INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Multi-Site Constraint Programming Model of Alternative Splicing Regulation

Damien Eveillard — Delphine Ropers — Hidde de Jong — Christiane Branlant — Alexander Bockmayr

N° 4830

Mai 2003

THÈMES 2 et 3

ISSN 0249-6399 ISRN INRIA/RR--4830--FR+ENG

A Multi-Site Constraint Programming Model of Alternative Splicing Regulation

Damien Eveillard [*] [†], Delphine Ropers [†], Hidde de Jong [‡], Christiane Branlant[†], Alexander Bockmayr [*§]

Thèmes 2 et 3 — Génie logiciel
et calcul symbolique — Interaction homme-machine,
images, données, connaissances
Projets Modbio et Helix

Rapport de recherche n° 4830 — Mai 2003 — 29 pages

Abstract: Alternative splicing is a key process in post-transcriptional regulation, by which several kinds of mature RNA can be obtained from the same premessenger RNA. The resulting combinatorial complexity contributes to biological diversity, especially in the case of the human immunodeficiency virus HIV-1. Using a constraint programming approach, we develop a model of the alternative splicing regulation in HIV-1. Our model integrates different scales (single site vs. multiple sites), and thus allows us to exploit several types of experimental data available to us.

Key-words: computational biology, constraint programming, modeling, hybrid system, alternative splicing, HIV-1

[*] Equipe Modbio, LORIA (UMR 7503 – CNRS, INPL, INRIA, Université Henri Poincaré Nancy 1, Université Nancy 2), BP 239, 54506 Vandœuvre-lès-Nancy, France

[†] Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, BP 239, 54506 Vandœuvre-lès-Nancy, France

[‡] Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier, France

[§] Part of this work was done within the ARC INRIA "Process Calculi and Biology of Molecular Networks", `http://contraintes.inria.fr/cpbio`

Modèle multi-site en programmation par contraintes de la régulation de l'épissage alternatif

Résumé : L'épissage alternatif est un processus clef parmi ceux qui composent la régulation post-transcriptionnelle. Il permet d'obtenir différents ARN matures provenant du même ARN prémessager. La complexité combinatoire qui en résulte contribue à la diversité biologique. Ceci est particulièrement explicite au cours du cycle de vie du virus d'immunodéficience humaine de type 1 (VIH-1). En utilisant la programmation par contraintes, nous proposons un modèle de régulation de l'épissage alternatif de VIH-1. Ce dernier intègre différentes échelles (un unique site d'épissage contre plusieurs sites d'épissage) et nous permet ainsi d'exploiter des données expérimentales de natures différentes à notre disposition.

Mots-clés : bioinformatique, biologie computationnelle, programmation par contraintes, modélisation, système hybride, épissage alternatif, VIH-1

1 Introduction

Alternative splicing is a biological process occurring in post-transcriptional regulation of RNA. Through the elimination of selected introns, alternative splicing allows the generation of several kinds of mature RNA from the same premessenger RNA. The resulting combinatorial complexity contributes to biological diversity, especially in the case of the human immunodeficiency virus (HIV-1). Recent biological studies show the impact that SR proteins have on the dynamics of post-transcriptional regulation via the control of the splicing process [9]. SR proteins can be divided into two functional classes: they may either activate or inhibit splicing. Due to the complexity of alternative splicing regulation, the knowledge that can be gained from experiments is limited. Each experiment focuses on one splicing site. In a first approach, we model SR regulation in this restricted context. Using differential equations, we develop a model for the regulation of the A3 splicing site in HIV-1. The qualitative behavior depends on the values of the reaction kinetic parameters. Experimental results available to us validate this first approach in the equilibrium phase. Our second approach aims at validation on a higher scale. The ultimate goal is to obtain a model that can be validated qualitatively both on the scale of a single splicing site and on the scale of the whole HIV-1, in order to represent the global effect of alternative splicing in the HIV-1 cycle.

Our models are developed in a constraint programming framework [2, 3]. Constraint programming seems well-suited for modeling biological systems because it allows one to handle partial or incomplete information. Each constraint gives one piece of information on the system that is studied. The overall knowledge is accumulated in the constraint store. The constraint engine available in constraint programming systems operates on the constraint store. It may add new information to the store or check whether some property is entailed by the information present in the store. While a constraint model may be refined whenever additional biological knowledge becomes available, it allows one to make useful inferences even from partial and incomplete information. Therefore, constraint programming seems to be a natural computational approach to face the current situation in systems biology as it is described by [17]: "Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviors. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behavior precisely."

The organisation of the paper is as follows: we start in Section 2 with a description of the biological process of alternative splicing regulation. Based on a number of biological hypotheses, we develop in Section 3 a continuous model of the regulation at one splicing site. This model includes competition and compensation of different proteins on two binding sites, ESE and ESS2. The single-site model is validated in a qualitative way by extracting from the model a splice efficiency function, which can be measured in experiments. In Section 4, we briefly present the hybrid concurrent constraint programming language `Hybrid cc` [10, 11], and explain how it can be used for modeling dynamic biological systems. In Section 5, we first simulate the single-site continuous model in this language. Then we derive a more global

model involving three generic splicing sites, which may be generalized to multiple sites. This means that we model at two different scales, using the splice efficiency function as a time-scale abstraction of the local model of one site in the more global context of different sites. The three-site model uses the constraint solving and default reasoning facilities of `Hybrid cc`. This allows us to make predictions on the global behavior even in absence of detailed local information on some of the splicing sites. This report extends earlier work presented in [8].

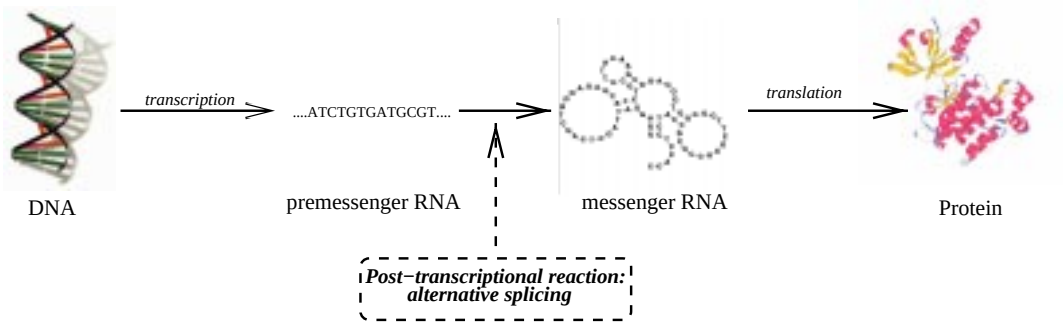

Figure 1: Information flow in molecular biology

2 Alternative splicing: A biological problem for formals methods

2.1 The biological problem of alternative splicing regulation

Molecular biology studies the information flow in the *transcription* of DNA into RNA, and the *translation* of RNA in proteins, see Fig. 1. In eukaryotes and viruses, the transcriptional process, where information contained in the DNA molecule is extracted and transcribed into an RNA molecule, is followed by another process, which is *alternative splicing* [15]. In fact, the DNA molecule first yields a *premessenger* RNA molecule. The premessenger RNA can be decomposed in exons and introns. During splicing, introns are removed. Only the exons are kept for the *messenger* or *mature* RNA (mRNA).

