A Gentle Introduction to Stochastic (Poly)Automata Collectives and the (Bio)Chemical Ground Form *

Gianluigi Zavattaro

Dip. Scienze dell'Informazione, Università di Bologna, Italy.
Email: zavattar@cs.unibo.it

Abstract. We present uniformly four related models for the representation of biochemical systems recently proposed in the literature in different publications. Namely, we consider Stochastic Automata Collectives (SAC) [2], Stochastic Polyautomata Collectives (SPC) [2], Chemical Ground Form (CGF) [3], and Biochemical Ground Form (BGF) [4].

1 Introduction

The aim of this paper is to provide a unified introduction to four related models for the representation of biochemical systems recently proposed in the literature in three different papers. Namely, we present Stochastic Automata Collectives (SAC) [2], Stochastic Polyautomata Collectives (SPC) [2], the Chemical Ground Form (CGF) [3], and the Biochemical Ground Form (BGF) [4]. The first pair of models are based on a graphical automata-based notation, while the second pair of models have been defined with a formal syntax and semantics similar to those of traditional (stochastic) process algebras.

We unify the presentation of the four models presenting for all of them both a graphical and a process algebraic notation. For the sake of readability, we do not report the definition of the formal semantics of the calculi that can be found in [4]. Moreover, we gently introduce the four models using several examples that allow us to focus on the specific differences and similarities among the four different models.

The remainder of the paper is divided in four Sections, one for each of the considered models.

2 Stochastic Automata Collectives

In this section we introduce Stochastic Automata Collectives (SAC), the notation for the representation of chemical systems presented in [2]. In that paper, SAC are

* This paper is an introductory material to the lecture "Expressiveness Issues in Calculi for Artificial Biochemistry" given by the author at the summer school SFM-08:Bio. More precisely, the calculi considered in the lecture are gently and uniformly introduced.

only informally presented. In order to equip this model with a formal semantics, we simply observe that this model is a fragment of CGF, a process algebra whose formal syntax and semantics have been defined in [3]. We characterize the precise fragment defining a syntax for SAC as a subset of the syntax of CGF. The reader interested in the definition of the formal semantics of SAC can then refer to [3] where the semantics of the whole CGF is reported.

Before presenting SAC, we introduce the running example for this section.

Example 1 (Two-stations rotaxane). We consider two-stations rotaxanes [10] (simply called rotaxanes in the following), which are supramolecular systems composed of an axle surrounded by a ring-type molecule. Bulky chemical moieties ("stoppers") are placed at the extremities of the axle to prevent the disassembly of the system. In rotaxanes containing two different recognition sites on the axle ("stations"), it is possible to switch the position of the ring between the two stations by an external energy input (called the "stimulus") as illustrated in Figure 1. The part (a) of the figure represents the structure of the rotaxane,

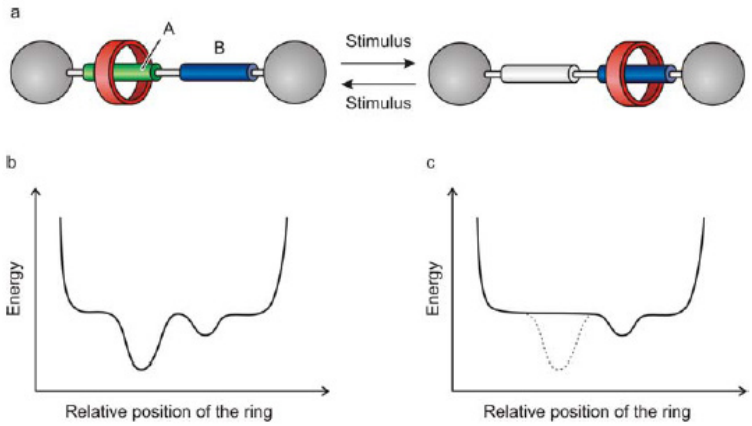

Fig. 1. Representation of a rotaxane with stations A and B (a) and its energy curves before (b) and after (c) the stimulus activating the ring movement from A to B.

while in part (b) and (c) the energy curves before and after the stimulus are depicted: in the former the energy minimum corresponds to station A, while in the latter it corresponds to station B. For this reason, the stimulus triggers the shuttling of the ring from station A to station B.

It is worth mentioning that several rotaxanes of this kind, known as *molecular shuttles*, have been already developed (see [5] and the references therein) and used for building more complex systems [8, 7, 1].

We complete the example presenting a modeling of the behavior of the rotaxane given in chemical reaction style. We call the two stations of the rotaxane

A and B, respectively, and we call S the species of the molecules that stimulates the movement of the ring from station A to station B. We consider four distinct species for the representation of the rotaxane: R^A (resp. R^B) representing the *nonstimulated* rotaxane with the ring in position A (resp. B), and R_s^A (resp. R_s^B) representing the *stimulated* rotaxane with the ring in position A (resp. B).

The chemical reactions are as follows (here we abstract away from the rates of the reactions that will be discussed in the next Example 2):

$$\begin{aligned} R^A + S &\to R_s^A \\ R^B + S &\to R_s^B \\ R_s^A &\leftrightarrow R_s^B \\ R^A &\leftrightarrow R^B \end{aligned}$$

We consider two bi-molecular reactions and two mono-molecular invertible reactions. The first two represents the reaction between the stimulus and the rotaxane. As the rotaxane has two nonstimulated species R^A and R^B, we need to consider two distinct reactions, one for each of these species. The two mono-molecular bidirectional reactions model the movement of the ring. We need to consider two distinct reactions because by Brownian motion we can assume that the ring can move from station A to station B, and vice versa, both when the rotaxane is stimulated and when it is not stimulated.

