Analyzing Pathways using SAT-based Approaches[⋆]

Ashish Tiwari, Carolyn Talcott, Merrill Knapp, Patrick Lincoln, and Keith Laderoute

SRI International, Menlo Park, CA 94025

Abstract. A network of reactions is a commonly used paradigm for representing knowledge about a biological process. How does one understand such generic networks and answer queries using them? In this paper, we present a novel approach based on translation of generic reaction networks to Boolean *weighted MaxSAT*. The Boolean weighted MaxSAT instance is generated by encoding the equilibrium configurations of a reaction network by weighted boolean clauses. The important feature of this translation is that it uses reactions, rather than the species, as the boolean variables. Existing weighted MaxSAT solvers are used to solve the generated instances and find equilibrium configurations. This method of analyzing reaction networks is generic, flexible and scales to large models of reaction networks. We present a few case studies to validate our claims.

1 Introduction

A network of reactions is a convenient way to represent knowledge about a biological process. Each reaction converts some reactants into products in the presence of certain other molecules. There is no single universal meaning, or a single formal semantics, that can be ascribed to the various reaction networks and pathways in the literature. Consequently, it is unclear how to build computational support for understanding and reasoning about large reaction networks.

A reaction network can be interpreted in various ways. They are often mapped onto a continuous dynamical system, where the dynamics are given by ordinary differential equations. These differential equations can be generated using different kinetic laws, such as Mass Action and Michaelis-Menten. However, it is not easy to experimentally determine, especially for biochemical reactions, the rate constants required to build the continuous dynamical system. As a result, fully specified and experimentally validated continuous dynamical system models are rarely available. Moreover, it has also been argued that the assumptions used to arrive at the differential equations may not be valid inside a biological compartment, where certain molecules may be few.

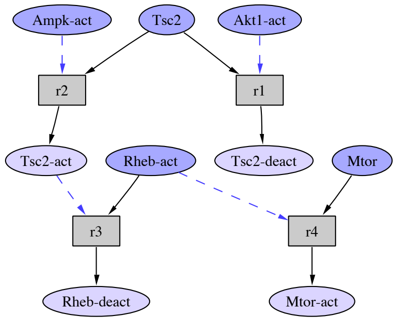
[⋆] This work was supported in part by Public Health Service grant GM068146-03 from the National Institute of General Medical Sciences and by the National Science Foundation under grants IIS-0513857 and CCR-0326540.

A reaction network can also be interpreted as a dynamical system over a discrete state space. In this case, the state space consists of mappings from the set of species to the natural numbers that specifies the number of molecules of each species. The dynamics over this state space can be defined either in continuous time (using a stochastic model) as a Chemical Master Equation, or in discrete time as a (standard or stochastic) Petri net. While such models are considered to be more accurate, they are difficult to analyze because of the horrendously huge state space. For example, when analyzing systems containing just 100 total molecules of 4 different species, the state space size is 4^{100}.

All continuous time models require reaction rates in some form. To overcome this requirement, discrete time models are considered that abstract time to a before-after relationship. When considered over the discrete state space mentioned above, a reaction network simply maps to a Petri net. Analyzing Petri nets is not easy. For instance, while Petri net reachability is decidable, there is no known upper-bound.

To overcome the state space problem, the discrete-time discrete-space models are further simplified. For instance, boolean models abstract species to being either *present* or *absent*. Other qualitative abstractions, such as *absent*, *present in low quantities*, and *present in large quantities* are also possible. In the absence of accurate detailed models, these abstract models have been found to be highly useful for representing and understanding biological knowledge.

In this paper, we present a new scalable approach for analyzing large reaction networks interpreted in the discrete-time and abstract discrete-space domain. There are three main features in our approach. First, it is based on qualitatively abstracting the *reactions* into two states–*on* and *off*. This is dual to the more conventional approach where the presence or absence of molecular species, and not *reactions*, is used to define the state of the system [4, 8, 9]. Second, it uses a boolean MaxSat as its backend engine. There is a generic translation from reaction networks to boolean MaxSat instances. Third, it is flexible. Clauses and their weights can be adjusted for reaction networks encoding specific aspects, such as signaling pathways, or transcriptional regulation.



Consider, for example, the very simple network shown here (not necessarily biologically accurate). This network consists of 4 reactions:

$$\begin{array}{lrcl} r_1: & \text{Tsc2} & \xrightarrow{\text{Akt1-act}} & \text{Tsc2-deact} \\ r_2: & \text{Tsc2} & \xrightarrow{\text{Ampk-act}} & \text{Tsc2-act} \\ r_3: & \text{Rheb-act} & \xrightarrow{\text{Tsc2-act}} & \text{Rheb-deact} \\ r_4: & \text{Mtor} & \xrightarrow{\text{Rheb-act}} & \text{Mtor-act} \end{array}$$

We will use this network as a running example in the paper.

