Attenuation Regulation as a Term Rewriting System

Eugene Asarin, Thierry Cachat, Alexander Seliverstov, Tayssir Touili, Vassily Lyubetsky

Attenuation Regulation as a Term Rewriting System *

Eugene Asarin[1], Thierry Cachat[1], Alexander Seliverstov[2], Tayssir Touili[1], and Vassily Lyubetsky[2]

[1] LIAFA, CNRS and University Paris Diderot, asarin,txc,touili@liafa.jussieu.fr
[2] IITP, Russian Academy of Science, slvstv,lyubetsk@iitp.ru

Abstract The classical attenuation regulation of gene expression in bacteria is considered. We propose to represent the secondary RNA structure in the leader region of a gene or an operon by a term, and we give a probabilistic term rewriting system modeling the whole process of such a regulation.

1 Introduction

Modeling the mechanisms of regulation of gene expression, allowing prediction of quantitative characteristics of this expression (such as estimation of the level of expression and concentration of the substrate) is an important research challenge. In a previous work [LRSP06,LPRS07], a model of one particular kind of regulation, the classical attenuation regulation, has been suggested. In that model, the evolution of the secondary RNA structure in the leader region of a gene, and the progress of the ribosome and the polymerase along the RNA/DNA strands, are represented by a very special, elaborated in detail, Markov chain. In this chain the transition probability corresponding to the progress of the ribosome depends on a "control variable" — the concentration of charged tRNA molecules in the cell. All the other probabilities do not depend on the control variable, they can be determined from energy-based considerations. Termination and antitermination (of gene expression) correspond to particular random events in the Markov chain. In [LRSP06], a Monte-Carlo simulation of this Markov chain led to biologically realistic dependence of termination probability from the control variable. Due to a large size and a complex structure of the Markov chain, its simulation is a heavy computational task, but it was successfully solved, and a software tool called RNAMODEL simulates one trajectory in fractions of a second [LRSP06,RNA]. However, the approach based on the direct description of the Markov chain and its simulation has some limitations, especially for a theoretical analysis. Biologically, it would be nice to have a more structured and compact representation of the Markov chain and its instantaneous probability

* The support of CNRS-RAS cooperation agreement 19122 EVOLVER is gratefully acknowledged.

distributions over all states at every instant, or only for sufficiently large time, or only probabilities of the two biologically important events — termination and antitermination.

Note that the problem of modeling the classical attenuation regulation, as stated in [LRSP06] and in the current article, is related to the representation of the transient behavior of the secondary structure on a sliding window on the RNA strand between the ribosome and the polymerase (see below for details). This differs from the kinetics of the secondary RNA structure on a fixed nucleotide sequence for unlimited time, i.e. unlimited number of steps, investigated in many papers. The structure that appears after a large amount of time is called equilibrium secondary RNA structure, it corresponds to a minimum of energy, see e.g. [Zuk03,FFHS00]. The tool RNAMODEL has also the function of determining this equilibrium structure and its energy as a special part of the full model in [LRSP06]. However, real structures that appear on the RNA strand during the regulation process are far from the equilibrium and their energies are far from minimal.

In this article we discover a regular internal structure of the Markov chain describing the classical attenuation regulation. We show that it can be represented as a probabilistic term rewriting system for a particular type of terms. The set of rewriting rules can be large, but all of them are generated by a small set of (five) metarules. In fact we give the full description of the metarules and explain how to generate all the rules for the case of classical attenuation regulation.

Potential benefits of such a representation are multiple:

- easier and more precise modeling of regulation mechanisms depending on the dynamics of the secondary structure;
- compact description of such mechanisms, perhaps in dedicated languages, and hence a better biological understanding of regulation processes;
- convenient representation of secondary structures by terms;
- specific analysis and simulation methods for rewriting systems.

This article is structured as follows. In section 2 we describe shortly the biological phenomenon that we want to model: the mechanism of classical attenuation regulation (CAR). In section 3 we introduce a class of terms and probabilistic term rewriting systems. In section 4 we represent a qualitative metamodel of the biological mechanism of CAR by a term rewriting system. In section 5 we refine the previous system and decorate its transitions with rates, thus obtaining a representation of the Markov chain by a probabilistic term rewriting system. In section 6 we show some simulation results. In section 7 we discuss some related work on term rewriting and its applications. In section 8 we conclude with a discussion of perspectives of the rewriting approach to modeling the mechanisms involving RNA secondary structures, especially regulation.

2 Classical attenuation regulation

To begin with, we recall some well-known biological facts about the biological phenomenon playing the central role in this article.

