Erschienen in: Journal of Mathematical Biology ; 69 (2014), 3. - S. 767-797
https://dx.doi.org/10.1007/s00285-013-0738-7

Markov chain aggregation and its applications to combinatorial reaction networks

Arnab Ganguly · Tatjana Petrov · Heinz Koeppl

Abstract We consider a continuous-time Markov chain (CTMC) whose state space is partitioned into aggregates, and each aggregate is assigned a probability measure. A sufficient condition for defining a CTMC over the aggregates is presented as a variant of weak lumpability, which also characterizes that the measure over the original process can be recovered from that of the aggregated one. We show how the applicability of de-aggregation depends on the initial distribution. The application section is devoted to illustrate how the developed theory aids in reducing CTMC models of biochemical systems particularly in connection to protein-protein interactions. We assume that the model is written by a biologist in form of site-graph-rewrite rules. Site-graph-rewrite rules compactly express that, often, only a local context of a protein (instead of a full molecular species) needs to be in a certain configuration in order to trigger a reaction event. This observation leads to suitable aggregate Markov chains with smaller state spaces, thereby providing sufficient reduction in computational complexity. This is

A. Ganguly and T. Petrov contributed equally to this work.

A. Ganguly (✉)
Department of Mathematics, University of Louisville, 231 Natural Sciences Building, Louisville, KY, USA
e-mail: a0gang02@louisville.edu

T. Petrov
IST Austria, Am Campus I, 3400 Klosterneuburg, Austria
e-mail: tatjana.petrov@ist.ac.at

H. Koeppl
ETH Zurich, Automatic Control Lab, Physikstrasse 3, 8092 Zurich, Switzerland
e-mail: koepplh@ethz.ch

H. Koeppl
IBM Zurich Reserach Labs, Saeumerstrasse 4, 8092 Rueschlikan, Switzerland

further exemplified in two case studies: simple unbounded polymerization and early EGFR/insulin crosstalk.

Keywords Markov chain aggregation · Rule-based modeling of reaction networks · Site-graphs

Mathematics Subject Classification (2010) 60J10 · 60J20 · 60J27 · 60J28 · 92B · 92D20 · 68Q42 · 68Q45

1 Introduction

Continuous-time Markov chains (CTMCs) are major tools used in modeling the stochastic nature of signal transduction in cells (Wilkinson 2006). More precisely, a well-mixed reaction system containing several molecular species is appropriately modeled by a CTMC such that each state represents a vector of species' abundances, and the transitions between states are determined by a set of reactions. A species, for instance, can be a protein or its phosphorylated form or a protein complex that consists of several proteins bound to each other. Especially in protein interactions that perform signal transduction in cells, the number of different such species can be combinatorially large, due to the rich internal structure of proteins and their mutual binding (Hlavacek et al. 2005; Walsh 2006). Consequently, even the simplest networks with only a few interacting proteins can result in very long reaction lists and the analysis of the underlying Markov chains becomes computationally inefficient or prohibitive. In these cases, model reduction is a major challenge.

In this work, we assume that the model of protein interactions is written by a biologist in form of site-graph-rewrite rules. Site-graph-rewrite rules compactly document hypotheses on signal transduction mechanisms, and *rule-based languages*, such as Kappa (Danos and Laneve 2003; Krivine et al. 2009) or BNGL (Blinov et al. 2004), are designed for the purpose of automated analysis of such models. A *site-graph* is a generalization of a graph where each node contains different types of sites, and edges emerge from these sites. Molecular species are suitably represented by site-graphs, where nodes are proteins and their sites are the protein binding-domains or modifiable residues; the edges indicate bonds between proteins. Every species is a connected site-graph, and in accordance with the traditional model, a state of a network is a multi-set of connected site-graphs.

We next model the dynamics of the protein interactions by constructing a Markov chain $\{X_t\}$ on an appropriate space of site-graphs, which essentially tells us how the 'reaction soup' looks like at different points of time. The usual species-based Markov chain turns out to be a particular aggregation of $\{X_t\}$. But more importantly, there exist other aggregations which lead to Markov chains living on much smaller state spaces, and they can be detected directly from the rule-set description, by a single scan of the set of rules. The idea for directly obtaining the smaller CTMC is the following. The site-graph description of the species' structure allows to describe interactions locally, by enumerating only parts of molecular species. For instance, it can be stated by a single site-graph rewrite rule, that any species containing a protein of type A can

have that protein A phosphorylated. In this case, the event of phosphorylation of A is independent of the rest of the species' context, i.e. A can equally be part of a dimer (complex of two proteins) or of a very large protein complex. If, moreover, the whole rule set never conditions other property except that A is not already phosphorylated at a particular domain, it can be proven that the whole stochastic process can be represented faithfully by tracing the *total* abundance of all species that conform to the motif of unphosphorylated A. In other words, all states of the original CTMC that count, for example, two unphosphorylated A, will be aggregated into a single state in the reduced CTMC (no matter whether these two unphosphorylated A's constitute two different monomers, or, for example, a single dimer). In such a case, the motif of unphosphorylated A is termed *fragment*, and the process in a lower dimension can be read-out directly from the rule set by simply annotating groups of protein domains that are updated independently. In Feret et al. (2012, 2013), the authors elaborate the described idea towards a procedure that detects a set of fragments for *any* given rule-set written in a Kappa language; The modeler can work on a reduced model directly, without ever considering the original CTMC, because the relation between the smaller and the original model is guaranteed. In Feret et al. (2012, 2013), these guarantees are provided through a weighted labeled transition system (WLTS). However, the relation between the concrete WLTS (the one obtained by counting species' abundances), and the abstract WLTS (the one obtained by counting fragments' abundances) is guaranteed for a particular class of initial distributions; Moreover, the WLTS is designed to capture the operational semantics of Kappa, thus making the actual relation between the underlying CTMC's implicit and not amenable to the usual Markov chain analysis techniques.

In the present article, we study the relation between the CTMC's arising from the original (species-based) and the reduced (fragment-based) model. To this end, we first develop a theory on aggregation of Markov chains whose application covers but is not limited to the class of rule-based models. Then, instead of operational semantics of Kappa, we present a general and comprehensive framework based on site-graph-rewrite models; The mathematical treatment of the rule-based models is carried out efficiently by the tools of graph theory. Finally, we extend the reduction framework of Feret et al. (2012, 2013), with a criterion on the rule-set for claiming the asymptotic possibility of reconstruction of the species-based dynamics.

Aggregation or lumping of a Markov chain is a technique instrumental in reducing the size of the state space of the chain and in modeling of a partially observable system. Typically, the original state space, S, of the Markov chain $\{X_n\}$ is partitioned into a set of equivalence classes, $\tilde{S} = \{A_1, \ldots, A_i\}$ and a process, $\{Y_n\}$, is defined over $\tilde{S}$. More precisely, let π be an initial distribution on S for the chain $\{X_n\}$. For a given partition $\tilde{S}$ of S, let the aggregated chain $\{Y_n\}$ be defined by

$$\{Y_n = A_i\} \quad \text{if and only if } \{X_n \in A_i\}.$$

Observe that $\{Y_n\}$ is not necessarily Markov, nor homogeneous. Conditions are imposed on the transition matrix of the Markov chain $\{X_n\}$ to ensure that the new process $\{Y_n\}$ is also Markov (see Buchholz 1994; Rubino and Sericola 1991, 1993; Sokolova and de Vink 2003; Tian and Kannan 2006 and references therein). In this

context, *strong lumpability* refers to the property of $\{X_n\}$, when the aggregated process $\{Y_n\}$ (associated with a given partition) is Markov with respect to any initial distribution π. If P denotes the transition matrix of $\{X_n\}$, then it has been shown that a necessary and sufficient condition for $\{X_n\}$ to be *strongly lumpable* with respect to the partition $\tilde{S}$ is that for every A_k, A_l, $\sum_{s\in A_l} P(s', s) = \sum_{s\in A_l} P(s'', s)$ for any $s', s'' \in A_k$. Tian and Kannan (2006) extended the notion of strong lumpability to continuous time Markov chains. A more general situation is when $\{X_n\}$ is *weakly lumpable* (with respect to a given partition), that is, when $\{Y_n\}$ is Markov for a subset of initial distributions π. The notion first appeared in Kemeny and Snell (1960) and subsequent papers (Ledoux 1995; Rubino and Sericola 1991, 1993) focussed toward developing an algorithm for characterizing the desired set of initial distributions. More general work focussed on studying Markovian nature of a functional of a Markov chain (see Gurvits and Ledoux 2005; Ledoux 2004). The characterization is done through some kind of recursive equations which sometimes might be hard to read to a non-practicioner.

The sufficient condition, that we provide in the current paper, for $\{X_n\}$ to be weakly lumpable with respect to partition $\tilde{S}$ is easy to implement and is geared toward applications in combinatorial reaction networks. We believed that in order to make this model reduction approach accessible to modelers and theoretical biologists, it was important to state direct proofs of the results used in our paper rather than digging it up from the general theory which turned out quite indirect and non-trivial. Toward this end, we decided to state the results for a generic Markov chain in the first half of our paper with the hope that they might find use in some other areas, and then moved on to the application section. In particular, our condition enables the user to recover information about the original Markov chain from the smaller aggregated one (see Theorems 2, 6, 9, 11). This 'invertibility' property is particularly useful for modeling protein networks and is not addressed explicitly for weakly lumpable chains in previous literature. A variant of our condition can be found in Buchholz (2008) where the author considered backward bisimulation over a class of weighted automata (finite automata where weights and labels are assigned to transitions). For each i, let α_i be a probability measure over A_i. The condition that we impose requires that for every i and j, $\sum_{s\in A_i} \alpha_i(s)P(s, s')/\alpha_j(s'), s' \in A_j$ is constant over A_j. The condition can be interpreted as follows: Suppose that the Markov chain is at the state $s' \in A_j$ and the user looks back and tries to compute the probability of the previous position being somewhere in A_i. Then, the above condition implies that this probability is independent of the specific location in A_j. Loosely speaking, this preserves the rate of flow in different states of the aggregate (equivalence) class. In particular, the above condition generalizes the notion of exact lumpability which corresponds to the case when the measures α_i are uniform (Buchholz 1994). Interestingly, if the initial distribution 'π respects α_i' in the sense that $\pi(s)/\pi(A_i) = \alpha_i(s)$, then the conditional probability $P(X_t = s \mid Y_t = A_i) = \alpha_i(s)$, for all $t > 0$. In fact, we proved that even if the initial distribution does not respect the α_i, the above result holds asymptotically. These convergence results established in the article are particularly useful for modeling purposes and to the best of our knowledge have not been discussed before. They imply that the modeler can run the 'smaller', aggregated process $\{Y_t\}$ and can still extract information about the 'bigger' process $\{X_t\}$ if the need arises. This is further illustrated in the application section.

The paper is organized as follows. In Sect. 2, we describe conditions on the transition matrix and initial distribution of the Markov chain $\{X_n\}$ which will ensure that the aggregated chain $\{Y_n\}$ is also Markov. The conditions described are tailor-made for our applications to biochemical reaction networks. We also prove convergence properties of the transition probabilities of the aggregated chain when the initial distribution does not satisfy the required conditions. The case of continuous time chains has been treated in Sect. 3. Section 4 first discusses the traditional Markov chain modeling of biochemical reaction systems using reactions and species. Next, the mathematical definition of site-graphs is introduced and the formal description of site-graph based modeling of protein-protein interaction is given. Section 5 is devoted to applications. We describe the criteria for testing the aggregation conditions on the CTMCs which underly rule-based models. Illustrative case studies are given at the end.

2 Discrete time case

Let $\{X_n\}$ be a Markov chain taking values in a finite set S with transition matrix P and initial probability distribution π (see Norris 1998, Chapter 1, for a nice introduction to discrete-time Markov chains). Let $\tilde{S} = \{A_1, \ldots, A_m\}$ be a finite partition of S. Moreover, let $\{\alpha_i : i = 1, \ldots, m\}$ be a family of probability measures on S, such that $\alpha_i(s) = 0$ for $s \notin A_i$. Define $\delta : \tilde{S} \times S \to \mathbb{R}_{\geq 0}$ by

$$\delta(A_i, s) = \frac{\sum_{s' \in A_i} \alpha_i(s') P(s', s)}{\alpha_j(s)}, \quad \text{where } s \in A_j.$$

Assume that the following condition holds.

(Cond1) For any $A_i, A_j \in \tilde{S}$ and $s, s' \in A_j$, $\delta(A_i, s) = \delta(A_i, s')$.

Fix $s \in A_j$ and let $\tilde{P}(A_i, A_j) := \delta(A_i, s)$. Notice that $\tilde{P}$ is unambiguously defined under (Cond1).

Remark 1 Throughout this section we will assume that (Cond1) holds.

Theorem 1 *$\tilde{P}$ is a probability transition matrix.*

Proof Notice that by (Cond1),

$$\alpha_j(s) \tilde{P}(A_i, A_j) = \sum_{s' \in A_i} \alpha_i(s') P(s', s).$$

Summing over $s \in A_j$, we have

$$\tilde{P}(A_i, A_j) = \sum_{s' \in A_i} \sum_{s \in A_j} \alpha_i(s') P(s', s).$$

It follows that

$$\begin{aligned}\sum_{j=1}^{m} \tilde{P}(A_i, A_j) &= \sum_{j=1}^{m} \sum_{s' \in A_i} \sum_{s \in A_j} \alpha_i(s') P(s', s) \\ &= \sum_{s' \in A_i} \alpha_i(s') \sum_{s \in S} P(s', s) = \sum_{s' \in A_i} \alpha_i(s') \\ &= 1.\end{aligned}$$

□

Definition 1 For any probability distribution π on S, define the probability distributions $\pi|_{A_i}$ on A_i and $\tilde{\pi}$ on $\tilde{S}$ by

$$\pi|_{A_i}(s) := \begin{cases} \frac{\pi(s)}{\sum_{s' \in A_i} \pi(s')}, & \sum_{s' \in A_i} \pi(s') > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{\pi}(A_i) := \sum_{s' \in A_i} \pi(s').$$

Definition 2 We say that a probability distribution π **respects** $\{\alpha_i : i = 1, \ldots, m\}$ if $\pi|_{A_i}(s) = \alpha_i(s)$ for all $s \in A_i$, $i = 1, \ldots, m$.

