

Comparative Statistical Analysis of Qualitative Parametrization Sets

Adam Streck^(✉), Kirsten Thobe, and Heike Siebert

Freie Universität Berlin, Berlin, Germany
adam.streck@fu-berlin.de

Abstract. The problem of model parametrization is a core issue for all varieties of mathematical modelling in biology. This problem becomes more tractable when qualitative modelling is used, since the range of parameter values is finite and consequently it is possible to enumerate and evaluate all possible parametrizations of a model. If such an approach is undertaken, one usually obtains a vast set of parametrizations that are scored for various properties, e.g. fitness. The usual next step is to take the best scoring parametrization. However, as noted in recent works [1,4], there is knowledge to be gained from examining sets of parametrizations based on their scoring. In this article we extend this line of thought and introduce a comprehensive workflow for comparing such sets and obtaining knowledge from the comparison.

Keywords: Qualitative modelling · Statistical inference · Big data · Parameter identification · Data mining

1 Introduction

One of the key tasks in the field of systems biology is reverse engineering of regulatory and signalling networks [6]. A researcher is usually presented with sets of experimental data and observations and tries to design a model of the mechanics of the system. As the model can be constructed using various modelling frameworks, there is a zoo of methods for tasks like network inference, parameter identification etc., each having its particular set of pros and cons. In our work we are employing the so-called Thomas Networks [17] framework, whose main purpose is to provide insights into qualitative, high-level behaviour. A particular feature of this framework is that the values governing the behaviour of the model—its parameters—have a finite domain and thus it is possible to enumerate and evaluate all the options. At this point, two additional problems arise. Firstly, when evaluating a *parametrization* (a particular set of parameter values) of the network, one is usually focusing again only on some abstract, qualitative feature, e.g. whether the system is stable or not. While this is useful information, its binary nature means that the set of all possible parametrizations is simply split in two, one part having the feature, the other not. This poses a problem if one is aiming to pick an optimal parametrization, as all the members in each of

the two sets are between themselves indistinguishable. Secondly, considering all the options leads to a rapid combinatorial explosion and while quite huge sets can still be easily manipulated and stored by the computer, it becomes swiftly infeasible for the researcher to keep a mental insight into the structure of the parametrization pool.



Fig. 1. Our workflow starts with the enumeration (1) of all possible parametrizations that fit the expectations of the modeller about the structure. Each of the parametrizations is evaluated for certain properties (2) like dynamical behaviour, and the result of the evaluation is stored with the parametrization as its *label*. After parametrizations are labelled the user can select (3) a subset of these that seem of special interest. This selection can then be analysed using various tools (4) and compared to other selections (5). The selection or analysis can then be refined based on the newly gained knowledge.

Similarly to recent works of other authors in the field [1,4] we propose to shift the focus from individual parametrizations to sets thereof. Following on our previous work [5,16], we introduce a unified workflow for parameter identification, illustrated in Fig. 1. This workflow combines formal and statistical methods with the aim of maximizing the amount of qualitative knowledge obtained from data. All the methods presented in the article have been implemented in the tool TREMPPI, whose preliminary version is available at [13]. All of the methods are illustrated on a toy running example and later the functionality of our workflow is demonstrated on a case study of Hepatocyte Growth Factor (HGF) signalling, based on data of [2].

2 Background

In this section we define the notions necessary for our workflow. Most of the terms are illustrated in Fig. 2 on a toy example.

The topology of a biological system is encoded as a directed *regulatory graph* (RG) $G = (V, \rho, E)$ where V is a set of named *components*, $\rho : V \rightarrow \mathbb{N}$ is the *maximum activity* label s.t. each component can adapt an integer from $[0, \rho(v)]$, denoting its current *activity level*, and $E \subseteq V \times \mathbb{N} \times V$ is a set of *regulations* s.t. for each $(u, t, v) \in E$ it holds that $t \leq \rho(u)$. For $(u, t, v) \in E$ the value t denotes a *threshold* i.e. the lowest *activity level* of u at which the regulation can affect v . Additionally, we introduce a threshold function $\theta : V \times V \rightarrow 2^{\mathbb{N}}$ s.t. $\theta(u, v) = \{t \mid (u, t, v) \in E\}$ and its extended version Θ s.t. $\Theta(u, v) = \theta(u, v) \cup \{0, \rho(u) + 1\}$ for any pair $u, v \in V$. Moreover, if $(u, t, v) \in E$ then $t_-, t_+ \in \Theta(u, v)$ denote the closest lower and higher element of t , i.e. have \uparrow^{Θ} the ordinal successor function in Θ , then $\uparrow^{\Theta}(t_-) = t$ and $\uparrow^{\Theta}(t) = t_+$. A regulation becomes effective when the