The expression of a group of structural genes (that is synthesis of the corresponding proteins, which are ferments for a chemical reaction) can be regulated by a sequence of nucleotides placed on the DNA upstream inside the so called *leader region* of the genes [SB91]. This subsequence of the leader region is called the *regulatory region*. In this article we deal with one particular type of regulation, *classical attenuation regulation (CAR)* in bacteria. This regulation mechanism concerns structural genes (groups of genes — operons) that produce proteins which catalyze the synthesis of amino acids. The classical attenuation allows to activate such an operon when the cell contains a small concentration of the amino acid, to deactivate the operon whenever this concentration increases, and to do it fast. The mechanism of CAR involves several actors: the regulatory region on the DNA, its copy on the RNA, the ribosome, and a ferment called RNA polymerase (see Fig.1).

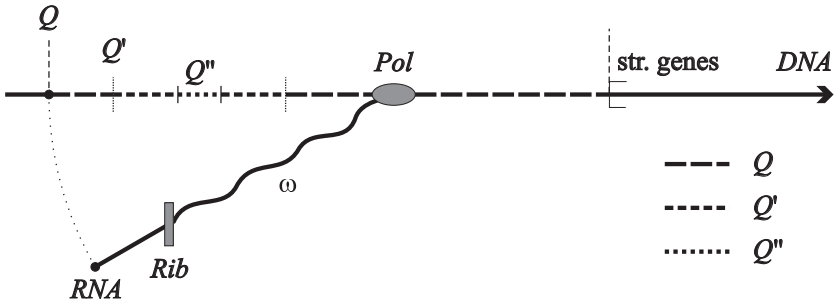

Figure1. Classical attenuation regulation. The RNA polymerase *Pol* transcribes the regulatory region Q, the ribosome *Rib* translates the leader peptide gene Q'. The movement of *Rib* on regulatory codons Q'' is controlled by the concentration of charged tRNA. The secondary RNA structure ω between *Rib* and *Pol* brakes *Pol* and pushes it off the chain. If *Pol* reaches the structural genes, then they are expressed, i.e. transcribed and then translated. Note that in both the DNA and the RNA, we use Q, Q' and Q'' to denote the regulatory region, the leader peptide gene, and the regulatory codons, respectively.

For structural genes to be expressed two concurrent processes should succeed: the regulatory region Q should be transcribed creating an RNA by RNA polymerase. At the same time the ribosome should be bound to the very beginning of the freshly created segment Q' (called the *leader peptide gene*) in the regulatory region Q on the RNA and starts translation of this leader peptide gene to an auxiliary protein. The essential part of the regulation process takes place when the ribosome moves on Q' on the RNA and the polymerase moves somewhere downstream of the ribosome on Q on the DNA.

The ribosome moves "rightwards" (formally speaking, in the direction from the $5'$ to the $3'$ end) on a segment Q' of the sequence Q. Its speed is constant

except on a subsequence Q'' (*regulatory codons*) where it depends directly on the concentration of the amino acid (via charged tRNA concentration). To the right of the ribosome and independently of it, the polymerase moves rightwards on Q. Between the ribosome and the polymerase a secondary structure ω is formed on the RNA. This structure consists in pairing of some nucleotides, and it changes very fast. An important effect of the secondary structure ω consists in slowing down the movement of the polymerase. There are two possible scenarios:

- When ω is strong enough, its "braking" action on the polymerase increases, and moreover, the polymerase can slip off the DNA (this can only happen on so-called *T-rich sequence*, where the connection of the polymerase and the DNA weakens). Such an event is called *termination*, and in this case the structural genes are not expressed: the transcription of the regulatory region is aborted, the structural genes are not transcribed and therefore not translated.
- Another possibility is that the ribosome moves fast enough to weaken or partly destroy most of the structure ω. In this case the polymerase safely traverses the T-rich sequence, and arrives to the end of the leader region Q. Next, the polymerase enters the structural genes, and their transcription, followed by translation are unavoidable. This event is called *antitermination* and in this case the structural genes are expressed.

In the rest of this article we build a qualitative and a quantitative models of the regulation process described above.

3 Terms and rewriting systems

3.1 Unranked unordered terms

Let Σ be a finite set of function symbols and $\mathcal{X}$ an enumerable set of *variables* (standing for sets of terms). The set $T_\Sigma[\mathcal{X}]$ of terms over Σ and $\mathcal{X}$ is the smallest set that satisfies:

- $\Sigma \subseteq T_\Sigma[\mathcal{X}]$,
- $\{f(x) \mid f \in \Sigma \wedge x \in \mathcal{X}\} \subseteq T_\Sigma[\mathcal{X}]$,
- if $f \in \Sigma$ and $s \subseteq T_\Sigma[\mathcal{X}]$ is a *set* of terms, then $f(s)$ is in $T_\Sigma[\mathcal{X}]$.

