; such places are called bulges. One-sided and two-sided bulges are distinguished; the latter are also called internal loops. As a result of the nucleotide pairing, a set of helices is formed in the mRNA, the so-called secondary structure. One mRNA macromolecule often allows multiple alternative secondary structures to be formed, including the antiterminator and terminator helices A and T (see Fig. 2). The helix A is formed by pairing two segments (denoted by A1 and A2 in Fig. 2a) within a regulation site, and the helix T is formed by pairing another two segments (denoted by T1 and T2

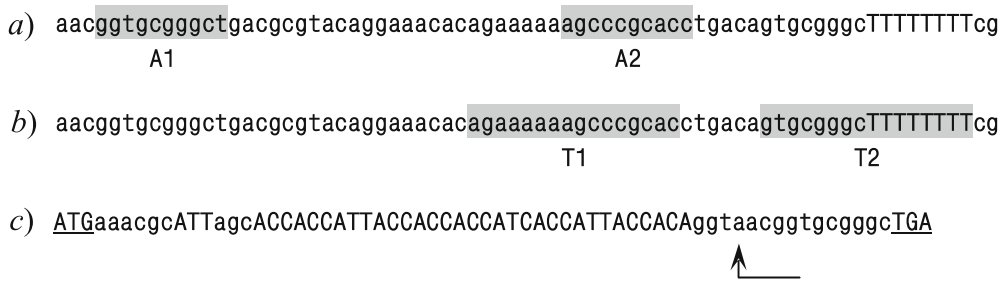

Fig. 2. Example of the regulatory signal for classical attenuation regulation: (a) A sequence with two shoulders of the antiterminator helix marked; the A state of the regulation site. (b) The same sequence with two shoulders of the terminator helix marked; the T state of the regulation site (the T-run is shown in capital letters). (c) The sequence of the leader peptide gene that contains start and stop codons (shown as underlined capital letters) and 12 regulatory codons (shown as capital letters). The sequences (a) and (b) continue the sequence (c) starting from the position marked by the arrow.

in Fig. 2b) within that site. It is important that the shoulders A2 and T1 substantially overlap because the RNA secondary structure therefore cannot be in the A and T states simultaneously: each nucleotide can participate in only one pair. Consequently, only one signal is given, either A or T. The regulation site itself or the sequence is also called the *primary structure*, and the site along with single or multiple helices is called the *secondary structure*. The helix T has an additional feature: its downstream sequence contains several nucleotides T, forming the T-run ("polyuracil").

For the classical attenuation regulation mechanism to exist, the leader peptide gene must appear upstream from the antiterminator helix as shown in Fig. 2c. Like any other gene, the leader peptide gene is a contiguous chain of codons. It always begins with the start codon, contains a number of regulatory codons encoding the amino acid whose intracellular concentration is regulated, and ends with the stop codon. In some cases, a group of biochemically related amino acids is regulated rather than a single amino acid. The leader peptide gene controls the regulatory site translation speed depending on the intracellular concentration of the regulated amino acid. If the amino acid is available, then the ribosome translates the leader peptide gene and destroys the helix A during its motion toward the stop codon. As a result, the helix T can be formed, which serves as a signal to terminate the mRNA synthesis. On the contrary, if there is little or no amino acid available, then the ribosome is delayed at the regulatory codons and cannot destroy the helix A. As a result, the helix T cannot appear, and the mRNA synthesis continues. Consequently, when the regulated amino acid is absent, mRNA of the structural genes is synthesized; these genes are responsible for biosynthesis of that amino acid. A feedback loop is thus established between intracellular molecules of the amino acid and biosynthesis of new molecules of the same amino acid. In the appendix, we describe how our model takes the leader peptide gene into account.

Other examples of the attenuation regulation are T-box regulation and regulations based on riboswitches, the LEU element, etc. Attenuation regulations principally serve to express the structural genes encoding enzymes used in the metabolism of amino acids, nucleosides, vitamins, and some others such as aminoacyl-tRNA-synthetase. Non-attenuation regulations also play important roles in cell functioning, for example, a protein-DNA regulation applied to a wider range of genes. We also note that the total number of different regulatory mechanisms of gene expression seems to be strongly bounded, and many of them are already known. This is why it seems reasonable to first model the evolution of each mechanism independently. The above general biological concepts, including a description of classical attenuation regulation, can be found in [4, ch. 3].

In this paper, we address the task of reconstructing the evolution of classical attenuation regulation sites. The reconstruction is based on the biologically reasonable principles of standard

evolution of the primary structure and conservation of the secondary structure. The phylogenetic tree is assumed to be known in this study (see Figs. 3 and 5 in Examples 1 and 2).

The properties of Markov processes and Gibbs fields on trees are well-studied in the case of a simple spin space, for example, for the Ising model on trees [8–10]. Markov processes on trees are mostly used as mathematical models of evolution, operating with a more complex spin in the form of a finite sequence. Those models work with an evolving sequence of nucleotides from the four-letter alphabet {A, C, T, G} or of amino acids from the alphabet of twenty letters corresponding to the twenty amino acids. There are several simple models of evolution (Jukes–Cantor, Kimura, etc.); each one is defined by a 4×4 matrix that determines the rates of nucleotide replacement by another nucleotide. Those models assume that the nucleotides in the sequence evolve independently (the models were reviewed, e.g., in [11]). In other words, it is assumed in those elementary probabilistic models that nucleotides change into each other (mutate) in the evolutionary process according to a fixed transition matrix. In addition to nucleotide substitution, other models also allow deletion (of a nucleotide or an entire segment of the evolving sequence), insertion (of a nucleotide or segment), segment duplication, and other genome modifications. These changes of the sequence (substitution, insertion, deletion, etc.) correspond to real intracellular processes, but they describe only changes of the primary structure. No interaction of distant regions within the sequence is taken into account, but such an interaction does exist as a result of the secondary structure folding.

As far as we know, all models that include some consideration of the secondary structure are similar: there is a set of independent Markov chains, each one modeling the evolution of a nucleotide (or a nucleotide pair) using different transition probability matrices (see, e.g., [12–15]). All positions in the sequence are split into two types: those participating in coupled pairs and individual free positions. The nucleotides at free positions evolve according to some model of letter substitution independently of other positions. The nucleotides at coupled positions evolve in pairs with so-called forced mutations being introduced for them: if one of the coupled nucleotides is subsituted, then the probability that the other nucleotide is substituted to preserve an allowable stacking pair is high. Each pair evolves independently of other pairs at coupled positions. A state of this model can be described by a sequence which contains one of four nucleotides at free positions and one of allowable pairs of nucleotides at coupled positions. Depending on the model, either the six allowable pairs that frequently occur in helix shoulders are considered, or one more state is added to represent all other rare pairs, or, finally, all 16 possible pairs are considered (see [12, 13]). These models thus take the secondary structure evolution into account in a very limited way because it seems difficult and unnatural to identify every position in the site as either free or coupled in advance.

Here, we propose a new approach to modeling the evolution of the gene expression regulation site together with its secondary structure along a given phylogenetic tree. The secondary structure is not associated with any positions within the site; it is specified by a sophisticated nonlocal interaction potential. The approach is based on a Gibbs-like posterior distribution, and we seek for configurations where the absolute minimum value of some "energy" functional H is attained. In this model, each tree node has an associated regulation site, and a configuration is the set of all such sites for the whole tree. Configurations that give the absolute minimum value to the energy functional H are called *minimal configurations*; we let $E_{\min}$ denote the set of them. To find an absolute minimum, we use the annealing procedure, which builds discrete trajectories $\{\sigma(n) \mid n \in \mathbb{N}\}$ converging to a minimal configuration $\widehat{\sigma} \in E_{\min}$. The algorithm is implemented as a stochastic-iterative scheme where a stochastic decision is made at each step from n to $n+1$ (the algorithm is described in Sec. 3).

Thus, our approach aims at building all minimal configurations that agree with the current data (known regulation sites assigned to the tree leaves) because of a special term in H. Note that we do not assume the existence of a secondary structure in extant sites assigned to the leaves, nor assume

that a multiple alignment of those sites is known. Conversely, in our tests, we only specified a primary structure at the leaves and then compared our computed secondary structures with those found in experiments or independently predicted by bioinformatic methods.

As an application of the method, we consider an important task of constructing the multiple alignment of sequences such that the presumptive common secondary structure is taken into account. The secondary structure in a minimal configuration induced by the evolutionary process allows selecting conserved helices in extant sites and then finding the multiple alignment that retains those conserved helices. The notion of multiple alignment is discussed below (also see, e.g., [11,16]).

