Symbolic dynamics of biochemical pathways as finite states machines

Ovidiu Radulescu[1], Satya Swarup Samal[2], Aurélien Naldi[1],
Dima Grigoriev[3], and Andreas Weber[2]

[1] DIMNP UMR CNRS 5235, University of Montpellier 2, Montpellier, France,

[2] Institut für Informatik II, Universität Bonn, Germany,

[3] CNRS, Mathématiques, Université de Lille, 59655, Villeneuve d'Ascq, France.

July 7, 2015

Abstract

We discuss the symbolic dynamics of biochemical networks with separate timescales. We show that symbolic dynamics of monomolecular reaction networks with separated rate constants can be described by deterministic, acyclic automata with a number of states that is inferior to the number of biochemical species. For nonlinear pathways, we propose a general approach to approximate their dynamics by finite state machines working on the metastable states of the network (long life states where the system has slow dynamics). For networks with polynomial rate functions we propose to compute metastable states as solutions of the tropical equilibration problem. Tropical equilibrations are defined by the equality of at least two dominant monomials of opposite signs in the differential equations of each dynamic variable. In algebraic geometry, tropical equilibrations are tantamount to tropical prevarieties, that are finite intersections of tropical hypersurfaces.

1   Introduction

Networks of biochemical reactions are used in computational biology as models of signaling, metabolism, and gene regulation. For various applications it is important to understand how the dynamics of these models depend on internal parameters and environment variables. Traditionally, the dynamics of biochemical networks is studied in the framework of chemical kinetics that can be either deterministic (ordinary differential equations) or stochastic (continuous time Markov processes). Within this framework, problems such as causality, reachability, temporal logics, are hard to solve and even to formalize. Concurrency models such as Petri nets and process algebra conveniently formalize these questions that remain nevertheless difficult. The main source of difficulty is the extensiveness of the set of trajectories that have to be analysed. Discretisation of the phase space does not solve the problem, because in multi-valued networks with m levels (Boolean networks correspond to $m = 2$) the number of the states is m^n and grows exponentially with the number of variables n. An interesting alternative to these approaches is symbolic dynamics which means replacing

the trajectories of the smooth system with a sequence of symbols. In certain cases, this could lead to relatively simple descriptions. According to the famous conjecture of Jacob Palis [11], smooth dynamical systems on compact spaces should have a finite number of attractors whose basins cover the entire ambient space. Compactness of ambient space is satisfied by networks of biochemical reactions because of conservation, or dissipativity. For high dimensional systems with multiple separated timescales it it reasonable to consider the following property: trajectories within basins of attraction consists in a succession of fast transitions between relatively slow regions. The slow regions, generally called metastable states, can be of several types such as attractive invariant manifolds, Milnor attractors or saddles. Because of compactness of the ambient space and smoothness of the vector fields defining the dynamics, there should be a finite number of such metastable states. This phenomenon, called itinerancy received particular attention in neuroscience [18]. We believe that similar phenomena occur in molecular regulatory networks. A simple example is the set of bifurcations of metastable states guiding the orderly progression of the cell cycle. In this paper we use tropical geometry methods to detect the presence of metastable states and describe the symbolic dynamics as a finite state automaton. The structure of the paper is the following. In the second section we compute the symbolic dynamics of monomolecular networks with totally separated constants. To this aim we rely on previous results [4, 12, 13]. In the third section we introduce tropical equilibrations of nonlinear networks. Tropical equilibrations are good candidates for metastable states. More precisely, we use minimal branches of tropical equilibrations as proxys for metastable states. In the forth section we propose an algorithm to learn finite state automata defined on these states.

2 Monomolecular networks with totally separated constants

Monomolecular reaction networks are the simplest reactions networks. The structure of these networks is completely defined by a digraph $G = (V, \mathcal{A})$, in which vertices $i \in V, 1 \leq i \leq n$ correspond to chemical species A_i, edges $(i, j) \in \mathcal{A}$ correspond to reactions $A_i \rightarrow A_j$ with kinetic constants $k_{ji} > 0$. For each vertex, A_i, a positive real variable c_i (concentration) is defined. The chemical kinetic dynamics is described by a system of linear differential equations

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_j k_{ij} c_j - (\sum_j k_{ji}) c_i, \tag{1}$$

where $k_{ji} > 0$ are kinetic coefficients. In matrix form one has : $\dot{c} = Kc$. The solutions of (1) can be expressed in terms of left and right eigenvectors of the kinetic matrix K:

$$c(t) = r^0(l^0, c(0)) + \sum_{k=1}^{n-1} r^k(l^k, c(0)) \exp(\lambda_k t), \tag{2}$$

where r^k, l^k are right and left eigenvectors of K, $Kr^k = \lambda_k r^k$, $l^k K = \lambda_k l^k$.

The system (1) has a conservation law $\frac{\mathrm{d}}{\mathrm{d}t}(c_1 + c_2 + \ldots + c_n) = 0$, and therefore there is a zero eigenvalue $\lambda_0 = 0$, $l^0 = (1, 1, \ldots, 1)$, $(l^0, c(0)) = c_1(0) + c_2(0) + \ldots + c_n(0)$. We say that the network constants are totally separated if for all $(i, j) \neq (i', j')$ one of the relations $k_{ji} \ll k_{j'i'}$, or $k_{ji} \gg k_{j'i'}$ is satisfied.

It was shown in [4, 12, 13] that the eigenvalues and the eigenvectors of an arbitrary monomolecular reaction networks with totally separated constants can be approximated with good accuracy by the eigenvalues of and the eigenvectors of a reduced monomolecular networks whose reaction digraph is acyclic (has no cycles), and deterministic (has no nodes from which leave more than one edge). Let us denote by $G_r = (V_r, \mathcal{A}_r)$ the reduced digraph, and by κ_i the kinetic constant of the unique reaction that leaves a node $i \in V_r$. The algorithm to obtain G from G_r can be found in [4, 12, 13] and will not be repeated here. Because G_r is deterministic it defines a flow (discrete dynamical system) on the graph: $\Phi(i) = j$, where j is the unique node following i on the digraph. Reciprocally, we define $\mathrm{Pred}(i) = \phi^{-1}(i)$ as the set of predecessors of the node i in the digraph G_r, namely $\mathrm{Pred}(i) = \{j \in V_r | (j, i) \in \mathcal{A}_r\}$.