It is not immediately obvious how to understand even this simple network. Using the approach described in this paper, all possible “steady-state” behaviors of the above network can be computed. For this example, the tool computes two possible behaviors. Either Akt1-act is present, deactivating Tsc2, while Mtor gets activated by Rheb-act (Reactions 1 and 4 are “on”); or, Ampk-act is present, activating Tsc2, which in turn deactivates Rheb-act (Reactions 2 and 3 are “on”). The important point here is that the steady-state behavior is thought of a subset of reactions that can be consistently “on”, as opposed to the traditional viewpoint where steady-state refers to species reaching some equilibrium concentrations.

As mentioned earlier, our approach is flexible and additional constraints can be added to specialize the search for certain steady-state configurations. We can specify an initial dish consisting of some of the species and search for most likely steady-state configurations resulting from the given initial dish. In the above example, if the initial dish only contains Tsc2, Ampk-act, Rheb-act and Mtor, then our tool identifies that the second and third reactions can be “on”, and that the other option, where reactions 1 and 4 are “on” is less likely. Similarly, target species can be specified, and the tool will generate paths (scenarios) that produce the target species. Each such scenario will be assigned a weight indicating its relative likelihood.

1.1 Motivation

The definition of “steady-state” behavior we use in this paper is nonstandard. Traditionally, a steady-state refers to all species in the network being at their equilibrium concentrations. In this paper, a steady-state refers to a subset of reactions that can be consistently “on”. This new definition is motivated by the observation that *signaling pathways* are best understood this way. More than the individual species concentrations, it is the chain of reactions that captures how information flows from the cell membrane to effect downstream activities in a cell. This chain of reactions corresponds directly to the notion of a steady-state in our approach.

The different reactions in the steady-state chain of reactions will, in reality, be temporally separated. While certain phosphorylation activity may occur in a few minutes after a cell is hit by ligands, other downstream activities may occur much later. In our approach, we identify the whole chain as one possible steady-state behavior of the reaction network. The complete chain of reactions may never simultaneously be “on” in reality. However, they are still useful in understanding the function of a given complex reaction network.

The approach based on translation to MaxSat is motivated by the need for flexibility. Reaction networks have slightly different meaning in different contexts. Metabolic pathways, signaling pathways, and transcriptional regulation networks work on different notions of species and reactions. Our basic semantics attempts to capture the minimal common meaning that can be ascribed to any such network. The weights on the MaxSat instance give flexibility in making certain constraints harder than others in different contexts.

Finally, it should be mentioned that the technology for solving SAT and MaxSAT problems has made significant advances in recent years and problems with thousands of boolean variables and even more clauses are routinely solved in a few seconds. We have used our tool on the HumanCyc database of metabolic pathways (containing over a thousand reactions) and we can answer queries in a few seconds.

2 Reaction Networks

In this section, we formalize our terminology. A *species* is a generic name used to denote any entity, such as a molecule, ion, protein, enzyme, ligand, receptor, complex, or a postranscriptionally modified form of a protein. We do not differentiate between these different roles and just formally identify a species with a unique name. The set of all species will be denoted by S. A *reaction* consists of a set of *reactants*, a set of *modifiers*, and a set of *products*. Thus, a reaction r is a 3-tuple $\langle R, M, P\rangle$, where R, M, P are pairwise disjoint subsets of S. Given a reaction r, we denote its set of reactants, modifiers, and products by $R(r)$, $M(r)$, and $P(r)$ respectively. Given a species s, the set of reactions in which s occurs as a reactant (modifier, product) is denoted by $R^{-1}(s)$ (respectively, $M^{-1}(s)$, $P^{-1}(s)$).

A *network* $\mathcal{N}$ is a collection of reactions. A *network instance* is a network together with an optional set of *input* species, a set of *forbidden* species, and a set of *target* species.

A *pathway* is a special kind of network. Informally, a pathway contains a *related* set of reactions that can be consistently switched "on". The following sections will formally define the constraints we impose to identify pathways.

2.1 Semantics of Reaction Networks

As mentioned in the introduction, motivated by the need to handle unknown model parameters while maintaining computational feasibility of analysis, we use a discrete-time abstract discrete-state semantics of reaction networks. The key aspect of our semantics is that we introduce a boolean variable for each reaction (and not for each species). Thus, the semantics of a biochemical network $\mathcal{N} = \{r_1, r_2, \ldots, r_n\}$ with n reactions is given as a state transition system defined over n boolean variables $b_1, \ldots, b_n$, where the i-th boolean variable b_i represents whether the i-th reaction r_i is "on" or "off".