Our approach to modeling a nucleotide sequence evolution with the secondary structure taken into account was presented in [17].

2. MODEL OF EVOLUTION FOR A SEQUENCE TOGETHER WITH THE SECONDARY STRUCTURE

Let a finite tree G of a nucleotide sequence evolution be given with a set of nodes V and edge lengths $\{t_j\}$, which are interpreted as the evolution time along the edges. We assume that the tree is binary here, although our model does not depend on this assumption. Let $V_1 \subset V$ be a set of leaves (terminal nodes) of the tree G. A function θ assigning a finite sequence in the four-letter alphabet to each leaf (the current data) is also given. A configuration σ is defined as a map from the set V to the set Q of all finite sequences in the four-letter alphabet. The set $\Sigma = Q^{|V|}$ is called the configuration set, $\sigma \in \Sigma$. In the Gibbs field context, a sequence at each node of the tree G can be called a spin; a position within a sequence (spin) is often called a site in the biological literature, of course, with a meaning different from the regulation site.

We assign each spin σ_k, $k \in V$, a set h_k of all possible helices $h_k = \{h_{k,m}\}$ in σ_k. This set usually includes only helices whose energy is less than some threshold (more stable states correspond to lower energy in our model) and which satisfy other constraints such as the shoulder and loop lengths being not less than three. The helix energy due to stacking is defined, for example, in [5]. The sequences in a configuration σ that are assigned to the leaves of the tree G are called the terminal sequences of that configuration.

We propose a functional $H(\sigma)$ below; its absolute minimum points (the arguments $\widehat{\sigma}$) describe possible paths of the sequence evolution along the tree when the secondary structure is constrained. The functional $H(\sigma)$ includes three constraints on the sought-for configuration $\widehat{\sigma}$: (1) The sequence σ_k at each node k undergoes independent modifications (mutations) at each position $i = 1, \ldots, n$ in accordance with the substitution rate matrix R and also insertions/deletions (indels); these modifications are reflected in the term H_1, the a priori pair interaction between two spins at neighboring nodes (connected by an edge); (2) The terminal sequences of the configuration σ are close to the corresponding extant sequences θ (the term H_2 reflects the influence of the data); (3) The sequences σ_k of the configuration σ retain the secondary structure along each edge and even along tree paths as much as possible, i.e., across many generations, and the longer and more numerous the paths are, the smaller the value of the functional (the term H_3 representing a nonlocal prior pair interaction reflects the required secondary structure conservation). The term H_1 describes the standard evolution of the primary structure, and the term H_3 describes the evolution of the secondary structure.

Because evolutionary changes of the primary structure include indels, sequences at the ends of an edge can differ in length, and the natural correspondence of positions within the sequences is thus broken. But even in the case of sequences of equal length, letters at the same position in the sequences can evolve inconsistently. Therefore, for each edge, we must establish a correspondence between positions in the sequences s and s' assigned to the ends of that edge. This is done by a so-called pairwise alignment procedure that inserts gaps in one or both sequences. The

resulting aligned sequences after this procedure are words in a five-letter alphabet and are always equal in length. A pair of aligned sequences is called an alignment. A procedure establishing a correspondence between positions in not two but n sequences ($n \geq 3$), is called multiple alignment. The procedure is much more complicated computationally for $n \geq 5$ than for the pairwise alignment. During the pairwise alignment, gaps are inserted such that the likeness function $\varphi(s, s')$, which compares new sequences $\bar{s}$ and $\bar{s}'$ obtained as a result of the alignment, is maximized. This function can be defined as

$$\varphi(s, s') = N_e a_e + N_t a_t + N_v a_v + \sum_k [a_d + a_g(\ell_k - 1)], \tag{1}$$

where N_e is the number of alignment positions where the letters of $\bar{s}$ and $\bar{s}'$ match; N_t is the number of positions where a "related" substitution occurs, i.e., A is replaced by G (or vice versa) or C is replaced by T (or vice versa), the so-called *transition*; and N_v is the number of positions where "crossed" substitutions occur, i.e., all other cases of letter replacement (the so-called *transversion*). The summation over k ranges over all contiguous segments of length $\ell_k \geq 1$ such that there is a gap in one sequence at each position in the segment. In the examples below, we used the following parameter values for the function $\varphi(s, s')$: $a_e = 1$, $a_t = -0.8$, $a_v = -1.2$, $a_d = -2$, and $a_g = -1$. In contrast to multiple alignment, fast algorithms of pairwise alignment are known; they are based on a dynamic programming procedure (see, e.g., [16]).

Let

$$H(\sigma) = H(\sigma, \theta) = H_1(\sigma) + H_2(\sigma, \theta) + \lambda H_3(\sigma) \tag{2}$$

be the energy functional. We recall that the term $H_1(\sigma)$ reflects the pair interaction energy in the spin system and, specifically, the a priori information about the tree G; the term $H_2(\sigma, \theta)$ reflects the dependence on the data θ given at the leaves; and the term $H_3(\sigma)$ reflects the secondary structure conservation requirement along each edge and entire paths in the tree.

We now describe the terms in $H(\sigma)$ in detail. Let σ_j and σ'_j denote sequences assigned to the respective starting (closer to the root) and ending points of the edge j in the configuration σ. To compute the term $H_1(\sigma)$, we must first align the sequences σ_j and σ'_j as described above. The result of the alignment are two sequences differing from the originals by some inserted gaps. These new sequences have the same length n_j, which depends on the edge j, and are denoted by $\bar{\sigma}_j$ and $\bar{\sigma}'_j$. Then

$$H_1(\sigma) = -\sum_j H_1(\bar{\sigma}_j, \bar{\sigma}'_j) = -\sum_j \Biggl(\ln \prod_{i=1}^{n_j}{}' (e^{\gamma_i t_j R})(\bar{\sigma}_{ji}, \bar{\sigma}'_{ji}) - \varkappa \sum_m (\ell_{j,m} + 1)^q \Biggr), \tag{3}$$

where the outer sum ranges over all edges j of the tree G, n_j is the length of alignment at the jth edge, the product Π' ranges over only the alignment positions that contain nucleotides in both sequences, R is the substitution rate matrix for letters of the nucleotide alphabet (we use the same matrix for every position at every node), and t_j is the length of the jth edge. The evolution rate γ_i at the ith position is usually considered a random variable distributed in accordance with the gamma law with two fixed parameters. Final results are then somehow averaged over this distribution. For simplicity, we used $\gamma_i = 1$ for every position i in Examples 1–3 below. Also, e^{cR} is a matrix-valued exponential with argument cR, and $(e^{\gamma_i t_j R})(\alpha, \beta)$ is the element of $e^{\gamma_i t_j R}$ corresponding to the transition from the letter α to the letter β.

The inner sum in (3) ranges over the segments of the aligned sequences $\bar{\sigma}_j$ and $\bar{\sigma}'_j$ for which a gap occurs in one of the sequences at every position of the segment. Here, m enumerates such segments, and $\ell_{j,m}$ is the length of the mth segment for the jth edge. The parameters $\varkappa$ and q determine the significance of indel evolutionary events as compared with letter substitutions. We use the values $\varkappa = 10$ and $q = 1$ in Examples 1–3.

The second term is defined such that it is minimal for configurations for which the terminal sequences coincide with the extant sequences θ at the leaves. It can be defined, for example, as

$$H_2(\sigma) = -\sum_{k\in V_1} \varrho(\varphi(\bar{\sigma}_k, \bar{\theta}_k)),$$

where the function $\varrho(\varphi(\bar{\sigma}_k, \bar{\theta}_k))$ has a unique maximum at the point $\varphi(\bar{\sigma}_k, \bar{\theta}_k) = n(\theta_k)$, and $n(\theta_k)$ is the length of the sequence θ_k. In Examples 1 and 2, we consider the limit case where

$$H_2(\sigma) = \begin{cases} 0 & \text{if } \varphi(\bar{\sigma}_k, \bar{\theta}_k) = n(\theta_k),\ \forall k \in V_1, \\ +\infty & \text{otherwise,} \end{cases} \tag{4}$$

i.e., the terminal sequences of every configuration coincide with the current data for the leaves.