We say that a node is a sink if it has no successors on the graph. For the sake of simplicity, we suppose that there is only one sink. For each one of the remaining $n-1$ nodes there is one reaction leaving from it. For a network with totally separated constants we have

$$\kappa_i \ll \kappa_j, \text{ or } \kappa_i \gg \kappa_j \text{ for all } i, j \in [1, n-1], \ i \neq j \tag{3}$$

For totally separated constants the following lemma is useful

Lemma 2.1 *If* (3) *is satisfied then, at lowest order, we have*

$$\frac{\kappa_i}{-\kappa_k + \kappa_j} = \begin{cases} 1, & \text{if } i = j \text{ and } \kappa_k < \kappa_i \\ -1, & \text{if } i = k \text{ and } \kappa_j < \kappa_i \\ 0, & \text{if } \kappa_i < \min(\kappa_k, \kappa_j) \\ \pm\infty, & \text{else} \end{cases} \tag{4}$$

The dynamics of the reduced model is given by

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j \in \mathrm{Pred}(i)} \kappa_j c_j - \kappa_i c_i, \tag{5}$$

where $\mathrm{Pred}(i)$ is the set of predecessors of the node i in the digraph G_r, namely $\mathrm{Pred}(i) = \{j \in V_r | (j, i) \in \mathcal{A}_r\}$.

As shown in [4] the eigenvectors of the approximated kinetic matrix satisfy

$$\sum_{j \in \mathrm{Pred}(i)} \kappa_j r_j = (\lambda + \kappa_i) r_i \tag{6}$$

$$\kappa_i l_{\Phi(i)} = (\lambda + \kappa_i) l_i, \tag{7}$$

where λ is the eigenvalue, r_i, l_i, $1 \leq i \leq n$ are the components of the right and left eigenvectors, respectively.

Eqs.(6) and (7) imply that the right and left eigenvectors can be computed by recurrence on the graph, in the direct direction and in the reverse direction, respectively. In order to have non-zero eigenvectors, $\lambda = -\kappa_i$ for some i not a sink, therefore the (non-zero) eigenvalues are $\lambda_k = -\kappa_k, 1 \leq k \leq n-1$. Taking into account the separation conditions (3) we get the following

Proposition 2.2 *Let us consider that $\kappa_k = 0$ when k is a sink in the graph G_r. Then, the eigenvalues of the kinetic matrix with totally separated constants are $\lambda_k = -\kappa_k$, with $\lambda_k = 0$ when k is a sink. The corresponding left eigenvectors*

are

$$l_j^k = \begin{cases} 1, & \text{if } \Phi^m(j) = k \text{ for some } m > 0 \text{ and } \kappa_{\Phi^l(j)} > \kappa_k \text{ for all } l = 0, \dots, m-1 \\ 0, & \text{otherwise} \end{cases},$$

(8)

and the right eigenvectors are

$$r_j^k = \begin{cases} 1, & \text{if } j = k \\ -1, & \text{if } j = \Phi^m(k) \text{ for some } m > 0 \text{ and } \kappa_{\Phi^m(k)} < \kappa_k < \kappa_{\Phi^l(k)}, \\ & \text{for all } l = 1, \dots, m-1 \\ 0, & \text{otherwise}. \end{cases}$$

The full proof of the Proposition 2.2 can be found in the appendices.

Let us now discuss the symbolic dynamics of the system. For each eigenvalue $\lambda_k = -\kappa_k$, $\kappa_k > 0$ we associate a transition time $t_k = \kappa_k^{-1}$. Without loss of generality we can consider that $t_1 \ll t_2 \ll \dots \ll t_{n-1}$. Any trajectory of the system is given by (2). At the time t_k one exponential term $exp(\lambda_k t)$ will vanish and the result will be a transition $c \to c - r^k(l^k, c(0))$, provided that $(l^k, c(0)) \neq 0$. In other works, a trajectory can be described as a discrete sequence of states $c(0), c(0) - r^1(l^1, c(0)), \dots$. Let us consider the following normalization $c_1(0) + c_2(0) + \dots + c_n(0) = 1$. Then c_i is the probability of presence in the node i of a particle moving through the reaction network. For monomolecular networks, particles are independent, therefore this simple picture is enough for understanding the dynamics. Let the index i_0 define the initial state of the system $c_{i_0}(0) = 1$, $c_j(0) = 0$ for $j \neq i_0$. i_0 represents the initial position of the particle. According to the Prop. 2.2 $(l^k, c(0)) = l_{i_0}^k = 1$ if the step κ_k is downstream of i_0 in the graph G_r and if all steps from i_0 to k are faster than κ_k. In this case the jump at t_k is $-r^k$. A jump $-r^k$ has two components different from zero, $-r_k^k = -1$ and $-r_j^k = 1$, where j is the first node downstream of k from which starts a step slower than κ_k. Thus, the jump $-r^k$ corresponds to displacing the particle from k to j. The set of right eigenvectors defines a symbolic flow on the reaction digraph. A particle starting in i_0 first jumps in i_1 where i_1 is the first node such that $\kappa_{i_1} < \kappa_{i_0}$, then continues to i_2 where i_2 is the first node such that $\kappa_{i_2} < \kappa_{i_1}$, and so one and so forth until it gets to the sink. Some nodes have negligible sojourn time, namely nodes such that $\kappa_i > \kappa_j$ for all $j \in \text{Pred}(i)$. This proves the main result of the section.

By transition graph of a finite state machine we mean the digraph $G_{rs} = (V_s, \mathcal{A}_s)$, where V_s is the set of states of the machine and $(i, j) \in \mathcal{A}_s$ if there are transitions from the state i to the state j. We have the following theorem:

Theorem 2.3 *The symbolic dynamics of a monomolecular network with totally separated constants can be described by a deterministic acyclic finite state machine. The transition graph $G_{rs} = (V_s, \mathcal{A}_s)$ of this machine can be obtained from the graph $G_r = (V_r, \mathcal{A}_r)$ in the following way: $V_s = V_r \setminus \{i \in V_r | \kappa_i > \kappa_j \text{ for all } j \in \text{Pred}(i)\}$, $\mathcal{A}_s = \{(i, j) | i, j \in V_s \text{ and there are } i_0 = i, i_1, \dots, i_m = j, \text{ such that } i_l \in V_r \setminus V_s, \text{ for } l = 1, \dots, m-1, \text{ and } (i_l, i_{l+1}) \in \mathcal{A}_r \text{ for } l = 0, \dots, m-1\}$.*