The regulation of the splicing process depends on several sites. The first one is the *donor site SD*, located at the end of one exon, see Fig. 2. Its main characteristic is the `GU` nucleic acid sequence motif, which contains the phosphate domain. The other site is the *acceptor site SA* located at the beginning of the next exon. It is characterized by an `AG` motif containing another phosphate domain. Together, they define the intron to be excised from the premessenger RNA. They permit the binding of a huge ribonucleoproteic complex: the spliceosome. This complex is partially activated by another motif, the *branching point BP*. This is another binding site contained inside the intron. The three sites permit the regulation of the splicing process by activation of the spliceosome complex. The key to the regulation is the choice of one acceptor and one donor site.

2.2 Combinatorial complexity

The splicing process is a key source of complexity in the cell. As illustrated by Fig. 3, various messenger RNA can be obtained from a unique premessenger RNA through the elimination of different introns, and the junction of the remaining exonic sequences. This process is regulated through the choice of the donor and acceptor sites. The complexity of

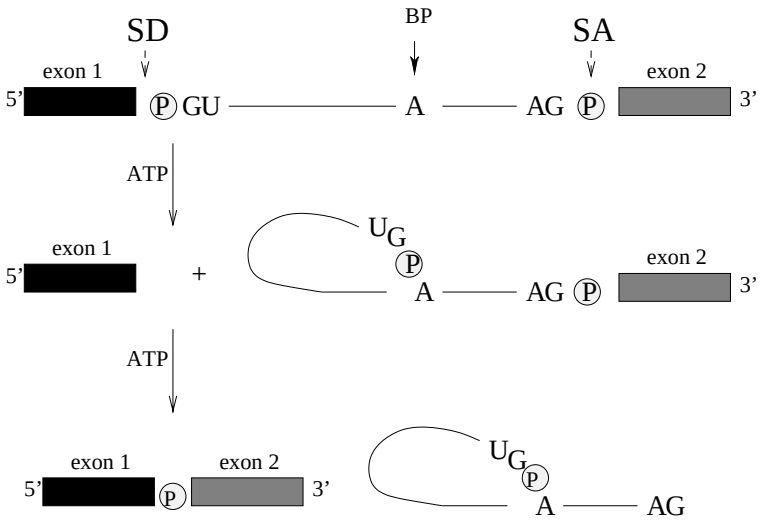

Figure 2: The splicing process operates in the intronic region of a premessenger RNA that lies between two exons. The exons are delimited by the SD and SA binding sites. A first reaction cuts the RNA at the SD binding site. A second reaction cuts in the phosphate domain of the SA binding site. Each reaction requires ATP energy.

the regulation increases with the capacity of an exon to become an intron in a given cellular condition.

In Fig. 3, seven types of intron elimination are shown. In each case, the dotted lines above and below the premessenger RNA give two concrete examples of intron elimination. The first case describes classical exons. The second case can be seen as a competition between different exons. In the third and the fourth case, regulation depends on the position of the acceptor and the donor site respectively. In the fifth case, an intron is partitioned between two exons. In the last two cases, the choice between different promoters or different polyadenylation sites increases the complexity of the alternative splicing.

2.3 Alternative splicing in the context of HIV-1

Eukaryotic and viral gene expression involves the production of RNA containing intronic sequences. The process of splicing allows for intron elimination and junction of exonic sequences. Splicing is a major biological process in the HIV-1 life cycle: the viral RNA either remains unchanged to serve as genomic RNA for new virions, or it is alternatively spliced to allow for the production of virion proteins [24]. The viral genome, initially RNA, is integrated into the host genome. In the HIV-1 case, splicing regulation is a complex phenomenon involving 4 donor sites (SD) and 8 acceptor sites (SA), which may yield 40 mature messenger RNAs [18]. This combinatorial complexity is achieved by regulating the

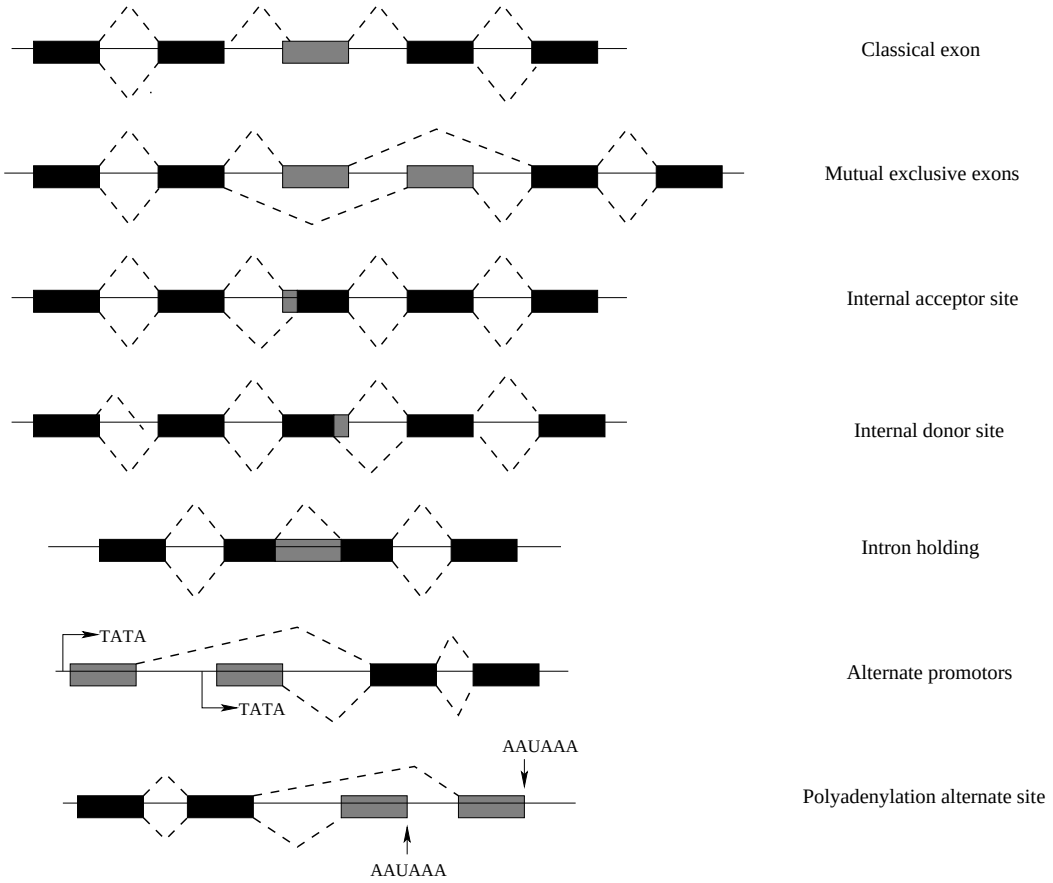

Figure 3: Complexity induced by alternative splicing: seven types of intron elimination

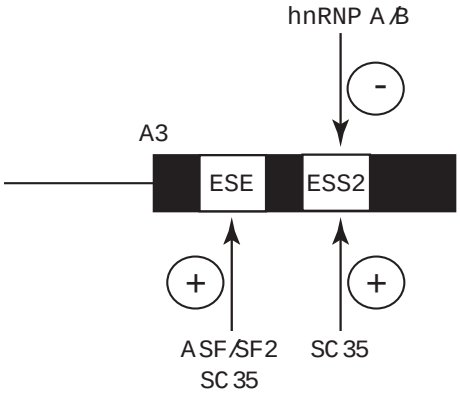

Figure 4: Regulatory elements of the A3 splicing site. The exon delimited by the A3 acceptor site contains the ESE and ESS2 binding sites, which bind ASF/SF2, SC35 and hnRNP A/B proteins. These regulatory elements activate or repress the splicing reaction on the A3 site.

selection of the acceptor site [16, 18]. Protein factors control the regulation via binding sites. We focus in our study on the acceptor site A3, see Fig. 4, where splicing can be repressed by hnRNP A/B proteins via the ESS2 binding site [4, 7]. Recent experimental studies carried out in our group [19] show that an ESE sequence can activate splicing at the A3 site when SC35 and ASF/SF2 proteins bind to it. SC35 can also bind to the ESS2 site. More specifically, the ratio of hnRNP A/B and SR proteins determines the splice efficiency at the A3 site.