We now define the third term as

$$H_3(\sigma) = H_3(\sigma, h) = -\sum_j \Phi(h_j, h'_j), \tag{5}$$

where $h = \langle h_j, h'_j\rangle$ and where $h_j = \{h_{jm}\}$ and $h'_j = \{h'_{jk}\}$ are two sets of helices with sufficiently low energy obtained from the two known sequences σ_j and σ'_j assigned to the ends of the jth edge. The potential Φ reflects the secondary structure conservation along the edges of the tree G. We also consider more complicated forms of the term $H_3(\sigma)$ in the Appendix.

The exact form of the potential Φ depends on the type of a desired secondary structure. In the case of classical attenuation regulation, the secondary structure contains the mutually exclusive terminator and antiterminator $T = (t_{m1}, t_{m2})$ and $A = (a_{m1}, a_{m2})$, where t_{m1}, t_{m2}, a_{m1}, and a_{m2} are the shoulders and m ranges over the set of all existing pairs (A, T). Here, the potential is naturally defined as

$$\Phi(h_j, h'_j) = \frac{1}{n_{mk}} \sum_{m,k} [\varphi(t_{m1}, t'_{k1}) + \varphi(t_{m2}, t'_{k2}) + \varphi(a_{m1}, a'_{k1}) + \varphi(a_{m2}, a'_{k2})]^{X+}, \tag{6}$$

where $[u]^{X+} = u$ for $u > X$ and $[u]^{X+} = 0$ for $u \le X$, X is a fixed threshold, φ is the function defined by equation (1), and n_{mk} is the number of nonzero terms in the sum. In other words, to compute $\Phi(h_j, h'_j)$, we first choose all possible antiterminator–terminator pairs in the two sets of helices h_j and h'_j. Then we compare pairs from the different sets by independently pairwise aligning the corresponding shoulders of the antiterminators and terminators. The likeness value is averaged over all pairs with a likeness greater than the predefined threshold X. We use the threshold $X = 0$ in Examples 1–3.

Remark. A potential $\Phi(h_j, h'_j)$ that characterizes the likeness of a secondary structure *as a whole* in the sequences σ_j and σ'_j assigned to the ends of the jth edge can be considered. Indeed, let h_j and h'_j be the respective sets of *all helices* in σ_j and σ'_j that have a sufficiently low energy. The potential $\Phi(h_j, h'_j)$ can be defined as

$$\Phi(h_j, h'_j) = \frac{1}{n_{mk}} \sum_{m,k} [\varphi(h_{m1}, h'_{k1}) + \varphi(h_{m2}, h'_{k2})]^{X+}.$$

The secondary structure is worse conserved along the tree paths in this case. Computing simulations for the model with such a potential demonstrates that several paths running from the leaves and retaining the secondary structure do not reach the tree root.

In what follows, we describe the algorithm for finding the configurations $\widehat{\sigma}$,

$$E_{\min} = \arg\min H(\sigma, \theta), \quad \widehat{\sigma} \in E_{\min}, \tag{7}$$

that give the absolute minimum of the energy functional $H(\sigma)$ defined by equations (2)–(6) and discuss the results of testing the proposed evolutionary model.

3. ALGORITHM BASED ON STOCHASTIC DYNAMICS

To avoid misunderstanding, we make our terminology more precise: we use the term "evolutionary dynamics" for the evolutionary process and the term "stochastic dynamics" for our simulations. In the evolutionary dynamics, a random process describes evolution in time from the tree root to the leaves. In the stochastic dynamics, a configuration on the tree as a whole is involved in a random process. The "time" of the stochastic dynamics has no relation to the "evolutionary time."

Energy functional (2) depends on very many interacting variables and has many minima, including local minima very close to each other. In this case, stochastic algorithms are effective for finding the global minima of the functional. Here we propose considering the annealing scheme based on the Metropolis–Hastings stochastic dynamics (see, e.g., [1, 18, 19]). Stochastic algorithms look through numerous minima of the functional H and escape from local minima, in contrast to deterministic algorithms that find the nearest local minimum. Physically, annealing is a slow cooling of the system to the zero temperature such that the limit distribution of the process is concentrated on configurations in the set $E_{\min}$ for any initial conditions:

$$\lim_{n\to\infty} P\big(X(n) \in E_{\min}\big) = 1. \tag{8}$$

The algorithm is realized as a nonhomogeneous Markov chain with transition probabilities dependent on the present configuration $\sigma(n)$ and the parameter β_n (the inverse system temperature). The algorithm constructs a sequence of configurations $\{\sigma(n)\}$ (a "trajectory of the dynamics") that starts with an arbitrary configuration $\sigma(0)$ and converges in probability to one of the minimal configurations for any initial condition.

We introduce the probability distribution on Σ in a Gibbs form:

$$\pi_\beta(\sigma) = \frac{1}{Z_\beta} e^{-\beta H(\sigma)} \tag{9}$$

with the normalizing constant $Z_\beta = \sum\limits_\sigma e^{-\beta H(\sigma)}$.

All possible modifications in one iteration step from the configuration $\sigma(n)$ to the configuration $\sigma(n+1)$ are letter substitution at one position, insertion, or deletion. We now describe these transitions in detail. At the next tree node k (in terms of a given linear order on the set of all inner nodes, which are sorted in the order of decreasing distance from the tree root), one position in the sequence σ_k is taken uniformly. Then the type of modification of the sequence at this position is chosen with the following probabilities: $P_s = 0.992$ is the substitution probability, $P_i = 0.004$ is the insertion probability, and $P_d = 0.004$ is the deletion probability (insertions and deletions are assumed equiprobable). If a letter substitution is taken as the modification, then a new letter is chosen in accordance with a 4×4 symmetric transition matrix with zeros on the main diagonal. The probabilities for transition and transversion mutations are, respectively, equal to 5/6 and 1/12, and the sum of elements over any column and any row equals 1. The choice of the transition matrix element depends on two model parameters. The first parameter is the ratio of the transition mutation frequency to the transversion mutation frequency (the so-called transition-to-transversion ratio). We assume that this parameter is equal to 5 in Examples 1–3 below. The second parameter is the ratio of the frequency of transitions to the frequency of deletions or insertions, which is taken to be 100. Our model is robust with respect to small parameter variations.

To make an insertion at a given position, we first choose an insertion length $\ell = 1, \ldots, 32$ with probability $2^{-\ell}/c$, where $c = 1 - 2^{-32}$, and then insert a word containing ℓ independent equiprobable letters. We make a deletion similarly with the chosen position as the "center" of the deleted segment. After a new sequence $\widetilde{\sigma}_k \in Q$ appears at a node $k \in V$ in the tree, a new

configuration $\widetilde{\sigma}$ is proposed. The configuration $\widetilde{\sigma}$ differs from the configuration σ in only one sequence assigned to the node k.

We have thus described possible modifications of sequences, which can be given by a symmetric transition matrix on the spin space Q. We say that this matrix defines the proposal distribution. The new configuration is accepted, i.e., $\sigma(n+1) = \widetilde{\sigma}$, with probability

$$q(\sigma, \widetilde{\sigma}) = \exp\{-\beta[H(\widetilde{\sigma}) - H(\sigma)]^+\}, \tag{10}$$

where $[u]^+ = u$ for $u \geq 0$ and $[u]^+ = 0$ for $u < 0$. Correspondingly, the new configuration is rejected, and the former configuration σ is kept, i.e., $\sigma(n+1) = \sigma$, with probability $1 - q(\sigma, \widetilde{\sigma})$.

The sequence length is changed after a deletion or insertion. Even after a letter substitution, the alignment of two sequences at the ends of edge j could totally differ from the former pair alignment; letters at the same position in the previous alignment could be at different positions in the new alignment of the sequences $\widetilde{\sigma}_j$ and σ'_j. Old couplings between letters at the same position would thus be destroyed, and new couplings would appear. As a result of this realignment, we obtain a "nonlocal" change of the energy H_1 for local configuration change. This situation is typical for models of evolution but differs drastically from models in statistical physics. In our case, to find (10), we must calculate both the terms $H_1(\widetilde{\bar{\sigma}}_j, \bar{\sigma}'_j)$ in $H_1(\widetilde{\sigma}_j)$ and $H_1(\bar{\sigma}_j, \bar{\sigma}'_j)$ in $H_1(\sigma'_j)$, while if the change of the energy H_1 were local, then the difference of H_1 for two sequences $\widetilde{\sigma}_j$ and σ_j could be obtained by calculating the energy difference only at the modified position. The differences of the energies H_2 and H_3 for local modifications of configurations could be found similarly.

We thus obtain a sequence of configurations

$$\sigma(0) \Rightarrow \sigma(1) \Rightarrow \ldots \Rightarrow \sigma(n) \Rightarrow \ldots.$$

The process $\sigma(n)$ on Σ is reversible with respect to the distribution π_β given by (9) for any given β; in particular, distribution (9) is the stationary distribution of the process. Consequently, we have the following statement.