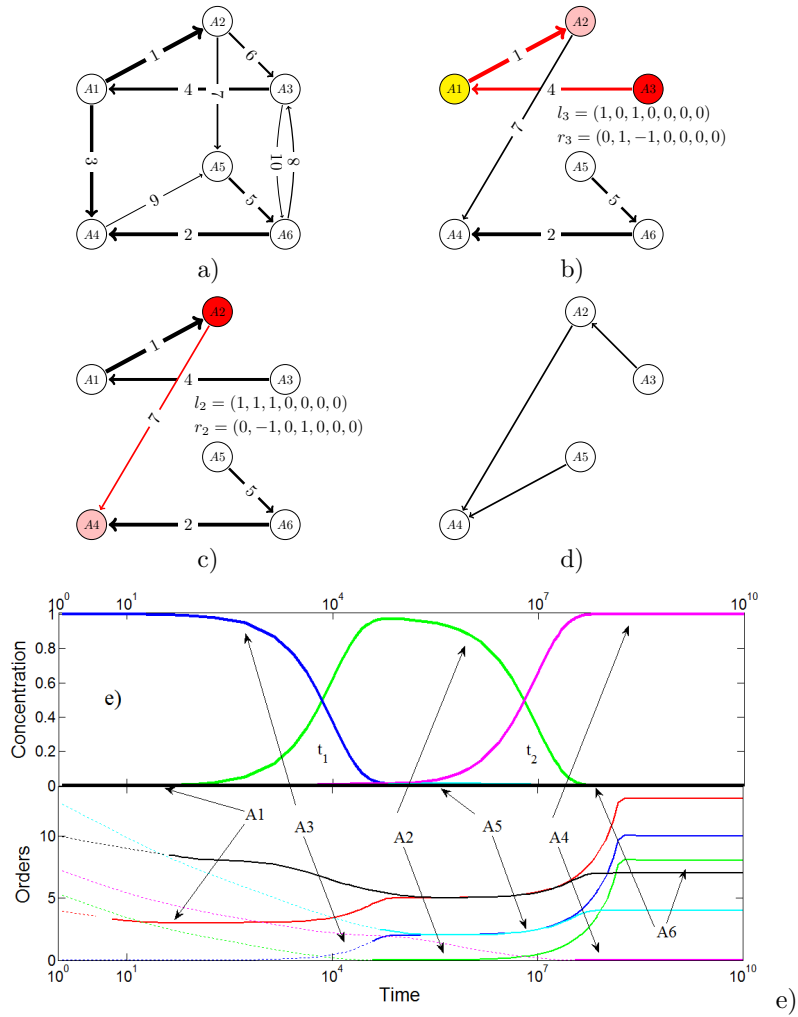**Remark 2.4** *An example is detailed in Figure 1.*

4

Figure 1: Symbolic dynamics of a monomolecular network with total separation. The integers γ_i labelling the reactions represent the orders of the kinetic constants, smaller orders meaning faster reactions. The model was reduced using the recipe described in [4, 13] (see appendices). a) full model; b-c) reduced model with active transitions and corresponding eigenvectors. During a transition the network behaves like a single step : the concentrations of some species (white) are practically constant, some species (yellow) are rapid, low concentration, intermediates, one species (red) is gradually consumed and another (pink) is gradually produced. The net result is the displacement of a particle one or several steps downstream; d) The transition graph of the finite state machine representing the symbolic dynamics of the network; e) Trajectory starting from $A3$ (at $t = 0$ the total mass is in $A3$), undergoing two transitions at t_1 and t_2. The simulation has been performed for kinetic constants $\kappa_i = \varepsilon^{\gamma_i}$, with $\varepsilon = 1/50$. On top, concentration of species (concentrations of $A1, A_4, A_6$ are negligible everywhere). At bottom, orders of concentrations (computed as $\log_\epsilon(x_i)$) with continuous lines if species is tropically equilibrated, dotted lines if not.

5

3 Tropical equilibrations of nonlinear networks with polynomial rate functions

In this section we consider nonlinear biochemical networks described by mass action kinetics

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_j k_j S_{ij} x^{\alpha_j}, \ 1 \le i \le n, \tag{9}$$

where $k_j > 0$ are kinetic constants, S_{ij} are the entries of the stoichiometric matrix (uniformly bounded integers, $|S_{ij}| < s$, s is small), $\alpha_j = (\alpha_1^j, \ldots, \alpha_n^j)$ are multi-indices, and $x^{\alpha_j} = x_1^{\alpha_1^j} \ldots x_n^{\alpha_n^j}$, where α_i^j are positive integers.

For chemical reaction networks with multiple timescales it is reasonable to consider that kinetic parameters have different orders of magnitudes. This can be conveniently formalized by considering that parameters of the kinetic models (9) can be written as $k_j = \bar{k}_j \varepsilon^{\gamma_j}$ (10). The exponents γ_j are considered to be integer or rational. For instance, the approximation $\gamma_j = \mathrm{round}(\log(k_j)/\log(\varepsilon))$ produces integer exponents, whereas $\gamma_j = \mathrm{round}(d\log(k_j)/\log(\varepsilon))/d$ produces rational exponents, where round stands for the closest integer (with half-integers rounded to even numbers) and d is a strictly positive integer. Kinetic parameters are fixed. In contrast, species orders vary in the concentration space and have to be calculated as solutions to the tropical equilibration problem. To this aim, the network dynamics is first described by a rescaled ODE system

$$\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t} = \sum_j \varepsilon^{\mu_j(a)-a_i} \bar{k}_j S_{ij} \bar{x}^{\alpha_j}, \tag{11}$$

where $\mu_j(a) = \gamma_j + \langle a, \alpha_j \rangle$ (12), and $\langle, \rangle$ stands for the dot product.

The r.h.s. of each equation in (11) is a sum of multivariate monomials in the concentrations. The orders μ_j indicate how large are these monomials, in absolute value. A monomial of order μ_j dominates another monomial of order $\mu_{j'}$ if $\mu_j < \mu_{j'}$.

The tropical equilibration problem consists in the equality of the orders of at least two monomials one positive and another negative in the differential equations of each species. More precisely, we want to find a vector a such that

$$\min_{j, S_{ij}>0}(\gamma_j + \langle a, \alpha_j \rangle) = \min_{j, S_{ij}<0}(\gamma_j + \langle a, \alpha_j \rangle) \tag{13}$$

Computing tropical equilibrations from the orders of magnitude of the model parameters is a NP-complete problem, cf. [17]. However, methods based on the Newton polytope [15] or constraint logic programming [16] exploit the sparseness and redundance of the system to effectively obtain sets of solutions. The equation(13) is related to the notion of *tropical hypersurface*. A *tropical hypersurface* is the set of vectors $a \in \mathbb{R}^n$ such that the minimun $\min_{j, S_{ij} \ne 0}(\gamma_j + \langle a, \alpha_j \rangle)$ is attained for at least two different indices j (with no sign conditions). *Tropical prevarieties* are finite intersections of tropical hypersurfaces. Therefore, our tropical equilibrations are subsets of tropical preverieties. The sign condition in (13) was imposed because species concentrations are real positive numbers. Compensation of a sum of positive monomials is not possible for real values of the variables.