2.1 Aggregation and de-aggregation

Remark 2 For the remainder of the section, we will assume that $\{Y_n\}$ is a Markov chain taking values in $\tilde{S}$ with transition matrix $\tilde{P}$ and initial distribution $\tilde{\pi}$.

Theorem 2 *Assume that π respects $\{\alpha_i : i = 1, \ldots, m\}$. Then for all $n = 0, 1, \ldots$*

(i) *(lumpability)* $\mathsf{P}(Y_n = A_i) = \mathsf{P}(X_n \in A_i)$;
(ii) *(invertibility)* $\mathsf{P}(X_n = s) = \mathsf{P}(Y_n = A_i)\alpha_i(s)$.

We need the following two lemmas to prove Theorem 2.

Lemma 1 *Assume that for all $i = 1, \ldots, m$, $\mathsf{P}(X_{n-1} = s | X_{n-1} \in A_i) = \pi P^{n-1}|_{A_i}(s) = \alpha_i(s)$. Then, $\mathsf{P}(X_n \in A_j | X_{n-1} \in A_i) = \tilde{P}(A_i, A_j)$.*

Proof Notice that

$$\begin{aligned}&\mathsf{P}\left(X_n \in A_j | X_{n-1} \in A_i\right) \\ &= \frac{\mathsf{P}\left(X_n \in A_j, X_{n-1} \in A_i\right)}{\mathsf{P}\left(X_{n-1} \in A_i\right)} \\ &= \frac{\sum_{s' \in A_i} \sum_{s \in A_j} \mathsf{P}\left(X_n = s, X_{n-1} = s'\right)}{\mathsf{P}\left(X_{n-1} \in A_i\right)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{s' \in A_i} \sum_{s \in A_j} \mathsf{P}\left(X_{n-1} = s'\right) \mathsf{P}\left(X_n = s | X_{n-1} = s'\right)}{\mathsf{P}\left(X_{n-1} \in A_i\right)} \\
&= \frac{\sum_{s' \in A_i} \sum_{s \in A_j} \mathsf{P}\left(X_{n-1} \in A_i\right) \mathsf{P}\left(X_{n-1} = s' | X_{n-1} \in A_i\right) P\left(s', s\right)}{\mathsf{P}\left(X_{n-1} \in A_i\right)} \\
&= \frac{\mathsf{P}\left(X_{n-1} \in A_i\right) \sum_{s' \in A_i} \sum_{s \in A_j} \alpha_i\left(s'\right) P\left(s', s\right)}{\mathsf{P}\left(X_{n-1} \in A_i\right)}, \quad \text{by the hypothesis} \\
&= \sum_{s' \in A_i} \sum_{s \in A_j} \alpha_i\left(s'\right) P\left(s', s\right) \frac{\alpha_j(s)}{\alpha_j(s)} \\
&= \sum_{s \in A_j} \tilde{P}\left(A_i, A_j\right) \alpha_j(s), \quad \text{by the definition of } \tilde{P} \\
&= \tilde{P}\left(A_i, A_j\right) \sum_{s \in A_j} \alpha_j(s) = \tilde{P}\left(A_i, A_j\right).
\end{aligned}
$$

Lemma 2 *Assume that* $\pi|_{A_i}(s) = \alpha_i(s)$ *for* $s \in A_i, i = 1, \ldots, m$. *Then* $\pi P^n|_{A_i}(s) = \mathsf{P}(X_n = s | X_n \in A_i) = \alpha_i(s)$.

Proof The first equality is of course by the definition. For the second, we use induction. The case $n = 0$ is given. Suppose that the statement holds for $k = n - 1$. First observe that if $s \notin A_i$, then both sides equal 0. So assume that $s \in A_i$. Then, by Lemma 1, we have that $\mathsf{P}(X_n \in A_j | X_{n-1} \in A_i) = \tilde{P}(A_i, A_j)$. Next note that

$$
\begin{aligned}
&P\left(X_n = s | X_n \in A_i\right) \\
&= \frac{\mathsf{P}\left(X_n = s\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{s' \in S} \mathsf{P}\left(X_{n-1} = s', X_n = s\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{s' \in S} \mathsf{P}\left(X_{n-1} = s'\right) \mathsf{P}\left(X_n = s | X_{n-1} = s'\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{j=1}^{m} \sum_{s' \in A_j} \mathsf{P}\left(X_{n-1} = s'\right) \mathsf{P}\left(X_n = s | X_{n-1} = s'\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{j=1}^{m} \sum_{s' \in A_j} \mathsf{P}\left(X_{n-1} \in A_j\right) \mathsf{P}\left(X_{n-1} = s' | X_{n-1} \in A_j\right) P\left(s', s\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{j=1}^{m} \mathsf{P}\left(X_{n-1} \in A_j\right) \left(\sum_{s' \in A_i} \alpha_j\left(s'\right) P\left(s', s\right) \cdot \frac{\alpha_i(s)}{\alpha_i(s)}\right)}{\mathsf{P}\left(X_n \in A_i\right)} \\
&= \frac{\sum_{j=1}^{m} \mathsf{P}\left(X_{n-1} \in A_j\right) \tilde{P}\left(A_j, A_i\right) \alpha_i(s)}{\mathsf{P}\left(X_n \in A_i\right)}
\end{aligned}
$$

$$= \alpha_i(s) \frac{\sum_{j=1}^{m} \mathsf{P}\left(X_{n-1} \in A_j\right) \mathsf{P}\left(X_n \in A_i | X_{n-1} \in A_j\right)}{\mathsf{P}\left(X_n \in A_i\right)}$$
$$= \alpha_i(s) \frac{\mathsf{P}\left(X_n \in A_i\right)}{\mathsf{P}\left(X_n \in A_i\right)} = \alpha_i(s).$$

□

We next proceed to prove Theorem 2.

Proof (Theorem 2) We use induction. Notice that both the statements hold for $n = 0$. Assume that (i) and (ii) hold for $n-1$. Then $\mathsf{P}(X_{n-1} = s | X_{n-1} \in A_i) = \alpha_i(s)$, and hence by Lemma 1 $\mathsf{P}(X_n \in A_j | X_{n-1} \in A_i) = \tilde{P}(A_i, A_j)$. Therefore,

$$\begin{aligned}\mathsf{P}(Y_n = A_i) &= \sum_{j=1}^{m} \mathsf{P}(Y_{n-1} = A_j)\tilde{P}(A_j, A_i) \\ &= \sum_{j=1}^{m} \mathsf{P}(X_{n-1} \in A_j)\mathsf{P}(X_n \in A_j | X_{n-1} \in A_i) \\ &= \mathsf{P}(X_n \in A_i).\end{aligned}$$

This proves (i). Next, notice that Lemma 2 implies

$$\begin{aligned}\mathsf{P}(X_n = s) &= \alpha_i(s)\mathsf{P}(X_n \in A_i) \\ &= \alpha_i(s)\mathsf{P}(Y_n = A_i), \text{ by (i).}\end{aligned}$$

This proves (ii). □

Remark 3 Notice that we have proved that under the assumption $\pi|_{A_i}(s) = \alpha_i(s)$, $\mathsf{P}(X_n \in A_j | X_{n-1} \in A_i) = \tilde{P}(A_i, A_j)$, for $n = 1, 2, \ldots$.

2.2 Convergence

In the previous section, we proved that if $\{X_n\}$ is a discrete time Markov chain on S with initial distribution π respecting $\{\alpha_i : i = 1, \ldots, m\}$, then the aggregate process $\{Y_n\}$ is an aggregated Markov chain satisfying lumpability and invertibility property. We now investigate the case when the initial distribution of $\{X_n\}$ doesn't respect $\{\alpha_i : i = 1, \ldots, m\}$. We start with the following theorem.

Theorem 3 $\tilde{P}^n(A_i, A_j) = \frac{\sum_{s' \in A_i} \alpha_i(s') P^n(s', s)}{\alpha_j(s)}$, *for any* $s \in A_j$.

Proof We use induction. Notice that for $n = 1$, the assertion is true by the definition of $\tilde{P}$. Assume that the statement holds for some n. Then,

$$\tilde{P}^{n+1}(A_i, A_j) = \sum_k \tilde{P}(A_i, A_k)\tilde{P}^n(A_k, A_j)$$

$$
\begin{aligned}
&= \sum_k \left(\sum_{s' \in A_i} \frac{\alpha_i(s') P(s', s_0)}{\alpha_k(s_0)} \right) \left(\sum_{s'' \in A_k} \frac{\alpha_k(s'') P^n(s'', s)}{\alpha_j(s)} \right), \text{ for any } s_0 \in A_k \\
&= \sum_k \sum_{s'' \in A_k} \left(\sum_{s' \in A_i} \frac{\alpha_i(s') P(s', s_0)}{\alpha_k(s_0)} \right) \left(\frac{\alpha_k(s'') P^n(s'', s)}{\alpha_j(s)} \right) \\
&= \sum_k \sum_{s'' \in A_k} \left(\sum_{s' \in A_i} \frac{\alpha_i(s') P(s', s'')}{\alpha_k(s'')} \right) \left(\frac{\alpha_k(s'') P^n(s'', s)}{\alpha_j(s)} \right), \text{ by (Cond1)} \\
&= \frac{1}{\alpha_j(s)} \sum_k \sum_{s'' \in A_k} \left(\sum_{s' \in A_i} \alpha_i(s') P(s', s'') P^n(s'', s) \right) \\
&= \frac{1}{\alpha_j(s)} \sum_{s' \in A_i} \alpha_i(s') \sum_k \sum_{s'' \in A_k} P(s', s'') P^n(s'', s) \\
&= \frac{\sum_{s' \in A_i} \alpha_i(s') P^{n+1}(s', s)}{\alpha_j(s)}
\end{aligned}
$$

We say that $s \to s'$, if for some $n \geq 0$, $P^n(s, s') > 0$. Recall that the Markov chain $\{X_n\}$ is irreducible if $s \to s'$ for any $s, s' \in S$. One corollary of Theorem 3 is that if for $s \in A_i, s' \in A_j$, $s \to s'$ then $A_i \to A_j$ for the Markov chain Y. In fact, we have the following result.

Theorem 4 *Let $\{X_n\}$ be a discrete time Markov chain on S with transition probability matrix P and $\{Y_n\}$ a Markov chain taking values in $\tilde{S}$ with transition matrix $\tilde{P}$. Then*

(i) *If the process $\{X_n\}$ is irreducible, then so is $\{Y_n\}$.*
(ii) *If $s \in A_i$ is recurrent for the process $\{X_n\}$, then so is A_i for the process $\{Y_n\}$.*
(iii) *If $s \in A_i$ has period 1, then the period of A_i is also 1.*

Proof (i) is immediate from the discussion. Notice that if $s' \in A_i, s \in A_j$, then by Theorem 3, $\tilde{P}^n(A_i, A_j) \geq \frac{\alpha_i(s')}{\alpha_j(s)} P^n(s', s)$. Therefore, it follows that if $s' \in A_i = A_j$, then $\tilde{P}^n(A_i, A_i) \geq P^n(s', s')$. Now $s \in A_i$ is recurrent if and only if $\sum_n P^n(s', s') = \infty$. Observe that

$$\sum_n \tilde{P}^n(A_i, A_i) \geq \sum_n P^n(s', s') = \infty.$$

It follows that for the process $\{Y_n\}$, A_i is recurrent. This proves (ii). Next observe that

$$\left\{ n : P^n(s', s') > 0 \right\} \subset \left\{ n : \tilde{P}^n(A_i, A_i) > 0 \right\},$$

and (iii) follows immediately.

For the following results we will assume that there exists a probability distribution π on S which respects $\{\alpha_i : i = 1, \ldots, m\}$.

Theorem 5 *Let $\{X_n\}$ be a discrete time Markov chain on S with transition probability matrix P and unique stationary distribution μ. Then μ respects $\{\alpha_i : i = 1, \ldots, m\}$.*

Proof Let π be a probability distribution on S which respects $\{\alpha_i : i = 1, \ldots, m\}$. Now since μ is unique, we have for any set A (see Hernández-Lerma and Lasserre 2003),

$$\frac{1}{n}\sum_{k=1}^{n} \pi P^k(A) \to \mu(A), \quad \text{as } n \to \infty.$$

By the choice of π, $\pi(s) = \alpha_i(s)\pi(A_i)$ for $s \in A_i$. By Theorem 2, $\pi P^k(s) = \alpha_i(s)\pi P^k(A_i)$, $s \in A_i$. Therefore, it follows that

$$\frac{1}{n}\sum_{k=1}^{n} \pi P^k(s) = \alpha_i(s)\frac{1}{n}\sum_{k=1}^{n} \pi P^k(A_i),\ s \in A_i.$$

Taking limit as $n \to \infty$, it implies that $\mu|_{A_i}(s) = \alpha_i(s)$, $s \in A_i$, $i = 1, \ldots, m$. □

For any set $A \subset S$, let $P^{(n)}(s, A) := \frac{1}{n}\sum_{k=1}^{n} P^k(s, A)$, and $\mathsf{P}^{(n)}(X_n \in A) := \frac{1}{n}\sum_{k=1}^{n} \mathsf{P}(X_k \in A)$.