For any initial configuration $\sigma(0)$ and any given β,

$$\lim_{n\to\infty} P\big(\sigma(n) = \eta \,|\, \sigma(0)\big) = \pi_\beta(\eta).$$

If the parameter β is not constant but sufficiently slowly increasing to infinity, $\beta_n \to \infty$ (the growth rate determines the so-called annealing regime), then the process constructed above is not reversible. This raises the question of defining the parameter increase rate (the annealing regime) to ensure that the limit measure is concentrated on the set $E_{\min}$. It was proved in [18, 19] that if $\beta_n \to \infty$ so that

$$\lim_{n\to\infty} \frac{\log n}{\beta_n} > C, \tag{11}$$

where the constant C depends on the function $H(\sigma)$, then relation (8) holds. Since this result comes from statistical physics, we should verify that the following two conditions are satisfied for our system of interacting spins: the values

$$H_{\min} = \min_\sigma H(\sigma), \qquad H_{\max} = \max_\sigma H(\sigma), \qquad \Delta = H_{\max} - H_{\min}$$

are bounded, and our system is a system with local interaction. Both conditions are satisfied because the tree has finitely many nodes and each spin interacts with at most three neighbors on the tree. Consequently, we have the following statement.

If the parameter β_n is taken according to (11)*, then for any initial configuration $\sigma(0)$ and any $\eta \in \Sigma$,*

$$\lim_{n\to\infty} P\big(\sigma(n) = \eta \,|\, \sigma(0)\big) = \widehat{\pi}(\eta),$$

where $\widehat{\pi}(E_{\min}) = 1$.

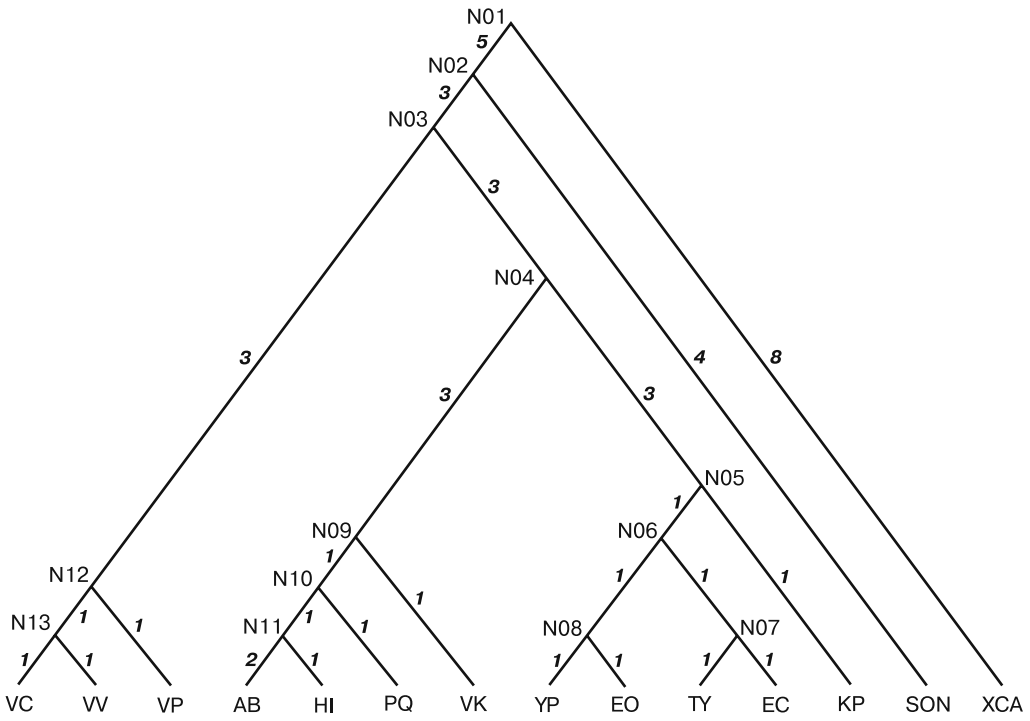

Fig. 3. Species tree in Example 1

The limit configurations $\widehat{\sigma}$ obtained as the result of the above stochastic annealing algorithm with $\beta_n \to \infty$ under inequality (11) form the set $E_{\min}$ of all minimal configurations (7).

4. RESULTS OF TESTING THE MODEL

We present results of testing the model in the case of classical attenuation regulation. In these examples, we use the values $\lambda \in [0.2, 0.3]$ and $\beta_n = C \ln^p(n+1)$, where $C = 0.01$ and $p = 1.5$ are model parameters and n is the iteration number. Depending on the starting point, 10^5–10^7 iterations of the algorithm usually suffice to reach one of the hypothetical absolute minima of the functional H. The computing time ranged from 10 hours to 3–5 days for a single processor implementation of the algorithm on a 3 GHz Pentium 4 PC. We also used a 12-node cluster provided by the Space Research Institute of the Russian Academy of Sciences, thereby speeding up the computation approximately 40 times. In the future, we will parallelize the algorithm to speed it up and decrease the starting point effect.

Example 1 (classical attenuation regulation of threonine biosynthesis in gamma-proteobacteria). Source regulation sites in the leaves are taken from [20] with gaps removed. We consider the standard species tree (Fig. 3). It has 27 nodes including 14 leaves, and each edge is assigned a phylogenetic length in conventional units. The leaves are marked with abbreviated species names as follows: VC – *Vibrio cholerae*, VV – *Vibrio vulnificus*, VP – *Vibrio parahaemolyticus*, AB – *Actinobacillus actinomycetemcomitans*, HI – *Haemophylus influenzae*, PQ – *Mannheimia haemolytica*, VK – *Pasterella multocida*, YP – *Yersinia pestis*, EO – *Erwinia carotovora*, TY – *Salmonella typhi*, EC – *Escherichia coli*, KP – *Klebsiella pneumoniae*, SON – *Shewanella oneidensis*, XCA – *Xanthomonas campestris*.

In Fig. 4, we show one of the algorithm results for $\lambda = 0.2$: a minimal configuration with $H = 1154$, $H_1 = 1352$, $H_2 = 0$, and $H_3 = -990$. The ancestral sequences of this minimal configuration are grouped in blocks. Each block defines a path from a leaf to the root, and the conserved

secondary structure is highlighted on the path. The terminator and antiterminator found by the algorithm are respectively marked by shading and underlining in all sequences. We recall that the terminator and antiterminator can include small one- and two-sided bulges, which appear in Fig. 4 as nonshaded or nonunderlined nucleotides within the shoulders. The highlighted regulatory structures for each leaf marked with a species name compose a *path* from the antiterminator–terminator pair in the leaf to the pair in the sequence N01 assigned to the tree root, for instance, the path from the leaf VC to the root N01. We note that such a path (possibly not unique) exists for each leaf. The paths with minimum total energy H_3 over all edges on each path are shown in Fig. 4. The name of a node is shown to the right of a sequence followed by the H_3 value on the corresponding edge (for inner nodes) or the total H_3 value over all edges of the path (for leaves).

Example 2 (classical attenuation regulation of leucine biosynthesis in gamma-proteobacteria). We consider a species tree with 23 nodes and 12 leaves (Fig. 5); it is a part of the tree shown in Fig. 3. In Fig. 6, we show one of the algorithm results for $\lambda = 0.25$; it is a minimal configuration with $H = 1718$, $H_1 = 1796$, $H_2 = 0$, and $H_3 = -310$. The blocks consisting of ancestral sequences in that minimal configuration are shown in Fig. 6. Each block represents a path from a leaf to the root with a highly conserved secondary structure along the path. The notation is the same as in Example 1.

5. CONCLUSIONS

The proposed algorithm models the evolution of classical attenuation regulation by building at ancestral nodes a reasonable regulatory secondary structure of the same type as in the primary structures given for leaves according to modern bioinformatic and experimental data. The secondary structure induced by the model into the source primary structures at the leaves coincides with or is close to the secondary structure predicted by independent data (see, e.g., [20]). Moreover, the primary structures at the leaves and inner nodes of the phylogenetic tree have good multiple alignment in the minimal configuration, and those structures are therefore well coordinated.