We now introduce `SAC`. It is an automata based notation in which each state of an automaton corresponds to a chemical species X, and each outgoing transition from state X represents one possible reaction in which the molecules of species X can be engaged. The transitions are labeled with one of three possible kinds of labels. The label $\tau_{(r)}$ indicates the possibility for one molecule to be engaged in a unary reaction with stochastic rate r. On the contrary, the transitions labeled with $?a_{(r)}$ and $!a_{(r)}$ models the complementary transitions executed by the two reacting molecules. The name a is a name used to identify the reaction, while r is a stochastic rate; both the name a and the rate r must match for the reaction to be enabled. For instance, if the states associated to the species X and Y have outgoing transitions labeled with $?a_{(r)}$ and $!a_{(r)}$, respectively, we have that one molecule of species X can react with one molecule of species Y, and the time needed for this reaction to occur is distributed according to exponential distribution with rate r. The target states of the transitions represent the species of the product of the reaction. For instance, if the two above transitions labeled with $?a_{(r)}$ and $!a_{(r)}$ have the species X' and Y' as target state, respectively, we have that the product of the reaction is given by two molecules, one of species X' and one of species Y'.

As a less trivial example of `SAC`, we model the rotaxane of the Example 1.

Example 2 (Modeling rotaxanes in `SAC` *–graphical notation–).* We present the modeling of rotaxanes in `SAC`. The main difference between this new modeling and the one proposed in the Example 1 is that it is molecular oriented instead of reaction oriented. In other words, the modeling approach of `SAC` is based on the description of the behavior of a molecule based on the sequence of reactions in

which a molecule can be engaged during its lifetime. Such behavior is depicted in Figure 2. It is worth noting that the SAC modeling is based on two distinct

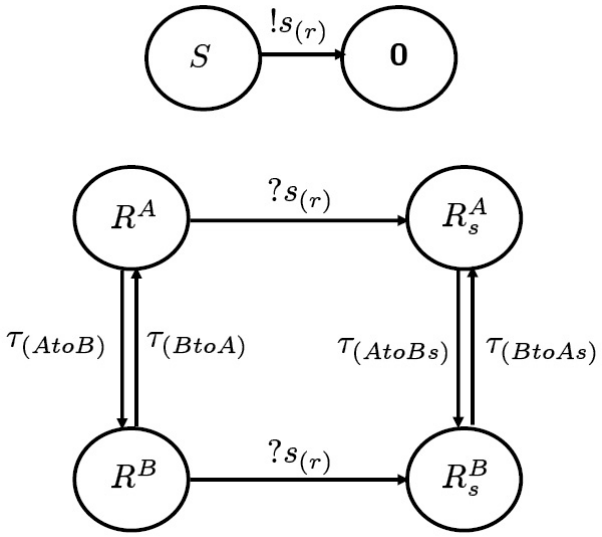

Fig. 2. Behavior of a rotaxane depicted as a stochastic automata collective.

automata; one for the description of the behavior of the stimulus S, and one for the behavior of the rotaxane. The stimulus can only be engaged in a reaction (that we call a) in order to stimulate the rotaxane. After this reaction, the molecule is "consumed" (i.e. it forms a complex with the rotaxane). Consumed molecules, are modeled with a special species that we denote with $\mathbf{0}$. On the contrary, the modeling of the rotaxane includes bi-directional transitions for the ring shuttling, and the complementary transitions for the reaction with the stimulus.

We conclude this example reporting a discussion about the rates that are included in the SAC modeling as symbolic (i.e. we use names instead of positive real numbers) subscripts of the transition labels. The rate r is the stochastic rate for the bi-molecular reaction between the rotaxane and the stimulus. As far as the ring shuttling is concerned, we recall that, by Brownian motion, we assume that the ring can move from station A to station B, and vice versa, both when the rotaxane is stimulated and when it is not stimulated. Different rates are considered for these movements: $AtoB$ (resp. $AtoBs$) for the movement from station A to station B when the rotaxane is nonstimulated (resp. stimulated), $BtoA$ (resp. $BtoAs$) for the opposite movement. According to the energy minimum in the two distributions in Figure 1 parts (b) and (c), we have that $AtoB < BtoA$ and $AtoBs > BtoAs$. Thus, according to the stochastic behavior

of these mono-molecular reactions, when the rotaxane is stimulated (resp. non-stimulated), the sojourn time of the ring on station A (resp. B) is longer than the sojourn time on station B (resp. A).

We complete this section describing a formal syntax for SAC. This is obtained as a fragment of a process algebra, called Chemical Ground Form (CGF) defined in [3]. According to this syntax, each species has an associated definition describing the possible actions for the molecules of that species. There are three kinds of actions that coincide with the possible labels for transitions in SAC. Namely, we have the action $\tau_{(r)}$ indicating the possibility for a molecule to be engaged in a unary reaction. For instance, the definition $A = \tau_{(r)}; B$ is used to specify the possibility for one molecule of species A to be engaged in a unary reaction that produces one molecule of species B. Binary reactions have two reactants. The two reactants perform two complementary actions $?a_{(r)}$ and $!a_{(r)}$, where a is a name used to identify the reaction; both the name a and the rate r must match for the reaction to be enabled. For instance, given the definitions $A = ?a_{(r)}; C$ and $B = !a_{(r)}; D$, we have that two molecules of species A and B can be engaged in a binary reaction that produces two molecules, one of species C and one of species D. If the molecules of one species can be engaged in several reactions, then the corresponding definition admits a choice among several actions. The syntax of choice is as follows: $A = \tau_{(r)}; B \oplus ?a_{(r')}; C$, meaning that molecules of species A can be engaged in either a unary reaction that produces a molecule of species B, or in a binary reaction with another molecule able to execute the complementary action $!a_{(r')}$. In the second case, the molecule of species A contributes to the reaction by producing a new molecule of species C.