Species timescales. The timescale of a variable x_i is given by $\frac{1}{x_i}\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{1}{\bar{x}_i}\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t}$ whose order is $\nu_i = \min\{\mu_j | S_{ij} \ne 0\} - a_i$ (14). The order ν_i indicates how fast is the variable x_i (if $\nu_{i'} < \nu_i$ then $x_{i'}$ is faster than x_i) .

Partial tropical equilibrations. It is useful to extend the tropical equilibration problem to partial equilibrations, that means solving (13) only for a subset of species. This is justified by the fact that slow species do not need to be equilibrated. In order to have a self-consistent calculation we compute the species timescales by (14). A partial equilibration is *consistent* if $\nu_i < \nu$ for all non-equilibrated species i. $\nu > 0$ is an arbitrarily chosen threshold indicating the timescale of interest.

Tropical equilibrations, slow invariant manifolds and metastable states. In dissipative systems, fast variables relax rapidly to some low dimensional attractive manifold called invariant manifold [3] that carries the slow mode dynamics. A projection of dynamical equations onto this manifold provides the reduced dynamics [8]. This simple picture can be complexified to cope with hierarchies of invariant manifolds and with phenomena such as transverse instability, excitability and itinerancy. Firstly, the relaxation towards an attractor can have several stages, each with its own invariant manifold. During relaxation towards the attractor, invariant manifolds are usually embedded one into another (there is a decrease of dimensionality) [2]. Secondly, invariant manifolds can lose local stability, which allow the trajectories to perform large phase space excursions before returning in a different place on the same invariant manifold or on a different one [7]. We showed elsewhere that tropical equilibrations can be used to approximate invariant manifolds for systems of polynomial differential equations [9, 10, 14]. Indeed, tropical equilibration are defined by the equality of dominant forces acting on the system. The remaining weak non-compensated forces ensure the slow dynamics on the invariant manifold. Tropical equilibrations are thus different from steady states, in that there is a slow dynamics. In this paper we will use them as proxies for metastable states.

Branches of tropical equilibrations and connectivity graph. For each equation i, let us define $M_i(a) = \operatorname*{argmin}_{j}(\mu_j(a), S_{ij} > 0) = \operatorname*{argmin}_{j}(\mu_j(a), S_{ij} < 0)$ (15), in other words M_i denotes the set of monomials having the same minimal order μ_i. We call *tropically truncated system* the system obtained by pruning the system (11), i.e. by keeping only the dominating monomials.

$$\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t} = \varepsilon^{\mu_i - a_i} \big(\sum_{j \in M_i(a)} \bar{k}_j \nu_{ji} \bar{x}^{\alpha_j} \big), \tag{16}$$

The tropical truncated system is uniquely determined by the index sets $M_i(a)$, therefore by the tropical equilibration a. Reciprocally, two tropical equilibrations can have the same index sets $M_i(a)$ and truncated systems. We say that two tropical equilibrations a_1, a_2 are equivalent iff $M_i(a_1) = M_i(a_2)$, for all i. Equivalence classes of tropical equilibrations are called *branches*. A branch B with an index set M_i is *minimal* if $M_i' \subset M_i$ for all i where M_i' is the index set B' implies $B' = B$ or $B' = \emptyset$. Closures of equilibration branches are defined by a finite set of linear inequalities, which means that they are polyhedral complexes. Minimal branches correspond to maximal dimension faces of the polyhedral complex. The incidence relations between the maximal dimension faces ($n - 1$ dimensional faces, where n is the number of variables) of the polyhedral complex define the *connectivity graph*. More precisely, minimal branches are the vertices of this graph. Two minimal branches are connected if the corresponding faces of the polyhedral complex share a $n - 2$ dimensional face. In terms of index sets, two minimal branches with index sets M and M' are con-

nected if there is an index set M'' such that $M_i' \subset M_i''$ and $M_i \subset M_i''$ for all i.

Tropical equilibrations and monomolecular networks. Eqs.(13) have a simpler form in the case of monomolecular networks

$$\min_{j \in \mathrm{Pred}(i)} (\gamma_{ij} + a_j) = \min_{j \in \mathrm{Succ}(i)} (\gamma_{ji} + a_i) \tag{17}$$

where $\mathrm{Pred}(i) = \{j | (j, i) \in \mathcal{A}\}$, $\mathrm{Succ}(i) = \{j | (i, j) \in \mathcal{A}\}$ are the sets of predecessors and successors of the node i in the digraph G.

Let us recall that by min-plus algebra we understand the semi-ring $(\mathbb{R} \cup \{\infty\}, \oplus, \otimes)$ where the two operations are defined as $x \oplus y = \min\{x, y\}$ and $x \otimes y = x + y$. In other words the addition and the min operation play the role of min-plus multiplication and addition, respectively. Therefore Eqs.(17) are linear in the unknowns a_i. Computing tropical equilibrations of monomolecular networks boils down to solving linear equations in min-plus algebra. For linear tropical systems there are fast algorithms [5, 6].

We have tested the tropical equilibration conditions (17) for the trajectories of the monomolecular network presented in Figure 1 by checking if the absolute value of the difference between the r.h.s and l.h.s of (17) is smaller than a threshold. The result is illustrated in Fig. 1e). For this model, the tropical equilibration solutions are changing along the trajectory. This can been seen by following the orders of the concentrations along the trajectories. These orders change by integers at transition points. Furthermore, at transition points some of the variables that where not previously equilibrated, become equilibrated. The analysis of the tropical equilibrations finds the transitions previously detected in Section 2 from the approximated eigenvalues and eigenvectors (t_1 and t_2 for this example) but adds some more. For instance, species $A1$ equilibrates at the timescale $1/\kappa_1 = 10$. This was not taken into account in the description of the automaton in Figure 1d) because the species $A1$ is fast and can not accumulate.

4 Learning a finite state machine from a nonlinear biochemical network

We are using the algorithm based on constraint solving introduced in [16] to obtain all rational tropical equilibration solutions $a = (a_1, a_2, \dots, a_n)$ within a box $|a_i| < b$, $b > 0$ and with denominators smaller than a fixed value d, $a_i = p_i/q$, p_i, q are positive integers, $q < d$. The output of the algorithm is a matrix containing all the tropical equilibrations within the defined bounds. A post-processing treatment is applied to this output consisting in computing truncated systems, index sets, and minimal branches. Tropical equilibrations minimal branches are stored as matrices $A_1, A_2, \dots, A_b$, whose lines are tropical solutions within the same branch. Here b is the number of minimal branches.