Our analysis of the composition of minimal configurations (ground states of our model) versus the parameter λ shows that in the domain of "moderate" values $\lambda \in [0.2, 1]$, a strong regulatory structure of one type is retained along the entire evolutionary tree. When $\lambda = 0$, i.e., only the primary structure of the evolving sequence is considered, all tests show the absence of paths with a secondary structure conserved from the extant regulation sites at the leaves to a regulation site at the root. A similar situation is observed for sufficiently small $\lambda \leq 0.1$. If λ enters the domain of moderate values, then the composition of minimal configurations changes: from almost every leaf to the root, a path that conserves a secondary structure appears. Finally, for large $\lambda > 2$, when the primary structure interaction becomes less significant in our functional, the composition of minimal configurations again changes. The number of long paths with a conserved secondary structure from the leaves to the root sharply decreases: only pieces of such paths, energetically favored for the term H_3, remain. In contrast to the case of a small λ, the primary structures of the basal sequences now differ greatly. Various secondary structures, which are only significant for isolated edges, can be identified in them, but they do not contribute to forming the long path. Therefore, for large values of λ, the evolutionary process tends to form short isolated periods of secondary structure conservation; this turns out to be energetically favored because of the term H_3.

The term H_3, first introduced in this paper and responsible for the secondary structure conservation in the energy functional H, thus essentially changes properties of minimal configurations. It turns out to be important for finding configurations where a fixed regulatory structure is retained along the entire evolutionary tree.

To compare the results of our algorithm with those of standard ones, we input the same sequences at the leaves to known computer programs that can reconstruct ancestral sequences given only the

```
gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13 -47.5
tGTTGGGGCAGGCTgctgagcgcaaaatttcacAAAAAAGGCCTGTATCCCAACcGATACAGGCCTTTTTTta       VC -234.3

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaacaaatttcAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa   VV -248.1

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -57.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcacAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTta      VP -182.6

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      NO9 -39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10 -51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata       N11  -6.2
aaTGGGGCGGGctagtgcgttgaagaatagaattcatGAACCCGCaTTTCCCGAGaGCGGGTTTttttatg          AB -240.5

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      NO9 -39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10 -51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata       N11 -35.0
aaTGGTGCGGGTTagtgcagcaaaaacaagatacAGAAAACCCGCGATTCAactGAATaGCGGGTTTTTTtata      HI -269.3

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -38.8
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -27.8
aGAtgGTGCGGGTTagtgctgacaaaaaaaATGAAcaaAAAACCCGCACTCaacaaaaaGCGGGTTTTTTTATa      NO9 -46.0
aaTGGTGCGGGTTagtactggcaaaaaaaATGAAcaaAAAACCCGCAacTCAactaaaaGCGGGTTTTTTTATa      N10  -8.6
catagtGCGGGTTTTaatTGGctgaaataatgaaagaATAAACCGAAAACCCGCTacaAGCGGGTTTTTTTGTa      PQ -201.9

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      NO9 -19.4
atagtGTGCGGGTTagtgcgtaacaaaaagatcgaattccacAAAACCCGTACTgaataaaaAGTGCGGGTTTTtttatg VK -163.6

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -38.2
taacGGTGCGGGCTgacgcgtacaggaaacacAGAAAAAAGCCCGCACCTgaacAGTGCGGGCTTTTTTTTa        NO5 -55.1
taacGGTGCGGGCTgacgcgtacaggaaacacAGAAAAAAGCCCGCACCTgaacAGTGCGGGCTTTTTTTTt        NO6 -55.1
taacGGTGCGGGCTgacgcgtacaggaaatacAGAAAAAAGCCCGCACCTgaacAGTGCGGGCTTTTTTTTt        NO8 -44.8
ttacGGgGCGGGCTgacgcgtacaggaaacaatAGAAAAAAGCCCGCACCtagacaGTGCGGGCTTTTTTTTt      YP -312.8
```

Fig. 4. Algorithm results in Example 1

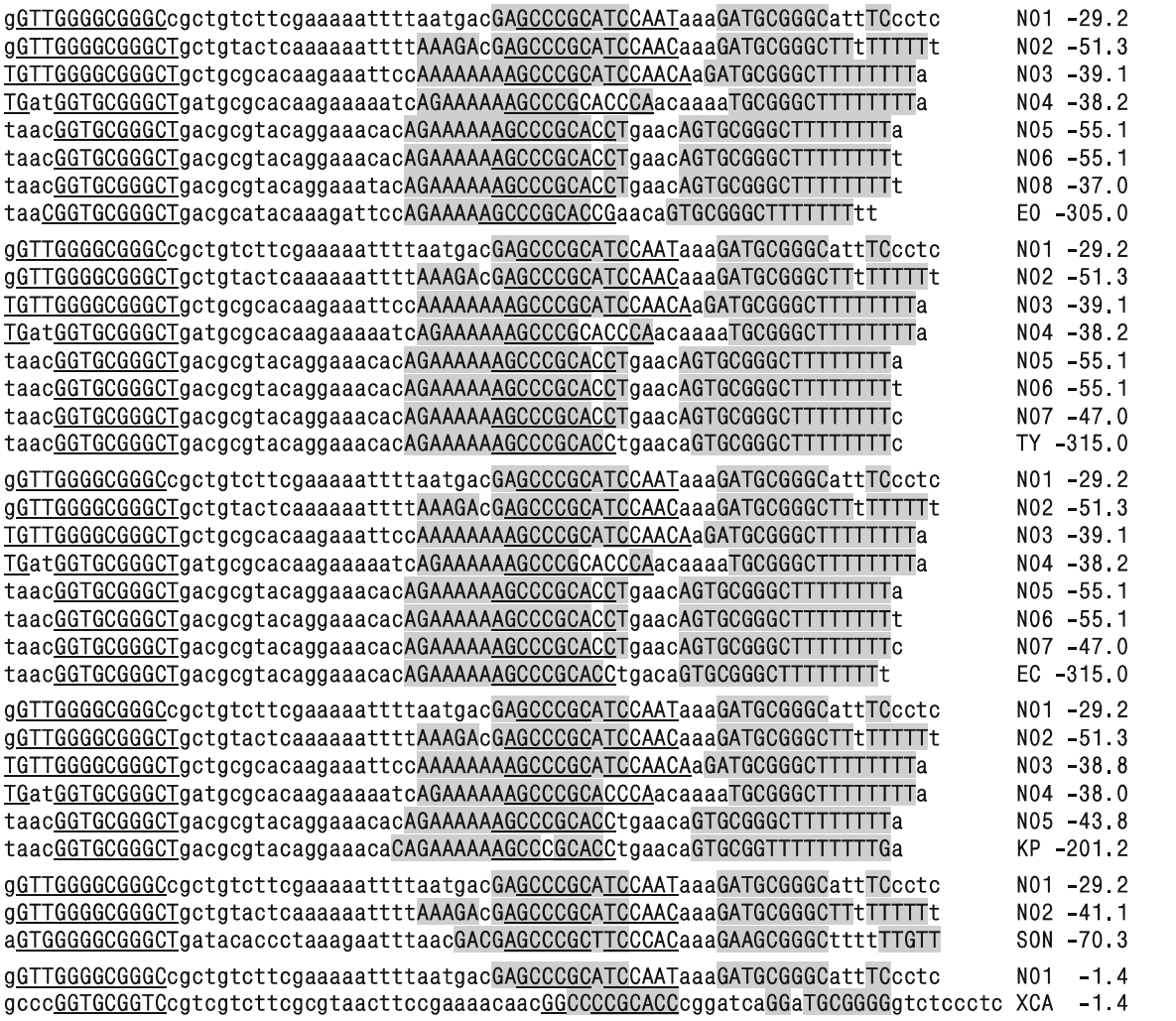

Fig. 4. (continued)

primary structure (PAML, PAUP, etc.; see a list of software available at `http://evolution.genetics.washington.edu/phylip/software.serv.html`). Those programs require a multiple alignment as input, and we therefore supplied them with sequences *prealigned according to the known secondary structure* given in [20]. Even so, the PAML program did not construct conserved secondary structures of the desired type in ancestral sequences at all. The PAUP program did construct such structures, apparently because of the secondary structure given at the leaves. But the absolute values of H_3 in the constructed configuration were approximately half the corresponding H_3 values in our minimal configurations.

The model was also tested by adding noise in artificial and biological examples, and it produced stable results.

To verify that the ancestral signals can function, we tested them using the corresponding regulation model. In the case of classical attenuation regulation, we used the model in [5] and the website based on it, `http://lab6.iitp.ru/rnamodel`. Such a simulation requires the presence of the leader peptide gene, and we therefore applied the model to longer sequences, which include the leader peptide genes (absent in Examples 1 and 2), at the leaves. We briefly discuss the results at the end of the Appendix.