Theorem 6 *Let $\{X_n\}$ be an irreducible Markov chain taking values in S with transition matrix P. Let μ be the stationary distribution of P. Let $\{Y_n\}$ be a Markov chain on $\tilde{S}$ with transition matrix $\tilde{P}$. Then $\tilde{\mu}$ is the stationary distribution for $\tilde{P}$. Also for all $n = 0, 1, \ldots$,*

(i) $\mathsf{P}^{(n)}(Y_n = A_i) - \mathsf{P}^{(n)}(X_n \in A_i) \to 0$;
(ii) $\mathsf{P}^{(n)}(X_n = s)/\mathsf{P}^{(n)}(Y_n = A_i) \to \alpha_i(s)$.

Proof We first show that $\tilde{\mu}$ is a stationary distribution of $\tilde{P}$. Towards this end, first observe that by Theorem 5 μ respects $\{\alpha_i : i = 1, \ldots, m\}$. Take μ as the initial distribution of $\{X_n\}$. Then by (i) of Theorem 2,

$$\tilde{\mu}(A_i) = \mu(A_i) = \mu P^n(A_i) = \tilde{\mu}\tilde{P}^n(A_i).$$

It follows that $\tilde{\mu}$ is stationary for $\tilde{P}$. Since $\tilde{P}$ is irreducible by Theorem 4, $\tilde{\mu}$ is unique. Now let π be any initial distribution for P. Since $\tilde{\mu}$ is the unique stationary distribution for $\tilde{P}$, $\pi P^{(n)}(A_i) \to \tilde{\mu}(A_i)$. Hence (i) and (ii) follow. □

The above result can be improved if we assume in addition that the Markov chain $\{X_n\}$ is aperiodic.

Theorem 7 *Let $\{X_n\}$ be an irreducible, aperiodic Markov chain taking values in S with transition matrix P. Let μ be the stationary distribution of P. Let $\{Y_n\}$ be a Markov chain on $\tilde{S}$ with transition matrix $\tilde{P}$. Then*

(i) $\mathsf{P}(Y_n = A_i) - \mathsf{P}(X_n \in A_i) \to 0$;

(ii) $\mathsf{P}(X_n = s)/\mathsf{P}(Y_n = A_i) \to \alpha_i(s)$.

Proof By Theorem 4, the Markov chain $\{Y_n\}$ is also aperiodic and irreducible. Moreover by the previous theorem, $\tilde{\mu}$ is the unique stationary distribution for $\{Y_n\}$. The result follows by noting that for any aperiodic, irreducible Markov chain $\{Z_n\}$ with a stationary distribution η, $\mathsf{P}(Z_n \in A) \to \eta(A)$.

3 Continuous time case

We now consider a continuous time Markov chain, $\{X_t\}_{t\in[0,\infty)}$, taking values in a countable set S (see Norris 1998, Chapters 2, 3, for an introduction to continuous-time Markov chains). Let Q be the generator matrix for $\{X_t\}$. As before, let $\tilde{S} = \{A_1, \ldots, A_m\}$ be a finite partition of S and $\{\alpha_i : i = 1, \ldots, m\}$ a family of probability measures on S with $\alpha_i(s) = 0$, for $s \notin A_i$. Define $\Delta : \tilde{S} \times S \to \mathbb{R}_{\geq 0}$ by

$$\Delta(A_i, s) = \frac{\sum_{s' \in A_i} \alpha_i(s') Q(s', s)}{\alpha_j(s)}, \text{ where } s \in A_j. \tag{1}$$

Assume the following condition holds.

(Cond2) For any $A_i, A_j \in \tilde{S}$ and $s, s' \in A_j$, $\Delta(A_i, s) = \Delta(A_i, s')$.

Fix $s \in A_j$ and let $\tilde{Q}(A_i, A_j) := \Delta(A_i, s)$. Notice that $\tilde{Q}$ is unambiguously defined under (Cond2).

Remark 4 Throughout the present section we assume that (Cond2) holds.

Theorem 8 *$\tilde{Q}$ is a generator matrix.*

Proof We only need to prove that $\sum_{j=1}^{m} \tilde{Q}(A_i, A_j) = 0$. The proof proceeds almost exactly in the same way as that of Theorem 1. □

For any generator matrix $Q = (q_{ij})$, define

$$q_i = -q_{ii} = \sum_{j \neq i} q_{ij}.$$

3.1 Aggregation and de-aggregation

We next prove the analogue of Theorem 2.

Theorem 9 *Let $\{X_t\}$ be a continuous time Markov chain taking values in a countable set S with generator matrix Q and initial probability distribution π. Let $\{Y_t\}$ be a continuous time Markov chain taking values in $\tilde{S}$ with generator matrix $\tilde{Q}$ and initial distribution $\tilde{\pi}$. Assume that π respects $\{\alpha_i : i = 1, \ldots, m\}$. Also assume that there exists an $r > 0$ such that $\sup_i q_i < r$. Then for all $t \geq 0$*

(i) *(lumpability)* $\mathsf{P}(Y_t = A_i) = \mathsf{P}(X_t \in A_i)$;

(ii) *(invertibility)* $\mathsf{P}(X_t = s) = \mathsf{P}(Y_t = A_i)\alpha_i(s)$.

We prove the above theorem by constructing a uniformized discrete time Markov chain out of $\{X_t\}$ (see Ibe 2009, Chapter 4). For any matrix $A = ((a_{ij}))_{i,j\in S}$, we use the norm $\|A\| = \sup_i \sum_j |a_{ij}|$. Note by the assumptions in Theorem 9, $\|Q\| < r < \infty$. If P denotes the transition probability matrix of $\{X_t\}$, then P satisfies the Kolmogorov forward equation

$$P'(t) = P(t)Q, \quad t > 0.$$

Since $\|Q\| < \infty$, the solution to the above equation is given by

$$P(t) = e^{Qt} = \sum_{k=0}^{\infty} (Qt)^k/k!.$$

Define the transition matrix M by $M = I + Q/r$. Writing $Q = r(M - I)$ we have

$$P(t) = e^{r(M-I)t} = e^{-rt} \sum_{k=0}^{\infty} \frac{(rt)^k}{k!} M^k. \tag{2}$$

Let $\{Z_n\}$ be a Markov chain on S with transition probability matrix M. Let ξ be a Poisson process with intensity r independent of $\{Z_n\}$. Then (2) implies that $\{X_t\} \stackrel{d}{=} \{Z(\xi(t))\}$. We will need to consider the aggregate Markov chain $\{\tilde{Z}_n\}$ on $\tilde{S}$ with the transition matrix defined by

$$\tilde{M}(A_i, A_j) = \frac{\sum_{s'\in A_i} \alpha_i(s')M(s', s)}{\alpha_j(s)}, \quad s \in A_j. \tag{3}$$

Lemma 3 *$\tilde{M}$ is well-defined.*

Proof We need to show that $\tilde{M}(A_i, A_j)$ does not depend on the choice of $s \in A_j$. Let $\rho(A_i, s)$ denote the right side of (3) for $s \in A_j$. We will use (Cond2). First assume that $i \neq j$. Then,

$$\rho(A_i, s) = \frac{1}{r} \frac{\sum_{s'\in A_i} \alpha_i(s')Q(s', s)}{\alpha_j(s)} = \frac{1}{r}\tilde{Q}(A_i, A_j), \quad s \in A_j. \tag{4}$$

by the definition of the $\tilde{Q}$ matrix. If $i = j$, then

$$\begin{aligned}
\rho(A_i, s) &= \frac{\sum_{s'\neq s, s'\in A_i} \alpha_i(s')Q(s', s)/r + \alpha_i(s)(1 + Q(s, s)/r)}{\alpha_i(s)} \\
&= \frac{\sum_{s'\in A_i} \alpha_i(s')Q(s', s)/r - \alpha_i(s)Q(s, s)/r + \alpha_i(s)(1 + Q(s, s)/r)}{\alpha_i(s)} \\
&= 1 + \frac{1}{r} \frac{\sum_{s'\in A_i} \alpha_i(s')Q(s', s)}{\alpha_i(s)} = 1 + \frac{1}{r}\tilde{Q}(A_i, A_i).
\end{aligned} \tag{5}$$

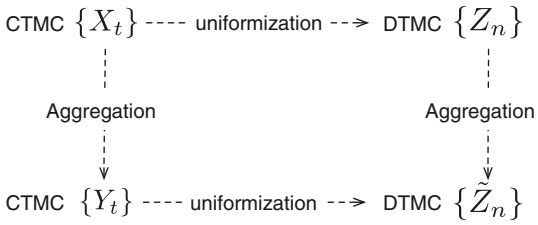

Fig. 1 Commutative diagram showing that the order of carrying out the process of aggregation and uniformization does not matter

It follows $\rho(A_i, s)$ is independent of the choice of $s \in A_j$, $j = 1, \ldots, m$. □

The next theorem proves the commutativity diagram depicted in Fig. 1. In other words, it essentially proves that it does not matter whether starting from a CTMC X we do uniformization first and then aggregation or aggregation first and then uniformization; the resulting DTMC is the same. The importance of Theorem 10 lies in paving the way for usage of the DTMC results in proving Theorem 9.

Theorem 10 *Let Q be a generator matrix with $\sup_i q_i < r$ for some r, and let*

(i) *$\{X_t\}$ be a continuous time Markov chain taking values in a countable set S with generator matrix Q and initial probability distribution π.*
(ii) *$\{Y_t\}$ be a continuous time Markov chain taking values in $\tilde{S}$ with generator matrix $\tilde{Q}$ and initial distribution $\tilde{\pi}$.*
(iii) *$\{Z_n\}$ be the uniformized discrete time Markov chain (corresponding to $\{X_t\}$) on S with transition matrix $M = I + Q/r$ and initial distribution π.*
(iv) *$\{\tilde{Y}_n\}$ be the uniformized discrete time chain (corresponding to $\{Y_t\}$) on $\tilde{S}$ with transition matrix $\bar{M} = I + \tilde{Q}/r$ and initial distribution $\tilde{\pi}$.*
(v) *$\{\tilde{Z}_n\}$ be the discrete time Markov chain on $\tilde{S}$ with transition matrix $\tilde{M}$ and initial distribution $\tilde{\pi}$.*

Then $\{\tilde{Z}_n\} \stackrel{d}{=} \{\tilde{Y}_n\}$.

Proof We only need to show that $\tilde{M} = \bar{M}$. But this readily follows from (4) and (5).

Proof (Theorem 9) Note that (2) implies.

$$\begin{aligned}
\mathsf{P}(Y_t = A_i) &= \sum_{k\geq 0} \mathsf{P}\left(\tilde{Y}_k = A_i\right) \frac{e^{-rt}(rt)^k}{k!} \\
&= \sum_{k\geq 0} \mathsf{P}\left(\tilde{Z}_k = A_i\right) \frac{e^{-rt}(rt)^k}{k!} \\
&= \sum_{k\geq 0} \mathsf{P}(Z_k \in A_i) \frac{e^{-rt}(rt)^k}{k!} \\
&= \sum_{s\in A_i} \left(\sum_{k\geq 0} \mathsf{P}(Z_k = s) \frac{e^{-rt}(rt)^k}{k!} \right) \\
&= \sum_{s\in A_i} \mathsf{P}(X_t = s) = \mathsf{P}(X_t \in A_i).
\end{aligned}$$

Here, the second equality is by Theorem 10 while the third is by (i) of Theorem 2. This proves (i) and (ii) follows similarly. □

3.2 Convergence

Let μ be a stationary distribution of the continuous time Markov chain $\{X_t\}$, that is μ satisfies $\mu Q = 0$. Then we have the corresponding analogue of Theorem 6.

Theorem 11 *Let $\{X_t\}$ be an irreducible Markov chain taking values in S with generator matrix Q. Assume that $\sup_i q_i < r$, for some $r > 0$. Let μ be the stationary distribution of Q. Let $\{Y_t\}$ be a Markov chain on $\tilde{S}$ with generator matrix $\tilde{Q}$. Then $\tilde{\mu}$ is the stationary distribution for $\tilde{Q}$. Moreover,*

(i) $\mathsf{P}(Y_t = A_i) - \mathsf{P}(X_t \in A_i) \to 0$;
(ii) $\mathsf{P}(X_t = s)/\mathsf{P}(Y_t = A_i) \to \alpha_i(s)$.

Proof We first consider the uniformized chain $\{Z_n\}$ corresponding to $\{X_t\}$ with transition matrix $M = I + Q/r$. Note that μ is the stationary distribution for $\{M\}$. It follows by Theorem 6, that $\tilde{\mu}$ is the stationary distribution for $\{\tilde{Z}_n\}$, hence for $\{\tilde{Y}_n\}$. It follows that $\tilde{\mu}\tilde{Q} = 0$. Next $\sup_i q_i < \infty$ guarantees that the chain does not explode. The result follows by noting that for any irreducible, non-exploding continuous time Markov chain $\{Z_t\}$ with a stationary distribution η, $\mathsf{P}(Z_t \in A) \to \eta(A)$ as $t \to \infty$. □

4 Formalism

The standard model of biochemical networks is typically based on counting chemical species (complexes). However, for our purpose it is useful to consider a *site-graph* based description of the model. We start by briefly outlining the Markov chain formulation of a species-based model of a biochemical reaction system, and then move on to the concept of site-graph.