Let $\mathit{present4i}(s, i)$ denote the formula $\bigvee_{r_j \in P^{-1}(s)} b_j \;\wedge\; \bigwedge_{r_j \in R^{-1}(s), j \neq i} \neg b_j$, which means some reaction that produces s is "on" and every reaction other than r_i that consumes s is "off". Intuitively, $\mathit{present4i}(s, i)$ represents the availability of species s for reaction r_i. The transitions of the state transition system are given by nondeterministically applying one of the following $2n$ guarded commands:

$$\neg b_i \wedge \bigwedge_{s \in R(r_i) \cup M(r_i)} \mathit{present4i}(s, i) \longrightarrow b_i' := \mathit{true}$$

$$b_i \wedge \bigvee_{s \in R(r_i) \cup M(r_i)} \neg \mathit{present4i}(s, i) \longrightarrow b_i' := \mathit{false}$$

The first guarded command says that if a reaction r_i is "off", but each of its reactants and modifiers is "present" (for r_i), then it can be turned "on". The second guarded command says that if a reaction r_i is "on", but one of its reactants or modifiers is not present (for r_i), then it can be turned "off".

3 Biochemical Networks to Boolean SAT

In this section, we describe the procedure that generates a set of boolean constraints from a network. The boolean constraints represent the equilibrium configurations of the network in the semantics given above. Later in this section, we describe the additional constraints that are generated from a network instance.

An equilibrium state is defined as a state in which none of the $2n$ guarded transitions are enabled. Hence, if a state $(b_1, \ldots, b_n)$ is an equilibrium state of the above state transition system, then it should be the case that, for all i,

$$\neg(\neg b_i \ \wedge \bigwedge_{s \in R(r_i) \cup M(r_i)} \mathit{present4i}(s,i)) \ \wedge \ \neg(b_i \ \wedge \bigvee_{s \in R(r_i) \cup M(r_i)} \neg \mathit{present4i}(s,i))$$

This is equivalent to saying that for all i,

$$b_i \ \Leftrightarrow \bigwedge_{s \in R(r_i) \cup M(r_i)} \mathit{present4i}(s,i) \tag{1}$$

Any boolean assignment that satisfies these constraints is an equilibrium state of the given reaction network. In the implementation (Section 5), we break up the constraint in Formula 1 into the following constraints to enable the MaxSat solver to partially satisfy these constraints.

$$b_i \Rightarrow \bigwedge_{s \in R(r_i) \cup M(r_i)} \ \bigvee_{r_j \in P^{-1}(s)} b_j \tag{2}$$

$$b_i \Rightarrow \bigwedge_{s \in R(r_i)} \ \bigwedge_{r_j \in R^{-1}(s), j \neq i} \neg b_j \tag{3}$$

$$b_i \Rightarrow \bigwedge_{s \in M(r_i)} \ \bigwedge_{r_j \in R^{-1}(s), j \neq i} \neg b_j \tag{4}$$

$$\neg b_i \Rightarrow \neg \bigwedge_{s \in R(r_i) \cup M(r_i)} \mathit{present4i}(s,i) \tag{5}$$

Formula 2 captures the rule that if a reaction is "on", then each of its reactants and modifiers is produced by some "on" reaction. Formula 3 encodes the inhibitory effect that a reaction may have on another that shares a reactant with it by saying that if a reaction is "on", then none of its reactants is consumed (used as a reactant) by any *other* reaction. Formula 4 encodes the competitive inhibition between reactions through a species that is a reactant in one reaction and a modifier in another. Note that if two reactions share a modifier, then they do not inhibit each other. Finally, Formula 5 encodes that if all reactants and modifiers of a reaction are present, then it should be "on".

3.1 Completing the Network

Biological databases of biochemical networks are often incomplete. They often use species that are not created by any reaction in the network. In the running example, Tsc2, Akt1-act, Ampk-act, Rheb-act, and Mtor are all species with no producers. The presence of such species is a problem for our encoding since, to be "on", a reaction requires all of its reactants (and modifiers) to be produced by some other reaction. If there are no producers of certain species, then reactions using that species can never be turned on.

We solve this problem by adding dummy reactions that create species that have no producers. Specifically, for each species s such that $P^{-1}(s) = \emptyset$, we add a new reaction $r = \langle R, M, P \rangle$, where $R = \emptyset$, $M = \emptyset$, and $P = \{s\}$. We perform this step as a preprocessing step. As a result, these additional dummy reactions are taken into account when the constraints given in Formula 1 are generated.

We also encode the fact that these dummy reactions are different from other reactions by adding boolean constraints that force these dummy reactions to be "off". For each dummy reaction r, if b is the corresponding boolean variable, then we add the following clause

$$\neg b \tag{6}$$

This constraint says that the dummy reaction, and hence the corresponding species, should *preferably* not be used. In Section 4, we will discuss how this preference is effected by means of weights.