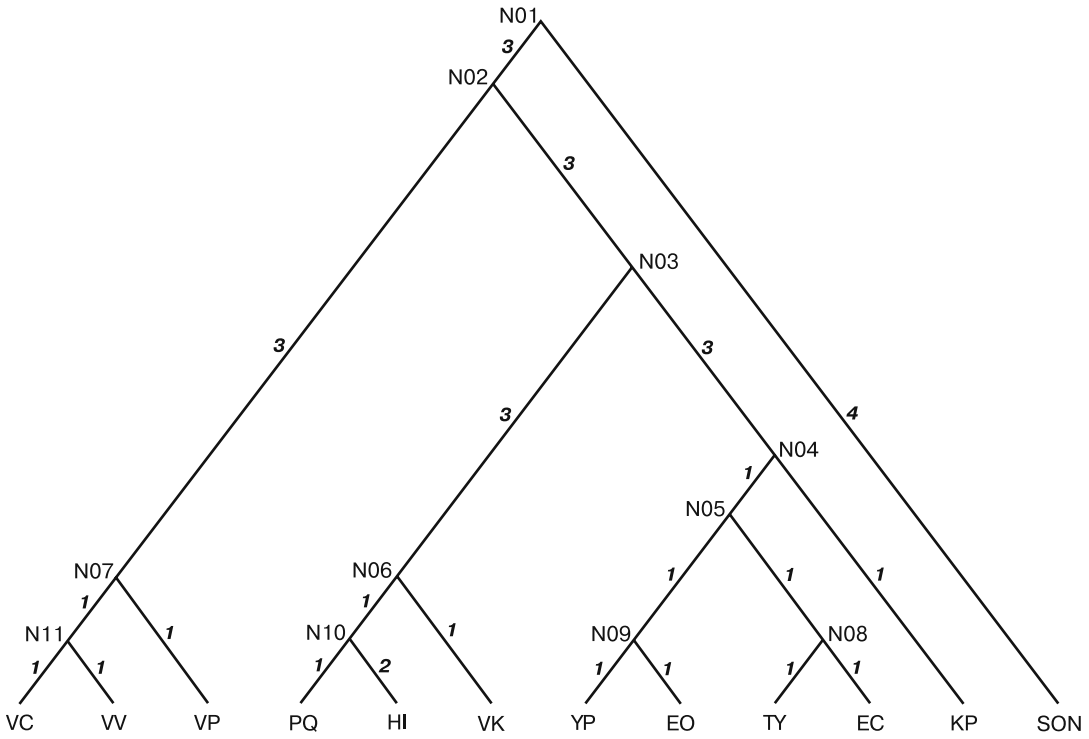

Fig. 5. Species tree in Example 2

The model proposed here can also be naturally formulated in the case of an infinite tree. It is reasonable to use approaches and methods from the theory of the Gibbs random fields to study such models. Because both the structure of the spin space and spin interactions are complicated, the resulting system should have quite nontrivial properties that reflect the evolution of certain regions of the genome, the regulatory signals.

We note several specific features of the proposed model. The set of all minimal configurations $E_{\min}$ has a large cardinality: computational results for a single trajectory differ greatly in their primary structures, especially in the basal domain of the tree, depending on the chosen starting point. Nevertheless, every minimal configuration contains the same set of paths (up to a small shift in few positions) composed of the secondary structures corresponding to the regulation type; those paths run from every leaf to the root with the secondary structure highly conserved along each path.

Sometimes the regulation becomes weaker at some nodes or along some paths in the constructed minimal configuration, most often closer to the root. It can be supposed that the regulation does not function in such places of the tree; therefore, the model can point out evolutionary periods when one type of regulation changed to another. At some nodes, the model predicts an ensemble of several antiterminators and terminators. For instance, one terminator at node N04 in Example 1 has a loop of length 6 and a four-letter bulge in the left shoulder, and another terminator has a loop of length 10 and no bulges. In the minimal configuration shown in Fig. 4, the paths through node N04 run to the root as follows: from the leaves AB, HI, VK, YP, EO, TY, and EC via the terminator with a loop of 6 nucleotides, and from PQ and KP via the terminator with a loop of 10 nucleotides (both terminator options are grey-shaded in the corresponding blocks). In the same example, two terminators exist at node N05, and two by two terminators and antiterminators exist at node N09, i.e., there are four putative secondary structures in all. This can indicate the role of the secondary structure ensembles and also the evolutionary preference for different structures during subsequent evolution along the species tree after the ensemble has appeared.

```
attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -19.6
ttcgcGCGGGTgggctgtggaagaaaactaaaccacaccaaatatccAAAAAACCCGCAcaatgaTGCGGGTTTTTTtata            NO7 -36.0
atcgcGCGGGTaggctgtggaagaaaactaaaccacacaaaataacAAAAAACCCGCAcactgaTGCGGGTTTTTTtata            N11 -36.0
atcacGCGGGTaggctgtggacaaaaacaccacacaagatAAAAAACCCGCAgctgaTGCGGGTTTTTTtata                   VC -125.7

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -19.6
ttcgcGCGGGTgggctgtggaagaaaactaaaccacaccaaatatccAAAAAACCCGCAcaatgaTGCGGGTTTTTTtata            NO7 -36.0
atcgcGCGGGTaggctgtggaagaaaactaaaccacacaaaataacAAAAAACCCGCAcactgaTGCGGGTTTTTTtata            N11  -5.4
atcgcGCGGGTaGGctgtggaagaaaaataaaccacacagaataacaaCTAGCCCGCAcatcgaTGCGGGCTttttttata           VV  -95.1

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -19.6
ttcgcGCGGGTgggctgtggaagaaaactaaaccacaccaaatatccAAAAAACCCGCAcaatgaTGCGGGTTTTTTtata            NO7 -12.2
TTCGcGCGGGTaggctgtggaagaaaaataaccacaccaatttcttAGAAACCCGCATGAAaATGCGGGTTTTTttata              VP  -65.9

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -12.6
ttttTGTGCGGctagattgtggatgaaaaaaagaaaagtaaccccaaaactAAAACCCGCACActataaTGCGGGTTTTtttttta       NO6 -22.0
attttGTGCGGctaaattgtggttaaaaaaatcaaatgaaatacccaaattcAAACCCGCACcctataaTGCGGGTTTttttttt        N10 -11.4
attttGTGCGaGGataaagattgattggaaaaagtaaatggactatccacattcttGCCCGCACcttaaaaTGCGGGCttttttttat     PQ -114.1

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -12.6
ttttTGTGCGGctagattgtggatgaaaaaaagaaaagtaaccccaaaactAAAACCCGCACActataaTGCGGGTTTTtttttta       NO6 -22.0
attttGTGCGGctaaattgtggttaaaaaaatcaaatgaaatacccaaattcAAACCCGCACcctataaTGCGGGTTTttttttt        N10 -21.0
tttgtGTGCGGctaagttgtggataaaaaaacagtcagatgtaaatacccaattttAAACCCGCACTTtatAAGtTGCGGGTTTtttatct  HI -123.7

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -12.6
ttttTGTGCGGctagattgtggatgaaaaaaagaaaagtaaccccaaaactAAAACCCGCACActataaTGCGGGTTTTtttttta       NO6 -11.4
tCATtGTGCGGctaggttgtggataaaaaacagtaaaatatcccacaaattAGACCCGCACATGtaaGCGGGTCTtttttta          VK  -92.1

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -35.8
cttttGCGCGGtaagctggtgggcgacattcagaattaagtcagctcaaaactAAAAAACCCGCGCcatgGTGCGGGTTTTTTtatg     NO4 -28.2
tttttGCGCGGtaagctggtgggcggcattcagaattaagtcagctcaaaATtAAAAAACCCGCGCcattGCGCGGGTTTTTTtATg     NO5 -22.2
ttcTTGCGCGGtaagttggtgggcggaattcagaactaagttagccgtcaaaattAAAAAGCCCGCGCAAatGCGCGGGTTTTTTtttta  NO9 -39.8
ctcTTGCGCGGtatgttggtggacggaaatcagaactgattcagccatcaagattAAcAAGCCCGCGCAAatGCGCGGGTTTtTTttgt   YP -194.2

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -35.8
cttttGCGCGGtaagctggtgggcgacattcagaattaagtcagctcaaaactAAAAAACCCGCGCcatgGTGCGGGTTTTTTtatg     NO4 -28.2
tttttGCGCGGtaagctggtgggcggcattcagaattaagtcagctcaaaATtAAAAAACCCGCGCcattGCGCGGGTTTTTTtATg     NO5 -22.2
ttcTTGCGCGGtaagttggtgggcggaattcagaactaagttagccgtcaaaattAAAAAGCCCGCGCAAatGCGCGGGTTTTTTtttta  NO9 -26.0
ttcttGCGCGGtaggtttgtgggcagacttcagaactaagtttctcgccacacgatacAAAAACCCGCGCtgatGCGCGGGTTTTTtttta  EO -180.4

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -35.8
cttttGCGCGGtaagctggtgggcgacattcagaattaagtcagctcaaaactAAAAAACCCGCGCcatgGTGCGGGTTTTTTtatg     NO4 -28.2
tttttGCGCGGtaagctggtgggcggcattcagaattaagtcagctcaaaATtAAAAAACCCGCGCcattGCGCGGGTTTTTTtATg     NO5 -31.8
gttttGCGCGGtagaccggtgggcggcattcagcattaagtcagctcgaagtcaAAcAAAACCCGCGCcattGCGCGGGTTTTtTTatg   NO8 -38.2
attgtGCGCGGtagaccggtgggcggcattcaacattaagtcagctcgaagtcaAAcAAAACCCGCGCcgtGTGCGGGTTTTtTTatg   TY -202.2

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -35.8
cttttGCGCGGtaagctggtgggcgacattcagaattaagtcagctcaaaactAAAAAACCCGCGCcatgGTGCGGGTTTTTTtatg     NO4 -28.2
tttttGCGCGGtaagctggtgggcggcattcagaattaagtcagctcaaaATtAAAAAACCCGCGCcattGCGCGGGTTTTTTtATg     NO5 -31.8
gttttGCGCGGtagaccggtgggcggcattcagcattaagtcagctcgaagtcaAAcAAAACCCGCGCcattGCGCGGGTTTTtTTatg   NO8 -34.0
tcttgGCGCGGtagacgagtgagcggcatccagcattaagccagcacgcagtcaaacAAAAAACCCGCGCcattGCGCGGGTTTTTTtatg  EC -198.0

attaCGCGGGtgtcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtgta          NO1 -34.0
tttGCGCGGGtggattgtggacgaaaactagaaaagtaaaccaaaaaccAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata          NO2 -34.0
cttttGCGCGGctagattgtggacgaaaataagaaaagtaaacccaaaactAAAAAcCCCGCGCcatgGTGCGGGtTTTTTtata       NO3 -35.8
cttttGCGCGGtaagctggtgggcgacattcagaattaagtcagctcaaaactAAAAAACCCGCGCcatgGTGCGGGTTTTTTtatg     NO4 -26.2
catttGCGCGGtaggctgttgggcgacgttcagcgttaagtcatcttccagcaagactatAAAACCCGCGCcttgGCGCGGGTTTTtttatg KP -130.1

attacgCGGGTGTcttatggttgcccactcgaaaggtgaacaaaacactAAAAAcCCCGCGCCaTGGTGCGGGtTTTTTtgta          NO1  -4.4
aaaacGCGGaGGTcttagtgttgctcgctcgatagataggcaaaacactcatAAACCCCGCActaatgtTGCGGGGTTTtttgta        SON  -4.4
```