Our method computes numerical approximations of the tropical prevariety. Given a value of ϵ, this approximation is better when the denominator bound d is high. At fixed d, the dependence of the precision on ϵ follows more intricate rules dictated by Diophantine approximations. For this reason, we systematically test that the number b and the truncated systems corresponding to minimal branches are robust when changing the value of ϵ.

Trajectories $x(t) = (x_1(t), \ldots, x_n(t))$ of the smooth dynamical system are generated with different initial conditions, chosen uniformly and satisfying the conservation laws, if any. For each time t, we compute the Euclidian distance $d_i(t) = \min_{y \in A_i} \|y - log_\varepsilon(x(t))\|$, where $\|*\|$ denotes the Euclidean norm and $\log_\varepsilon(x) = (\log x_1 / \log(\varepsilon), \ldots, \log x_n / \log(\varepsilon))$. This distance classifies all points of the trajectory as belonging to a tropical minimal branch. The result is a symbolic trajectory $s_1, s_2, \ldots$ where the symbols s_i belong to the set of minimal branches. In order to include the possibility of transition regions we include an unique symbol t to represent the situations when the minimal distance is larger than a fixed threshold. We also store the residence times $\tau_1, \tau_2, \ldots$ that represent the time spent in each of the state.

The stochastic automaton is learned as a homogenous, finite states, continuous time Markov process, defined by the lifetime (mean sojourn time) of each state T_i, $1 \leq i \leq b$ and by the transition probabilities $p_{i,j}$ from a state i to another state j. We use the following estimators for the lifetimes and for the transition probabilities:

$$T_i = (\sum_n \tau_n \mathbb{1}_{s_n=i})/(\sum_n \mathbb{1}_{s_n=i}) \tag{18}$$

$$p_{i,j} = (\sum_n \mathbb{1}_{s_n=i,s_{n+1}=j})/(\sum_n \mathbb{1}_{s_n=i}), \ i \neq j \tag{19}$$

As a case study we consider a nonlinear model of dynamic regulation of Transforming Growth Factor beta TGF-β signaling pathway proposed in [1]. This model has a dynamics defined by $n = 18$ polynomial differential equations and 25 biochemical reactions. The paper [1] proposes three versions of the mechanism of interaction of TIF1γ (Transcriptional Intermediary Factor 1 γ) with the Smad-dependent TGF-β signaling. We consider here the version in which TIF1 interacts with the phosphorylated Smad2–Smad4 complexes leading to dissociation of the complex and degradation of Smad4. The results are similar for the other versions of this model. The example was chosen because it is a medium size model based on polynomial differential equations. The computation of the tropical equilibrations for this model shows that there are 9 minimal branches of full equilibrations (in these tropical solutions all variables are equilibrated). The connectivity graph of these branches and the learned automaton are shown in Figure 2. The study of this example shows that branches of tropical equilibration can change on trajectories of the dynamical system. Furthermore, all the observed transitions between branches are contained in the connectivity graph resulting from the polyhedral complex of the tropical equilibration branches.

The transition probabilities of the automaton are coarse grained properties of the statistical ensemble of trajectories for different initial conditions. Given a state and a minimal branch close to it, it will depend on the actual trajectory to which other branch the system will be close to next. However, when initial data and the full trajectory are not known, the automaton will provide estimates of where we go next and with which probability. For the example studied, the branch B1 is a globally attractive sink: starting from anywhere, the automaton will reach B1 with probability one. This branch contains the unique stable steady state of the initial model. Figure 2 bottom right shows the structure of most probable branches, the ones in which the systems spends most of his time. The branches B1, B3 and B2 correspond to different compositions of the membrane and of the endosome, rich in the receptor RI, rich in the receptor RII and
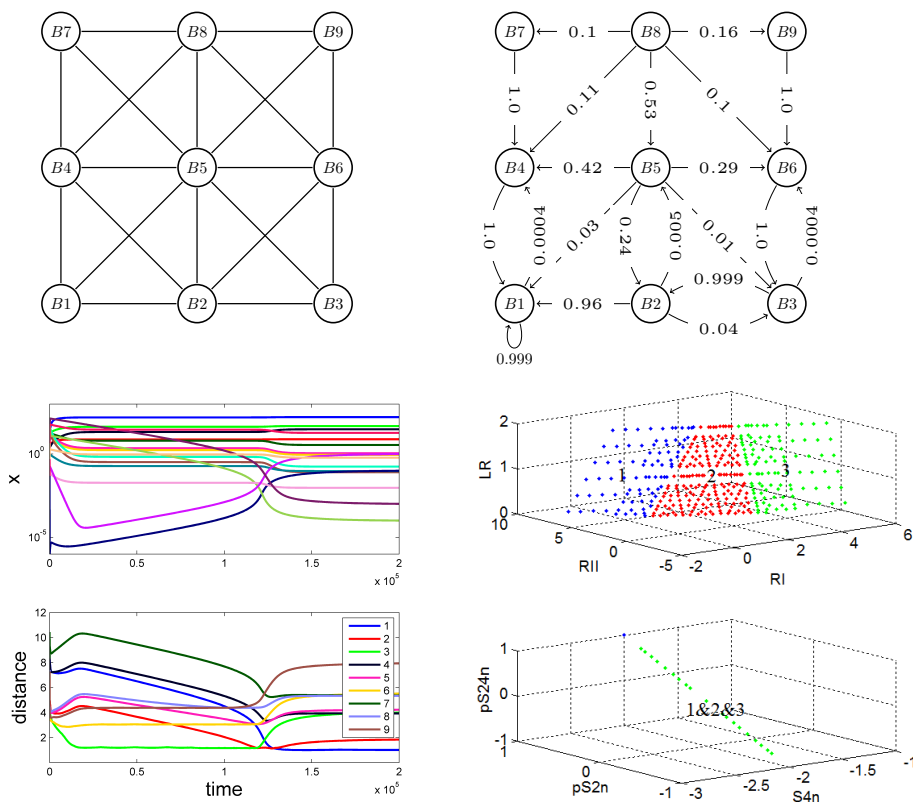
Figure 2: TGFβ model. Upper left: Connectivity graph of tropical minimal branches; upper right: finite state automaton; bottom left: trajectories with jumps and distances to minimal branches; the closest branch changes with time along the trajectory; bottom right: first three tropical equilibrations minimal branches in various projections in concentration orders space. The variables RI, RII, LR are membrane receptors (signaling input layers) concentration orders, whereas pS2n, S4n, pS24n are nuclear transcription factors and complexes (effectors) concentration orders. The structure tropical branches shows that composition of input layers is more flexible (varies on planes) than the concentrations of effectors (vary on lines).

rich in both types of receptors, respectively. Even if this compositon is changed on wide domains of orders (planes in the space of orders), the concentrations of effectors are robust (are more constained than the concentrations of receptors).