4.1 Modeling biochemical networks by a CTMC

A biochemical reaction system involves multiple chemical reactions and several species. In general, chemical reactions in single cells occur far from thermodynamic equilibrium and the number of molecules of chemical species is often low (Keizer 1987; Guptasarma 1995). Recent advances in real-time single cell imaging, microfluidic techniques and synthetic biology have testified to the random nature of gene expression and protein abundance in single cells (Yu et al. 2006; Friedman et al. 2010). Thus a stochastic description of chemical reactions is often mandatory to analyze the behavior of the system. The dynamics of the system is typically modeled by a continuous-time Markov chain (CTMC) with the state being the number of molecules of each species. Anderson and Kurtz (2011) is a good reference for a review of the tools of Markov processes used in the reaction network systems.

Consider a biochemical reaction system consisting of n species and v reactions, and let $X(t)$ denote the state of the system at time $t \in [0, \infty)$. If the k-th reaction occurs

at time t, then the system is updated as $X(t) = X(t-) + \nu_k^+ - \nu_k^-$, where $X(t-)$ denotes the state of the system just before time t, and $\nu_k^-, \nu_k^+ \in \mathbb{Z}_+^n$ represent the vector of number of molecules consumed and created in one occurrence of reaction k, respectively. For convenience, let $\nu_k = \nu_k^+ - \nu_k^-$. Let $a_k(x)$ denote the propensity of the reaction k, which captures the likelihood of an occurrence of reaction k per unit of time. Specifically, the evolution of the process X is modeled by

$$\mathsf{P}[X(t + \Delta t) = x + \nu_k | X(t) = x] = a_k(x)\Delta t + o(\Delta t).$$

The propensity functions a_k are often calculated by using the *law of mass action* (Wilkinson 2006; Gillespie 2007). The generator matrix or the Q-matrix of the CTMC X is given by $q_{x,x+\nu_k} = a_k(x)$. The CTMC X will have an *invariant measure* π if $\pi Q \equiv 0$.

Example 1 (Simple scaffold) Assume that a scaffold protein B can bind to protein A and protein C. Moreover, assume that the binding of B and A occurs at rate c_1 (and the respective reverse reaction at rate c_1^-), no matter whether protein B is already bound to C or not, and that the binding of B and C occurs at rate c_2 (and the respective reverse reaction at rate c_2^-), no matter whether protein B is already bound to A or not. Denoting the respective species by $S_A,\ S_B,\ S_C,\ S_{AB},\ S_{BC},\ S_{ABC}$, the described model is given by four reversible reactions:

$$S_A, S_B \underset{c_1^-}{\overset{c_1}{\rightleftharpoons}} S_{AB}, \qquad S_A, S_{BC} \underset{c_1^-}{\overset{c_1}{\rightleftharpoons}} S_{ABC},$$
$$S_B, S_C \underset{c_2^-}{\overset{c_2}{\rightleftharpoons}} S_{BC}, \qquad S_{AB}, S_C \underset{c_2^-}{\overset{c_2}{\rightleftharpoons}} S_{ABC}.$$

If the reaction mixture counts three copies of species S_B, and one copy of each of the species $S_A,\ S_C$, the state of the process is $x = (1, 3, 1, 0, 0, 0)$. Upon dimerization between S_A and S_B, the system is in the state $x' = (0, 2, 1, 1, 0, 0)$, and the rate of the transition between x and x' is $q_{x,x'} = 3c_1$, where the coefficient 3 reflects that three different collisions lead from the state x to the state x'.

By explicitly encoding the two different binding domains of scaffold B, the equivalent CTMC can be specified with only two reversible site-graph-rewrite rules, depicted in Fig. 2a. In the following Section, the site-graph-rewrite rules are formally introduced.

4.2 Site-graphs

The notion of a site-graph is a generalization of that of a standard graph. A site-graph consists of nodes and edges; Each node is assigned a set of sites, and the edges are established between two sites of (different) nodes. The nodes of a site-graph can be interpreted as protein names, and sites of a node stand for protein binding domains. Let $\mathcal{S}$ denote the set of all the sites in a site-graph, and let $\mathcal{P}(\mathcal{S})$ denote the class of all subsets of $\mathcal{S}$.

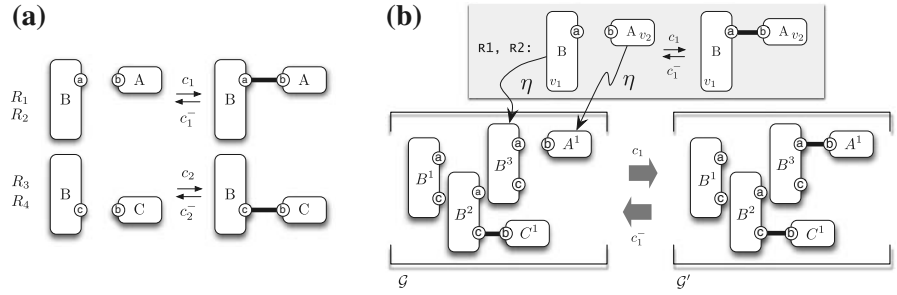

Fig. 2 Case study 1: simple scaffold. **a** The model consists of two reversible rules: a scaffold B has two binding sites, a and c, which serve for binding nodes A and C, respectively. **b** The application of rule R_1 to the reaction mixture $\mathcal{G}$ via node renaming function η results in a reaction mixture $\mathcal{G}'$, which is equivalent to $\mathcal{G}$ except in the sub-site-graph captured by node renaming η

Definition 3 A **site-graph** $G = (V, \Sigma, E)$ is defined by a set of nodes V, an interface function $\Sigma : V \to \mathcal{P}(\mathcal{S})$, and a set of edges $E \subseteq \{\{(v, s), (v', s')\} | v, v' \in V, v \neq v', s \in \Sigma(v), s' \in \Sigma(v')\}$.

The function Σ in the above definition tracks the sites corresponding to a particular node of a site-graph.

Definition 4 Given a site-graph $G = (V, \Sigma, E)$, a sequence of edges $(e_1, \ldots e_k) \in E^k$, $e_i = \{(v_i, s_i), (v_i', s_i')\}$, such that $v_i' = v_{i+1}$ and $s_i' \neq s_{i+1}$ for $i = 1, \ldots k-1$, is called a *path* between nodes v_1 and v_k. If there exists a path between every two nodes $v, v' \in V$, a site-graph $G = (V, \Sigma, E)$ is *connected*.

Definition 5 Let $G = (V, \Sigma, E)$ be a site-graph. A site graph G' is a **sub-site-graph** of G, written $G' \subseteq G$, if $V' \subseteq V$, for all $v \in V'$, $\Sigma'(v) \subseteq \Sigma(v)$, and $E' \subseteq E$.

4.3 Site-graph-rewrite rules

Definition 6 Let $G = (V, \Sigma, E)$ be a site-graph. We introduce two elementary site-graph transformations: adding/deleting an edge.

- $\delta_{ae}(G, e) = (V_{new}, \Sigma, E_{new})$: $V_{new} = V$, $E_{new} = E \cup \{e\}$,
- $\delta_{de}(G, e) = (V_{new}, \Sigma, E_{new})$: $V_{new} = V$, $E_{new} = E \backslash \{e\}$,

The interface function Σ is unaltered under any of the above transformations. Let $G' = (V', \Sigma, E')$ be a site-graph derived from $G = (V, \Sigma, E)$ by a finite number of applications of δ_{ae}, δ_{de}. Let $c \in \mathbb{R}_{\geq 0}$ be a non-negative real number denoting the rate of the transformation. The triple (G, G', c), also denoted by $G \xrightarrow{c} G'$, is called a **site-graph-rewrite rule**.

4.4 Rule-based model

Suppose that $\mathcal{R} \equiv \{R_1, \ldots, R_n\}$ is a collection of site-graph rewrite rules such that for $i = 1, \ldots, n$, $R_i \equiv (G_i, G_i', c_i)$ and $G_i = (V_i, \Sigma_i, E_i)$. From now on, for a given set of rules $\mathcal{R}$, we use the terminology

- the set of *node types* for $V := \cup_i V_i$,
- the set of *edge types* for $E := \cup_i E_i$,
- the *interface function* for $\Sigma : V \to \mathcal{P}(\mathcal{S})$, such that for $v \in V,\ \Sigma(v) := \cup_i \Sigma_i(v)$.

In order to model the reaction mixture (a collection of species present in the modeled cellular environment), for each node $v \in V$, we will consider n_v copies or instances of the node v, denoted by $v^1, v^2, \ldots, v^{n_v}$. Note that, in the Kappa rule-based models, the set of node types and edge types are predefined in the signature of the model; Here, it is deduced from the set of rules (a more detailed discussion to the relation with Kappa is given in Sect. 6).

Definition 7 A **reaction mixture** is a site-graph $\mathcal{G} = (\mathcal{V}, \hat{\Sigma}, \mathcal{E})$ where

- $\mathcal{V} = \{v^j | v \in V, j = 1, \ldots, n_v\}$;
- $\hat{\Sigma}(v^j) = \Sigma(v)$;
- $\mathcal{E} \subset \{\{(v_1^i, s_1), (v_2^j, s_2)\} | \{(v_1, s_1), (v_2, s_2)\} \in E, i = 1, \ldots n_{v_1}, j = 1, \ldots, n_{v_2}\}$

Definition 8 A **rule-based model** is a collection of rules $\mathcal{R}$, accompanied with the initial reaction mixture $\mathcal{G}_0$.

Remark 5 By definition, the site-graphs G_i and G_i' occurring in some rule (G_i, G_i', c_i), are such that a node $v \in V$, edge $e \in E$, and a site $s \in \Sigma(v)$ may be omitted. For example, notice that, in the simple scaffold example, the site c is omitted in rule R_1, telling that the rate of binding does not depend on whether site c is free or bound.

Definition 9 A rule (G_i, G_i', c_i) is reversible, if there exists a rule (G_j, G_j', c_j), such that $G_i = G_j'$ and $G_i' = G_j$. A rule-based model is **reversible**, if all its rules are reversible.

Let $\mathbb{G}$ be the set of all reaction mixtures which can be reached by finite number of applications of rules from $\mathcal{R}$ to a reaction mixture $\mathcal{G}_0$. We will now describe a Markov chain taking values in $\mathbb{G}$. The following notion of renaming a site-graph will be used for the formal description.

Definition 10 Let $G = (V, \Sigma, E)$ be a site-graph, V' a set such that $|V'| \geq |V|$ ($|\cdot|$ denotes the set cardinality), and $\eta : V \to V'$ an injective function. Then the η-**induced node-renamed site-graph**, G^η, is given by $G^\eta = (\eta(V), \Sigma^\eta, E^\eta)$, where $\Sigma^\eta(\eta(v)) = \Sigma(v)$ and $E^\eta = \{\{(\eta(v_1), s_1), (\eta(v_2), s_2)\} \mid v_1, v_2 \in V\}$.

4.5 The CTMC of a rule-based model

Consider a reaction mixture $\mathcal{G} \in \mathbb{G}$, a rule $R_i = (G_i, G_i', c_i) \in \mathcal{R}$. Suppose that $\eta : V \to \mathcal{V}$ is a node renaming function such that $G_i^\eta \subseteq \mathcal{G}$. This implies that the rule R_i can be applied to a part of the reaction mixture $\mathcal{G}$. Let $\mathcal{G}'_{\eta,i}$ be the unique reaction mixture obtained after the application of the rule R_i. (For a more formal definition of $\mathcal{G}'_{\eta,i}$ see Danos et al. 2010.) Note that $G_i' \subseteq \mathcal{G}'$. Define the transition rate Q by

$Q(\mathcal{G}, \mathcal{G}'_{\eta,i}) = c_i$. More precisely,

$$Q(\mathcal{G}, \mathcal{G}') = \begin{cases} c_i & \text{if } \mathcal{G}' = \mathcal{G}'_{\eta,i} \text{ for some } \eta, i \\ 0 & \text{if } \mathcal{G}' \neq \mathcal{G}'_{\eta,i} \text{ for any } \eta \text{ and } i \\ -\sum_{\mathcal{G}' \neq \mathcal{G}} Q(\mathcal{G}, \mathcal{G}') & \text{if } \mathcal{G}' = \mathcal{G} \end{cases} \tag{6}$$

Let $\{X_t\}$ be a CTMC with state-space $\mathbb{G}$ and generator matrix Q.

Example 2 (Simple scaffold, cont'd) For a site-graph-rewrite model $\mathcal{R} \equiv \{R_1, R_2, R_3, R_4\}$ depicted in Fig. 2a, $V = \cup_{i=1}^{4} V_i = \{A, B, C\}$, $\Sigma(A) = \{b\}$, $\Sigma(B) = \{a, c\}$, $\Sigma(C) = \{b\}$, and $E = \{\{(A, b), (B, a)\}, \{(C, b), (B, c)\}\}$. In Fig. 2b, we show the application of rule R_1 to the reaction mixture $\mathcal{G} = (\mathcal{V}, \Sigma, \mathcal{E})$, such that $\mathcal{V} = \{A^1, B^1, B^2, B^3, C^1\}$, and $\mathcal{E} = \{((B^3, c), (C^1, b))\}$, $\Sigma(A^1) = \{b\}$, $\Sigma(B^1) = \Sigma(B^2) = \Sigma(B^3) = \{a, c\}$, $\Sigma(C^1) = \{b\}$.

5 Application

The main practical limitation when it comes to aggregating and de-aggregating a Markov chain is to construct the appropriate partition and, in case of de-aggregation, to find the distribution over the aggregates. We here illustrate how the results from Sect. 3 (lumpability, invertability, convergence) can be efficiently used in the scenario of modeling biochemical systems. We start by illustrating the idea of the reduction intuitively on the example of a simple scaffold. Then, the observations are generalized and discussed for two other case studies: unbounded polymerization and EGF/insulin crosstalk.