Fig. 6. Algorithm results in Example 2

In addition to the model described above, other models have been developed for reconstructing the evolution of the regulatory signal and its characteristics under constraints on the secondary structure. The first of them [21] reconstructs the secondary structure at inner nodes from the leaves to the root based on the minimum evolution principle and simultaneously constructs a multiple alignment of all sequences. In the second model [22], an algorithm was developed for reconstructing the matrices of nucleotide frequencies from the leaves to the root of the species tree. Results obtained with these models agree with each other and also with those of the model proposed here.

APPENDIX

Other ways to describe secondary structure conservation. The most difficult question is what kind of conservation of the secondary structure to express in the term $H_3(\sigma)$. As is mentioned above, biologically reasoned considerations could include the edge lengths in $H_3(\sigma)$ and also conservation of the secondary structure along entire paths in G, i.e., during many generations. The corresponding functional leads to a model where the spin interaction along an edge of the tree G becomes nonlocal because entire paths are taken into account. This distinctly complicates the model analysis. The effect of the edge length is determined by the description of the evolution environment. For example, we considered a simple modification in the form $U(\Phi) = t^g\Phi$, where $g = 1.5$ is a model parameter.

We considered the following two candidates for the term $H_3(\sigma)$ involving paths instead of the sum over isolated edges. The first candidate had the form

$$H_3(\sigma) = -\sum_{k\in V_1}\max_{p_k}\sum_{m\in p_k}[\varphi(t_{m1},t_{m'1})+\varphi(t_{m2},t_{m'2})+\varphi(a_{m1},a_{m'1})+\varphi(a_{m2},a_{m'2})]^{X_+},$$

where p_k is a path from the leaf $k \in V_1$ to the root composed of the shoulders a_{m1} and a_{m2} of the antiterminators and the shoulders t_{m1} and t_{m2} of the terminators taken from sequences σ_m at nodes m along that path. Modeling with such a term $H_3(\sigma)$ leads to results similar to those shown in Figs. 4 and 6. The secondary structure along the path from every leaf to the root is conserved even more in this case, but the primary structure conservation decreases.

The second candidate for $H_3(\sigma)$ is more consistent with the biological standpoint and is defined by the formula $H_3(\sigma) = -\sum\limits_{j\in G}\sum\limits_{p(j)\in G}U_{p(j)}(\Phi,\{t_j\})$, where $p(j)$ are paths along the tree G that run from the structures at the edge j to the root and Φ is defined by formula (6), where we take

$$U_{p(j)}(\Phi,\{t_j\}) = \prod_{\ell\in p(j)}\Phi(h_\ell,h'_\ell)\frac{1}{(1+rt_\ell)}$$

with a certain parameter r.

Example 3 (taking the leader peptide gene into account). We consider the same sequences as in Example 1 but now taken as beginning at the start codon of the leader peptide gene. Because the complete leader peptide gene for the sequence PQ was not available, we used the sequence AS (*Actinobacillus succinogenes*) instead. To account for presence of the leader peptide gene in a sequence, we supplemented the energy functional H with the term $H_4(\sigma)$, which reduces the energy if a sequence includes the leader peptide in the proper place, i.e., near the antiterminator upstream from its loop. The term depends linearly on the number of regulatory codons in the leader peptide up to certain threshold m:

$$H_4(\sigma) = \begin{cases} -\mu r, & \text{for } r \le m,\\ -\mu m, & \text{for } r > m,\end{cases}$$

where μ and m are model parameters and r is the number of regulatory codons. In all instances,

Table 1. Percentage of the premature termination events vs. the amino acid concentration (data for leaves of the species tree)

c	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0	Q
AB	9	13	20	25	27	35	37	42	44	48	48	50	51	52	54	53	53	55	6.1
EC	16	14	17	17	21	26	28	34	40	48	48	52	54	57	61	63	65	66	4.7
EO	14	12	11	13	19	25	28	35	40	44	50	52	57	62	60	66	68	67	6.2
HI	16	18	20	20	21	23	23	27	25	24	26	26	28	32	29	30	31	32	2.0
KP	21	22	20	25	28	33	36	39	41	47	51	53	58	58	64	61	65	68	3.4
AS	22	22	28	32	35	40	45	50	54	55	58	62	64	64	65	67	65	67	3.0
SON	18	21	23	32	36	41	46	53	56	58	60	63	66	69	69	69	70	74	4.1
TY	21	17	19	23	24	30	33	41	44	48	48	52	58	59	62	60	66	66	3.9
VC	10	14	16	24	34	39	48	51	57	63	64	69	69	70	71	72	75	75	7.5
VK	27	29	32	38	45	50	53	59	61	63	67	70	69	72	73	70	72	72	2.7
VP	48	49	52	51	59	61	64	65	68	68	71	74	72	73	76	74	75	78	1.6
VV	47	46	51	53	56	57	62	65	66	67	69	73	72	74	73	74	75	75	1.6
XCA	26	27	27	28	33	35	37	39	41	39	41	46	43	44	44	46	48	47	1.8
YP	48	51	53	53	59	61	62	65	67	68	69	72	70	72	77	73	74	76	1.6