5   Conclusion

We have presented a method to coarse grain the dynamics of a smooth biochemical reaction network to a discrete symbolic dynamics of a finite state automaton. The coarse graining was obtained by two methods, approximated eigenvectors for mono-molecular networks and minimal branches of tropical equilibrations for more general mass action nonlinear networks. The two methods are compatible one to another, because when applied to monomolecular networks the

method based on tropical geometry detects all the transitions indicated by approximated eigenvectors. For both methods the automaton has a small number of states, less than the number of species in the first method and the number of minimal tropical branches in the second method. The coarse grained automaton can be used for studying statistical properties of biochemical networks such as occurrence and stability of temporal patterns, recurrence, periodicity and attainability problems. The coarse graining can be performed in a hierarchical way. For the nonlinear example studied in the paper we computed only the full tropical equilibrations that stand for the lowest order in the hierarchy (coarsest model). As discussed in Section 3 we can also consider partial equilibrations when slow variables are not equilibrated and thus refine the automaton. Our approach extends the notion of steady states of a network and propose a simple recipe to characterize and detect metastable states. Most likely metastable states have biological importance because the network spends most of its time in these states. The itinerancy of the network, described as the possibility of transitions from one metastable state to another is paramount to the way neural networks compute, retrieve and use information [18] and can have similar role in biochemical networks.

References

[1] Andrieux, G., Fattet, L., Le Borgne, M., Rimokh, R., Théret, N.: Dynamic regulation of Tgf-B signaling by Tif1γ: a computational approach. PloS one 7(3), e33761 (2012)

[2] Chiavazzo, E., Karlin, I.: Adaptive simplification of complex multiscale systems. Physical Review E 83(3), 036706 (2011)

[3] Gorban, A., Karlin, I.: Invariant manifolds for physical and chemical kinetics, Lect. Notes Phys. 660. Springer, Berlin, Heidelberg (2005)

[4] Gorban, A., Radulescu, O.: Dynamic and static limitation in reaction networks, revisited . In: Guy B. Marin, D.W., Yablonsky, G.S. (eds.) Advances in Chemical Engineering - Mathematics in Chemical Kinetics and Engineering, Advances in Chemical Engineering, vol. 34, pp. 103–173. Elsevier (2008)

[5] Grigoriev, D.: Complexity of solving tropical linear systems. Computational complexity 22(1), 71–88 (2013)

[6] Grigoriev, D., Podolskii, V.V.: Complexity of tropical and min-plus linear prevarieties. Computational Complexity 24, 31–64 (2015)

[7] Haller, G., Sapsis, T.: Localized instability and attraction along invariant manifolds. SIAM Journal on Applied Dynamical Systems 9(2), 611–633 (2010)

[8] Maas, U., Pope, S.B.: Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space. Combustion and Flame 88(3), 239–264 (1992)

[9] Noel, V., Grigoriev, D., Vakulenko, S., Radulescu, O.: Tropical geometries and dynamics of biochemical networks application to hybrid cell cycle models. In: Feret, J., Levchenko, A. (eds.) Proceedings of the 2nd International Workshop on Static Analysis and Systems Biology (SASB 2011), Electronic Notes in Theoretical Computer Science, vol. 284, pp. 75–91. Elsevier (2012)

[10] Noel, V., Grigoriev, D., Vakulenko, S., Radulescu, O.: Topical and Idempotent Mathematics and Applications, vol. 616, chap. Tropicalization and tropical equilibration of chemical reactions. American Mathematical Society (2014)

[11] Palis, J.: A global view of dynamics and a conjecture on the denseness of finitude of attractors. Astérisque 261, 339–351 (2000)

[12] Radulescu, O., Gorban, A.N., Zinovyev, A., Lilienbaum, A.: Robust simplifications of multiscale biochemical networks. BMC systems biology 2(1), 86 (2008)

[13] Radulescu, O., Gorban, A.N., Zinovyev, A., Noel, V.: Reduction of dynamical biochemical reactions networks in computational biology. Frontiers in Genetics 3(131) (2012)

[14] Radulescu, O., Vakulenko, S., Grigoriev, D.: Model reduction of biochemical reactions networks by tropical analysis methods. Mathematical Model of Natural Phenomena, in press (2015)

[15] Samal, S.S., Radulescu, O., Grigoriev, D., Fröhlich, H., Weber, A.: A Tropical Method based on Newton Polygon Approach for Algebraic Analysis of Biochemical Reaction Networks. In: 9th European Conference on Mathematical and Theoretical Biology (2014)

[16] Soliman, S., Fages, F., Radulescu, O.: A constraint solving approach to model reduction by tropical equilibration. Algorithms for Molecular Biology 9(1), 24 (2014)

[17] Theobald, T.: On the frontiers of polynomial computations in tropical geometry. Journal of Symbolic Computation 41(12), 1360–1375 (2006)

[18] Tsuda, I.: Chaotic itinerancy as a dynamical basis of hermeneutics in brain and mind. World Futures: Journal of General Evolution 32(2-3), 167–184 (1991)

Appendix 1: Proof of Proposition 2.2. Let us consider that $r_k^k = 1$. Taking $r_k^j = 0$ for all predecessors j of k and for all other nodes that lead to k by the flow Φ satisfy Eq.(6)(main body text) with $\lambda = -\kappa_k$. The same is valid for all the nodes that do not lead to k and are not accessible from k. Remain the nodes that are accessible from k. Let j be such a node. Then $j = \Phi^m(k)$ for some $m > 0$. Eq.(6)(main body text) implies that

$$\kappa_{\Phi^{l-1}(k)} r_{\Phi^{l-1}(k)}^k = (-\kappa_k + \kappa_{\Phi^l(k)}) r_{\Phi^l(k)}^k, \ \text{ for } 1 \leq l \leq m.$$

Thus $r_{\Phi^m(k)}^k = \frac{\kappa_k}{-\kappa_k + \kappa_{\Phi(k)}} \times \frac{\kappa_{\Phi(k)}}{-\kappa_k + \kappa_{\Phi^2(k)}} \times \ldots \times \frac{\kappa_{\Phi^{m-1}(k)}}{-\kappa_k + \kappa_{\Phi^m(k)}}$. Suppose that $\kappa_k < \kappa_{\phi^l(k)}$ for $l = 1, \ldots, m-1$ and $\kappa_{\phi^m(k)} < \kappa_k$. Using Lemma2.1(main body text) it follows $r_{\Phi^m(k)}^k = -1$. If any of the previous inequality does not hold then at least one factor in the expression of $r_{\Phi^m(k)}^k$ vanishes and the remaining factors are finite, thus $r_{\Phi^m(k)}^k = 0$. Consider now that $l_k^k = 1$. Taking $l_k^j = 0$ for all the nodes j that can be obtained from k and for all other nodes that do not lead to k by the flow Φ satisfy Eq.(7)(main body text) with $\lambda = -\kappa_k$. The remaining nodes are all leading to k. Let j be such a node. Then $k = \Phi^m(j)$ for some $m > 0$. Eq.(7) (main body text) implies that

$$\kappa_{\Phi^{l-1}(j)} l_{\Phi^l(j)}^k = (-\kappa_k + \kappa_{\Phi^{l-1}(j)}) l_{\Phi^{l-1}(j)}^k, \ \text{ for } 1 \leq l \leq m.$$