For each case study, we first define a trivial aggregation of $\{X_t\}$, denoted by $\{Y_t\}$, which corresponds to the usual population-based description with mass-action kinetics. We then show that there exists another aggregation of $\{X_t\}$, denoted by $\{Z_t\}$, with

Table 1 Summary of the reduction for the presented case studies

	Lumping	# rules	dim.	Estimated # of states
Simple scaffold	Species	8	3	$(n+1)(n+2)(n+3)/6$
(3 node types)	Fragment	4	2	$(n+1)^2$
Polymerization	Species	–	n	$> 3P(n)$
(2 node types)	Fragment	4	2	$(n+1)^2$
	Fragment 2	2	1	2n+1
EGF/insulin	Species	42956	2768	–
(8 node types)	Fragment	38	609	–

In case study 1, for $n_A = n_B = n_C = n$, the number of states is reduced from $O(n^3)$ to $O(n^2)$. The number of partitions of n is denoted by $P(n) \approx \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}$ (Hardy and Ramanujan 1918). In case study 2, for $n_A = n_B = n$, there is an exponential reduction in the number of states from standard to the aggregated CTMC. In case study 3 (a crosstalk between the epidermal growth factor, EGF, and insulin pathway), the dimension of the state vector is reduced from 2768 to 609, and we did not estimate the size of the state space

much smaller state space. Finally, we show that $\{Z_t\}$ is an aggregation of $\{Y_t\}$, and that each of the properties—lumpability, invertability and convergence, can be established between $\{Y_t\}$ and $\{Z_t\}$. In particular, we outline how the conditional distribution of $\{Y_t\}$ can be retrieved from $\{Z_t\}$. The summary of all presented reductions is given in Table 1.

Before continuing, we clarify the relation to the related works. The advantage of using the rule-based model is that the process $\{Z_t\}$ can be read-out directly from the site-graph-rewrite rules, without first constructing the process $\{Y_t\}$. To this end, the reduction approach presented in this paper differs from other reduction techniques, for example, separating time-scales (Kang and Kurtz 2013), exploiting conservation laws (Borisov et al. 2006), or finite state projection (Munsky and Khammash 2006). The idea of directly constructing the process $\{Z_t\}$ was first shown in Feret et al. (2012, 2013), where the authors sought for a procedure that, for *any* rule-set written in Kappa language, provides immediately the aggregated process $\{Z_t\}$, such that $\{Z_t\}$ meets condition (Cond1) or (Cond2) with respect to $\{Y_t\}$ (finally, the latter was chosen because it led to a procedure which works for any rule-set, while (Cond1) would need the classification of rule-sets depending on the presence of deletion events).

Here, we will use the aggregated process $\{Z_t\}$ proven to satisfy (Cond2) in Feret et al. (2012, 2013), and we illustrate how the general results shown in Sect. 3 extend modeler's insights, when only the process $\{Z_t\}$ is available: based on Theorem 11, the modeler now knows that the reduction is applicable in the limit, even when the initial distribution does not respect α. In particular, in the outlined examples, we choose the process Z_t so to satisfy (Cond2), by the algorithm suggested in Feret et al. (2012, 2013); As each of the presented case studies contains only reversible rules, so the CTMC of each rule-set is trivially irreducible, and the convergence result holds. The validity of condition (Cond1) in all of the given examples is a consequence to that no deletion events occur in the considered networks. The general discussion on the applicability of (Cond1) is left to the future work.

Example 3 (Simple scaffold: cont'd) **Counting species.** In site-graph terminology, a molecular species is a class of connected reaction mixtures that are isomorphic up to renaming of the nodes of same type. Then, the species S_A is a site-graph of a particular type, denoted by, for example, type (A), the species S_{AB} is a site-graph of type (AB) etc. While the state space of $\{X_t\}$ is $\mathbb{G}$, the state space of the aggregated process $\{Y_t\}$ is $\mathbb{N}^6$. Formally, two reaction mixtures $\mathcal{G}$ and $\mathcal{G}'$ are aggregated by relation $\sim_1 \subseteq \mathbb{G} \times \mathbb{G}$ if they count the same number of each of the species:

$$\mathcal{G} \sim_1 \mathcal{G}' \text{ iff } \phi_1(\mathcal{G}) = \phi_1(\mathcal{G}'),$$

where $\phi_1 : \mathbb{G} \to \mathbb{N}^3$, with $\phi_1(\mathcal{G}) = (m_{AB}, m_{BC}, m_{ABC})$, if contains m_{AB} different site-graphs of type (AB) (species S_{AB}), m_{BC} different site-graphs of type (BC) (species S_{BC}), and m_{ABC} different site-graphs of type (ABC) (species S_{ABC}) are sub-site-graphs of $\mathcal{G}$. The effective dimension of the state space of $\{Y_t\}$ is three (instead of six), since the total number of instances of each of the proteins A, B and C does not change over time, so for each $\mathcal{G} \in \mathbb{G}$, the value $\phi_1(\mathcal{G})$ uniquely determines the number of the

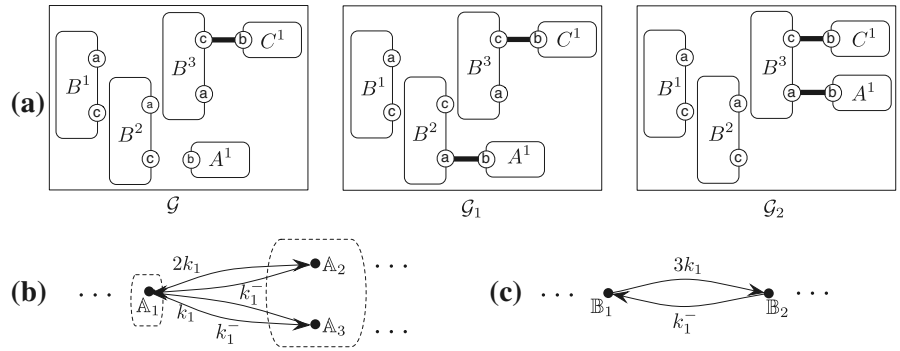

Fig. 3 Interpreting the case study 1 (simple scaffold). **a** Examples of reaction mixtures—$\mathcal{G}$, $\mathcal{G}_1$ and $\mathcal{G}_2$; **b** a part of the CTMC $Y_t \in \{\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3, \ldots\}$, such that $\mathcal{G} \in \mathbb{A}_1$, $\mathcal{G}_1 \in \mathbb{A}_2$, $\mathcal{G}_2 \in \mathbb{A}_3$; **c** a part of the CTMC $Z_t \in \{\mathbb{B}_1, \mathbb{B}_2, \ldots\}$, such that $\mathcal{G} \in \mathbb{B}_1$, $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{B}_2$. The state $\mathbb{B}_2$ is lumping of states $\mathbb{A}_2$ and $\mathbb{A}_3$. Then, by Theorem 9, we have that $\mathsf{P}(Z_t = \mathbb{B}_2) = \mathsf{P}(Y_t \in \{\mathbb{A}_2, \mathbb{A}_3\})$ (lumpability), and $\mathsf{P}(Y_t = \mathbb{A}_2) = 2/3\mathsf{P}(Z_t = \mathbb{B}_2)$, $\mathsf{P}(Y_t = \mathbb{A}_3) = 1/3\mathsf{P}(Z_t = \mathbb{B}_2)$, whenever $\mathsf{P}(Y_0 = \mathbb{A}_2) = 2\mathsf{P}(Y_0 = \mathbb{A}_3)$ (invertability). Moreover, by Theorem 11, $\mathsf{P}(Y_t = \mathbb{A}_2) \to 2\mathsf{P}(Y_t = \mathbb{A}_3)$, as $t \to \infty$ (convergence)

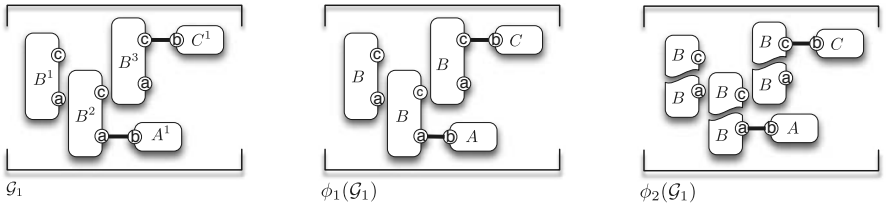

Fig. 4 Case study 1: simple scaffold. A reaction mixture $\mathcal{G}_1$ (*left*) and its graphical representation in aggregation ϕ_1 (*center*) or ϕ_2 (*right*)

remaining species S_A, S_B and S_C (for example, in Fig. 3, $\phi_1(\mathcal{G}_1) \neq \phi_1(\mathcal{G}_2)$, because $\mathcal{G}_1$ counts zero species S_{ABC}).

Counting fragments. Since the values of sites a and c of nodes of type B are updated without testing each-other, let the relation $\sim_2 \subseteq \mathbb{G} \times \mathbb{G}$ lump all reaction mixtures that count the same number of free sites c and the same number of free sites a:

$$\mathcal{G} \sim_2 \mathcal{G}' \text{ iff } \phi_2(\mathcal{G}) = \phi_2(\mathcal{G}'),$$

where $\phi_2 : \mathbb{G} \to \mathbb{N}^2$ is such that $\phi_2(\mathcal{G}) = (m_{AB*}, m_{*BC})$, if $\mathcal{G} \in \mathbb{G}$ has m_{AB*} nodes B bound to A and m_{*BC} nodes B that bound to C. In related literature, the connected site-graph patterns that are used for determining aggregated states, are called *fragments*. We do not introduce fragments (nor species) formally, since we do not aim at a general procedure for finding the set of fragments for a given rule-set. For example, in Fig. 3, $\phi_2(\mathcal{G}_1) = \phi_2(\mathcal{G}_2)$. The aggregation of $\{X_t\}$ by ϕ_2 results in a CTMC $\{Z_t\}$, which takes values in $\mathbb{N}^2$, and it therefore provides a better reduction than $\{Y_t\}$. A way to visualize the states of CTMC's $\{X_t\}$, $\{Y_t\}$ and $\{Z_t\}$ is shown in Fig. 4.

It can be inspected that the process $\{X_t\}$ satisfies both (Cond1) and (Cond2) with respect to aggregation with $\sim_1$. For example, if $\{Y_t\}$ starts in the state, say, $x_0 = (1, 3, 1, 0, 0, 0)$, then the process $\{X_t\}$ is in the unique corresponding site-graph

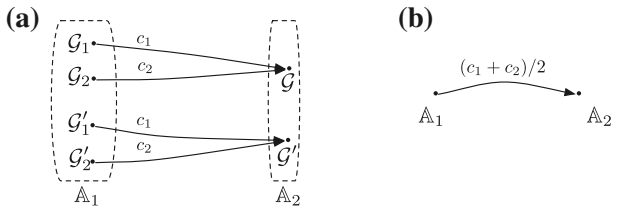

Fig. 5 Illustration for testing (Cond3) and its relation to (Cond2): Let $\mathbb{A}_1 = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}'_1, \mathcal{G}'_2\}$, $\mathbb{A}_2 = \{\mathcal{G}, \mathcal{G}'\}$. For $\mathcal{G}, \mathcal{G}' \in \mathbb{A}_2$, the permutation $\sigma(\mathcal{G}_1) = \mathcal{G}'_1$, $\sigma(\mathcal{G}'_1) = \mathcal{G}_1$, $\sigma(\mathcal{G}_2) = \mathcal{G}'_2$ and $\sigma(\mathcal{G}'_2) = \mathcal{G}_2$ proves that the predecessors of $\mathcal{G}$ and those of $\mathcal{G}'$ via application of rule with rate c_1 (respectively c_2), and inside class $\mathbb{A}_1$, are in bijection, that is, (Cond3) holds. (Cond2) follows, since $\tilde{Q}(\mathbb{A}_1, \mathbb{A}_2) = \delta(\mathbb{A}_1, \mathcal{G}) = \delta(\mathbb{A}_2, \mathcal{G}')$, which is because $Q(\mathcal{G}_1, \mathcal{G}) + Q(\mathcal{G}_2, \mathcal{G}) = Q(\sigma(\mathcal{G}_1), \mathcal{G}') + Q(\sigma(\mathcal{G}_2), \mathcal{G}') = c_1 + c_2$, and the rate in the aggregated chain is $Q(\mathbb{A}_i, \mathbb{A}_j) = \frac{|\mathbb{A}_j|}{|\mathbb{A}_i|}(c_1 + c_2) = \frac{1}{2}(c_1 + c_2)$

representation of x_0. Upon applying the rule, say R_1, in $\{X_t\}$, three different rule-applications will occur with equal probability, and they lead to three different states (reaction mixtures) in $\{X_t\}$, but to the same state in $\{Y_t\}$, $x = (0, 2, 1, 1, 0, 0)$. By proving the same argument for all rules and all states reachable from the initial state, whenever $\{Y_t\}$ is in the state $x = (0, 2, 1, 1, 0, 0)$, the process $\{X_t\}$ is in either of the three reaction mixtures (corresponding site-graph representations of x) with equal probability.

However, this observation brings no additional insight with respect to the usual stochastic modeling, because the aggregated process $\{Y_t\}$ coincides with the population-based model given by classical chemical kinetics (as introduced in Sect. 4.1), and the assumption of spatial homogeneity and mass-action are trivially equivalent to the consequences of Theorem 9 and Theorem 11. The better aggregation, $\sim_2$ is a result of the procedure outlined in Feret et al. (2012, 2013), and the process $\{X_t\}$ satisfies both (Cond1) and (Cond2) with respect to aggregation with $\sim_2$. The resulting process $\{Z_t\}$ is also termed *fragment-based*, because instead of species' populations, fragments' populations are considered instead. Even though the proof is a consequence of the works in Feret et al. (2012, 2013), for the sake of intuition, we next sketch an algorithmic criterion for testing a particular case of (Cond2), when the distributions over aggregates are uniform. The criterion's validity is obvious from (7). An illustration is given in Fig. 5.