By definition we also put $f(\emptyset) = f$ for $f \in \Sigma$. For convenience we write $f(g, h(e))$ instead of $f(\{g, h(\{e\})\})$. However one should remember that the coma-separated terms are unordered.

Example 1. Let $\Sigma = \{e, f, g, h\}$ and $\mathcal{X} = \{x, y, z, \dots\}$, then the followings are terms in $T_\Sigma[\mathcal{X}]$: $f(g, h(e))$, $f(f(x))$ and $e(g, f)$.

Note that we consider function symbols of variable arity. T_Σ stands for $T_\Sigma[\emptyset]$. Terms in T_Σ are called *ground terms*. Variables are used only to define substitution and rewriting rules. The "real" terms are ground terms. A *substitution* σ is a mapping from $\mathcal{X}$ to $2^{T_\Sigma[\mathcal{X}]}$, written as $\sigma = \{x_1 \to T_1, \dots, x_n \to T_n\}$, where

T_i, $1 \leq i \leq n$, is a finite set of terms that substitutes the variable x_i. The term obtained by applying the substitution σ to a term t is written $t\sigma$. We call it an *instance* of t.

Let R be a rule of the form $l \to r$, where l and r are terms in $T_\Sigma[\mathcal{X}]$. For ground terms t, t' we write $t \to_R t'$ if there exists a substitution σ such that t' can be obtained from t by replacing an occurrence of the subterm $l\sigma$ by $r\sigma$. $\to_R$ defines a relation between ground terms. Let $\to_R^*$ be the reflexive transitive closure of $\to_R$.

Example 2. Let $R = l \to r$ with $l = f(x, e)$, $r = f(g(x), e)$ and $t = e(f(h, e))$, then $t \to_R t'$ where $t' = e(f(g(h), e))$.

A *term rewriting system (TRS)* is a finite set of rules of the form $l \to r$. Given a TRS $\mathcal{R}$ and a set of terms $I \subset T_\Sigma$, the language $\mathcal{R}^*(I)$ is defined as the set of all ground terms that can be obtained from the terms in I by applying a finite number of times the rules from $\mathcal{R}$, i.e., $\mathcal{R}^*(I) = \{t \in T_\Sigma \mid \exists t' \in I, t' \to_{\mathcal{R}}^* t\}$.

Example 3. Let $\mathcal{R} = \{f(x) \to g(f(x))\}$ and $I = \{f(e, h)\}$, then

$$\mathcal{R}^*(I) = \{g^n(f(e, h)) \mid n \in \mathbb{N}\}.$$

3.2 Probabilistic Term Rewriting Systems

A *Continuous Time Markov Chain* is a pair (S, ρ), where S is a finite or enumerable set of states and $\rho : S \times S \to [0, \infty)$ is the rate matrix. For $s, s' \in S$, $\rho(s, s') > 0$ means that there is a transition between states s and s', and that the probability for moving from s to s' within t time units is equal to $1 - e^{-\rho(s,s') \cdot t}$. If a state s has more than one outgoing transition (i.e., if there exist more than one state s' for which $\rho(s, s') > 0$) there exists a *race* between these transitions and the probability for moving from s to s' within t time units is equal to $\frac{\rho(s,s')}{E(s)}\left(1 - e^{-E(s) \cdot t}\right)$, where $E(s) = \sum_{s' \in S} \rho(s, s')$.

A (continuous time) *Probabilistic term rewriting system (PTRS)* over $\Sigma \cup \mathcal{X}$ is a (finite) set of rules of the form $l \xrightarrow{\Lambda} r$, where l and r are terms in $T_\Sigma[\mathcal{X}]$, and $\Lambda \in (0, \infty)$ is a rate.

A PTRS $\mathcal{R}$ over $\Sigma \cup \mathcal{X}$ defines a continuous time Markov chain on ground terms $M = (T_\Sigma, \rho)$, where $\rho(t, t') = \Lambda$ iff there exists a rule $l \xrightarrow{\Lambda} r \in \mathcal{R}$ such that $t \to_R t'$, where R is the "non probabilistic" rule $l \to r$.

Remark 1. If there are several rules (or several instances of the same rule) that lead from t to t', then $\rho(t, t') = \sum \Lambda$, where the sum is taken over all such rules or instances.

4 Metamodel

We want to model the phenomenon of the classical attenuation regulation described in section 2.