We are now ready to define formally the syntax for Stochastic Automata Collectives.

Definition 1 (Stochastic Automata Collectives (SAC)). *Consider the following denumerable sets:* Species *ranged over by variables X, Y, $\cdots$,* Channels *ranged over by a, b, $\cdots$, Moreover, let r, s, $\cdots$ be rates (i.e. positive real numbers).*
The syntax of SAC *is as follows (where the big $\mid$ separates syntactic alternatives while the small $\mid$ denotes parallel composition):*

$$
\begin{array}{rcll}
E & ::= & \mathbf{0} \mid X = M, E & \textit{Reagents} \\
M & ::= & \mathbf{0} \mid \pi; X \oplus M & \textit{Molecule} \\
P & ::= & \mathbf{0} \mid X | P & \textit{Solution} \\
\pi & ::= & \tau_{(r)} \mid ?a_{(r)} \mid !a_{(r)} & \textit{Internal, Input, Output prefix} \\
\texttt{SAC} & ::= & (E, P) & \textit{Reagents and initial Solution}
\end{array}
$$

Given a SAC (E, P)*, we assume that all variables occurring in P occur also in E. Moreover, for every variable X occurring in E, there is exactly one definition $X = M$ in E.*

In the following, trailing $\mathbf{0}$ are usually left implicit, and we use $\mid$ also as an operator over the syntax: if P and P' are $\mathbf{0}$-terminated lists of variables,

according to the syntax above, then $P|P'$ means appending the two lists into a single $\mathbf{0}$-terminated list. Therefore, if P is a solution, then $\mathbf{0}|P$, $P|\mathbf{0}$, and P are syntactically equal. Moreover, the solution composed of k instances of X is denoted with $\prod_k X$.

As an example of exploitation of the SAC syntax, we report the syntax for the modeling of the rotaxane graphically depicted in the Figure 2 and discussed in the Example 2.

Example 3 (Modeling rotaxanes in SAC *–formal syntax–).* We can consider the following definitions for the species R^A, R^B, R^A_s, R^B_s, and S used in the previous examples.

$$\begin{array}{lcl} R^A & = & \tau_{(AtoB)}; R^B \oplus ?s_{(r)}; R^A_s \\ R^B & = & \tau_{(BtoA)}; R^A \oplus ?s_{(r)}; R^B_s \\ R^A_s & = & \tau_{(AtoBs)}; R^B_s \\ R^B_s & = & \tau_{(BtoAs)}; R^A_s \\ S & = & !s_{(r)}; \mathbf{0} \end{array}$$

where $\mathbf{0}$ specifies reactions which have no product. Let E be the sequence of definitions of the species R^A, R^B, R^A_s, R^B_s, and S are defined above. A solution with one instance of non-stimulated rotaxane with the ring on station A and 2 instances of stimulus, is represented by the SAC $(E, R^A|S|S)$.

As already discussed, the syntax of SAC is obtained as a fragment of the process algebra CGF defined in [3]. More precisely, the fragment is simply obtained imposing that after an action π only one molecule can be produced, i.e. using the syntax $\pi; X$ instead of the more general syntax $\pi; (X_1|\cdots|X_n)$ of CGF. In [3] also the formal semantics for CGF is defined; here we simply recall how the semantics is defined without reporting the full definition (the interested reader can refer to [3]).

The semantics is obtained associating to each term of the process algebra a Continuous Time Markov Chain (CTMC). Such CTMC is obtained in two steps. First, a labeled transition graph (LTG) is defined which represents all possible actions that can be executed by the molecules in the considered solution. Second, a CTMC is extracted from such labeled transition graph by collapsing those transitions which share the same source and target solutions in one CTMC transition, whose rate is the sum of the rates of the collapsed transitions.

More precisely, the labeled transition graph is a labaled transition system among solutions that consider two possible kinds of labels: $i : r$ and $i, j : r$ representing, respectively, mono-molecular reactions with rate r involving the i-th molecule and bi-molecular reactions with rate r involving the i-th and the j-th molecules. As an example of labeled transition graph, we consider the SAC $(E, R^A|S|S)$ defined in the example 3.

Example 4 (LTG of a rotaxane). As an example of labeled transition graph, we show in Figure 3 the LTG of the SAC $(E, R^A|S|S)$ defined the Example 3. It is worth noting that due to the presence of two stimulating molecules there exist two pairs of transitions sharing the same source and target solutions.

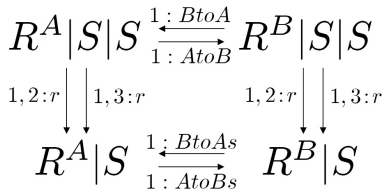

Fig. 3. Labeled Transition Graph of the SAC $(E, R^A|S|S)$.

As reported above, the extraction of the CTMC from one labeled transition graph simply requires the collapsing of those transitions which share the same source and target solutions in one CTMC transition, whose rate is the sum of the rates of the collapsed transitions. As an example, we discuss the CTMC of the solution considered in the Example 4.