Lemma 4 *Let $\tilde{\mathbb{G}} = \{\mathbb{A}_1, \ldots, \mathbb{A}_n\}$ be a partitioning of $\mathbb{G}$ induced by an equivalence relation $\sim \subseteq \mathbb{G} \times \mathbb{G}$. Let α_i be the uniform probability measure on $\mathbb{A}_i$, that is, for any $\mathcal{G} \in \mathbb{A}_i$, $\alpha_i(\mathcal{G}) = |\mathbb{A}_i|^{-1}$. Note that in this case* (1) *reduces to,*

$$\Delta(\mathbb{A}_i, \mathcal{G}) = \frac{|\mathbb{A}_j|}{|\mathbb{A}_i|} \sum_{\mathcal{G}_1 \in \mathbb{A}_i} Q(\mathcal{G}_1, \mathcal{G}), \quad \mathcal{G} \in \mathbb{A}_j. \tag{7}$$

Then, the following condition implies (Cond2):

(Cond3) For all $\mathbb{A}_i, \mathbb{A}_j \in \tilde{\mathbb{G}}$, there exists a permutation of states in $\mathbb{A}_i$, $\sigma : \mathbb{A}_i \to \mathbb{A}_i$, such that for any pair $\mathcal{G}, \mathcal{G}' \in \mathbb{A}_j$, $Q(\mathcal{G}_1, \mathcal{G}) = Q(\sigma(\mathcal{G}_1), \mathcal{G}')$ (Fig. 6).

Definition 11 If the equivalence relation $\sim \subseteq \mathbb{G} \times \mathbb{G}$ satisfies (Cond3) and for each $i = 1, \ldots, m$, α_i is a uniform probability measure on $\mathbb{A}_i$, then the corresponding Markov chain $\{Y_t\}$ (with generator matrix $\tilde{Q}(\mathbb{A}_i, \mathbb{A}_j) \equiv \Delta(\mathbb{A}_i, \mathcal{G}), \mathcal{G} \in \mathbb{A}_j$) is a **uniform aggregation** of $\{X_t\}$.

Notice that the uniform aggregation implies that the condition (Cond2) holds for uniform α_i. In that sense, uniform aggregation implies (Cond2) (weak lumpability). On the other hand, the condition (Cond1) is neither a sufficient, nor necessary condition for (Cond2), as shown in Feret et al. (2012), and it does not relate to uniform aggregation either.

By the outlined criterion, it can be shown that $\{Z_t\}$ and $\{Y_t\}$ are uniform aggregations of $\{X_t\}$. If $\{Z_t\}$ is an aggregation of $\{Y_t\}$ that satisfies (Cond2) (not necessarily uniform), the modeler can retrieve the conditional distribution of $\{Y_t\}$ given $\{Z_t\}$. This is possible by the following result.

Theorem 12 *Let $\sim_1$ and $\sim_2$ be two equivalence relations of $\mathbb{G}$, such that $\mathbb{G}_1 = \{\mathbb{A}_1, \mathbb{A}_2, \ldots\}$ and $\mathbb{G}_2 = \{\mathbb{B}_1, \mathbb{B}_2, \ldots\}$ are the corresponding sets of equivalence classes. If $\sim_1$ is coarser than $\sim_2$ (that is, $\sim_1 \subseteq \sim_2$), then $\mathbb{G}_2$ can be obtained by partitioning $\mathbb{G}_1$ as follows.*

$$\mathbb{A}_i \sim \mathbb{A}_j \text{ iff there exist } \mathcal{G} \in \mathbb{A}_i, \mathcal{G}' \in \mathbb{A}_j, \text{ such that } \mathcal{G} \sim_2 \mathcal{G}'.$$

Equivalently,

$$\mathbb{A}_i \sim \mathbb{A}_j \text{ iff there exists } \mathbb{B}_k \text{ such that } \mathbb{A}_i \cup \mathbb{A}_j \subset \mathbb{B}_k. \tag{8}$$

Assume that $\{Y_t\}$ and $\{Z_t\}$ with generator matrices Q_1 and Q_2 are two uniform aggregations of the Markov chain $\{X_t\}$ induced by $(\sim_1, \{\alpha_i\})$ and $(\sim_2, \{\beta_i\})$, where α_i and β_j are uniform over $\mathbb{A}_i$ and $\mathbb{B}_j$ respectively. Define

$$\alpha'_j(\mathbb{A}_i) := \begin{cases} \frac{|\mathbb{A}_i|}{|\mathbb{B}_j|}, & \text{if } \mathbb{A}_i \subseteq \mathbb{B}_j \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Then $\{\alpha'_j\}$ satisfies (Cond2) and hence $\{Y_t\}$ is an aggregation of the Markov chain $\{Z_t\}$.

Proof It is trivial to check that $\sim$ defined by (8) is a well-defined equivalence relation.

Assume now that $\mathbb{B}_j, \mathbb{B}_{j'} \in \mathbb{G}_2$ and $\mathbb{A}_{i'} \subseteq \mathbb{B}_{j'}$. We have to show that $\Delta(\mathbb{B}_j, \mathbb{A}_{i'})$ is constant for all $\mathbb{A}_{i'} \subseteq \mathbb{B}_{j'}$. Toward this end notice that

$$\begin{aligned}
\Delta(\mathbb{B}_j, \mathbb{A}_{i'}) &= \frac{\sum_{i:\mathbb{A}_i \subseteq \mathbb{B}_j} \alpha_j(\mathbb{A}_i) Q_1(\mathbb{A}_i, \mathbb{A}_{i'})}{\alpha_j(\mathbb{A}_{i'})} = \frac{\sum_{i:\mathbb{A}_i \subseteq \mathbb{B}_j} |\mathbb{A}_i|/|\mathbb{B}_j| Q_1(\mathbb{A}_i, \mathbb{A}_{i'})}{|\mathbb{A}_{i'}|/|\mathbb{B}_{j'}|} \\
&= \frac{\sum_{\mathbb{A}_i \subseteq \mathbb{B}_j} |\mathbb{A}_i|/|\mathbb{B}_j| \sum_{\mathcal{G}' \in \mathbb{A}_i} Q(\mathcal{G}', \mathcal{G}) |\mathbb{A}_{i'}|/|\mathbb{A}_i|}{|\mathbb{A}_{i'}|/|\mathbb{B}_{j'}|}, \quad \text{for some } \mathcal{G} \in \mathbb{A}_{i'} \\
&= \frac{\sum_{\mathbb{A}_i \subseteq \mathbb{B}_j} \sum_{\mathcal{G}' \in \mathbb{A}_i} Q(\mathcal{G}', \mathcal{G}) |\mathbb{A}_i|/|\mathbb{B}_j| |\mathbb{A}'_i|/|\mathbb{A}_i|}{|\mathbb{A}_{i'}|/|\mathbb{B}_{j'}|} \\
&= \sum_{\mathcal{G}' \in \mathbb{B}_j} Q(\mathcal{G}', \mathcal{G}) |\mathbb{B}_{j'}|/|\mathbb{B}_j| \\
&= Q_2(\mathbb{B}_j, \mathbb{B}_{j'}).
\end{aligned}$$

Here the third and the last equalities are because by the assumption $\{Y_t\}$ and $\{Z_t\}$ are uniform aggregations of $\{X_t\}$. □

Example 4 (Simple scaffold: Cont'd) We show that both relations $\sim_1$ and $\sim_2$ induce uniform aggregations of $\{X_t\}$. Moreover, $\sim_1 \subseteq \sim_2$, that is, $\sim_2$ is coarser than $\sim_1$. Consider lumping by $\sim_2$. Let $\mathcal{G}_1, \mathcal{G}_2$ be two reaction mixtures such that $\mathcal{G}_1 \sim_2 \mathcal{G}_2$, and let $\phi_2(\mathcal{G}_1) = \phi_2(\mathcal{G}_2) = (m_{AB*}, m_{*BC})$. If $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{B}_j$, by Theorem 12, it is enough to show that for any $\mathbb{B}_i \in \mathbb{G}_2$, and any $\mathcal{G} \in \mathbb{B}_i$, there is a permutation $\sigma : \mathbb{B}_i \to \mathbb{B}_i$, such that $Q(\mathcal{G}, \mathcal{G}_1) = Q(\sigma(\mathcal{G}), \mathcal{G}_2)$. Choose some $\mathbb{B}_i \in \mathbb{G}_2$ and $\mathcal{G} \in \mathbb{B}_i$. Then, $\phi_2(\mathcal{G}) \in \{(m_{AB*} - 1, m_{*BC}), (m_{AB*} + 1, m_{*BC}), (m_{AB*}, m_{*BC} + 1), (m_{AB*}, m_{*BC} - 1)\}$. We analyze the case $\phi_2(\mathcal{G}) = (m_{AB*} - 1, m_{*BC})$; the other three cases are analogous. Let $\mathcal{G}_1 = (\mathcal{V}, \hat{\Sigma}, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}, \hat{\Sigma}, \mathcal{E}_2)$. Since $\phi_2(\mathcal{G}_1) = \phi_2(\mathcal{G}_2)$, there exists a bijective renaming function $\eta : \mathcal{V} \to \mathcal{V}$, such that $\mathcal{G}_2 = \mathcal{G}_1^\eta$, that is, $\mathcal{G}_2$ is η-induced node-renamed site-graph $\mathcal{G}_1$. It is easy to inspect that $Q(\mathcal{G}, \mathcal{G}_1) = c_1$ if and only if $Q(\mathcal{G}^\eta, \mathcal{G}_2) = c_1$. So, the bijection over the reaction mixtures aggregated to $\mathbb{B}_i$ is the one induced by renaming η. For showing that $\sim_2$ is coarser than $\sim_1$, it is enough to observe that the map $\phi : \mathbb{N}^3 \to \mathbb{N}^2$ defined by $\phi(m_{AB}, m_{BC}, m_{ABC}) = (m_{AB} + m_{ABC}, m_{BC} + m_{ABC})$ is such that $\phi_2 = \phi \circ \phi_1$.

Consequently, Theorem 12 applies. Then, the process $\{Z_t\}$ is also lumpable with respect to $\{Y_t\}$, and Theorem 9 applies. The convergence result is applicable in the following situation. Imagine that it is possible to experimentally synthesize only the complexes of type (AB) and of (BC), but not a complex of type (A), (B), (C) or (ABC). Then, the initial distribution does not respect α_i, as soon as $n_A \geq 1$, $n_B \geq 2$, $n_C \geq 1$. However, since each reversible rule-based model trivially has an irreducible CTMC, the Theorem 11 holds.

A concrete example is demonstrated in Fig. 3. The details for the calculation for Table 1, de-aggregation, as well the discussion for $n_A = n_C = 1$, $n_B = 2$ can be found in the "Appendix".

5.1 Case study 2: Two-sided polymerization

The two-sided polymerization case study illustrates the drastic advantage of using the fragment-based CTMC, because it shows to have exponentially smaller state space than the species-based CTMC.

Consider a site-graph-rewrite model $\mathcal{R}$ depicted in Fig. 6: proteins A and B can polymerize by forming bonds of two kinds: between site b of protein A and site a of protein B, or between site r of protein A and site l of protein B. Assume that there are n_A nodes of type A and n_B nodes of type B. Let $\mathbb{G}$ be the set of all reaction mixtures. All connected site-graphs occurring in a reaction mixture can be categorized into two types: *chains* and *rings*. *Chains* are the connected site-graphs having two free sites, and *rings* are those having no free sites. We say that a chain or a ring is of length i if it has i bonds in total. Chains can be classified into four different kinds, depending on which sites are free.

Species. Let $\phi_1 : \mathbb{G} \to \mathbb{N}^{5m}$ be such that

$$\phi_1(\mathcal{G}) = (x_{11}, \ldots, x_{1m}, x_{21}, \ldots, x_{2m}, x_{31}, \ldots, x_{3m}, x_{41}, \ldots, x_{4m}, x_{51}, \ldots, x_{5m}),$$

if $\mathcal{G} \in \mathbb{G}$ has

x_{1i} chains of type $(A..B)_i$, that is, of length $2i - 1$, with free sites b and a,
x_{2i} chains of type $(B..A)_i$, that is, of length $2i - 1$, with free sites l and r,
x_{3i} chains of type $(A..A)_i$, that is, of length $2i$, with free sites b and a,
x_{4i} chains of type $(B..B)_i$, that is, of length $2i$, with free sites l and r,
x_{5i} rings of type $(.A..B.)_i$, that is, of length $2i$.

The two states $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are aggregated by the equivalence relation $\sim_1 \subseteq S \times S$ if $\phi_1(\mathcal{G}) = \phi_1(\tilde{\mathcal{G}})$.

Fragments. Let $\phi_2 : \mathbb{G} \to \mathbb{N}^2$ be such that $\phi_2(\mathcal{G}) = (m_{rl}, m_{ba})$, if $\mathcal{G} \in \mathbb{G}$ has m_{rl} bonds between sites r and l, and m_{ba} bonds between sites b and a. The two states $\mathcal{G}$ and $\mathcal{G}'$ are aggregated by the equivalence relation $\sim_2 \in \mathbb{G} \times \mathbb{G}$ if $\phi_2(\mathcal{G}) = \phi_2(\mathcal{G}')$. Alternatively, since the rates of forming and releasing bonds do not depend on the type of the bond, let $\phi_3 : \mathbb{G} \to \mathbb{N}$ be such that $\phi_3(\mathcal{G}) = m$, if $\mathcal{G} \in \mathbb{G}$ has in total m bonds. The two states $\mathcal{G}$ and $\mathcal{G}'$ be aggregated by equivalence relation $\sim_3 \in \mathbb{G} \times \mathbb{G}$ if $\phi_3(\mathcal{G}) = \phi_3(\mathcal{G}')$.