We suppose that a *regulatory region* Q (see Fig. 1) is given and fixed in the sequel, it is a sequence (word) $Q \in \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}^*$, the letters of this alphabet are called *nucleotides*. We denote by $|x|$ the length of any word x and x_i the ith letter of x, so $x = x_1 x_2 \dots x_{|x|}$. The sequence Q can be folded[1] in a way that some nucleotides of Q are paired: $\mathbf{A}$ with $\mathbf{T}$ and $\mathbf{C}$ with $\mathbf{G}$. The complement of a nucleotide is written using a bar: $\mathbf{A} = \overline{\mathbf{T}}$, $\mathbf{T} = \overline{\mathbf{A}}$, $\mathbf{C} = \overline{\mathbf{G}}$, $\mathbf{G} = \overline{\mathbf{C}}$. We look in Q for subwords ("stems") of the form

$$\begin{array}{l} Q_A Q_{A+1} \dots Q_B \quad \text{and} \quad Q_C Q_{C+1} \dots Q_D \quad \text{such that} \\ B - A = D - C,\ A + 3 \le B,\ B + 3 \le C \\ Q_A = \overline{Q_D},\ Q_{A+1} = \overline{Q_{D-1}}, \dots\ Q_B = \overline{Q_C}\ . \end{array} \tag{1}$$

Any pair of such stems forms a *hypohelix* (see Figure 2, where the labels A_i, B_i, C_i and D_i are positions in the word Q).

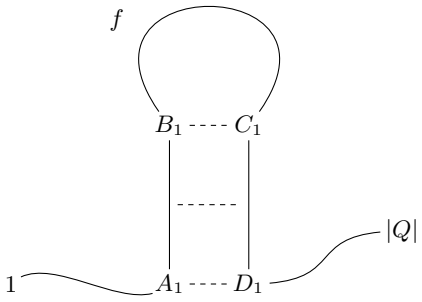

Figure2. One hypohelix f.

We describe a hypohelix f by a tuple of its stems' extremities $f = (A, B, C, D)$, and we introduce the following notations:

$$\mathit{stem}(f) = [A, B] \cup [C, D],\ \mathit{loop}(f) = [B + 1, C - 1],\ \mathit{supp}(f) = [A, D].$$

There is a *ribosome* at some position on Q' and an *RNA polymerase* somewhere to the right of it. Both move to the right, in one step the ribosome moves by three successive nucleotides and the polymerase by one nucleotide. The *window* $w = (R, P)$ represents the segment of RNA from the first position R after the end of the ribosome to the last position P before the beginning of the polymerase. In fact the folding of the RNA sequence Q can only happen within the current window, i.e. between positions R and P. When the ribosome advances to the right, it can destroy the leftmost hypohelix of a current configuration, because it consumes the first three letters of the window. On the other hand any polymerase move adds one new letter to the window.

[1] only on its "active" part called *window*, as we will see below

Formally a window has the form $w = (R, P)$ with $R, P \in \mathbb{N}$. The following constraints should be satisfied:

$$13 \leq R \leq P \leq |Q| \tag{2}$$

Thus, the window is moving and changing its length.

Let $W = \{w = (R, P) \mid$ conditions (2) are satisfied$\}$ be the alphabet of all windows. We define

$$\mathit{stem}(w) = \emptyset,\ \mathit{loop}(w) = [R, P],\ \mathit{supp}(w) = [R, P]\ .$$

We will write terms over the alphabet Σ of all hypohelices and all windows:

$$\Sigma = H \cup W \text{ where } H = \{f = (A, B, C, D) \mid \text{conditions (1) are satisfied}\}.$$

We consider only terms of the form $w(\dots)$ for some $w \in W$ (rooted by some window w). According to the conditions that we will define next, a symbol $f = (A, B, C, D)$ can appear in a term $w(\dots)$ only if $R \leq A$ and $D \leq P$, where $w = (R, P)$.

We say that a hypohelix f is *embedded* in g (which can be a hypohelix or a window), written $f \prec g$, if $\mathit{supp}(f) \subseteq \mathit{loop}(g)$. Two hypohelices f and g are *disjoint*, written $f \bowtie g$, if $\mathit{supp}(f) \cap \mathit{supp}(g) = \emptyset$. We call f and g *unknotted* if either one of them is embedded in the other or they are disjoint. We say that $g = (A_2, B_2, C_2, D_2)$ is an *extension* of $f = (A_1, B_1, C_1, D_1)$, denoted $f \sqsubseteq g$, if $[A_1, B_1] \subseteq [A_2, B_2]$ and $B_2 - B_1 = C_1 - C_2$, hence $[C_1, D_1] \subseteq [C_2, D_2]$, and the pairing in g is an extension of that in f. See Figure 3.