Example 5 (CTMC of a rotaxane). As an example of Continuous Time Markov Chain extracted from a Labeled Transition Graph, we show in Figure 4 the CTMC obtained from the LTG in Figure 3. It is worth noting that the CTMC

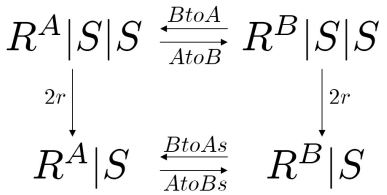

Fig. 4. Continuous Time Markov Chain of the SAC $(E, R^A|S|S)$.

has the same states of the corresponding LTG. There are two differences: the transitions are labeled only with the stochastic rates, and the transitions sharing the same source and target solutions collapse in a unique transition, with rate equal to the sum of the rates of the collapsed transitions.

The CTMC semantics allows us to interpret the behavior of a SAC (E, P) as follows. Given any state T of the CTMC of (E, P), if it has n outgoing transitions labeled with $r_1, \cdots, r_n$, then the probability that the sojourn time in T is less than t is exponentially distributed with rate $\sum_i r_i$, i.e. $Prob\{delay < t\} = 1 - e^{-t \sum_i r_i}$, and the probability that the j-th transition is taken is $r_j/(\sum_i r_i)$.

3 Chemical Ground Form

One of the main feature of SAC is that the number of molecules in a modeled solution is an invariant, in fact when a molecule engage a reaction it produces exactly one new molecule. This is guaranteed by the syntax of molecule definitions $\pi; X$, according to which an actions π is always followed by one and only one species X. In [3], an extension of the model is considered in which the product can be a multiset of species, namely, the new syntax of action execution is $\pi; (X_1 | \cdots | X_n)$. The new model is called Chemical Ground Form (CGF). The motivation for the definition of CGF is to obtain a process algebraic modeling of basic chemistry. As in basic chemistry there is no limitation to the number of molecules in the product of one reaction, it is necessary to admit more than one molecule as the product of one action.

The syntax and semantics of CGF can be found in [3]. We recall the syntax.

Definition 2 (Chemical Ground Form (CGF)). *Consider the following denumerable sets:* Species *ranged over by variables X, Y, $\cdots$,* Channels *ranged over by a, b, $\cdots$, Moreover, let r, s, $\cdots$ be rates (i.e. positive real numbers). The syntax of* CGF *is as follows:*

$$
\begin{array}{rcll}
E & ::= & \mathbf{0} \mid X = M, E & \textit{Reagents} \\
M & ::= & \mathbf{0} \mid \pi; P \oplus M & \textit{Molecule} \\
P & ::= & \mathbf{0} \mid X | P & \textit{Solution} \\
\pi & ::= & \tau_{(r)} \mid ?a_{(r)} \mid !a_{(r)} & \textit{Internal, Input, Output prefix} \\
\texttt{CGF} & ::= & (E, P) & \textit{Reagents and initial Solution}
\end{array}
$$

Given a CGF *(E, P), we assume that all variables occurring in P occur also in E. Moreover, for every variable X occurring in E, there is exactly one definition $X = M$ in E.*

It is worth observing that the difference between the syntax of SAC and the syntax of CGF is that after the execution of one action π a solution, i.e. a multiset of molecules, can be specified as the product of the action. We call *molecule splitting* this possibility for one reactant to produce more than one molecule.

In [3] CGF is proved to be equivalent to basic chemistry both for discrete state and continuous state semantics. Discrete state semantics describe a solution as a multiset of molecules (i.e. for each molecule the exact number of instances is known) while continuous state semantics model a solution indicating the concentration of each species of interest (i.e. each species has an associated real number quantifying the concentration). By basic chemistry we mean systems modeled by a finite set of mono-molecular and bi-molecular reactions. To prove this equivalence result, CGF is equipped with both a discrete state semantics defined in terms of CTMC and a continuous state semantics defined in terms of ordinary differential equations. In this paper, we consider only the discrete state semantics.

We now present the running example for this section.

Example 6 (Counting the number of reactions). This example is not inspired by a specific chemical system, but it is proposed on purpose to focus on the increment of expressive power of CGF with respect to SAC. The idea is to consider two kinds of bi-molecular reactions, the first one called a and the second one called b. We present a system in which an arbitrary number of reactions of kind a are executed, and then a corresponding number of reactions of kind b occurs. In order to define such a system, we need the ability to "count" the number of occurrences of the reaction of kind a.

We can define such system in CGF considering two pairs of species: A and A' as the reactants of the reaction a and B and B' as the reactants of the reaction b:

$$\begin{aligned} A &= !a_{(h)}; (A|B) \oplus \tau_{(l)}; B' \\ A' &= ?a_{(h)}; A' \\ B &= !b_{(h)}; \mathbf{0} \\ B' &= ?b_{(h)}; B' \end{aligned}$$

We assume that the rate h is greater than the rate l. We consider, as initial solution, one instance of species A and one of species A': formally, we consider the CGF $(E, A|A')$ where E includes the definitions of the species A, A', B, and B' as reported above.

As done in the previous section for SAC, we do not report the formal definition of the semantics that can be found in [3]. We simply recall that it is defined in terms of CTMCs obtained in two steps: first a labeled transition graph is associated to a CGF, then a CTMC is extracted from this labeled transition graph. As an example, we discuss the CTMC of the CGF $(E, A|A')$ defined in the Example 6.