In the running example, for each of the 5 species that have no producers, we add one new dummy reaction. Thus, we have new dummy reactions $r_5, \ldots, r_9$ that respectively produce Tsc2, Akt1-act, Ampk-act, Rheb-act, and Mtor. Thus the complete network has 9 reactions, and hence, the boolean encoding will be over 9 boolean variables $b_1, \ldots, b_9$. The constraints given by Formula 1 will be:

$$\begin{array}{ll} b_1 \Leftrightarrow (b_5 \wedge \neg b_2) \wedge (b_6) & \quad b_3 \Leftrightarrow (b_8) \wedge (b_2) \\ b_2 \Leftrightarrow (b_5 \wedge \neg b_1) \wedge (b_7) & \quad b_4 \Leftrightarrow (b_9) \wedge (b_8 \wedge \neg b_3) \end{array}$$

Additionally, we will also get boolean constraints $\neg b_5, \neg b_6, \ldots, \neg b_9$ coming from Formula 6. Note that Reaction r_3 requires the modifier Tsc2-act, which is produced by Reaction r_2. This gets reflected as $b_3 \Rightarrow b_2$ above. As an example of competitive inhibition, note that Reaction r_1 and Reaction r_2 share a common reactant, namely Tsc2. This shows up as $b_1 \Rightarrow \neg b_2$ and $b_2 \Rightarrow \neg b_1$. Similarly, Reaction r_3 and Reaction r_4 compete for Rheb-act—Reaction r_3 uses it as a reactant, whereas Reaction r_4 requires it as a modifier. This generates the constraint $b_4 \Rightarrow \neg b_3$.

3.2 Optional Clauses

In case of analyzing a network *instance*, we may optionally have additional information about the input species, forbidden species, and target species. We now show how these are incorporated into the constraints.

Initial Species The set of species specified as initial are assumed to be present. If a set of initial species is specified, then the preprocessor adds a dummy reaction that produces all the initial species. Specifically, if S_{init} is the set of initial species, then the preprocessor will add a dummy reaction $r = \langle R, M, P \rangle$, where $R = M = \emptyset$ and $P = S_{init}$. Furthermore, the boolean variable b corresponding to this reaction is forced to be "on" by simply adding a clause b in the generated set of boolean constraints. If some initial species are specified, then the initial dummy reaction is added to the network *before* the network is completed (Section 3.1). Hence, fewer dummy reactions get added in the network completion phase if some of the species with no producers in the network are assumed to be in the initial soup.

Target Species The set of target species is a list of species that should be *present* in the equilibrium configurations generated by the tool. If a set of target species is specified, then the boolean constraint generator adds additional constraints that say that for each target species, there is at least one producer of it turned "on".

For each species s in the set of target species, we add the constraint,

$$\bigvee_{r_i \in P^{-1}(s)} b_i \tag{7}$$

Forbidden Species The set of forbidden species specifies the set of species that should not be used in any equilibrium configuration generated by the system. If this set is provided, then the following additional boolean constraint is generated for each species s in this forbidden set,

$$\bigwedge_{r_i \in P^{-1}(s)} \neg b_i \tag{8}$$

3.3 Mode Based Constraints

Given the above constraints, we can try to turn "on" as many reactions as possible, or turn "on" as few reactions as possible. These two possibilities are encoded as two different sets of constraints.

If we wish to turn "on" as many reactions as possible, then, for each reaction $r_i \in \mathcal{N}$, we add the clause

$$b_i \tag{9}$$

to the set of constraints. This clause simply says that reaction r_i is "on".

If we wish to turn "on" as few reactions as possible (say to find minimal pathways), then, for each reaction $r_i \in \mathcal{N}$, we add the clause

$$\neg b_i \tag{10}$$

to the set of constraints.

4 Biochemical Pathway to Boolean Max-SAT

The constraints outlined above are not all equally important. This is captured by adding a weight (number) to each constraint that indicates its relative importance.

In particular, constraints obtained by instantiating Formula 2, Formula 3, and Formula 5 are each given a very large weight W. In the current implementation, W is equal to the total number of reactions in the completed network. The constraint represented in Formula 4 is given weight equal to $W/2$ since competitive inhibition between reactions via a species that is a modifier in one reaction and a reactant in another is intuitively weaker than the inhibition via shared reactants. The constraint saying that species with no producers should not be used (Formula 6) is given intermediate weight (approximately $W/(k+1)$, where k is the total number of species with no producers). Whenever present, the constraint for creation of target species (Formula 7) is given weight W. The constraints that specify the hints (Formula 9 and Formula 10) are given weight 1.

The choice of weights for each constraint gives additional flexibility that can be used, in the future, to encode other biologically relevant information that is not generic to all biochemical processes.

4.1 Weighted MaxSAT

A *solution* is a mapping from the boolean variables to $\{true, false\}$. In our context, a solution maps reactions to either "on" or "off". Under a given solution, constraints also evaluate to either *true* or *false*.

Each solution can be associated with a weight: the sum of the weights of all the constraints that are made *true* by that solution. A *weighted MaxSAT solver* finds a solution that has the maximum weight.