We call a term t over Σ *well-formed* if it satisfies the following conditions:

(compatibility) any f and g appearing in t are unknotted, in particular any f can appear at most once,

(ordering) if f and g occur in t, then $f \prec g$ iff f is in the scope of g.

The combination of two hypohelices in Figure 4 is biologically feasible, but according to our rules these hypohelices are incompatible. We believe that this restriction (crucial for representation by terms) does not undermine significantly the accuracy of the model.

Notice, that a well-formed term of the form $w(\dots)$ (rooted by some window w) contains only hypohelices from

$$\Sigma_w = \{f \in H \mid f \prec w\}.$$

This simple observation greatly simplifies the simulation process.

In [LRSP06] an additional *maximality* condition is imposed. Using the terminology of this article, it requires that no hypohelix f in t can be replaced by its proper extension without creating an overlapping. Here we do not impose this restriction.

Each well-formed term represents a possible secondary RNA structure in a window in Q: the set of hypohelices that are present in this window. It could be

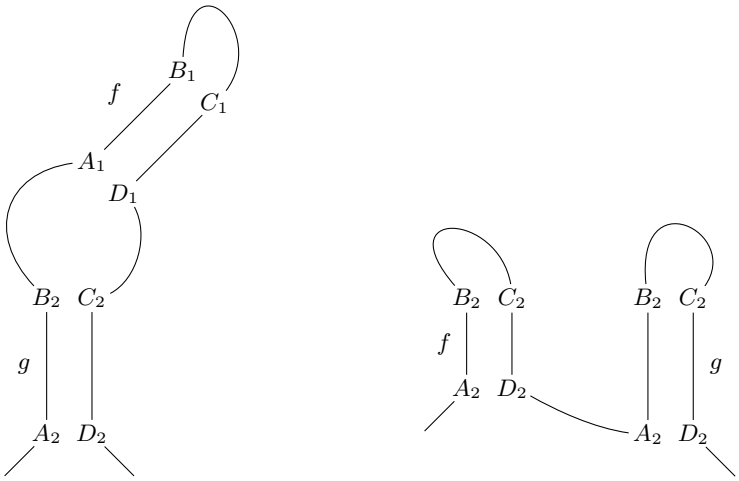

Figure3. Relative positions of two hypohelices f and g: $f \prec g$ and $f \bowtie g$. Here $f = (A_1, B_1, C_1, D_1)$ and $g = (A_2, B_2, C_2, D_2)$. On the left $B_2 < A_1$ and $D_1 < C_2$, on the right $D_2 < A_2$.

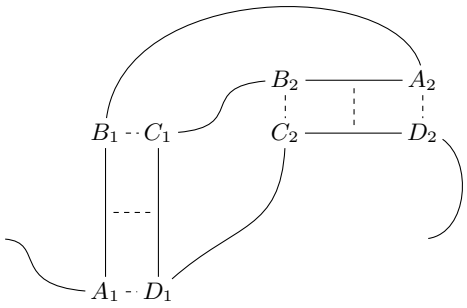

Figure4. Pseudo-knot: $A_1 < B_1 < A_2 < B_2 < C_1 < D_1 < C_2 < D_2$. Such configurations are not allowed in our model.

possible to allow knotted hypohelices, and hypohelices of length less than 3, but here we do not consider them.

We extend the definitions of $\bowtie$ and $\prec$: let f be a term and $\boldsymbol{c}$ a set of terms,

$$\boldsymbol{c} \bowtie f \text{ iff } \forall g \in \boldsymbol{c}\,(g \bowtie f),$$
$$\boldsymbol{c} \prec f \text{ iff } \forall g \in \boldsymbol{c}\,(g \prec f).$$

In the former case we say that f and $\boldsymbol{c}$ are *disjoint*, in the latter that $\boldsymbol{c}$ is embedded into f.

We start from a sequence Q without any pairing of nucleotides, this structure is described by a term $w()$ — "an empty window", where $w = (13, 13)$. Our aim is to represent the evolution of the secondary structure in the window, as well as the progress of the ribosome and the polymerase, through rewriting terms starting from $w()$. Our rewriting system will generate only well-formed terms.