A concrete example is demonstrated in Fig. 7. The details for the calculation for Table 1, and on de-aggregation can be found in the Appendix.

5.2 Case study 3: EGF/insulin pathway

We take a model of the network of interplay between insulin and epidermal growth factor (EGF) signaling in mammalian cells from literature (Conzelmann et al. 2008). The original model suffers from the huge number of feasible multi-protein species and the high complexity of the related reaction networks. It contains 42956 reactions and 2768 different molecular species, i.e. connected reaction mixtures which differ up to node identifiers. The reactions can be translated into a Kappa model of only 38 transition rules.

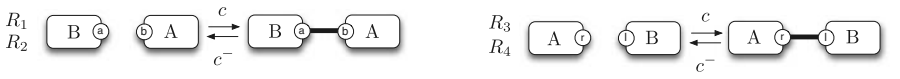

Fig. 6 Case study 2: two-sided polymerization

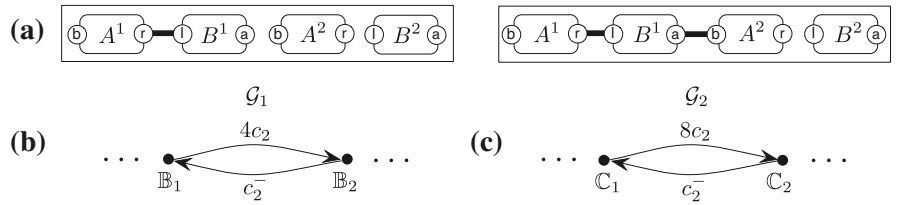

Fig. 7 Case study 2: two-sided polymerization. **a** Examples of reaction mixtures; **b** a part of the CTMC $Z_t \in \{\mathbb{B}_1, \mathbb{B}_2, \ldots\}$, such that $\mathcal{G}_1 \in \mathbb{B}_1$, $\mathcal{G}_2 \in \mathbb{B}_2$, **c** a part of the CTMC $Z'_t \in \{\mathbb{C}_1, \mathbb{C}_2, \ldots\}$, such that $\mathcal{G}_1 \in \mathbb{C}_1$, $\mathcal{G}_2 \in \mathbb{C}_2$

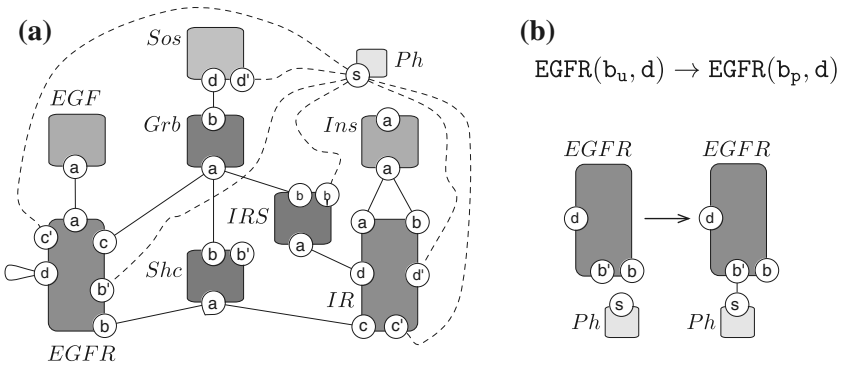

Fig. 8 **a** Summary of interactions between nodes in case study 3. The *dotted lines* represent phosphorylation, and *solid lines* denote standard bindings. The self-loop at the site d of node $EGFR$ means that it can bind to another node $EGFR$, i.e. receptor dimerization. **b** An example of a Kappa rule, and a corresponding site-graph rewrite rule

The bases for the framework of site-graph-rewrite models used in this paper is a rule-based modeling language Kappa (Danos and Laneve 2003). A Kappa rule and an example of the corresponding site-graph-rewrite rule are shown in Fig. 8b. The general differences to Kappa are detailed in Sect. 6. In Fig. 8a, we show the summary of protein interactions for this model, adapted to the site-graph-rewrite formalism used in this paper. The procedure in Feret et al. (2012, 2013) proves that it is enough to track the copy number of 609 partially defined complexes, that are named *fragments*. Intuitively, this is due to the independent updates of values of sites a and b of protein Grb. Consequently, the dimension of the state vector in the reduced system is 609, instead of 2768 in the concrete system.

Species. Two reaction mixtures $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are aggregated by relation $\sim_1 \subseteq \mathbb{G} \times \mathbb{G}$ if they contain the same number of molecular species.

Fragments. Let a fragment be a part of a molecular species that either does not contain protein Grb, or it contains only a site a of protein Grb, or it contains only

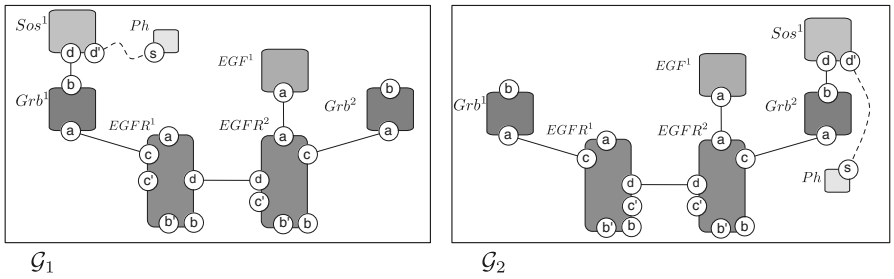

Fig. 9 Case study 3: reaction mixtures $\mathcal{G}_1$, $\mathcal{G}_2$, such that they are aggregated in the fragment description—both states contain one protein Grb that is free on site b, one protein Grb that is bound to a site d of protein Sos, and one species containing a dimer of $EGFR$ proteins, such that each $EGFR$ protein is bound to one Grb protein, and one of them is bound to an EGF protein. Let $\mathcal{G}_1 \in \mathbb{A}_1 \subseteq \mathbb{B}_1$ and $\mathcal{G}_2 \in \mathbb{A}_2 \subseteq \mathbb{B}_1$. Then, by Theorem 9, we have that $\mathsf{P}(Z_t = \mathbb{B}_1) = \mathsf{P}(Y_t \in \{\mathbb{A}_1, \mathbb{A}_2\})$ (lumpability), and $\mathsf{P}(Y_t = \mathbb{A}_1) = 0.5\mathsf{P}(Z_t = \mathbb{B}_1)$ whenever $\mathsf{P}(Y_0 = \mathbb{A}_1) = \mathsf{P}(Y_0 = \mathbb{A}_2)$ (invertability). Moreover, by Theorem 11, $\mathsf{P}(Y_t = \mathbb{A}_1) \to 0.5\mathsf{P}(Z_t = \mathbb{B}_1)$, when $t \to \infty$ (convergence)

a site b of protein Grb. Two reaction mixtures $\mathcal{G}$ and $\mathcal{G}'$ are aggregated by relation $\sim_2 \subseteq \mathbb{G} \times \mathbb{G}$ if they contain the same number of fragments.
A concrete example is demonstrated in Fig. 9.

6 Conclusion

In this paper, we have studied model reduction for a Markov chain using aggregation techniques. We provided a sufficient condition for defining a CTMC over the aggregates, a *lumpable* reduction of the original one. Moreover, we characterized sufficient conditions for *invertability*, that is, when the measure over the original process can be recovered from that of the aggregated one. We also established *convergence* properties of the aggregated process and showed how lumpability and invertability depend on the initial distribution.

Three case studies demonstrated the usefulness of the techniques discussed in the paper. The application is inspired by the works in Feret et al. (2012, 2013): when the model of protein interactions is given by a biologist in form of site-graph-rewrite rules, then the usual, species-based CTMC model can be avoided, and the smaller, aggregated CTMC can be analyzed directly. In general, the less dependency there is between the modifications of proteins' domains, the more states are potentially aggregated by the method of detecting fragments. The applicability of the method on real case studies depends on the level of detail incorporated in the model. For small models focusing on a single signaling unit (for example, MAPK cascade), there is typically no reduction. However, on larger, more realistic case studies, such as the EGF/insulin crosstalk presented in this paper, the reduction can be significant. This is because the larger, more detailed case studies, contain proteins with many domains, and these domains can be exploited for different signaling pathways, independently of each-other. The dimension reduction renders the computational analysis more efficient: a small enough aggregated CTMC may be amenable to integrating the chemical master equation, or,

in other cases, the simulation of the population model becomes faster. To that end, an interesting topic for future work is how simulating the rule-based model over fragments' populations instead of over species' populations can be used for efficient estimation of kinetic parameters of a rule-based model.

Acknowledgments A. Ganguly and H. Koeppl acknowledge the support from the Swiss National Science Foundation, grant number PP00P2 128503/1. T. Petrov is supported by SystemsX.ch—the Swiss Initiative for Systems Biology. The authors would like to thank Prof. James Ledoux for his useful comments and for bringing to their attention some of the general work done in the area.

Appendix

De-aggregation: simple scaffold

Assume that $\mathcal{G} \in \mathbb{G}$ is such that $\phi_1(\mathcal{G}) = (m_{AB}, m_{BC}, m_{ABC})$. Let $m_A := n_A - m_{AB} - m_{ABC}$, $m_B := n_A - m_{AB} - m_{BC} - m_{ABC}$ and $m_C := n_C - m_{BC} - m_{ABC}$. If $\mathcal{G} \in \mathbb{A}_i$, then $\alpha_{1i}(\mathcal{G}) = |\mathbb{A}_i|^{-1}$, where

$$|\mathbb{A}_i| = \frac{n_A! n_B! n_C!}{m_{AB}! m_{BC}! m_{ABC}! m_A! m_B! m_C!}. \tag{10}$$

The explanation is as follows. The m_A free nodes of type A, m_B free nodes of type B and m_C free nodes of type C can be chosen in $\binom{n_A}{m_A}\binom{n_B}{m_B}\binom{n_C}{m_C}$ possible ways. Among the remaining nodes, m_{AB} nodes of type A and m_{AB} nodes of type B can be chosen in $\binom{n_A - m_A}{m_{AB}}\binom{n_B - m_B}{m_{AB}}$ ways. There are $m_{AB}!$ different ways to establish bonds between m_{AB} identified nodes A and m_{AB} identified nodes B. In the same way, we choose m_{BC} complexes of type (BC) among the $n_B - m_B - m_{AB}$ nodes of type A, and $n_C - m_C$ nodes of type B. Finally, there is exactly one way to choose m_{ABC} complexes of type (ABC) among the $n_A - m_A - m_{AB}$, $n_B - m_B - m_{AB} - m_{BC}$ and $n_C - m_C - m_{BC}$ nodes of type A, B and C respectively. Connecting the bonds can be done in $(m_{ABC}!)^2$ different ways (for each node B^j, there are exactly $m_{ABC}!$ ways to choose the A^i and $m_{ABC}!$ ways to choose C^k). The final expression follows.

Moreover, if $\phi_2(\mathcal{G}) = (m_{AB*}, m_{*BC})$ and $\mathcal{G} \in \mathbb{B}_j$, then $\alpha_{2j}(\mathcal{G}) = |\mathbb{B}_j|^{-1}$, where

$$|\mathbb{B}_j| = \binom{n_A}{m_{AB*}}\binom{n_B}{m_{AB*}} m_{AB*}! \binom{n_C}{m_{*BC}}\binom{n_B}{m_{*BC}} m_{*BC}!. \tag{11}$$

We first choose the m_{AB*} nodes of type A and m_{AB*} nodes of type B; There are $m_{AB*}!$ different ways to establish the bonds; In total, it makes $\binom{n_A}{m_{AB*}}\binom{n_B}{m_{AB*}} m_{AB*}!$ choices. Independently, the m_{*BC} bonds between B and C can be chosen in $\binom{n_B}{m_{*BC}}\binom{n_C}{m_{*BC}} m_{*BC}!$ ways.

De-aggregation: two-sided polymerization

Assume that s is a site-graph such that

$$\phi_1(s) = (x_{11}, \ldots, x_{1m}, x_{21}, \ldots, x_{2m}, x_{31}, \ldots, x_{3m}, x_{41}, \ldots, x_{4m}, x_{51}, \ldots, x_{5m}).$$

We do not give the analytic expression for $\alpha_{1i}(s)$. For computing it, it is enough to use the following:

- choosing a chain of type $(A..B)_i$ among m_A nodes A and m_B nodes B can be done in $f_1(m_A, m_B, i) = \binom{m_A}{i}\binom{m_B}{i}(i!)^2$ ways; there are $(m_A - i)$ nodes A, and $(m_B - i)$ nodes B left. The same is used for choosing a chain of type $(B..A)_i$;
- choosing a chain of type $(A..A)_i$ among m_A nodes A and m_B nodes B can be done in $f_2(m_A, m_B, i) = \binom{m_A}{i}\binom{m_B}{i-1}i!(i-1)!$ ways; there are $(m_A - i)$ nodes A, and $(m_B - (i-1))$ nodes B left. The same is used for choosing a chain of type $(B..B)_i$;
- choosing a chain of type $(.A..B.)_i$ among m_A nodes A and m_B nodes B can be done in $f_3(m_A, m_B, i) = \binom{m_A}{i}\binom{m_B}{i}(i!)^2/i$ ways; there are $(m_A - i)$ nodes A, and $(m_B - i)$ nodes B left. Division by i is done because of symmetries—every ring of type $(.A..B.)_i$ is determined by choosing i nodes of type A, i nodes of type B, ordering nodes A in one of $i!$ ways, ordering nodes B in one of $i!$ ways, but every ordering $(A_{j1} - B_{k1} - A_{j2} - B_{k2} - \ldots A_{ji} - B_{ki})$ defines the same ring as $(A_{j2} - B_{k2} - A_{j3} - B_{k3} - \ldots A_{j1} - B_{k1})$ etc. (i of them in total).