Example 7. We present in Figure 5 the CTMC that, following the technique already described in the previous section (and formlized in [3]), is associated to the CGF $(E, A|A')$ of the Example 6. As we assume that the rate h is greater than

$$\begin{array}{ccccccccccc} A|A' & \xrightarrow{h} & A|A'|B & \xrightarrow{h} & A|A'|B|B & \xrightarrow{h} & A|A'|B|B|B & \xrightarrow{h} & \dots \\ \downarrow l & & \downarrow l & & \downarrow l & & \downarrow l & & \\ B'|A' & \xleftarrow{h} & B'|A'|B & \xleftarrow{h} & B'|A'|B|B & \xleftarrow{h} & B'|A'|B|B|B & \xleftarrow{h} & \dots \end{array}$$

Fig. 5. Continuous Time Markov Chain of the CGF $(E, A|A')$.

l, the initial solution more probably will select the reaction of kind a depicted horizontally in the Figure. Due to the fairness implicit in stochastic systems, the transition with the lower rate l cannot be delayed indefinitely, thus eventually one of the states of the second row will be reached with probability one. At this

point of the computation, a number of transitions of kind b will be executed that coincides with the number of transitions of kind a already executed.

We complete the section showing how the graphical notation of SAC can be extended to cope also with molecule splitting of CGF. The idea is to separate, in case of splitting, the transitions in two parts adding an intermediary state. This new intermediary state is graphically represented with a line. We use one transition from the state representing the species of the reactant leading to the new intermediary transition. This transition is labeled with the executed action. Then, we use as many (unlabeled) transitions as the number of produced molecules. Each transition is from the new intermediary state to the state representing the species of one of the product. As an example, we show the graphical representation of the system described in the Example 6.

Example 8. The definitions of the species A, A', B, and B' reported in the Example 6 can be graphically rendered as in the Figure 6. The only novelty with

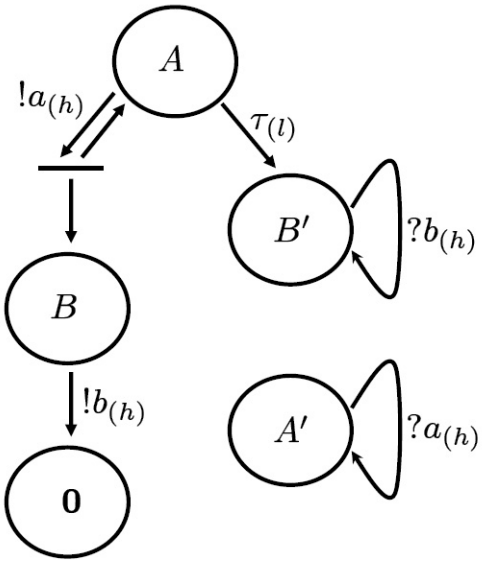

Fig. 6. Graphical representation of the CGF described in the Example 6.

respect to the graphical notation of SAC is in one splitting that occur when a molecule of species A is engaged in one reaction with one molecule of species A': in this case, the molecule splits and produces one molecule of species A and one of species B.

4 Stochastic Polyautomata Collectives

Stochastic Polyautomata Collectives (SPC) have been proosed in [2] as an extension of SAC able to capture the essential primitives of biochemistry. Biochemistry is obviously based on chemistry, and in principle one can always express the behavior of a biochemical systems by a collections chemical reactions. But there is a major practical problem with that approach: the collection of reactions for virtually all biochemical systems is an infinite one. For example, just to express the chemical reactions involved in linear polymerization, we need to have a different chemical species for each length n of polymer P_n, with reactions to grow the polymer: $P_n + M \rightarrow P_{n+1}$. While each polymer is finite, the set of possible polymerization reactions is infinite.

Nature adopts a more modular solution: the act of joining two molecules is called *complexation*, and polymers are made by iteratively complexing monomers. Each monomer obeys a *finite* simple set of rules that leads to the formation of polymers of any length; therefore, it seems that there should be a finite way of describing such systems. One can start by writing pseudo-reactions like $P + M \rightarrow P : M$, where $P : M$ is meant to represent a P(olymer) molecule attached to an extra M(onomer), yielding a longer polymer. However, there are in general many possible ways (that is, many different patches on the surface of a molecule) by which one molecule can exclusively form a complex with other molecules, and soon one needs to describe the *interface* of each molecule. This situation, while not commonly found in basic chemistry, is particularly acute in biochemistry, where virtually all reactions are governed by enzymes and molecular machines, which are themselves often built by complexation, and which usually operate by complexing with their reactants.

Both SAC and CGF have been extended to model also a minimalistic form of complexation. In this section we present Stochastic Polyautomata Collectives (SPC), the extension of the former with primitives for association (i.e. the creation of a complex) and dissociation (i.e. the separation of parts of the complex). This model has been presented in [2]. More precisely, two additional pairs of complementary prefixes, $\&?a_{(r)}, \&!a_{(r)}$ for association and $\%?a_{(r)}, \%!a_{(r)}$ for dissociation are added. Before presenting the formal syntax of SPC, we introduce the new primitives informally by means of examples. To simplify the notation, in the examples we abstract away from the stochastic rates, e.g., we write $\&?a$ instead of $\&?a_{(r)}$.

Example 9 (Linearly growing polymer). Each complexation event involves exactly two partners. We imagine that the partners have two complementary surface patches that can interlock. If c represents a surface shape (say, a paraboloid), then $!c$ indicates one of the two patches (say, the convex one) and $?c$ indicates the complementary patch (the concave one). Then, $\&!c$ is the action that presents the convex patch, and $\&?c$ is the action that presents the concave patch. When two such *association* actions meet, an actual complexation event can take place, joining the two complementary surfaces.