On the whole, there are five rewriting *Meta*-rules:

- Binding and decomposition of a hypohelix f:

$$\big(\omega = g(\boldsymbol{c}, \boldsymbol{d})\big) \longleftrightarrow \big(\omega' = g(\boldsymbol{c}, f(\boldsymbol{d}))\big) \quad \text{with } \boldsymbol{c} \bowtie f,\ \boldsymbol{d} \prec f,\ f \prec g, \quad (3)$$

  where $\boldsymbol{c}$ and $\boldsymbol{d}$ are sequences of terms. The *concrete* rewriting rules — and their rates — depend on $\boldsymbol{c}$ and $\boldsymbol{d}$, as explained below.
- Extension and reduction of a hypohelix

$$\big(\omega = f\big) \longleftrightarrow \big(\omega' = g\big) \quad \text{with } f \sqsubseteq g. \quad (4)$$

- The window movement can be described by the following rules, where $w = (R, P)$:

$$(R, P)(\omega) \longrightarrow (R + 3, P)(\omega')\ , \quad (5)$$
$$(R, P)(\omega) \longrightarrow (R, P + 1)(\omega)\ , \quad (6)$$
$$w(\omega) \longrightarrow \bot\ . \quad (7)$$

In the last rule, $\bot$ is a special symbol denoting termination. Rules (5) describe the movement of the ribosome. In these rules, ω' is obtained from ω by removing only the possible symbol that is incompatible with the new window $(R + 3, P)$, or replacing it by a "shorter" hypohelix. Indeed, if the leftmost hypohelix in ω starts at a position between R and $R + 3$, then the movement of the ribosome by three positions to the right will destroy this hypohelix. More formally, if $\omega \prec (R + 3, P)$, then $\omega' = \omega$. Otherwise the ribosome destroys the leftmost hypohelix. In this case, there is a single symbol f in ω such that $f \not\prec (R+3, P)$. Suppose the subterm rooted by f is $f(\boldsymbol{c})$. Then, ω' is obtained by replacing in ω $f(\boldsymbol{c})$ by either $f'(\boldsymbol{c})$ or $\boldsymbol{c}$, depending on the size of f, where $f' \sqsubseteq f$.

Rules 6 describe the movement of the polymerase. Note that if the polymerase reaches a position $P+1$ where the structural genes are expressed, then we reach antitermination and the gene is expressed.

5 Quantitative model

Now, we introduce the rates of the five rewriting rules.

Let $h(f_1(*), \ldots, f_n(*))$ be a term. Then the *free loop length* of the hypohelix h in this term is

$$l_h = |\, loop(h)| - \sum_{i=1}^{n} |\, supp(f_i)| \ .$$

This numeric characteristic corresponds to the number of nucleotides in the loop of the hypohelix h that do not participate in inner hypohelices.

In order to define the rate, we have to consider the concrete rule corresponding to the Metarule (3). For any $f, g, \boldsymbol{c} = c_1(x_1), \ldots, c_m(x_m)$ and $\boldsymbol{d} = d_1(y_1), \ldots, d_n(y_n)$ such that $\boldsymbol{c} \bowtie f,\ \boldsymbol{d} \prec f,\ f \prec g$ there is a concrete rule

$$\begin{aligned} & \big(\omega = g(c_1(x_1), \ldots, c_m(x_m),\ d_1(y_1), \ldots, d_n(y_n))\big) \\ \longleftrightarrow\ & \big(\omega' = g(c_1(x_1), \ldots, c_m(x_m),\ f(d_1(y_1), \ldots, d_n(y_n)))\big) \end{aligned} \tag{8}$$

Recall that the subterms are unordered. Similarly the concrete rule corresponding to (4) is

$$\begin{aligned} & \big(\omega = a(c_1(x_1), \ldots, c_m(x_m),\ f(d_1(y_1), \ldots, d_n(y_n)))\big) \\ \longleftrightarrow\ & \big(\omega' = a(c_1(x_1), \ldots, c_m(x_m),\ g(d_1(y_1), \ldots, d_n(y_n)))\big) \end{aligned} \tag{9}$$

Note that this transformation can change the free loop length of the hypohelix a. The rate of the rules (8-9) is denoted $K(\omega \to \omega')$, given by

$$K(\omega \to \omega') = \kappa \cdot \exp\left(\frac{1}{2}\left(E(\omega) - E(\omega')\right)\right), \tag{10}$$

where the energy $E(\omega) = G_{hel}(\omega) + G_{loop}(\omega)$, κ is a parameter — usually $\kappa = 10^3$ — and

$$G_{hel}(\omega) = \frac{1}{RT} \cdot \sum_h E_h \ \text{ and } \ G_{loop}(\omega) = \sum_h 1.77 \cdot \ln(l_h + 1) + B \ , \tag{11}$$

and h varies over all hypohelices from ω. E_h represents the total stacking energy along the hypohelix h. It is the sum of stacking bond energies of the adjacent base pairs of h. B can take three different values depending on the three possible types of the loop of the hypohelix g: terminal loop, single-strand bulge and double-strand bulge.