  Moreover, if s is such that $\phi_2(s) = (m_{rl}, m_{ba})$, then

$$\alpha_{2i}(s) = \binom{n}{m_{rl}}^2 m_{rl}! \binom{n}{m_{ba}}^2 m_{ba}!.$$

If s is such that $\phi_2(s) = m$, then

$$\alpha_{3i}(s) = \sum_{i=0}^{m} \binom{n}{i}^2 i! \binom{n}{m-i}^2 (m-i)!.$$

We choose m_{rl} nodes of type A among n of them, and the same number of nodes of type B. There is $m_{rl}!$ different ways to connect them. We independently choose the m_{ba} bonds in the same way.

To compute $\alpha_{3i}(s)$, since all of the m bonds can be either of type m_{rl} or m_{ba}, we choose i bonds of type m_{rl} and $(m - i)$ bonds of type m_{ba}, for $i = 0, \ldots, m$.

Figure 3

The CTMC $\{X_t\}$, for given one node A, three nodes B and one node C contains different reaction mixtures over the set of nodes $\{A^1, B^1, B^2, B^3, C^1\}$. For example, let $\mathcal{G}$ be the reaction mixture with the set of edges $\{\{(A^1, b), (B^3, a)\}\}$. There are three ways to apply the rule R_2 on $\mathcal{G}$: by embedding via function $\eta_1 = \begin{pmatrix} B & C \\ B^1 & C^1 \end{pmatrix}$, $\eta_2 = \begin{pmatrix} B & C \\ B^2 & C^1 \end{pmatrix}$, or $\eta_3 = \begin{pmatrix} B & C \\ B^3 & C^1 \end{pmatrix}$. If $\mathcal{G}_1$ is a mixture with a set of

edges $\{\{(B^3, a), (A^1, b)\}, \{(B^2, c), (C^1, b)\}\}$ and $\mathcal{G}_2$ is a mixture with a set of edges $\{\{(B^3, a), (A^1, b)\}, \{(B^3, c), (C^1, b)\}\}$, then $Q(\mathcal{G}, \mathcal{G}_1) = Q(\mathcal{G}, \mathcal{G}_2) = c_2$.

Note that $\phi_1(\mathcal{G}) = (1, 0, 0)$, $\phi_1(\mathcal{G}_1) = (1, 1, 0)$, $\phi_1(\mathcal{G}_2) = (0, 0, 1)$. Let $\mathcal{G} \in \mathbb{A}_1$, $\mathcal{G}_1 \in \mathbb{A}_2$, $\mathcal{G}_2 \in \mathbb{A}_3$. By applying the Equation (10), we have $\alpha_{11}(\mathcal{G}) = (\frac{1!3!1!}{1!0!0!0!2!1!})^{-1}$=1/3, $\alpha_{12}(\mathcal{G}_1) = (\frac{1!3!1!}{1!0!1!0!2!0!})^{-1} = 1/3$, and $\alpha_{13}(\mathcal{G}_2) = (\frac{1!3!1!}{1!1!0!0!1!0!})^{-1} = 1/6$.

Moreover, since $\phi_2(\mathcal{G}) = (1, 0)$, and $\phi_2(\mathcal{G}_1) = \phi_2(\mathcal{G}_2) = (1, 1)$, let $\mathbb{B}_1, \mathbb{B}_2 \in \mathbb{G}_2$ be such that $\mathcal{G} \in \mathbb{B}_1$ and $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{B}_2$. Then, $\alpha_{21}(\mathcal{G}) = (\binom{1}{1}\binom{3}{1}1!\binom{1}{0}\binom{3}{0}0!)^{-1} = 1/3$ and $\alpha_{22}(\mathcal{G}_1) = \alpha_{22}(s_2) = (\binom{1}{1}\binom{3}{1}1!\binom{1}{1}\binom{3}{1}1!)^{-1} = 1/9$.

Finally, observing the aggregation from $\mathbb{G}_1$ to $\mathbb{G}_2$, we have that $\alpha_1(\mathbb{A}_1) = \frac{\alpha_{21}(\mathcal{G})}{\alpha_{11}(\mathcal{G})} = 1$, $\alpha_2(\mathbb{A}_2) = \frac{\alpha_{22}(\mathcal{G}_1)}{\alpha_{12}(\mathcal{G}_1)} = 1/3$, and $\alpha_2(\mathbb{A}_3) = \frac{\alpha_{22}(\mathcal{G}_2)}{\alpha_{13}(\mathcal{G}_2)} = 2/3$.

Table 1

In order to illustrate how powerful the presented reduction method is in comparison to the standard, species-based models, we compare the size of the state space in the species-based model, $\mathbb{G}_1$, and in the fragment-based model, $\mathbb{G}_2$.

Simple scaffold. The size of $\mathbb{G}_2$ is $(n + 1)^2$: there are $n + 1$ possible situations between A and B nodes with $0, 1, \ldots, n$ bonds between them. The same holds for possible configurations between nodes of type B and C. Let $f(k)$ denote the number of states with k copies of each of the nodes A, B and C, and with no complexes of type (ABC). If there is $0 \leq i \leq k$ complexes of type (AB), there can be $0 \leq j \leq (k - i)$ complexes of type (BC), and we thus have $f(k) = \sum_{i=0}^{k}(k - i + 1) = \frac{(k+1)(k+2)}{2}$. The number of complexes of type (ABC) can vary from 0 to n, and thus we have the total number of states in $\mathbb{G}_1$ to be $\sum_{k=0}^{n} f(k) = \frac{1}{2}\sum_{k=0}^{n}(k^2 + 3k + 2) = \frac{1}{2}(\sum_{k=0}^{n} k^2 + 3\sum_{k=0}^{n} k + 2\sum_{k=0}^{n} 1) = \frac{1}{2}(n(n+1)(2n+1)/6 + 3n(n+1)/2 + 2(n+1)) = (n+1)(n+2)(n+3)/6$.

Two-sided polymerization. We first estimate the size of $\mathbb{G}_2$. The value of m_{rl} varies between 0 and n, and the same holds for the value of m_{ba}. Each state $(i, j) \in \{0, \ldots, n\} \times \{0, \ldots, n\}$ is reachable, since the bonds are created independently of each-other. The size of the state space $\mathbb{G}_2$ is thus $(n+1)^2$. The size of $\mathbb{G}_2$ is $2n+1$, because the value of m varies between 0 and $2n$. Let $P(n)$ denote the number of partitions of number n—number of ways of writing n as a sum of positive integers. One of the well-known asymptotics is $P(n) \approx \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}$ (Hardy and Ramanujan 1918). Consider one partition $n = n_1 + \cdots + n_k$, $n_1 \leq \cdots \leq n_k$, and a state $s_1 \in \mathbb{G}_1$ that counts one chain of type $(A..B)_{n_1}$, one chain of type $(A..B)_{n_2}$ etc. It is in $\mathbb{G}_1$, because it has exactly n nodes A and n nodes B. Therefore, the set $\mathbb{G}_1$ counts at least $P(n)$ states. This approximation can be improved by factor three: think of the states $\mathcal{G}_2$ and $\mathcal{G}_3$, which are constructed of chains of type $(B..A)_i$, or $(.A..B.)_i$ instead of $(A..B)_i$.

Relation between site-graph-rewrite rules and Kappa

Since the main purpose of this paper is not to formally present the reduction procedure for a general rule-set, we described the rule-based model directly as a collection of site-graph-rewrite rules, which is a simplification with respect to standard site-graph framework of Kappa (Danos et al. 2010). The simplification arises in three aspects.

First, the site (protein domain) in Kappa may be *internal*, in the sense that they bear an internal state encoding, for instance, post-translational modification of protein-residues such as phosphorylation, methylation, ubiquitylation—to name a few. Moreover, one site can simultaneously serve as a binding site, and as an internal site. We omit the possibility of having internal sites, but, it can be overcome: for example, the phosphorylation of a site can be encoded by a binding reaction to a node with a new name, for example, Ph. In order to mimic the standard unimolecular modification process by this bimolecular one, we need to ensure that the nodes of type Ph are always highly abundant, that is, are not rate limiting at any time. As a side remark, we point out that in reality it takes a binding event (e.g. binding of ATP) for a modification to happen. If a site is both internal and binding site, another copy of the site is created, so that one site bears an internal state, and another one is a binding state. A Kappa rule and an example of the corresponding site-graph-rewrite rule are shown in Fig. 8b.

Second, each Kappa program has a predefined signature of site types and agent types, where the agent type consists of a name, and a predetermined interface (set of sites). Each node of a 'Kappa' site-graph is assigned a unique name. On top of that, a type function partitions all the nodes according to their agent type. We instead embed the information about the node type (and we also abandon the use of term 'agent' in favor of 'node') directly in the name of the node: a node v^i, $i \in \mathbb{N}$ is of type v; The rules are accordingly written with these generative node names. The interface of a node type v is read from the collection of site-graph-rewrite rules, as a union of all the sites which are assigned to v along the rules. Our formalism cannot specify a rule which operates over a connected site-graph with more than one node of a certain type, but the examples which we present here do not contain such rules.

Third, we restrict to the conserved systems—only edges can be modified by the rules, while Kappa can specify agent birth or deletion.

Finally, it is worth noting that we define the notion of embedding in a non-standard way, through a combination of node-renaming function and sub-site-graph property.

References

Anderson DF, Kurtz TG (2011) Continuous time markov chain models for chemical reaction networks. In: Koeppl H, Setti G, di Bernardo M, Densmore D (eds) Design and analysis of biomolecular circuits. Springer, Berlin

Blinov M, Faeder JR, Goldstein B, Hlavacek WS (2004) Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. Bioinformatics (Oxford, England) 20(17):3289–3291

Borisov NM, Markevich NI, Hoek JB, Kholodenko BN (2006) Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains. BioSystems 83:152–166

Buchholz P (1994) Exact and ordinary lumpability in finite Markov chains. J Appl Probab 31(1):59–75

Buchholz P (2008) Bisimulation relations for weighted automata. Theoret Comput Sci 393(1–3):109–123

Conzelmann H, Fey D, Gilles ED (2008) Exact model reduction of combinatorial reaction networks. BMC Syst Biol 2(78):342–351
Danos V, Laneve C (2003) Core formal molecular biology. Theoret Comput Sci 325:69–110
Danos V, Feret J, Fontana W, Harmer R, Krivine J (2010) Abstracting the differential semantics of rule-based models: exact and automated model reduction. In: Symposium on logic in computer science, pp 362–381
Feret J, Henzinger T, Koeppl H, Petrov T (2012) Lumpability abstractions of rule-based systems. Theoret Comput Sci 431:137–164
Feret J, Koeppl H, Petrov T (2013) Stochastic fragments: a framework for the exact reduction of the stochastic semantics of rule-based models. Int J Softw Inf 4 (to appear)
Friedman N, Cai L, Sunney XX (2010) Stochasticity in gene expression as observed by single-molecule experiments in live cells. Israel J Chem 49:333–342
Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58(1):35–55
Guptasarma P (1995), Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli? BioEssays: news and reviews in molecular, cellular and, developmental biology 17(11):987–997
Gurvits L, Ledoux J (2005) Markov property for a function of a Markov chain: a linear algebra approach. Linear Algebra Appl 404:85–117
Hardy GH, Ramanujan S (1918) Asymptotic formula in combinatory analysis. Proc Lond Math Soc S2–17(1):75–115
Hernández-Lerma O, Lasserre J-B (2003) Markov chains and invariant probabilities. In: Progress in mathematics, vol 211. Birkhäuser Verlag, Basel
Hlavacek WS, Faeder JR, Blinov ML, Perelson AS, Goldstein B (2005) The complexity of complexes in signal transduction. Biotechnol Bioeng 84:783–794
Ibe OC (2009) Markov processes for stochastic modeling. Elsevier, Amsterdam
Kang H-W, Kurtz TG (2013) Separation of time-scales and model reduction for stochastic reaction networks. Ann Appl Probab 23(2):529–583
Keizer J (1987) Statistical thermodynamics of nonequilibrium processes, 1st edn. Springer, Berlin
Kemeny J, Snell JL (1960) Finite Markov chains. Van Nostrand
Krivine J, Danos V, Benecke A (2009) Modelling epigenetic information maintenance: a kappa tutorial. CAV, pp 17–32
Ledoux J (1995) On weak lumpability of denumerable Markov chains. Statist Probab Lett 25(4):329–339
Ledoux J (2004) Linear dynamics for the state vector of Markov chain functions. Adv Appl Probab 36(4):1198–1211
Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys 124:044104
Norris JR (1998) Markov chains. Cambridge university press, Cambridge
Rubino G, Sericola B (1991) A finite characterization of weak lumpable Markov processes. part II: The continuous time case. Stochast Process Appl 38(2):195–204
Rubino G, Sericola B (1993) A finite characterization of weak lumpable Markov processes. part I: The discrete time case. Stochast Process Appl 45(1):115–125
Sokolova A, de Vink EP (2003) On relational properties of lumpability. In: Proceedings of the 4th PROGRESS
Tian JP, Kannan D (2006) Lumpability and commutativity of Markov processes. Stochast Anal Appl 24(3):685–702
Walsh CT (2006) Posttranslation modification of proteins: expanding nature's inventory. Roberts and Co Publisher, Englewood
Wilkinson DJ (2006) Stochastic modelling for systems biology. Chapman & Hall, Boca Raton
Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. Science 311(5767):1600–1603