3 Modeling one splicing site

3.1 Biological hypotheses

We model the regulation by SR proteins in the restricted context of the A3 splicing site under the following hypotheses, see Fig. 4:

- We study only one splicing site. Thus, we consider regulation at the scale corresponding to our experimental results, which are measurements of the splice efficiency given as the ratio of mature RNA over premessenger RNA.
- We suppose that the splicing process involves two reactions, relating three functional classes of RNA, see Fig. 2: immature RNA (`rna`), intermediate RNA (`irna`), and mature RNA (`mrna`). Intermediate RNA corresponds to immature RNA activated by proteins. Mature RNA corresponds to mature RNA and introns in lariat.
- The protein concentration in experiments is saturated. Therefore, we assume that it is constant, despite the binding of proteins to the RNA during regulation.
- SR proteins regulate the splicing process by initialization of the splicing machinery.
- Regulation is controlled by the ESE and ESS2 binding sites, which are independent.
- The SR proteins ASF/SF2 and SC35 may activate the first splicing reaction by binding to the site ESE. We assume that these two proteins compensate each other.
- The hnRNP A/B proteins may inhibit the first splicing reaction by binding to the site ESS2. On the other hand, if the SC35 proteins bind to ESS2, this activates the first splicing reaction. Therefore we have a competition between hnRNP A/B and SC35.
- Due to a lack of experimental results, we assume Michaelis-Menten kinetics for the reactions involved in the splicing process.

The biological hypotheses are summarized in Fig. 5

3.2 Mathematical model

Our biological hypotheses can be represented by a system of ordinary differential equations based on the Michaelis-Menten relation [23]. The single-site model that we obtain will later be integrated into a larger multi-site model, see Section 5. We will describe the splicing process by seven kinetic reactions. The symbols used are given in Tab. 1.

The reaction r_1 represents the transformation of premessenger RNA (rna) to intermediate RNA (irna). It requires cooperation between ASF/SF2 and SC35 proteins for the regulation of ESE. Since we assume compensation, only the sum of the activator proteins is important.

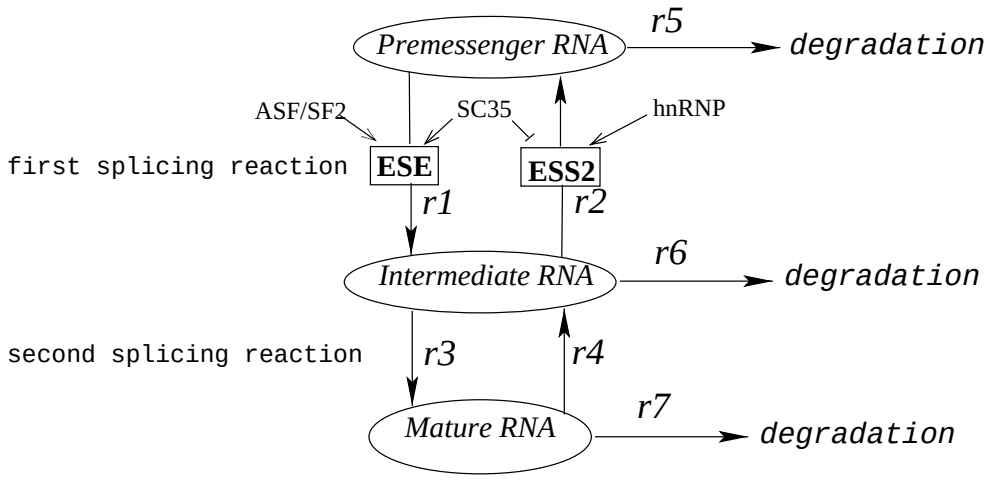

Figure 5: Schematic representation of the splicing site regulation. The two splicing reactions are composed by 7 kinetic reactions.

We represent this interaction by a Michaelis-Menten function depending on the quantity of immature RNA, and controlled by the sum of the proteins ASF/SF2 and SC35 [14]

$$r_1 = \frac{\varphi_{ESE}(ASF + SC)}{k_{ESE} + ASF + SC}\, rna$$

The reaction r_2 represents the transformation of intermediate RNA to premessenger RNA. It captures the antagonistic function of hnRNP A/B and SC35 proteins on the site ESS2. In this case, the Michaelis-Menten function represents the inhibitive competition between two proteins: hnRNP A/B and SC35 [25]. It depends on the quantity of intermediate RNA:

$$r_2 = \frac{\varphi_R \cdot R}{k_R(1 + \frac{SC}{k_{SC}+R})}\, irna$$

The reaction r_3 represents the transformation of intermediate RNA to mature RNA (mrna). We assume for this reaction a simple first-order kinetics with a constant parameter κ. Similarly, r_4 represents the reaction which transforms mature RNA to intermediate RNA:

$$r_3 = \kappa \cdot irna, \qquad r_4 = \kappa' \cdot mrna.$$

r_5, r_6 and r_7 respectively represent the degradation reaction of immature RNA, intermediate RNA and mature RNA. Different RNAs decrease proportional to the same degradation factor λ:

$$r_5 = \lambda \cdot rna, \quad r_6 = \lambda \cdot irna, \quad r_7 = \lambda \cdot mrna.$$

Table 1: Symbols and units for the biological variables and parameters

Symbol	Variables and Parameters	unit
rna	Immature RNA	μM
$irna$	Intermediate RNA	μM
$mrna$	Mature RNA	μM
ASF	Protein ASF/SF2	μM
SC	Protein SC35	μM
R	Protein hnRNP A/B	μM
φ_{ESE}	Maximal affinity for the enhancer	s^{-1}
φ_R	Maximal affinity of hnRNP A/B	s^{-1}
k_{ESE}	Half saturation coefficient for the enhancer	μM
k_{SC}	Half saturation coefficient for SC35	μM
k_R	Half saturation coefficient for hnRNP A/B	μM
κ	Reaction rate	s^{-1}
κ'	Reaction rate	s^{-1}
λ	Degradation coefficient	s^{-1}

We formalize the splicing process at site A3 by the system of differential equations:

$$\begin{aligned}
\frac{d(rna)}{dt} &= r_2 - r_1 - r_5, \\
\frac{d(irna)}{dt} &= r_1 + r_4 - r_2 - r_3 - r_6, \\
\frac{d(mrna)}{dt} &= r_3 - r_4 - r_7,
\end{aligned}$$

which corresponds to

$$\begin{aligned}
\frac{d(rna)}{dt} &= \frac{\varphi_R \cdot R}{k_R(1+\frac{SC}{k_{SC}+R})}\, irna - \frac{\varphi_{ESE}(ASF+SC)}{k_{ESE}+ASF+SC}\, rna - \lambda \cdot rna, \\
\frac{d(irna)}{dt} &= \frac{\varphi_{ESE}(ASF+SC)}{k_{ESE}+ASF+SC}\, rna - \frac{\varphi_R \cdot R}{k_R(1+\frac{SC}{k_{SC}+R})}\, irna \\
&\quad -\kappa \cdot irna + \kappa' \cdot mrna - \lambda \cdot irna, \\
\frac{d(mrna)}{dt} &= \kappa \cdot irna - \kappa' \cdot mrna - \lambda \cdot mrna.
\end{aligned}$$