In our running example, using the above rule for assigning weights (we do not break Formula 1 into smaller parts and assign it a weight $W = 9$ for simplicity here), we get the following weighted basic constraints:

c_1 :	$b_1 \Leftrightarrow (b_5 \wedge \neg b_2) \wedge (b_6)$	$w_1 = 9$	c_2 :	$b_3 \Leftrightarrow (b_8) \wedge (b_2)$	$w_2 = 9$
c_3 :	$b_2 \Leftrightarrow (b_5 \wedge \neg b_1) \wedge (b_7)$	$w_3 = 9$	c_4 :	$b_4 \Leftrightarrow (b_9) \wedge (b_8 \wedge \neg b_3)$	$w_4 = 9$
c_5 :	$\neg b_5$	$w_5 = 1$	c_6 :	$\neg b_6$	$w_6 = 1$
c_7 :	$\neg b_7$	$w_7 = 1$	c_8 :	$\neg b_8$	$w_8 = 1$
c_9 :	$\neg b_9$	$w_9 = 1$	c_{10} :	b_1	$w_{10} = 1$
c_{11} :	b_2	$w_{11} = 1$	c_{12} :	b_3	$w_{12} = 1$
c_{13} :	b_4	$w_{13} = 1$			

Note that the 5 constraints, $c_5, \ldots, c_9$, encode the fact that the five species with no producers can be used by paying a small penalty; and the last 4 constraints say that each reaction should preferably be turned "on".

For this set of constraints, the solution in which all b_i are *false* has a weight 41 (since only $c_{10}, \ldots, c_{13}$ are violated). The solution $b_1 = b_4 = b_5 = b_6 = b_8 =$

$b_9 = \mathit{true}$ (and the rest *false*) has weight 39; and the solution $b_2 = b_3 = b_5 = b_7 = b_8 = \mathit{true}$ has weight 40. These three solutions are the top three maximum weight solutions. The latter two correspond exactly to the two scenarios described in Section 1. The first solution captures the scenario where no reaction is "on", which can be eliminated by using a nonempty initial set of species that includes (some of) the 5 species with no producers.

5 Implementation and Case Studies

We have implemented a tool based on the technique described in this paper. As a backend MaxSAT solver, we use Yices [14, 3], which is a more general *satisfiability modulo theory* solver. The input format for our tool is a network or network instance described in a very simple intermediate language. We also have several front-ends that convert from other formats to our intermediate language format. For example, we have front-ends for Pathway Logic [12, 11] and BioCyc [5, 7].

In this section, we describe the results obtained using this tool on some specific networks.

5.1 Sporulation Initiation in *B.Subtilis*

Bacillus subtilis is considered a model organism for Gram-positive bacteria and has been extensively studied in the laboratory. It is an endospore-forming bacteria most commonly found in the soil. Endospore formation is initiated when nutrients become limiting and is an adaptive response of the bacteria to their environment.

Sporulation is a one-way decision and once the decision is made, the cell undergoes changes which take 6 to 8 hours in most organisms. If conditions improve in the meantime, then the cell will be at a disadvantage. Hence the decision to initiate sporulation is important to the organism and is subject to a variety of control.

The formation of spores in Bacillus subtilis is a developmental process under genetic control. The decision to either grow vegetatively or sporulate is regulated by the state of phosphorylation of the Spo0A transcription factor [10, 6]. Spo0A obtains its phosphate through a phosphorylation pathway (see Figure 1), the so-called *phosphorelay*, in which at least three histidine protein kinases transfer phosphate to the relay protein, Spo0F, then to Spo0B, and finally to Spo0A (represented by ReactionIDs $r17$, $r19$, and $r20$ in Table 1). In addition, the phosphorylation state of Spo0A is modulated by specific phosphatases, such as Spo0E, which dephosphorylates Spo0A-P, and RapA, which dephosphorylates Spo0F-P (ReactionIDs $r18$, $r21$).

The *SinI* and *SinR* pair is a regulatory operon in the sporulation initiation network. While SinR is a transcriptional regulator that represses *spo0A* transcription, SinI disrupts the SinR tetramer through the formation of a SinI-SinR

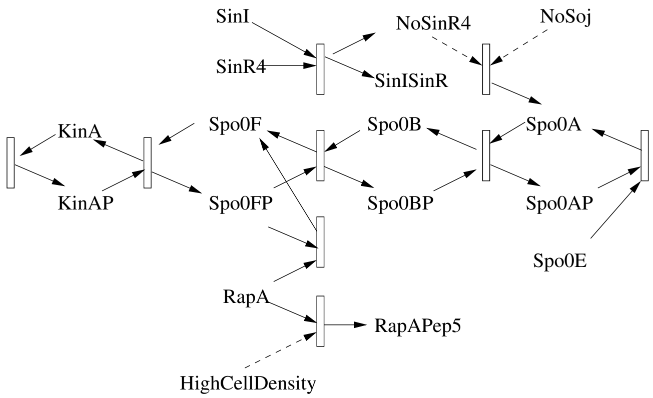


Fig. 1. Selected reactions from the sporulation initiation network of *B. Subtilis*. The reactions are represented using standard Petri net notation and show the main phosphorelay.

heterodimer. This aspect, along with the logic regulating SinI transcription, is encoded in ReactionIDs $r1$, $r2$, $r3$, and $r4$.