Table 2. Percentage of the premature termination events vs. the amino acid concentration (data for inner nodes of the species tree)

c	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	Q
N01	15	17	17	16	18	24	21	24	24	23	25	26	24	27	2.0
N02	18	21	26	31	31	37	43	46	50	51	55	54	58	58	3.9
N03	27	29	31	35	40	44	46	47	50	52	52	57	57	59	2.1
N04	30	33	36	40	42	46	50	53	55	58	58	57	61	60	2.1
N05	28	34	39	42	48	56	58	62	62	66	69	71	68	74	3.6
N06	49	50	53	55	59	65	67	67	69	73	73	75	71	76	3.6
N07	24	27	36	36	43	46	53	58	60	63	66	67	69	73	3.4
N08	54	54	58	59	64	65	67	70	68	74	73	76	77	78	2.6
N09	37	41	45	51	55	60	63	66	68	68	72	71	70	71	2.8
N10	54	57	55	57	61	59	63	63	67	66	68	71	71	71	1.4
N11	18	24	24	27	32	35	38	41	44	46	45	51	52	50	2.4
N12	53	53	55	57	59	63	65	68	69	70	70	72	73	73	1.4
N13	46	50	56	56	58	60	64	67	64	70	70	69	72	75	1.5

the minimal configuration constructed by our algorithm contains the leader peptide gene at every ancestral node, while without the term $H_4(\sigma)$, the leader peptide gene can be found in only 3–4 nodes of all 13 inner nodes at best. The complete example of a minimal configuration built with $\mu = 5$ and $m = 12$ is given at `http://lab6.iitp.ru/docs/anneal/ex4_a.htm`.

To estimate the regulation quality in the reconstructed ancestral sequences independently, we used the model in [5] with default parameter values. Using this model for each ancestral sequence, we determined the dependence of the premature termination frequency $p(c)$ versus the concentration c of the regulating amino acids (threonine and isoleucine) in the range 0–1 with step 0.05. For each value of the concentration, we estimated the frequency of the issue (i.e., either termination or antitermination) based on 1000 different trajectories of the Monte Carlo procedure.

For the tree leaves, we found that in the frequency range $c \in [0.15, 1]$, all extant sequences demonstrate a monotonic growth (disregarding negligible variations) of the premature termination frequency with average ratio Q of the maximum frequency to the minimum frequency greater than 3.5 (see Table 1). As is explained, for example, in [5], such a monotonic growth over a sufficiently large range of c and with a large value of Q speaks in favor of the functionality of the regulatory structure.

A similar situation is observed at all inner nodes (see Table 2), but the range of monotonic growth for the dependence $p(c)$ narrows, $c \in [0.3, 0.95]$, and the average ratio Q equals 2.5 in this case.

Thus, the energy functional H together with the term H_4 allows one to successfully model the evolution of the entire regulation site including the leader peptide gene. This gene is restored in all ancestral sequences. Moreover, the model in [5] demonstrates the presence of an attenuation regulation signal of the type under consideration, and the signal improves (though weakly) as it approaches the current state.

REFERENCES

1. Ewens, W. and Grant, G., *Statistical Methods in Bioinformatics: An Introduction*, New York: Springer, 2001.

2. *Mathematics of Evolution and Phylogeny*, Gascuel, O., Ed., New York: Oxford Univ. Press, 2005.

3. Lyubetsky, V., Gorbunov, K., Rusin, L., and V'yugin, V., Algorithms to Reconstruct Evolutionary Events at Molecular Level and Infer Species Phylogeny, *Bioinformatics of Genome Regulation and Structure II*, Kolchanov, N., Hofestädt, R., and Milanesi, L., Eds., New York: Springer, 2006, pp. 189–204.

4. Singer, M., and Berg, P., *Genes & Genomes: A Changing Perspective*, Mill Valley: Univ. Science Book, 1991. Translated under the title *Geny i genomy*, Moscow: Mir, 1998.

5. Lyubetsky, V.A., Pirogov, S.A., Rubanov, L.I., and Seliverstov, A.V., Modeling Classic Attenuation Regulation of Gene Expression in Bacteria, *J. Bioinform. Comput. Biol.*, 2007, vol. 5, no. 1, pp. 155–180.

6. Lee, F. and Yanofsky, C., Transcription Termination at the trp Operon Attenuators of *Escherichia coli* and *Salmonella typhimurium*: RNA Secondary Structure and Regulation of Termination, *Proc. Natl. Acad. Sci. USA*, 1977, vol. 74, no. 10, pp. 4365–4369.

7. Mironov, A.A. and Kister, A.E., Theoretical Analysis of RNA Secondary Structure Formation Kinetics during Transcription and Translation. Accounting for Imperfect Helices, *Mol. Biol.* (Moscow), 1985, vol. 19, no. 5, pp. 1350–1357.

8. Bleher, P.M., Ruiz, J., and Zagrebnov, V.A., On the Purity of Limiting Gibbs State for the Ising Model on the Bethe Lattice, *J. Stat. Phys.*, 1995, vol. 79, nos. 1–2, pp. 473–482.

9. Evans, W., Kenyon, C., Peres, Y., and Schulman, L.J., Broadcasting on Trees and the Ising Model, *Ann. Appl. Probab.*, 2000, vol. 10, no. 2, pp. 410–433.

10. Martinelli, F., Sinclair, A., and Weitz, D., Glauber Dynamics on Trees. Boundary Conditions and Mixing Time, *Comm. Math. Phys.*, 2004, vol. 250, no. 2, pp. 301–334.

11. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge: Cambridge Univ. Press, 1998. Translated under the title *Analiz biologicheskikh posledovatel'nostei*, Moscow–Izhevsk: Regulyarnaya i haoticheskaya dinamika, 2006.

12. Muse, S.V., Evolutionary Analyses of DNA Sequences subject to Constraints on Secondary Structure, *Genetics*, 1995, vol. 139, pp. 1429–1439.

13. Savill, N.J., Hoyle, D.C., and Higgs, P.G., RNA Sequence Evolution with Secondary Structure Constraints: Comparison of Substitution Rate Models Using Maximum-Likelihood Methods, *Genetics*, 2001, vol. 157, no. 1, pp. 399–411.

14. Telford, M.J., Wise, M.J., and Gowri-Shankar, V., Consideration of RNA Secondary Structure Significantly Improves Likelihood-based Estimates of Phylogeny: Examples from the Bilateria, *Mol. Biol. Evol.*, 2005, vol. 22, no. 4, pp. 1129–1136.

15. Kosakovsky Pond, S.L., Mannino, F.V., Gravenor, M.B., Muse, S.V., and Frost, S.D., Evolutionary Model Selection with a Genetic Algorithm: A Case Study Using Stem RNA, *Mol. Biol. Evol.*, 2007, vol. 24, no. 1, pp. 159–170.

16. *Mathematical Methods for DNA Sequences*, Waterman, M.S., Ed., Boca Raton: CRC Press, 1989. Translated under the title *Matematicheskie metody dlya analiza posledovatel'nostei DNK*, Moscow: Mir, 1999.

17. Lyubetsky, V.A., Zhizhina, E.A., Gorbunov, K.Yu., and Seliverstov, A.V., Model of Evolution of Nucleotide Sequence, in *Proc. 13th All-Russia Conf. on Mathematical Methods of Pattern Recognition, Zelenogorsk, Russia, 2007*, Moscow: MAKS Press, 2007, pp. 605–609.

18. Geman, S. and Geman, D., Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images, *IEEE Trans. Pattern Anal. Machine Intelligence*, 1984, vol. 6, pp. 721–741.

19. Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., Optimization by Simulated Annealing, *Science*, 1983, vol. 220, no. 4598, pp. 671–680.

20. Vitreschak, A.G., Lyubetskaya, E.V., Shirshin, M.A., Gelfand, M.S., and Lyubetsky, V.A., Attenuation Regulation of Amino Acid Biosynthetic Operons in Proteobacteria: Comparative Genomics Analysis, *FEMS Microbiol. Lett.*, 2004, vol. 234, no. 2, pp. 357–370.

21. Gorbunov, K.Yu. and Lyubetsky, V.A., Model of Evolution of Nucleotide Sequence Considering Its Secondary Structure, in *Proc. Int. Sci. Conf. on Computational Phylogenomics and Genosystematics, Moscow, 2007*, Moscow: Mosk. Gos. Univ., 2007, pp. 68–71.

22. Gorbunov, K. and Lyubetsky, V., Reconstruction of Ancestral Regulatory Signals along a Transcription Factor Tree, *Mol. Biol.* (Moscow), 2007, vol. 41, no. 5, pp. 836–842.