3.3 Validation of the regulatory system

The mathematical model of regulation at a single site can be directly simulated in the constraint programming language `Hybrid cc`, as will be shown in Section 4. However, it should first be validated with respect to existing biological knowledge. In our model, the RNA concentrations do not reach an equilibrium, but continue to decrease until total degradation of RNA. However, we may assume that the splicing reactions quickly reach an equilibrium. In the equilibrium phase, we have $r_1 = r_2, r_3 = r_4$, which is equivalent to

$$\frac{\varphi_{ESE}(ASF + SC)}{k_{ESE} + ASF + SC}\, rna \;=\; \frac{\varphi_R \cdot R}{k_R(1 + \frac{SC}{k_{SC}+R})}\, irna, \qquad \kappa \cdot irna \;=\; \kappa' \cdot mrna.$$

If we define the *splice efficiency* by

$$\mathit{efficiency}(t) = \frac{mrna(t)}{rna(t)},$$

we obtain the following formula for the splice efficiency in the equilibrium phase:

$$\mathit{efficiency}_{eq} = \frac{\kappa \cdot \varphi_{ESE}(ASF + SC)(k_R \cdot k_{SC} + k_R \cdot SC + R \cdot k_{SC})}{\kappa'(k_{ESE} + ASF + SC) \cdot \varphi_R \cdot R \cdot k_{SC}}$$

According to our formula, the splice efficiency is

- an increasing function of the activators SC and ASF.
- a decreasing function of the inhibitor R.

Experimental results show that

- $(mrna/rna)_{eq}$ increases with an increase of activator proteins.
- $(mrna/rna)_{eq}$ decreases with an increase of inhibitor proteins.

Thus, the results of our model correlate with available experimental data. Therefore, we may consider the model to be qualitatively validated under the hypotheses described in Section 3.1. We next consider simulation in the concurrent constraint language `Hybrid cc`.

Table 2: Combinators of `Hybrid cc`

Agents	Propositions
c	c holds now
`if` c `then` A	if c holds now, then A holds now
`if` c `else` A	if c will not hold now, then A holds now
`new` X `in` A	there is an instance $A[T/X]$ that holds now
(A, B)	both A and B hold now
`hence` A	A holds at every instant after now
`always` A	same as $(A,$ `hence` $A)$
`unless(`c`)` A `else` B	same as `(if` c `then` B`, if` c `else` A`)`

4 Hybrid concurrent constraint programming

To model alternative splicing regulation, we will use hybrid concurrent constraint programming. The general idea of *constraint programming* is that the user specifies constraints on the behavior of the system that is being studied. Each constraint expresses some partial information on the system state. The constraint solver may check constraints for consistency or infer new constraints from the given ones. In *concurrent constraint programming* (`cc`), different computational processes may run concurrently. Interaction is possible via the *constraint store*. The store contains all the constraints currently known about the system. A process may *tell* the store a new constraint, or *ask* the store whether some constraint is entailed by the information currently available, in which case further action is taken [20]. One major difficulty in the original `cc` framework is that `cc` programs can detect only the presence of information, not its absence. To overcome this problem, [21] proposed to add to the `cc` paradigm a sequence of phases of execution. At each phase, a `cc` program is executed. At the end, absence of information is detected, and used in the next phase. This results in a synchronous reactive programming language, `Timed cc`. But, the question remains how to detect negative information instantaneously. `Default cc` extends `cc` by a negative ask combinator `if` c `else` A, which imposes the constraints of A unless the rest of the system imposes the constraint c. Logically, this can be seen as a default. Introducing phases as in `Timed cc` leads to `Timed Default cc` [22]. Only one additional construct is needed: `hence` A, which starts a copy of A in each phase after the current one.

`Hybrid cc` [10, 11] is an extension of `Default cc` over continuous time. First continuous constraint systems are allowed, i.e., constraints may involve differential equations that express initial value problems. Second, the `hence` operator is interpreted over continuous time. It imposes the constraints of A at every real time instant after the current one. The evolution of a system in `Hybrid cc` is piecewise continuous, with a sequence of alternating point and interval phases. All discrete changes take place in a point phase, where a simple `Default cc` program is executed. In a continuous phase, computation proceeds only through the evolution of time. The interval phase, whose duration is determined in the

previous point phase, is exited as soon as the status of a conditional changes [11]. Tab. 2 summarizes the basic combinators of `Hybrid cc`.

It has been argued in [2, 3] that `Hybrid cc` is well-suited for modeling dynamic biological systems. In addition to the general discussion in [3], we illustrate here by a number of small examples, how the basic combinators of `Hybric cc` can be applied naturally to the study of biological systems.

4.1 Interval constraints and continuous dynamics

The `Hybric cc` language that we are using is based on interval constraints [5]. This means that variables are defined over an interval of real numbers, and computations are done in interval arithmetic. This is very useful in biology, where typically parameters and values are not exactly known.

We illustrate this by a very simple example in `Hybrid cc` involving a single constraint on an interval variable x, see Fig. 6. Since we are reasoning about dynamical systems, we use the `always` A combinator, expressing that A holds at every time instant.

```
interval x;
x = [9.5,10];
always { x' = -(2*x)/(15+x);
       }
sample(x);
```

4.2 Parallel composition

`Hybrid cc` allows for parallel composition of constraints. (A, B) imposes the constraints of both A and B. Operationally, the program (A, B) behaves like the simultaneous execution of both A and B. A and B may share common variables, and thus communicate via the constraint store.

We illustrate parallel composition by a small `Hybrid cc` program specifying a Michaelis-Menten kinetics. Consider two molecular species X and Y with concentrations x and y, and suppose X is transformed into Y. The initial concentration of X lies in the interval $[14, 14.5]$. The production rate of Y depends on the concentration of X according to the formula $y' = (A_{max} * x)/(k_s + x)$, for some constants A_{max} and k_s. The concentration of X is reduced at the same rate. We add constraints $x, y \geq 0$ to say that concentrations are non-negative, and constraints $s = x + y, s' = 0$ to express conservation of matter. The constraint solver computes enclosures for x and y, see Fig. 7. We can observe that at the end of the experiment, the concentration of y will be greater than the concentration of x. The upper bound for the variable y follows from the non-negativity constraints and conservation of matter. Interval constraints are particularly useful in sensibility studies, where we can easily test the importance of one variable versus the others.

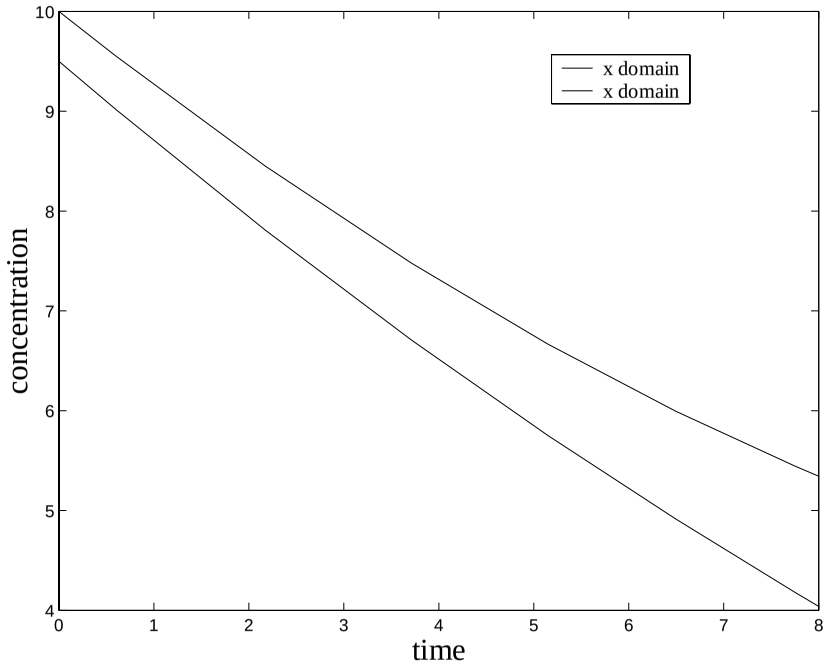

Figure 6: Enclosure for the dynamics of a molecular species with Michaelis-Menten kinetics