A linearly growing polymer could be represented as follows, using a seed S and a collection of equal monomers M. The seed starts the chain by presenting a concave patch $?c$: this is our initial, zero-length, polymer. Each monomer presents a convex patch $!c$, which can bind with an existing polymer on the complementary concave patch. After (and only after) such a binding, a bound monomer M' presents another concave patch $?c$, so that the polymer can keep growing. Both the seed and each monomer can have further behavior, S' and M''.

$$\begin{array}{lcl} S & = & \&?c; S' \\ M & = & \&!c; M' \\ M' & = & \&?c; M'' \end{array}$$

Each complexation event creates a unique bond between exactly the two molecules that are joined to each other. This bond needs to be represented somehow, to make sure that a molecule can bind with only one other molecule at a time on any given patch. We represent such a bond as a unique key k that is shared by the two complexed molecules (think of k as a fresh number, or as a fresh channel in π-calculus [9]). Such unique keys, and related information, are collected in the *association history* of each molecule. So, the first interaction of an S with an M, which initially have empty association histories ($\mathbf{0}$), proceeds as follows:

$$S_{\mathbf{0}} \mid M_{\mathbf{0}} \rightarrow S'_{\langle ?c,k1\rangle} \mid M'_{\langle !c,k1\rangle}$$

Interaction with a second monomer then introduces a second fresh key in the histories:

$$S_{\mathbf{0}} \mid M_{\mathbf{0}} \mid M_{\mathbf{0}} \rightarrow S'_{\langle ?c,k1\rangle} \mid M'_{\langle !c,k1\rangle} \mid M_{\mathbf{0}} \rightarrow S'_{\langle ?c,k1\rangle} \mid M''_{\langle ?c,k2\rangle::\langle !c,k1\rangle} \mid M'_{\langle !c,k2\rangle}$$

and so on. In any configuration, we can reconstruct from the association histories who is bound to whom, and on what surface the bond was formed. Note that the description of the system is finite (3 reagents, S, M, M'), but that polymers of any length can be assembled.

Example 10 (Branching polymer). After complexation, a molecule is still free to perform additional complexations or other interactions. That is, complexation places no restrictions on the behavior of the original molecules, except for the fact that new complexations cannot occur on surfaces that are already occupied, and that decomplexations must happen consistently with prior complexations (as we discuss shortly). To illustrate this freedom, let us modify the previous example and allow each bound monomer to offer a seed for growing a new polymer branch:

$$\begin{array}{lcl} S & = & \&?c; S' \\ M & = & \&!c; M' \\ M' & = & \&?c; S \end{array}$$

When an M' turns into a seed S, that is a seed with an non-empty association history that connects it to its current branch, but that can also start a new branch. If we do not wish to start a branch at every monomer, we can modify

M' to something like $M' = \&?c; S \oplus \tau; M''$, so that an M' has a temporary potential to act as a seed, but after some delay (τ) it may change to an M'' that is not a seed. By adjusting the stochastic rates of the delay and of c, we can produce different (stochastic) branching factors.

Example 11 (Actin-like polymer). Decomplexation is the inverse of complexation, that is, two formerly joined molecules can dissociate. We indicate by $\%!c$ the attempt to dissociate from the convex side, and $\%?c$ the attempt to dissociate from the concave side. When two complexed molecules attempt complementary dissociations, an actual decomplexation event can take place. To illustrate this situation, we describe a different kind of linear polymer: one that can grow only at one end, and can shrink only at the other end. There are four molecular states for each monomer: M^f (free monomer), M^l (monomer bound on the left), M^r (monomer bound on the right), and M^b (monomer bound on both sides). Each monomer has a left convex surface and a complementary right concave surface. A polymer should associate (grow) only on the right and should dissociate (shrink) only on the left.

$$\begin{aligned} M^f &= \&!c; M^l \oplus \&?c; M^r \\ M^l &= \%!c; M^f \oplus \&?c; M^b \\ M^r &= \%?c; M^f \\ M^b &= \%!c; M^r \end{aligned}$$

A free monomer M^f can either associate on the left convex surface and become bound on the left, or associate on the right concave surface and become bound on the right. A monomer M^l bound only on the left can either dissociate on the left (if allowed by its partner, which must in fact be an M^r in this case) and return free, or associate on the right (with an M^f) and become bound on both sides. A monomer M^r bound only on the right can only dissociate on the right: that is, a polymer cannot grow on the left. A monomer M^b bound on both sides can only dissociate on the left (with an M^r): that is, a polymer cannot shrink on the right or break in the middle. These rules cover also the base cases when a polymer of length 2 initially forms or finally dissolves.

A decomplexation should succeed only between a pair of molecules that were actually complexed in their past history, and this can be checked by inspecting the unique keys introduced during complexation. For example let us consider two M^f molecules that complex and then immediately decomplex:

$$M^f_{\mathbf{0}} \mid M^f_{\mathbf{0}} \to M^l_{\langle !c,k\rangle} | M^r_{\langle ?c,k\rangle} \to M^f_{\mathbf{0}} | M^f_{\mathbf{0}}$$

The second transition is allowed to happen because M^l offers $\%!c$, M^r offers the complementary $\%?c$, and the same key k appears in both association histories on the c interface (and with the correct convexity). As a consequence of decomplexation, the keys are removed from the histories.