The activity of protein RapA is modulated by quorum sensing, the process of sensing activity in neighboring cells and reacting in a cell-density-specific fashion. Under high population density, RapA is inhibited by PhrA pentapeptide (not modeled in the reactions). These aspects are captured in ReactionIDs $r13$, $r15$. The protein kinase KinA is a sensor that initiates the phosphorelay and is modeled here by ReactionIDs $r16$, $r17$. Most of the remaining reactions encode transcriptional regulation logic for different proteins.

On this simplified model of sporulation initiation, the tool implementing the approach described in this paper can find possible stable behaviors of the network. These behaviors are found as subsets of reactions in the network that can be consistently "on". The tool finds 3 different possibilities for the model above.

- SinI is produced, and it binds to SinR, thus preventing it from repressing *spo0A*. RapA is converted to RapAPep5, thus preventing it from dephosphorylating Spo0A-P. In the presence of stress signals, KipI is prevented from inhibiting KinA from self-kinasing. The self-kinasing of KinA triggers the phosphorelay, which leads to production of Spo0A-P, a precursor for sporulation.
- In the second stable state scenario, RapA dephosphorylates Spo0F-P, thus breaking the phosphorelay chain. Thus, there is no production of Spo0A-P.
- The third stable state scenario is similar to the first, except that Spo0E dephosphorylates the produced Spo0A-P, thus using up the produced Spo0A-P.

The three stable scenarios each make different assumptions about the environment. In our case, the environment consists of the species that are not created

ID	Reactants	+Modifiers	⟶Products
r1		+(Spo0AP, NoSinR4)	⟶SinI
r2		+(Spo0AP, NoAbrB6, NoHpr)	⟶SinI
r3	SinI, SinR4	+	⟶SinISinR, NoSinR4
r4	SinR	+	⟶SinR4
r5		+(NoSinR4, sigmaH, NoSoj)	⟶Spo0A
r6		+(NoAbrB6)	⟶Spo0E
r7	AbrB, AbrB6	+(Spo0AP)	⟶NoAbrB6
r8		+(NoSpo0AP)	⟶AbrB
r9		+(NoAbrB6)	⟶AbrB
r10	AbrB, NoAbrB6	+	⟶AbrB6
r11	NoHpr	+(AbrB6)	⟶Hpr
r12	Hpr	+(NoAbrB6)	⟶NoHpr
r13		+(ComAP)	⟶RapA
r14	RapA	+(Spo0AP, Hpr)	⟶
r15	RapA	+(HighCellDensity)	⟶RapAPep5
r16	KinA	+(NoKipI)	⟶KinAP
r17	KinAP, Spo0F	+	⟶Spo0FP, KinA
r18	Spo0FP, RapA	+	⟶Spo0F
r19	Spo0FP, Spo0B	+	⟶Spo0BP, Spo0F
r20	Spo0A, Spo0BP	+(NoSoj)	⟶Spo0AP, Spo0B
r21	Spo0AP, Spo0E	+	⟶Spo0A, NoSpo0AP
r22		+(sigmaH, sigmaA)	⟶Spo0F
r23		+(sigmaA)	⟶Spo0B
r24	KipI	+(NoFood, NoNitrogen)	⟶NoKipI

Table 1. The list of reactions modeling the sporulation initiation network.

by any of the reactions in the network. In the network above, HighCellDensity, and NoFood, are two examples of *input species.*

The tool can also be used in the mode in which a desired target set of species is specified (for example, Spo0A-P). In this case, the tool will generate the first stable scenario above to show how Spo0A-P could be produced.

5.2 MAPK Signaling Network

The Mitogen-Activated Protein kinase (MAPK) network regulates several cellular processes, including the cell cycle machinery. The MAPK cascade communicates signals from growth factors that bind receptor kinases to transcription and other cellular processes [2]. A simplified model of this network, taken from [2], can be encoded in our notation as shown in Table 2. The tool finds two stable sets of behavior for this network.

– The positive feedback loop is active. In this case, either Grb2, Sos1, or PKC* turns on Ras. This causes, in steps, the phosphorylation of Raf, MEK, and Erk. Activated Erk causes production of AA*, which stimulates PKC.