Hence $l_j^k = \frac{\kappa_j}{-\kappa_k + \kappa_j} \times \frac{\kappa_{\Phi(j)}}{-\kappa_k + \kappa_{\Phi(j)}} \times \ldots \times \frac{\kappa_{\Phi^{m-1}(j)}}{-\kappa_k + \kappa_{\Phi^{m-1}(j)}}$. Suppose that $\kappa_{\Phi^l(j)} > \kappa_k$, for all $l = 0, \ldots, m-1$. Using Lemma2.1(main body text) it follows $l_j^k = 1$. If one of these inequalities is not satisfied for a $l = 0, \ldots, m-1$ then the corresponding factor in the expression of l_j^k vanishes and $l_j^k = 0$.

The above formulas cover the zero eigenvalue case if we consider that $\kappa_k = 0$ for k being the sink. It follows that $r_k^0 = 1$ and $r_j^0 = 0$ elsewhere. Furthermore, $l_j^0 = 1$ for all j.

Appendix 2: Algorithm for reduction of monomolecular networks with total separation separation. This algorithm consists of three steps.
I. Constructing of an auxiliary reaction network: pruning.

For each A_i branching node (substrate of several reactions) let us define κ_i as the maximal kinetic constant for reactions $A_i \to A_j$: $\kappa_i = \max_j\{k_{ji}\}$. For correspondent j we use the notation $j = \phi(i)$: $\phi(i) = \arg\max_j\{k_{ji}\}$.

An auxiliary reaction network $\mathcal{V}$ is the set of reactions obtained by keeping only $A_i \to A_{\phi(i)}$ with kinetic constants κ_i and discarding the other, slower reactions. Auxiliary networks have no branching, but they can have cycles and confluences. The correspondent kinetic equation is

$$\dot{c}_i = -\kappa_i c_i + \sum_{\phi(j)=i} \kappa_j c_j, \tag{20}$$

If the auxiliary network contains no cycles, the algorithm stops here.
II gluing cycles and restoring cycle exit reactions

In general, the auxiliary network $\mathcal{V}$ has several cycles $C_1, C_2, \ldots$ with periods $\tau_1, \tau_2, \ldots > 1$.

These cycles will be "glued" into points and all nodes in the cycle C_i, will be replaced by a single vertex A^i. Also, some of the reactions that were pruned in the first part of the algorithm are restored with renormalized rate constants. Indeed, reaction exiting a cycle are needed to render the correct dynamics: without them, the total mass of the cycle is conserved, with them the mass

can also slowly leave the cycle. Reactions $A \to B$ exiting from cycles ($A \in C_i$, $B \notin C_i$) are changed into $A^i \to B$ with the rate constant renormalization: let the cycle C^i be the following sequence of reactions $A_1 \to A_2 \to ...A_{\tau_i} \to A_1$, and the reaction rate constant for $A_i \to A_{i+1}$ is k_i (k_{τ_i} for $A_{\tau_i} \to A_1$). For the limiting (slowest) reaction of the cycle C_i we use notation $k_{\lim i}$. If $A = A_j$ and k is the rate reaction for $A \to B$, then the new reaction $A^i \to B$ has the rate constant $kk_{\lim i}/k_j$. This rate is obtained using quasi-stationary distribution for the cycle. If kinetic constants are expressed as powers of a small positive parameter ϵ, i.e., if $k = \epsilon^\gamma$, then the order of the constant has to be changed according to the rule $\gamma \to \gamma + \gamma_{lim} - \gamma_j$, where γ, $\gamma_{lim\,i}$, γ_j are the orders of the constants k, $k_{\lim i}$ and k_j, respectively.

The new auxiliary network $\mathcal{V}^1$ is computed for the network of glued cycles. Then we decompose it into cycles, glue them, iterate until a acyclic network is obtained $\mathcal{V}^n$.

III Restoring cycles

The dynamics of species inside glued cycles is lost after the second part. A full multi-scale approximation (including relaxation inside cycles) can be obtained by restoration of cycles. This is done starting from the acyclic auxiliary network $\mathcal{V}^n$ back to $\mathcal{V}^1$ through the hierarchy of cycles. Each cycle is restored according to the following procedure:

For each glued cycle node A_i^m, node of $\mathcal{V}^m$,

- Recall its nodes $A_{i1}^{m-1} \to A_{i2}^{m-1} \to ...A_{i\tau_i}^{m-1} \to A_{i1}^{m-1}$; they form a cycle of length τ_i.

- Let us assume that the limiting step in A_i^m is $A_{i\tau_i}^{m-1} \to A_{i1}^{m-1}$

- Remove A_i^m from $\mathcal{V}^m$

- Add τ_i vertices $A_{i1}^{m-1}, A_{i2}^{m-1}, ...A_{i\tau_i}^{m-1}$ to $\mathcal{V}^m$

- Add to $\mathcal{V}^m$ reactions $A_{i1}^{m-1} \to A_{i2}^{m-1} \to ...A_{i\tau_i}^{m-1}$ (that are the cycle reactions without the limiting step) with correspondent constants from $\mathcal{V}^{m-1}$

- If there exists an outgoing reaction $A_i^m \to B$ in $\mathcal{V}^m$ then we substitute it by the reaction $A_{i\tau_i}^{m-1} \to B$ with the same constant, i.e. outgoing reactions $A_i^m \to ...$ are reattached to the beginning of the limiting steps

- If there exists an incoming reaction in the form $B \to A_i^m$, find its prototype in $\mathcal{V}^{m-1}$ and restore it in $\mathcal{V}^m$