After this gentle introduction to `SPC` by means of examples, we present the formal definition of its syntax. The main novelty deals with the association histories which are added to each molecule to keep track of the association keys

representing the bonds currently active between the molecule itself, and the other molecules to which it is complexed.

Definition 3 (Stochastic Polyautomata Collectives (SPC)). *Consider the following denumerable sets:* Species *ranged over by variables* X, Y, X^1, X^2, $\cdots$, Channels *ranged over by* a, b, $\cdots$, *a set of* Association keys *ranged over by* k, k', $\cdots$. *Moreover, let* r, s, $\cdots$ *be rates (i.e. positive real numbers).*
The syntax of SPC *is as follows:*

$$
\begin{array}{rll}
E ::= \mathbf{0} \mid X = M, E & & \textit{Reagents} \\
M ::= \mathbf{0} \mid \pi; X \oplus M & & \textit{Molecule} \\
\pi ::= \tau_{(r)} \mid ?a_{(r)} \mid !a_{(r)} & & \textit{Internal, Input, Output prefix} \\
\quad \mid \&?a_{(r)} \mid \&!a_{(r)} & & \textit{Association prefixes} \\
\quad \mid \%?a_{(r)} \mid \%!a_{(r)} & & \textit{Dissociation prefixes} \\
P ::= \mathbf{0} \mid X_H | P & & \textit{Solution} \\
H ::= \mathbf{0} \mid \langle ?a, k\rangle :: H \mid \langle !a, k\rangle :: H & & \textit{Association history} \\
\texttt{BGF} ::= (P, S) & & \textit{Reagents and initial Solution}
\end{array}
$$

Given a BGF (E, P), *we assume that all variables occurring in* P *occur also in* E. *Moreover, for every variable* X *occurring in* E, *there is exactly one definition* $X = M$ *in* E. *Moreover, each association key* k *in* P, *occur in exactly two complementary associations* $\langle ?a, k\rangle$ *and* $\langle !a, k\rangle$, *that appear in the association histories* H *and* H' *of two distinct molecules* X_H *and* $X'_{H'}$.

The syntax of SPC has been obtained as a fragment of the Biological Ground Form (BGF), a process algebra defined in [4]. More precisely, SPC is as the fragment of BGF without molecule splitting. In [4], the formal semantics of BGF is defined; clearly, this applies also to its fragment SPC.

We complete the section presenting an example of graphical notation for SPC, depicting the representation of the actin-like polymer described in the Example 11.

Example 12 (Graphical representation of the actin-like polymer). The graphical representation of SPC simply includes four new labels for the new actions $\&?a_{(r)}$, $\&!a_{(r)}$, $\%?a_{(r)}$ and $\%!a_{(r)}$. As an example, we depict in the Figure 7 the representation of the behavior of an actin-like polymer as described in the Example 11 (as done in that example, we abstract away from the rates).

5 Biochemical Ground Form

We now move to the last model considered in this paper, the Biochemical Ground Form (BGF). This model includes all mechanisms discussed in this paper, both molecule splitting and complexation. The main technical problem deals with the specification of the distribution of the associations in the association history of one reactant over the different products of a splitting. In fact, in case a molecule

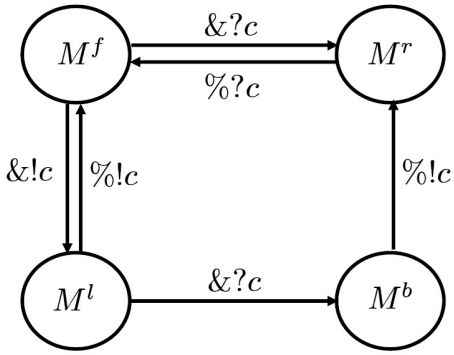

Fig. 7. Graphical representation of the actin-like polymer described in the Example 11.

forks it is necessary to specify how its associations are distributed over the produced molecules. This information is described by means of a new syntactic category called *association markers*. These are additional information associated to the produced molecule, that completely and uniquely define the distribution of associations, that is, all possible associations of one reactant should be reported in one and only one association marker of the product.

The formal syntax of BGF is defined as follows.

Definition 4 (Biochemical Ground Form (BGF)). *Consider the following denumerable sets:* Species *ranged over by variables* X, Y, X^1, X^2, $\cdots$, Channels *ranged over by* a, b, $\cdots$, *a totally ordered set of* Association keys *ranged over by* k, k', $\cdots$. *Moreover, let* r, s, $\cdots$ *be rates (i.e. positive real numbers).*
The syntax of BGF *is as follows:*

$$
\begin{array}{rll}
E & ::= \mathbf{0} \mid X = M, E & \textit{Reagents} \\
M & ::= \mathbf{0} \mid \pi; P \oplus M & \textit{Molecule} \\
\pi & ::= \tau_{(r)} \mid ?a_{(r)} \mid !a_{(r)} & \textit{Internal, Input, Output prefix} \\
 & \quad \mid \&?a_{(r)} \mid \&!a_{(r)} & \textit{Association prefixes} \\
 & \quad \mid \%?a_{(r)} \mid \%!a_{(r)} & \textit{Dissociation prefixes} \\
P & ::= \mathbf{0} \mid X_h | P & \textit{Product} \\
h & ::= \mathbf{0} \mid ?a :: h \mid !a :: h & \textit{Association markers} \\
S & ::= \mathbf{0} \mid X_H | S & \textit{Solution} \\
H & ::= \mathbf{0} \mid \langle ?a, k \rangle :: H \mid \langle !a, k \rangle :: H & \textit{Association history} \\
\texttt{BGF} & ::= (E, S) & \textit{Reagents and initial Solution}
\end{array}
$$