A *codon* is a triple of successive nucleotides. For a sequence Q', each codon is fixed to be either regulatory or non-regulatory. Analogously, each nucleotide in Q is fixed to be either non T-rich or T-rich [LRSP06]. Let s_0 be the "radius" of a ribosome — distance from P-site to the end of the ribosome — usually $s_0 = 12$, and let s_1 be the "radius" of a polymerase — distance from the $5'$ end of a polymerase to its transcription center — usually $s_1 = 9$. The rate of the rule (5) is denoted λ_{rib} and is constant when $R - s_0$ is a position of a non-regulatory codon, and otherwise λ_{rib} depends on an external parameter c — the

concentration of charged tRNA [SB91]. The rate of the rule (6) is denoted ν and depends on secondary structure ω in the window. The rule (7) applies only when $P + s_1$ is a position of a T-rich nucleotide and its rate is denoted μ.

In [LRSP06] the rate of the rule (5) was denoted λ_{rib} and

$$\lambda_{rib}(c) = \frac{45\, c}{1 + c} \ . \tag{12}$$

The rate of the rule (6) was denoted ν and

$$\nu = 40 - F(\omega) \ . \tag{13}$$

The rate of the rule (7) was denoted μ and

$$\mu = \frac{1}{4} F(\omega) \ . \tag{14}$$

The function $F(\omega)$ in (13-14) for $\omega = f_1(*), \ldots, f_n(*)$ depends only on functional symbols (hypohelices) $f_1, \ldots, f_n$, and not on the structure of their arguments denoted by $*$. More precisely $F(\omega) = \max_i F(f_i)$, where

$$F(f) = \frac{\delta \cdot \exp\left(-\frac{r(f)}{r_0}\right)}{(L_2)^2 \cdot (p(f) - p_0)^2 + 1} \ , \tag{15}$$

with $p(f) \approx \frac{\pi}{|\, supp(f)|}$, and $r(f)$ the "free distance" from f to the end P of the window: for $f = (A, B, C, D)$ and $w = (R, P)$, we have

$$r(f) = R - D - \sum_i |\, supp(f_i)| \ . \tag{16}$$

Other symbols in equation (15) denote constants: $r_0 = 1, \delta = 30, L_2 = 27.1, p_0 = 0.18$, see [LRSP06].

Note that the rates of the rules depend only on the local configuration as explained above and not on the outside context. In particular it does not depend on instantiations of $x_1, \ldots, x_m, y_1, \ldots, y_n$.

6 Simulation results

ATGAAAGCAATTTTCGTACTGAAAGGTTGGTGGCGCACTTCCTGAAACGGGCAGTGT
ATTCACCATGCGTAAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTG

Figure5. A regulatory region for *trpE* genes in *E. coli*.

We have adapted the simulator described in [LRSP06] and available at [RNA] to obtain sequences of terms. As an example in Figure 6 we give one (slightly shortened and simplified) terminating trajectory of the regulation process for the *trpE* genes (responsible for the synthesis of tryptophan) in *E. coli*. The regulatory region itself is presented in Figure 5.

7 Related Work

References to the literature on RNA regulation mechanisms can be found in [LRSP06,LPRS07].

Term rewriting systems have been used in the so called *Regular Model Checking* framework [KMM+01,BT02,AJMd02,ALdR05]. They have been successfully applied to the analysis of parameterized systems [BT02,AJMd02,ALdR05] and multithreaded programs [BT02,BT03,Tou05]. However, in the regular model checking framework, the rewriting rules are not probabilistic. This work constitutes the first step towards the extension of the regular model checking framework with probabilistic rewriting rules. This would allow for example the analysis of probabilistic parameterized systems and probabilistic multithreaded programs.

Rewriting systems have also been used in articles [BIK06,BCC+03] to model chemical reactions. Compared to our work, the rewriting systems considered in [BIK06,BCC+03] are not probabilistic. Moreover, these works consider the modeling of chemical reactions whereas we consider modeling of RNA secondary structure.

Finally, probabilistic term rewriting systems have also been considered in [BH03,BK02,KSMA03]. But in these works, the symbols are of fixed arities and the terms are ordered, whereas in our framework, the symbols have arbitrary arities and the terms are not ordered. Moreover, as far as we know, this is the first time that probabilistic term rewriting systems are used to model attenuation regulation.