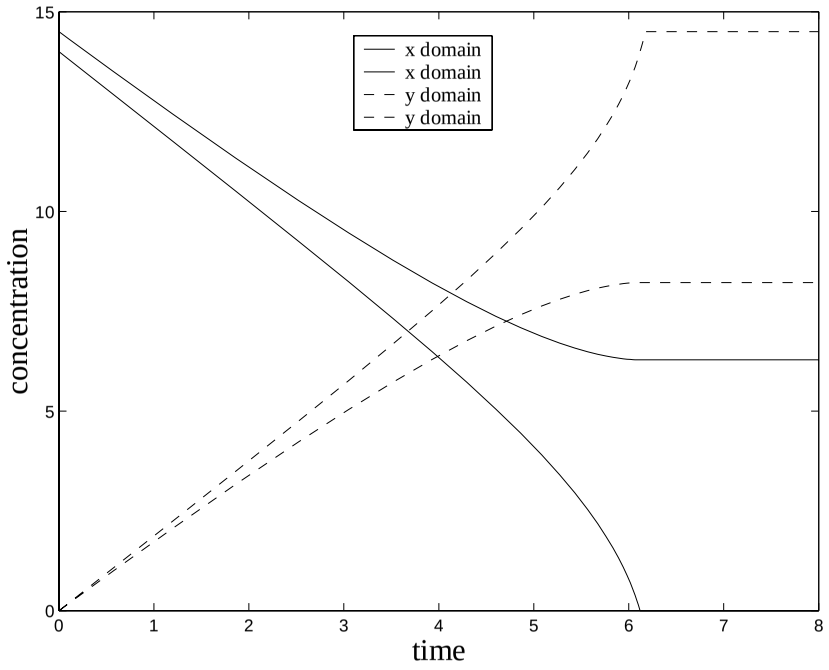

Figure 7: Enclosures for the dynamics of two molecular species with Michaelis-Menten kinetics

```
#define ks 1.5
#define Amax  2
interval x,y,s;
x = [14,14.5];             /* Initialization */
y = 0;
always {
   x' = -(Amax*x)/(ks+x); /* Michaelis-Menten kinetics */
   y' =  (Amax*x)/(ks+x);
   x >= 0;                /* Non-negative concentrations */
   y >= 0;
   s = x + y;             /* Conservation of matter */
   s' = 0;
}
sample(x, y);
```

4.3 Conditionals and discrete change

In general, the dynamics of a system will depend on conditions. In `Hybric cc`, we may use the combinator `if` c `then` A expressing that if c holds now, then A holds now. This allows one to make discrete changes to switch from one dynamics to another. The next program models the situation that the transformation of X to Y gets activated if a certain protein P reaches a threshold.

```
interval x, y, p;
x=[14,14.5];
y=0;
p=0.75;
always {
   p' = 1.5;
   if (p >= 3)
      { x' = -(Amax*x)/(ks+x);
        y' =  (Amax*x)/(ks+x);
   }
   if (p < 3)
      { x' = 0;
        y' = 0;
   }
}
sample(p, x ,y);
```

Here, we have assumed that there is no uncertainty on the initial value of p. Without this hypothesis, the constraint solver cannot decide between the two alternatives $p \geq 3$ and $p < 3$. In order to handle conditions in presence of uncertainty, we use default reasoning that we describe next.

4.4 Default behavior

The default combinator `if` c `else` A (or `unless(`c`)` A in the syntax of Hybrid cc) expresses that A holds now, if c will not hold now. Operationally, this means that the current store on quiescence does *not* entail c. Note that `unless`(c) A is not equivalent to `if` $\neg c$ `then` A. If A is executed, this may have two reasons:

- The current store entails $\neg c$ (in this case `unless`(c) A behaves like `if` $\neg c$ `then` A), or
- the current store neither entails c nor $\neg c$, i.e., it is not known whether or not c holds. In this case, A is executed *by default*.

A is not executed, if the current store entails c.

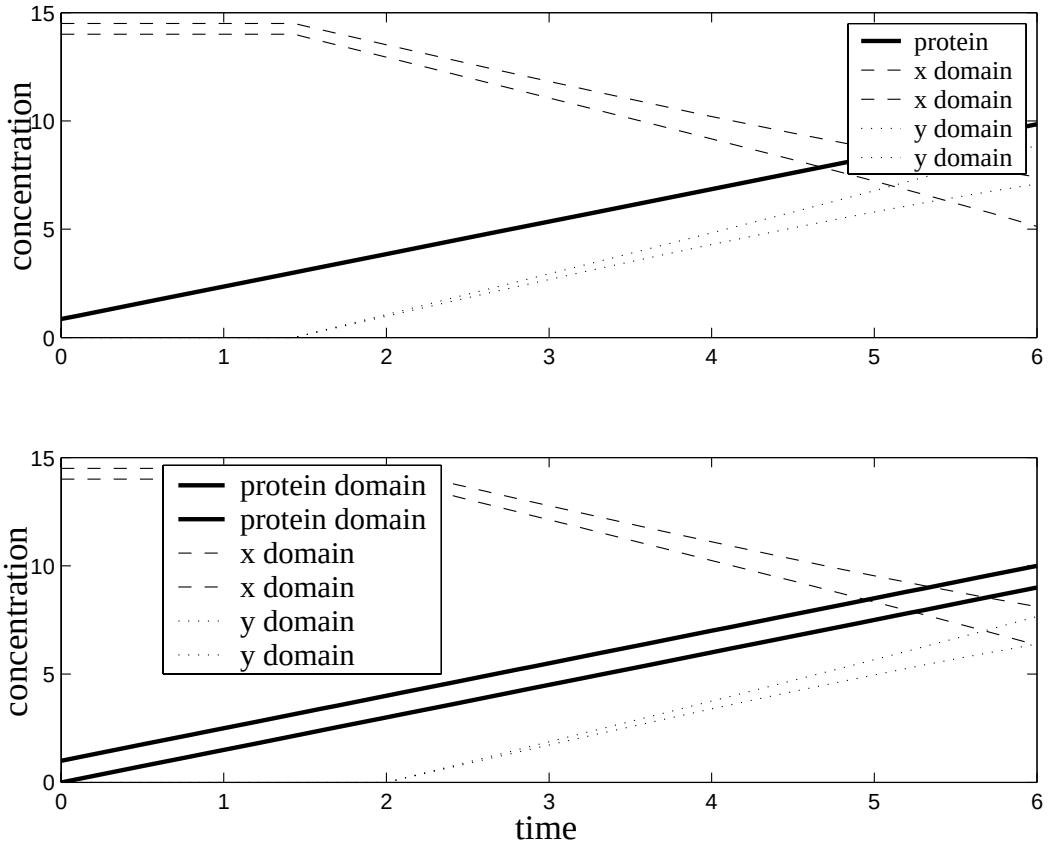

Figure 8: Switching behaviour for conditional (top) and default combinator (bottom).