Given a BGF (E, S), *we assume that all variables occurring in* S *occur also in* E. *Moreover, for every variable* X *occurring in* E, *there is exactly one definition* $X = M$ *in* E. *Moreover, each association key* k *in* P, *occur in exactly two*

complementary associations $\langle ?a, k\rangle$ *and* $\langle !a, k\rangle$*, that appear in the association histories* H *and* H' *of two distinct molecules* X_H *and* $X'_{H'}$*.*

As discussed above, a well formed BGF should be defined in such a way that every time a molecule splits, it is always possible to define the way in which the associations in the history of the reactants are distributed over the products. The reader interested in the formalization of this notion of well formed CGF can refer to [4], where also the formal definition of the semantics can be found.

We complete this section with an extension of the example of the actin-like polymer discussed in the Example 11. The idea is to allow a fully bound monomer to split into two independent monomers, each one inheriting one of the two bonds. In this way, the polymer breaks in two new independent polymers.

Example 13 (Breaking polymer). To illustrate complexation in combination with molecule splitting, we describe a linearly growing polymer similar to the actin-like polymer of the Example 11 in which each monomer, once bound on both sides, is free to split into two new monomers each one inheriting one of the two bonds. The definition is as follows:

$$\begin{aligned} M^f &= \&!c; M^l \oplus \&?c; M^r \\ M^l &= \%!c; M^f \oplus \&?c; M^b \\ M^r &= \%?c; M^f \\ M^b &= \%!c; M^r \oplus \tau; (M^l_{!c} | M^r_{?c}) \end{aligned}$$

It is worth observing that in case of splitting of the molecules of species M^b, it is necessary to indicate also how to split the associations among the two produced molecules of species M^l and M^r, respectively. This is obtained adding the association marker corresponding to the bonds to be split.

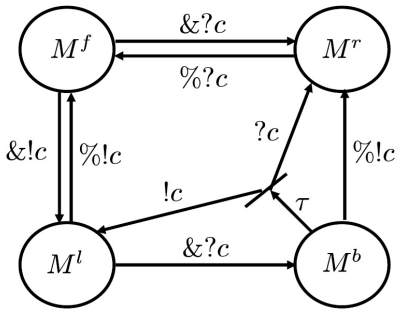

Fig. 8. Graphical representation of the breaking polymer described in the Example 13.

Also the graphical representation for CGF that we consider need to add graphical notation for dealing with association splitting. This is achieved adding the

association markers as labels of the transitions incoming into the species of the products of a splitting reaction. As an example, we depict the graphical representation of the breaking polymer of the Example 13.

Example 14 (Graphical representation of the breaking polymer). The graphical representation of `BGF` simply combine those of `CGF` and `SPC` with the addition of association markers as labels for the transitions representing the target states in case of splitting. As an example, we depict in the Figure 8 the representation of the behavior of an breaking polymer as described in the Example 13 (as done in that example, we abstract away from the rates).

Acknowledgements We thank Luca Cardelli for the discussions about the Chemical Ground Form and the association and dissociation mechanisms formalized in its extension Biochemical Ground Form. We thank also Alberto Credi, Marco Garavelli, Cosimo Laneve, Sylvain Pradalier, and Serena Silvi co-authors of the paper [6] from which the example of the rotaxane was taken.

References

1. J. D. Badjic, V. Balzani, A. Credi, S. Silvi, and J. F. Stoddart. A molecular elevator. *Science*, 303:1845–1849, 2004.
2. L. Cardelli. Artificial Biochemistry, 2007. Available at: http://lucacardelli.name.
3. L. Cardelli. On Process Rate Semantics. *Theoretical Computer Science*, in press, 2008. Available at http://dx.doi.org/10.1016/j.tcs.2007.11.012.
4. L. Cardelli and G. Zavattaro. On the Computational Power of Biochemistry, 2008. Available at: http://lucacardelli.name.
5. Benoit Champin, Pierre Mobian, and Jean-Pierre Sauvage. Transition metal complexes as molecular machine prototypes. *Chemical Society Reviews*, 36(2):358–366, 2006.
6. A. Credi, M. Garavelli, C. Laneve, S. Pradalier, S. Silvi, and G. Zavattaro. Modelization and Simulation of Nano Devices in nano-kappa Calculus. In *Proc. of Computational Methods in Systems Biology (CMSB07)*, volume 4695 of *LNCS*, pages 168–183, 2007.
7. T.-J. Huang, B. Brough, C.-M. Ho, Y. Liu, A.H. Flood, P.A. Bonvallet, H.-R. Tseng, J.F. Stoddart, M. Baller, and S. Magonov. A nanomechanical device based on linear molecular motors. *Applied Physics Letters*, 85(22):5391–5393, 2004.
8. M.C. Jimenez, C. Dietrich-Buchecker, and J.-P. Sauvage. Towards synthetic molecular muscles: Contraction and stretching of a linear rotaxane dimer. *Angew. Chem. Int. Ed.*, 39(18):3284–3287, 2000.
9. R. Milner. *Communication and Concurrency.* Prentice-Hall, 1989.
10. J.-P. Sauvage and C. O. Dietrich-Bucheker (eds.). *Molecular Catenanes, Rotaxanes and Knots.* Wiley-VCH, Weinheim, 1999.