ID	Reactants+Modifiers		⟶Products
r1	Ras	+(Grb2, Sos1)	⟶Ras*
r2	Ras	+(PKC*)	⟶Ras*
r3	Raf	+(Ras*)	⟶Raf*
r4	Raf*	+(PP2A)	⟶Raf
r5	Mek	+(Raf*)	⟶Mek*
r6	Mek*	+(PP2A)	⟶Mek
r7	Erk	+(Mek*)	⟶Erk*
r8	Erk*	+(MKP)	⟶Erk
r9		+(Erk*, MKPgene)	⟶MKP
r10	AA	+(Erk*, Ca)	⟶AA*
r11	PKC	+(DAG, Ca, AA*)	⟶PKC*

Table 2. The list of reactions modeling the MAPK signaling network.

– The negative feedback loops are active. In this case, protein phosphatase 2A (PP2A) dephosphorylates both Raf* and Mek*, and MKP dephosphorylates Erk*. MKP is created by transcription of *MKP* gene, and this is promoted by Erk*.

The two stable solutions clearly identify the positive cycle and the multiple negative cycles that break the positive cycle. The overall system behavior is seen to be a result of the close interaction between the positive and negative cycles.

We also used the detailed model of the MAPK signaling network from [1]. The total running time on the full network is of the order of a few seconds.

5.3 EGF Stimulation Network

In the Pathway Logic project [12, 11], a model of Egf stimulation is being developed by curating a network of biochemical reactions involved in mammalian cell signaling from the literature. When a cell is stimulated by Egf, certain species are experimentally observed to be present in the cell after its initial stimulation. These observations can be used to validate the model by checking whether the model predicts the observations. To carry out the validation, we started with a network of about 400 reactions and created a network instance by adding initial and target species. Specifically, we started with a set of about 250 *initial* species and 62 *target* species that are experimentally observed in response to EGF stimulation.

When this network instance is analyzed by our tool, our tool attempts to find a set of reactions that will create each of the target species using the initial species and the reactions in the network. A "–no-assume" option tells the tool to not assume any species not already specified in the initial set. (Recall that, by default, species that have no producers can be assumed, with a moderate penalty.)

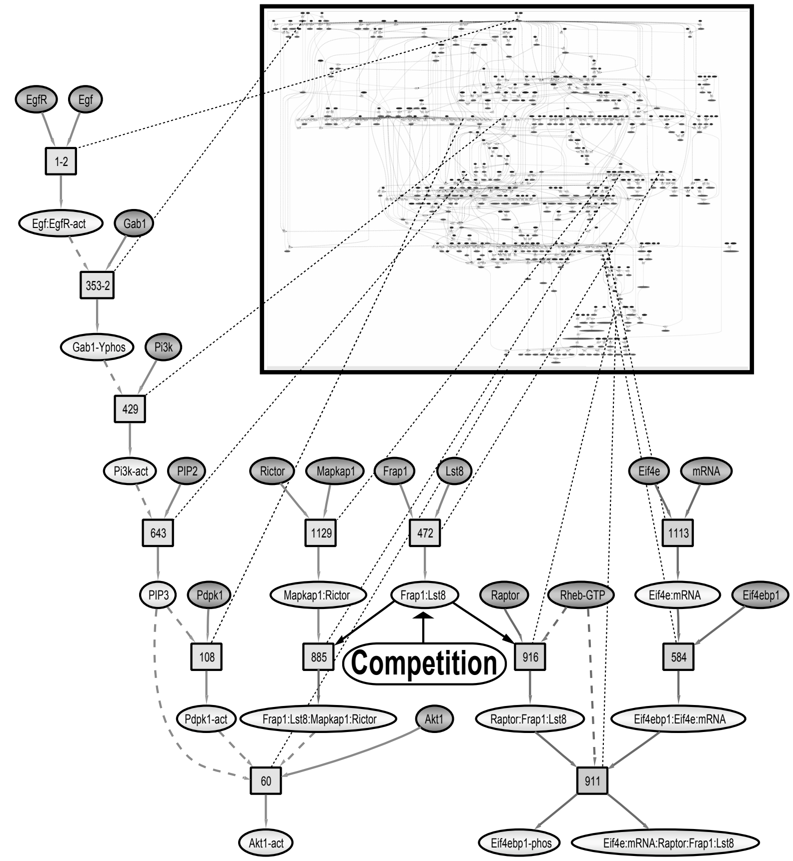


Fig. 2. A simple network with competing rules

The output of the tool indicated that it was not possible to find a solution without violating one Type 3 and one Type 4 *competitive inhibition* constraints. Specifically, the species (Frap1:Lst8)-CLc [1] is a reactant in two different reactions that are *both* required to be "on" to create the target species. This causes a Type 3 constraint to be violated. The Type 4 constraint that is violated is caused by the species Src-CLi, which is used as a reactant in a reaction to create Src-act-CLi, and it is also used as a modifier in the reaction that creates Cbl-Yphos-CLi. This violation pointed out a typing error in specifying the reaction rules which has been corrected. Figure 2 shows the pathways competing for (Frap1:Lst8)-CLc in the context of the larger network.