We use the same example as before. The only difference is that the variable p representing the protein concentration is initialized with the interval $[0, 1]$. As we can see in Fig. 8, the reaction gets activated when the *lower* bound for p reaches the threshold.

```
interval x, y, p;
x=[14,14.5];
y=0;
p=[0,1];
always {
   p' = 1.5;
   if ( p >= 3 )
      { x' = -(Amax*x)/(ks+x);
        y' =  (Amax*x)/(ks+x);
   }
   unless (p >= 3)
      { x' = 0;
        y' = 0;
   }
}
sample(p, x ,y);
```

The default combinator is a convenient way of handling incomplete knowledge in biology. In particular, we will use it in our multi-site model of alternative splicing regulation in Section 5.2.

5 Modeling the alternative splicing regulation with Hybrid cc

5.1 Single-site model: local modeling

The single-site model from Section 3.2 with experimental values can be expressed directly in `Hybrid cc`.

```
# define Pese 0.01              # define kr 0.01
# define Psc 0.2                # define k 0.19
# define Pr 0.4                 # define kk 0.01
# define kese 0.35              # define SC 2
# define ksc 2                  # define ASF 1.75
# define R 0.35

interval t, rna, irna, mrna;
t=0; rna = 0.06; irna = 0; mrna = 0;
always{
   rna'  =  (Pr*R*irna)/(kr*(1+(SC/ksc))+R)
           -(Pese*(ASF+SC)*rna)/(kese+ASF+SC)
           -delta*rna;
   irna' =  (Pese*(ASF+SC)*rna)/(kese+ASF+SC)
           -(Pr*R*irna)/(kr*(1+(SC/ksc))+R)
            -k*irna+kk*mrna-delta*irna;
   mrna' = k*irna-kk*mrna-delta*mrna;
}
sample(rna, irna, mrna);
```

During the simulation, we obtain the predicted equilibrium for the splice efficiency, see Fig. 9. Under our hypotheses, which include protein competition and compensation, the model correctly simulates the alternative splicing activity at site A3. This supports the hypotheses made in the model such as the role of the ESE and ESS2 binding sites.

5.2 Three-site model: global modeling

A realistic model of alternative splicing has to reflect the combinatorial complexity discussed in Section 2.3. Assuming that regulation is modular [13], the single-site model may be seen as one module inside a larger framework. The qualitative validation given in Section 3.3 justifies the introduction of the single-site model into a larger-scale model involving several splicing sites. To illustrate this, we consider the generic example of three acceptor sites (A3, A4 and A7) associated with one donor site (SD), see Fig. 10.

The behavior at one splicing site can be captured by a single function, the splice efficiency, which depends on the protein concentrations. This function is used in a larger-scale global

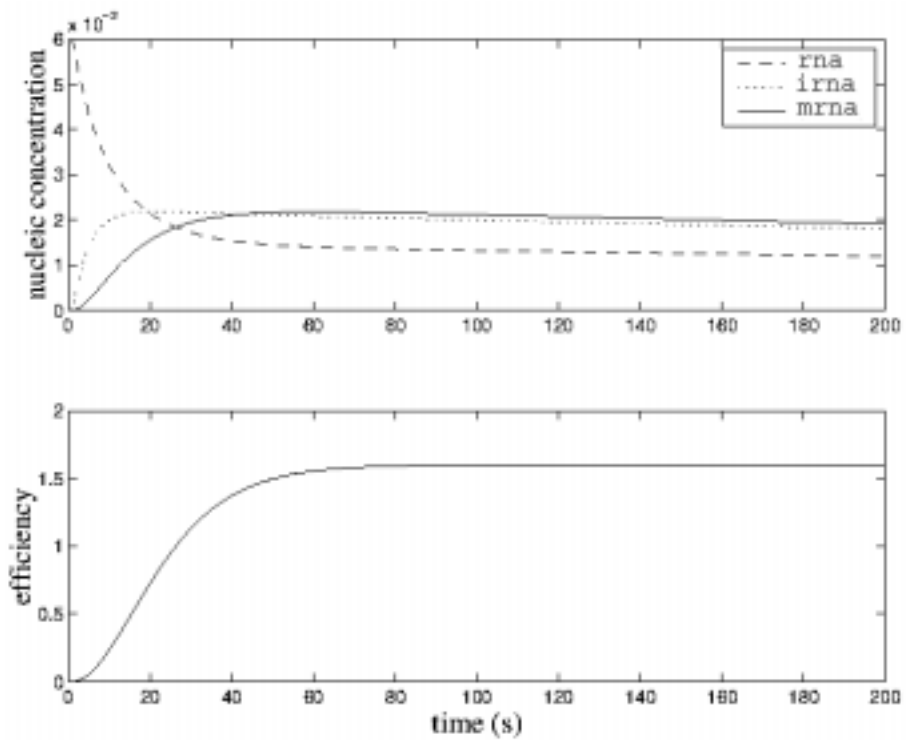

Figure 9: Variation of RNA pool and splice efficiency in the splicing reaction

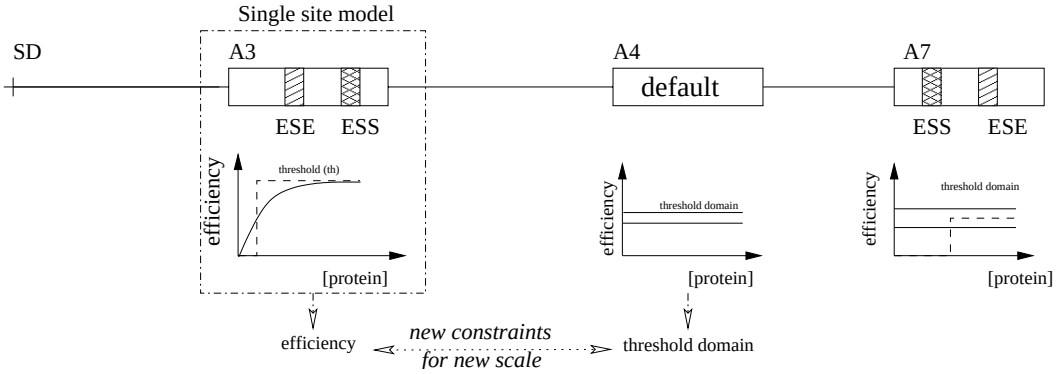

Figure 10: Single-site model inside a more general multi-site regulation model

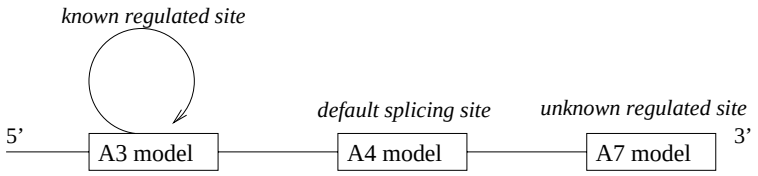

Figure 11: Biological information on three acceptor sites A3, A4, A7

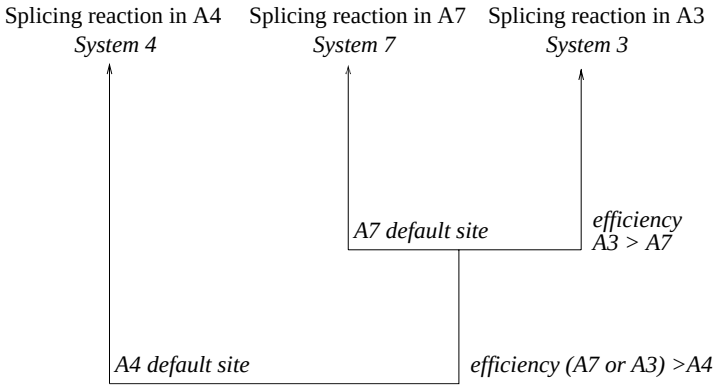

Figure 12: Choice of the acceptor site A3, A4 or A7 depending on the splice efficiency

model that describes the choice between three acceptor sites A3, A4 and A7. In the HIV-1 case, the A4 site is the default splicing site. Only if the efficiency of A3 (`effA3`) or A7 (`effA7`) gets larger than the efficiency of A4 (`effA4`), regulation switches to the other state. The sites A3, A4, and A7 exhibit three generic behaviors, see also Fig. 11:

- A3 is a regulated site with known behavior.
- A7 is a regulated site with unknown behavior.
- A4 is an unregulated site, i.e., the behavior does not depend on protein concentrations.