8 Conclusions and perspectives

We have established that the framework of probabilistic term rewriting systems provides compact and structured description of detailed models of RNA regulation.

We intend to continue exploration of this framework. The most important task consists in the development of adequate data structures and algorithms, as well as approximation and abstraction methods for analysis of this kind of models. The next step would be a massive computational experimentation, the biological interpretation of results and validation of results by real biological data.

Acknowledgments

The authors are thankful to Sergey Pirogov, Konstantin Gorbunov and Lev Rubanov for a valuable discussion. Lev Rubanov has also provided assistance in use of the RNAMODEL tool. Oleg Zverkov has helped us in preparing computer graphics for this article.

References

AJMd02. Parosh Aziz Abdulla, Bengt Jonsson, Pritha Mahata, and Julien d'Orso. Regular tree model checking. In *CAV'02*, volume 2404 of *Lecture Notes in Computer Science*, pages 555–568, 2002.

ALdR05. Parosh Aziz Abdulla, Axel Legay, Julien d'Orso, and Ahmed Rezine. Simulation-based iteration of tree transducers. In *TACAS'05*, volume 3440 of *Lecture Notes in Computer Science*, pages 30–44, 2005.

BCC+03. Olivier Bournez, Guy-Marie Côme, Valérie Conraud, Hélène Kirchner, and Liliana Ibanescu. A rule-based approach for automated generation of kinetic chemical mechanisms. In *RTA'03*, volume 2706 of *Lecture Notes in Computer Science*, pages 30–45. Springer, june 2003.

BH03. Olivier Bournez and Mathieu Hoyrup. Rewriting logic and probabilities. In *RTA'03*, volume 2706 of *Lecture Notes in Computer Science*, pages 61–75. Springer, June 2003.

BIK06. Olivier Bournez, Liliana Ibanescu, and Hélène Kirchner. From chemical rules to term rewriting. In *6th International Workshop on Rule-Based Programming*, volume 147(1) of *ENTCS*, pages 113–134, 2006.

BK02. Olivier Bournez and Claude Kirchner. Probabilistic rewrite strategies: Applications to ELAN. In *RTA'02*, volume 2378 of *Lecture Notes in Computer Science*, pages 252–266. Springer-Verlag, July 2002.

BT02. Ahmed Bouajjani and Tayssir Touili. Extrapolating tree transformations. In *CAV'02*, volume 2404 of *Lecture Notes in Computer Science*, pages 539–554, 2002.

BT03. Ahmed Bouajjani and Tayssir Touili. Reachability analysis of process rewrite systems. In *FSTTCS'03*, Lecture Notes in Computer Science, pages 73–87, 2003.

FFHS00. Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, 2000.

KMM+01. Yonit Kesten, Oded Maler, Monica Marcus, Amir Pnueli, and Elad Shahar. Symbolic model checking with rich assertional languages. *Theoretical Computer Science*, 256:93–112, 2001.

KSMA03. Nirman Kumar, Koushik Sen, José Meseguer, and Gul Agha. A rewriting based model for probabilistic distributed object systems. In *FMOODS'03*, volume 2884 of *Lecture Notes in Computer Science*, pages 32–46, 2003.

LPRS07. Vassily Lyubetsky, Sergey Pirogov, Lev Rubanov, and Alexander Seliverstov. Modeling classic attenuation regulation of gene expression in bacteria. *Journal of Bioinformatics and Computational Biology*, 5(1), 2007. in print.

LRSP06. Vassily Lyubetsky, Lev Rubanov, Alexander Seliverstov, and Sergey Pirogov. Model of gene expression regulation in bacteria via formation of RNA secondary structures. *Molecular Biology*, 40(3):440–453, 2006.

RNA. RNAmodel. Model of RNA-related regulation in bacteria. http://lab6.iitp.ru/rnamodel/rnamodee.html.

SB91. Maxine Singer and Paul Berg. *Genes & genomes*. University Science Books Mill Valley, Calif, 1991.

Tou05. Tayssir Touili. Dealing with communication for dynamic multithreaded recursive programs. In *1st VISSAS workshop*. IOS Press, 2005.

Zuk03. Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

Figure6. A simulation result: one typical terminating trajectory for classical attenuation regulation of *trpE* genes in *E. coli*. **Notations:** $\rightarrow$ means one rewriting; $\stackrel{*}{\rightarrow}$ means several similar rewritings; repeated window positions (e.g. repetitions of $\langle 40, 51\rangle$)are replaced by a $\cdot$ symbol; $\perp$ means termination. There are 24 helices, denoted by letters from a to x.