- If in the initial $\mathcal{V}^m$ there existed a "between-cycles" reaction $A_i^m \to A_j^m$ then we find the prototype in $\mathcal{V}^{m-1}$, $A \to B$, and substitute the reaction by $A_{i\tau_i}^{m-1} \to B$ with the same constant, as for $A_i^m \to A_j^m$ (again, the beginning of the arrow is reattached to the head of the limiting step in A_i^m)
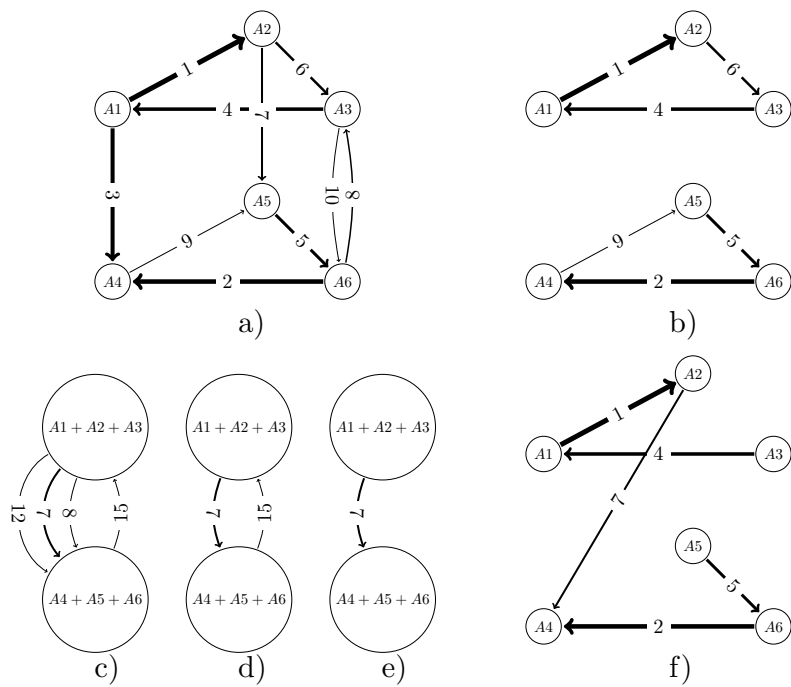
Figure 3: The successive steps of the reduction algorithm, illustrated for the prism model used in the paper. a) is the initial model; b) is the auxiliary network resulting from step I, pruning; c) is the result of gluing 3 species cycles and renormalizing the exit reactions (the constants of orders $3, 7, 10, 8$ are renormalized to $3 + 6 - 1 = 8, 7 + 6 - 6 = 7$, $10 + 6 - 4 = 12$, and $8 + 9 - 2 = 15$, respectively); d) is the auxiliary network after one more iteration; e) results from gluing and then restoring the 3 species cycles without the limiting step (constant of order 15); f) results from restoring the single species cycles without their limiting steps.

Appendix 3: Description of the TGFb model used in this paper. The model is described by the following system of differential equations

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = k_2 x_2 - k_1 x_1 - k_{16} x_1 x_{11}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = k_1 x_1 - k_2 x_2 + k_{17} k_{34} x_6$$

$$\frac{\mathrm{d}x_3}{\mathrm{d}t} = k_3 x_4 - k_3 x_3 + k_7 x_7 + k_{33} k_{37} x_{18} - k_6 x_3 x_5$$

$$\frac{\mathrm{d}x_4}{\mathrm{d}t} = k_3 x_3 - k_3 x_4 + k_9 x_8 - k_8 x_4 x_6$$

$$\frac{\mathrm{d}x_5}{\mathrm{d}t} = k_5 x_6 - k_4 x_5 + k_7 x_7 + 2k_{11} x_9 - 2k_{10} x_5^2 - k_6 x_3 x_5 + k_{16} x_1 x_{11}$$

$$\frac{\mathrm{d}x_6}{\mathrm{d}t} = k_4 x_5 - k_5 x_6 + k_9 x_8 + 2k_{13} x_{10} - 2k_{12} x_6^2 - k_{17} k_{34} x_6 + k_{31} k_{36} x_8 - k_8 x_4 x_6$$

$$\frac{\mathrm{d}x_7}{\mathrm{d}t} = k_6 x_3 x_5 - x_7 (k_7 + k_{14})$$

$$\frac{\mathrm{d}x_8}{\mathrm{d}t} = k_{14} x_7 - k_9 x_8 - k_{31} k_{36} x_8 + k_8 x_4 x_6$$

$$\frac{\mathrm{d}x_9}{\mathrm{d}t} = k_{10} x_5^2 - x_9 (k_{11} + k_{15})$$

$$\frac{\mathrm{d}x_{10}}{\mathrm{d}t} = k_{15} x_9 - k_{13} x_{10} + k_{12} x_6^2$$

$$\frac{\mathrm{d}x_{11}}{\mathrm{d}t} = k_{23} x_{14} - k_{30} x_{11}$$

$$\frac{\mathrm{d}x_{12}}{\mathrm{d}t} = k_{18} - x_{12} (k_{20} + k_{26}) + k_{30} x_{11} + k_{27} x_{15} - k_{22} k_{35} x_{12} x_{13}$$

$$\frac{\mathrm{d}x_{13}}{\mathrm{d}t} = k_{19} - x_{13} (k_{21} + k_{28}) + k_{30} x_{11} + k_{29} x_{16} - k_{22} k_{35} x_{12} x_{13}$$

$$\frac{\mathrm{d}x_{14}}{\mathrm{d}t} = k_{22} k_{35} x_{12} x_{13} - x_{14} (k_{23} + k_{24} + k_{25})$$

$$\frac{\mathrm{d}x_{15}}{\mathrm{d}t} = k_{26} x_{12} - k_{27} x_{15}$$

$$\frac{\mathrm{d}x_{16}}{\mathrm{d}t} = k_{28} x_{13} - k_{29} x_{16}$$

$$\frac{\mathrm{d}x_{17}}{\mathrm{d}t} = k_{31} k_{36} x_8 - k_{32} x_{17}$$

$$\frac{\mathrm{d}x_{18}}{\mathrm{d}t} = k_{32} x_{17} - k_{33} k_{37} x_{18}$$

These variables are as follows:

- Receptors on membrane: $x_{12} = \text{RI}$, $x_{13} = \text{RII}$, $x_{14} = \text{LR}$.

- Receptors in the endosome: $x_{11} = \text{LRe}$, $x_{15} = \text{RIe}$, $x_{16} = \text{RIIe}$.

- Transcription factors and complexes in cytosol: $x_1 = \text{S2c}$, $x_3 = \text{S4c}$, $x_5 = \text{pS2c}$, $x_7 = \text{pS24c}$, $x_9 = \text{pS22c}$, $x_{18} = \text{S4ubc}$.

- Transcription factors and complexes in the nucleus: $x_2 = \text{S2n}$, $x_4 = \text{S4n}$, $x_6 = \text{pS2n}$, $x_8 = \text{pS24n}$, $x_{10} = \text{pS22n}$, $x_{17} = \text{S4ubn}$.