Current biological experiments give information on the local behavior at one site. However, modeling the local behavior is not enough. In order to understand the global splicing process, we must integrate several types of knowledge. One the one hand, we have information on the local behavior at individual acceptor sites. On the other hand, we have some information on the global behavior, like the default role of A4 or the competition between different acceptor sites. Constraint programming allows us to integrate this information, and to produce a global model.

Recent work [6] shows the linearity of the splicing kinetics. Thus, on the larger scale, we may consider splicing as a linear process described by three systems of ordinary differential

equations. For each acceptor site Ai, $i \in \{3, 4, 7\}$, we introduce one system with four differential equations:

- r_{i1} represents the consumption of immature RNA if Ai is dominating.
- r_{i3} represents the production of mature RNA at A3.
- r_{i4} represents the production of mature RNA at A4.
- r_{i7} represents the production of mature RNA at A7.

k_{ij} is the kinetic constant for reaction r_{ij}.

A4 is the default splicing site. It is dominating unless the splice efficiency of A3 or A7 gets larger than the splice efficiency of A4. If this happens, A7 becomes the default splicing site unless the efficiency of A3 gets larger than the efficiency of A7, see Fig. 12. The local behavior at A3 has been described by the single-site model given in Section 5.1. This model predicts the splice efficiency of A3 depending on the protein concentrations.

In the `Hybrid cc` program given below, the concentration of SC35 is increased linearly. Depending on the corresponding variation of the splice efficiency at A3, the three-site model exhibits different behaviors, characterized by the choice of one of the three differential equation systems. The default behavior discussed before can be expressed naturally in `Hybrid cc` using the combinator `unless(`c`)` A.

```
#define c1 2
#define c2 0.5
#define c3 0.8
#define c4 1
#define c5 0.9
#define c6 0.1
#define R 3.5

interval t, prot, effA3, effA4, effA7, rna, mrnaA3,
    mrnaA4, mrnaA7;
     t = 0;
   rna = 10;
mrnaA3 = 0; /*known regulated acceptor site */
mrnaA4 = 0; /*unregulated acceptor site*/
mrnaA7 = 0; /*unknown regulated acceptor site*/

always { t' = 10;
prot = 0.1*t;                /* protein variation */

/* A3 efficiency depends on the protein concentration
   A3 efficiency represents the local behavior of A3
   (observer of A3) */
```

```
effA3 = c1*(prot+c2)*(c3*prot+c4)/(c5*(prot+c6));

6 <= effA4; effA4 <= 8;  /* effA4 : efficiency domain of A4*/
7 <= effA7; effA7 <= 9;  /* effA7 : efficiency domain of A7*/
}

/* The behavior depends on the efficiency of
   the 3 acceptor sites*/

always {
/* if A3 or A7 dominant */
if (effA3 >= effA4 || effA7 >= effA4) {
      if (effA7 <= effA3)  {        /* splicing on A3 */
          rna' = -0.51 * rna - 0.01*rna;
       mrnaA3' =  0.4  * rna - 0.1*mrnaA3;  /* A3 kinetics */
       mrnaA4' =  0.01 * rna - 0.1*mrnaA4;  /* A4 kinetics */
       mrnaA7' =  0.1  * rna - 0.1*mrnaA7;  /* A7 kinetics */
      };
      unless ((effA7 <= effA3)) { /*default splicing on A7*/
          rna' = -0.51 * rna - 0.01*rna;
       mrnaA3' =   0.1 * rna - 0.1*mrnaA3; /* A3 kinetics */
       mrnaA4' =  0.01 * rna - 0.1*mrnaA4; /* A4 kinetics */
       mrnaA7' =   0.4 * rna - 0.1*mrnaA7; /* A7 kinetics */
      };
   }
/* default splicing on A4 */
unless (effA3 >= effA4 || effA7 >= effA4) {
     rna' = -0.32 * rna - 0.01*rna;
  mrnaA3' = -0.01 * rna - 0.1*mrnaA3; /* A3 kinetics */
  mrnaA4' =  0.3  * rna - 0.1*mrnaA4; /* A4 kinetics */
  mrnaA7' = -0.01 * rna - 0.1*mrnaA7; /* A7 kinetics */
   }
}
sample(prot, effA3, rna, mrnaA3, mrnaA4, mrnaA7)
```

According to the semantics of the default combinator, the A4 site will be chosen if the solver cannot deduce that (effA3 $\geq$ effA4) or (effA7 $\geq$ effA4). This may have *two* reasons:

- (effA3 $\geq$ effA4) or (effA7 $\geq$ effA4) is false, i.e., (effA3 < effA4) and (effA7 < effA4), or
- it is not known whether (effA3 $\geq$ effA4) or (effA7 $\geq$ effA4) holds (default behavior).

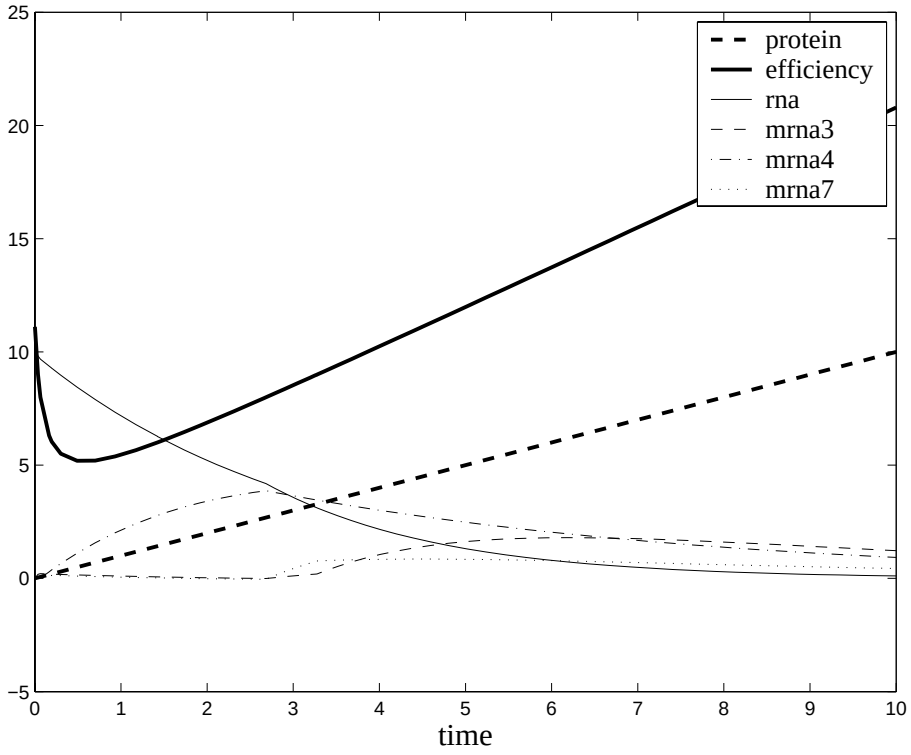

Figure 13: Variation of mRNA production depending on variation of SR proteins

Thus the A4 site is the default site if the splice efficiency of A3 and A7 is not sufficiently high. If A3 or A7 dominate A4, then A7 is the default splicing site, unless A3 dominates A7.