Using our tool provided two valuable forms of feedback to the model developer. One was a form of meta analysis or type-checking that detected syntactic

[1] a complex containing Frap1 and Lst8 located in the cytoplasm,CLc

problems with the model. (The first pass detected a number of inconsistencies that were easily repaired.) The second was the identification of the point of competition. Using the Pathway Logic Assistant [13] one can check whether a given set of observations is predicted, singly or jointly. However if a prediction fails there is no feedback as to the cause of failure. Using MaxSAT, candidate conflicting constraints can be identified to guide the modeler.

Starting with the discovered Type 3 violation and studying the subnetwork connected to this reaction lead to two hypotheses: (1) (Frap1:Lst8)-CLc splits into two *populations* one for each of the two competing reactions; (2) there is a feedback loop that can reset the state of (Frap1:Lst8)-CLc and the system oscillates between the two pathways. Experiments are ongoing to test these hypotheses.

6 Related Work

We compare here with work that is closer in spirit to our work, and do not mention all the literature devoted to building various kinds of models and improving understanding of specific biological phenomena, such as sporulation and MAPK signaling.

Senachak et. al. [8] give a generic interpretation to a reaction network by translating it to a graph. Strongly-connected components of the graph are related to the pathways. The construction of the graph has some unusual steps, such as *cascading*, that arise primarily because the authors use species as defining the nodes of the graph. The main difference in our approach is that, in our approach, the boolean variables correspond to reactions in the network. We believe this leads to a much simpler and natural encoding of the "cascading"-style constraints of [8].

7 Conclusion

We presented a new approach for analyzing biochemical reaction networks using MaxSAT. The novelty here is that we make reactions central to the notion of a steady-state behavior. A steady-state behavior is a subset of reactions that can be mutually and consistently "on".

The attractiveness of our approach is that it is generic and applies to networks coming from different kinds of biological networks. Additionally, it is also flexible and allows encoding of knowledge specific to certain kinds of networks via suitable manipulation of the weights on the generic constraints.

The analysis approach is promising. Even for the largest networks we have studied, the analysis takes at most a few seconds to compute answers.

Possible future work include studying quantitative variants of the boolean constraints. Fortunately, our backend tool, Yices, supports reasoning over linear arithmetic constraints. We can replace the use of boolean MaxSAT with MaxSAT over arbitrary combination of boolean and linear arithmetic constraints.

Acknowledgments: We thank the referees for helpful suggestions.

References

1. U. S. Bhalla and R. Iyengar. Robustness of the bistable behavior of a biological signalling feedback loop. *Chaos*, 11(1), 2001.
2. U. S. Bhalla, P. T. Ram, and R. Iyengar. MAP kinase phosphatase as a locus of flexibility in a Mitogen-Activated Protein kinase signaling network. *Science*, 297, 2002.
3. B. Dutertre and L. M. de Moura. A fast linear-arithmetic solver for dpll(t). In *Computer Aided Verification, 18th International Conference*, volume 4144 of *LNCS*, pages 81–94. Springer, 2006.
4. F. Fages, S. Soliman, and N. Chabrier-Rivier. Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. *Journal of Biological Physics and Chemistry*, 4(2):64–73, 2004.
5. I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp. EcoCyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33:D334–7, 2005.
6. L. M. Prescott, D. A. Klein, and J. P. Harley. *Microbiology.* McGraw-Hill, 2002.
7. P. Romero, J. Wagg, M.L. Green, D. Kaiser, M. Krummenacker, and P.D. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(R2):1–17, 2004.
8. J. Senachak, M. Vestergaard, and R. Vestergaard. Rewriting game theory and protein signalling in MAPK cascades. In *Proc. CMSB*, 2006.
9. Carron Shankland, Nam Tran, Chitta Baral, and Walter Kolch. Reasoning about the ERK signal transduction pathway using BioSigNet-RR. In G. Plotkin, editor, *Proceedings of the Third International Conference on Computational Methods in System Biology*, 2005.
10. P. Stragier and R. Losick. Molecular genetics of sporulation in bacillus subtilis. *Annu. Rev. Genet.*, 30:297–341, 1996.
11. C. Talcott, S. Eker, M. Knapp, P. Lincoln, and K. Laderoute. Pathway logic modeling of protein functional domains in signal transduction. In *Proceedings of the Pacific Symposium on Biocomputing*, January 2004.
12. Carolyn Talcott. Symbolic modeling of signal transduction in pathway logic. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, editors, *2006 Winter Simulation Conference*, 2006.
13. Carolyn Talcott and David L. Dill. Multiple representations of biological processes. *Transactions on Computational Systems Biology VI*, 4220:221–245, 2006.
14. `http://yices.csl.sri.com/`. Yices home page.