Simulation in `Hybrid cc` yields the behavior shown in Fig. 13. First `mrnA4` is produced, i.e., the default site A4 is active. When `effA3` passes the upper threshold for `effA4`, site A7 gets activated, and `mrnA7` is produced. Finally, when `effA3` further increases and passes the upper threshold for `effA7`, site `A3` gets activated and we observe production of `mrnA3`, while the concentrations of `mrnA4` and `mrnA7` become stationary.

Basically, the model gives to the biologist three qualitative states : first a splicing at the A4 site, second a splicing at the A7 site, and finally a splicing at the A3 site. The constraint programming system can compute enclosures for the three biological states, despite the variation in the concentration of SR proteins. The enclosure is an important qualitative information to extend the single-site to a multi-site model. `Hybrid cc` permits a qualitative validation of the model, although the currently available information on the alternative splicing regulation in HIV-1 is incomplete.

6 Conclusion and further research

Our approach combines mathematical and computational methods. Mathematical analysis allows us to validate the single-site model in a qualitative way, based on the experimental data obtained in our group. The validation shows the consistency of our biological hypotheses. In a second step, we can extract the splice efficiency as a time-scale abstraction of the local behavior at one site inside a more global model involving different sites. For the experimental biologist, the single-site model may serve as a computational tool to evaluate his knowledge on a fine-grained biological process.

On the computational side, the constraint solving and default reasoning capabilities of `Hybrid cc` allow us to exploit as much as possible the incomplete knowledge of our system. Default behavior may compensate the lack of experimental data. Using constraint programming, we can delimit with our model the possible splicing behavior. This provides a powerful tool for qualitative validation.

Combining mathematical analysis and computational methods is the key to extending the single-site model to a multi-site model as described in this paper. It leads to the qualitative validation represented by the extraction of the splice efficiency function. The splice efficiency characterizes the modularity of the regulation. Thus, the one-site behavior is represented in the three-site model, based on the single-site splice efficiency. The extraction of a suitable criterion on the smaller scale is crucial to understanding an experimental process from a systems biology perspective. Furthermore, constraints can be used to handle the problem of missing data in time-scale abstraction of a single-site model in a more global multi-site model. Different scales usually correspond to biological experiments yielding different types of results. Despite the variety of possible experiments, these must be integrated into a global model in order to better understand the biological process.

Modeling alternative splicing requires a close interaction between biological and computational approaches. In the context of alternative splicing regulation, we are currently working on new experimental data for the quantitative validation of our models. On the computational side, we have integrated our model into a general HIV-1 model [12]. Preliminary results show that the modification of a splice constant may induce different behaviors in the HIV-1 life cycle model. Using the extended model, we may validate several biological hypotheses on the global effect of alternative splicing in the full HIV-1 life cycle.

Acknowledgement

The authors would like to thank Arnaud Courtois for his comments on a draft of this paper.

References

[1] O. Bernard and J.-L. Gouzé. Nonlinear qualitative signal processing for biological systems: application to the algal growth in bioreactors. *Mathematical Biosciences*, 157:357–372, 1999.

[2] A. Bockmayr and A. Courtois. Modeling biological systems in hybrid concurrent constraint programming (Abstract). In *2nd Int. Conf. Systems Biology, ICSB'01, Pasadena, CA*, page 106, 2001.

[3] A. Bockmayr and A. Courtois. Using hybrid concurrent constraint programming to model dynamic biological systems. In *18th International Conference on Logic Programming, ICLP'02, Copenhagen*, pages 85–99. Springer, LNCS 2401, 2002.

[4] M.A. Caputi, M.A. Mayeda, A.R. Krainer, and A.M Zahler. hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. *EMBO*, 18(14):4060–4067, 1999.

[5] B. Carlson and V. Gupta. Hybrid cc and interval constraints. In *Hybrid Systems: Computation and Control, HSCC'98*, pages 80 – 95. Springer, LNCS 1386, 1998.

[6] F. Dautry and D. Weill. Kinetic analysis of mRNA metabolism. In *Interdisciplinary School on Imaging, Modelling and Manipulating Transcriptional Regulatory Networks*, page 20, Ambleteuse, 2002.

[7] F. Del Gatto-Konczak, M. Olive, M.C. Gesnel, and R. Breathnach. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. *Mol. Cell. Biol.*, 19(1):251–260, 1999.

[8] D. Eveillard, D. Ropers, H. de Jong, C. Branlant, and A. Bockmayr. Multiscale modeling of alternative splicing regulation. In *Computational Methods in Systems Biology, CMSB'03, Rovereto, Italy*, pages 75–87. Springer, LNCS 2602, 2003.

[9] B.R Graveley. Sorting out the complexity of SR protein functions. *RNA*, 6:1197–1211, 2000.

[10] V. Gupta, R. Jagadeesan, and V. Saraswat. Computing with continuous change. *Science of computer programming*, 30(1-2):3–49, 1998.

[11] V. Gupta, R. Jagadeesan, V. Saraswat, and D. G. Bobrow. Programming in hybrid constraint languages. In *Hybrid Systems II*, pages 226–251. Springer, LNCS 999, 1995.

[12] B.J Hammond. Quantitative study of the control of HIV-1 gene expression. *J. Theor. Biol.*, 163:199–221, 1993.

[13] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

[14] R. Heinrich and S. Schuster. *The regulation of cellular systems.* Thomson Publishing, New York, 1996.

[15] M.J. Moore, C.C. Query, and P.A. Sharp. Splicing of precursors to mRNA by the spliceosome. In *The RNA World*, pages 303–357. Cold Spring Harbor Laboratory Press, 1993.

[16] M. O'Reilly, M.T. McNally, and K.L Beemon. Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 RNA. *Virology*, 213(2):373–385, 1995.

[17] B. Palsson. The challenges of in silico biology. *Nature Biotechnology*, 18:1147 – 1150, 2000.

[18] D.F. Purcell and M.A. Martin. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J. Virol.*, 67(11):6365–6378, 1993.

[19] D. Ropers, L. Ayadi, R. Gattoni, S. Jacquenet, L. Damier, C. Branlant, and J Stévenin. Differential effects of the SR proteins 9G8, SC35, ASF/SF2 and SRp40 on the utilization of the A1 to A5 splicing sites of HIV-1 RNA revealed by in vitro and ex vivo splicing experiments. *Submitted to J. Biol. Chemistry*, 2003.

[20] V. A. Saraswat. *Concurrent constraint programming.* MIT Press, 1993.

[21] V. A. Saraswat, R. Jagadeesan, and V. Gupta. Foundations of timed concurrent constraint programming. In *9th Symp. Logic in Computer Science, LICS'94, Paris*, pages 71 – 80. IEEE, 1994.

[22] V. A. Saraswat, R. Jagadeesan, and V. Gupta. Timed default concurrent constraint programming. *Journal of Symbolic Computation*, 22(5/6):475–520, 1996.

[23] L.A. Segel. *Modelling dynamic phenomena in molecular and cellular biology.* Cambridge University Press, 1984.

[24] H. Tang, K.L. Kuhen, and F. Wong-Staal. Lentivirus replication and regulation. *Annu. Rev. Genet*, 33:133–170, 1999.

[25] E. O. Voit. *Computational analysis of biochemical systems.* Cambridge Univ. Press, 2000.

Contents

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399