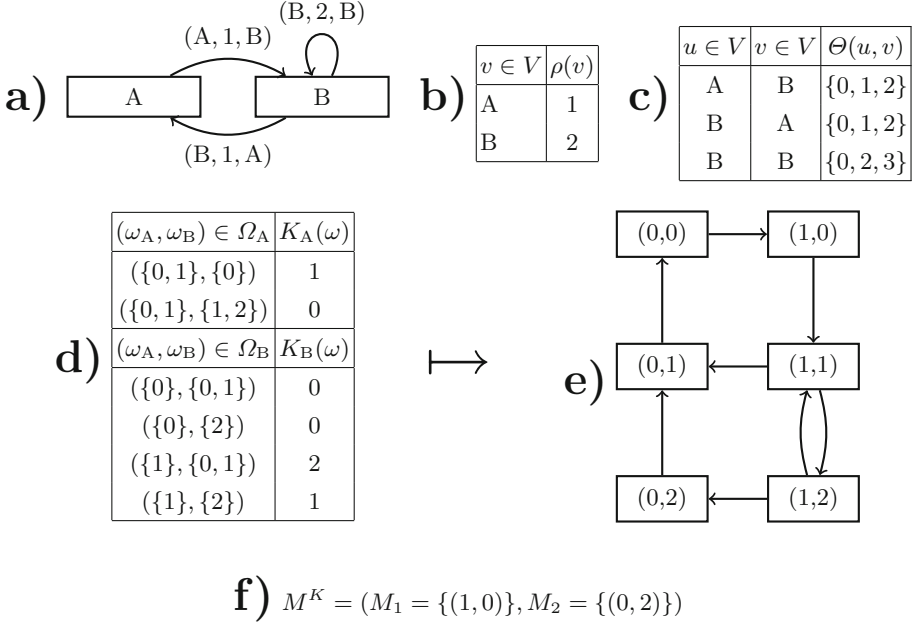


Fig. 2. A simple Thomas Network. **a)** The regulatory graph, **b)** its components, and **c)** regulators. **d)** One of the 324 possible parametrizations of the network. **e)** The asynchronous dynamics encoded by the parametrization. **f)** An example property. Note that the property is satisfied by K , i.e. $((1, 0), (1, 1), (1, 2), (0, 2)) \models M^K$.

activity of the respective regulator reaches the threshold value. Consequently, the thresholds divide the range of activity levels of a component into so-called activity intervals $I_v^u = \{[t, \uparrow^\Theta(t)) \mid t \in \theta(u, v) \cup \{0\}\}$. Note that if $\theta(u, v) = \emptyset$ then $I_v^u = [0, \rho(u) + 1)$. For each component we can then create a set of configurations of components of the system, called *regulatory contexts*, where the behaviour of a component $v \in V$ can qualitatively differ from the other contexts, denoted and defined $\Omega_v = \prod_{u \in V} I_v^u$. Note that $\omega \in \Omega_v$ is a vector of length $|V|$ and consequently we use the notation ω_u for its u -th element. The qualitative behaviour of a component is then fully described through a *partial parametrization* $K_v : \Omega_v \rightarrow [0, \rho(v)]$, as explained in the following paragraph. Note that for each $v \in V$ the set Ω_v is sufficient to obtain the set of regulators of v . The *parametrization* $K = (K_v)_{v \in V}$ therefore fully suffices to derive both the behaviour and the topology of a network. We will further use K as an identifier of a single model and \mathcal{K}^G to denote the set of all possible parametrizations of a regulatory graph G . If G can be arbitrary (but fixed) we use simply \mathcal{K} .

The asynchronous behaviour of a parametrized regulatory graph $G = (V, \rho, E)$ is then captured in a so-called *transition system* (TS), which is a pair (S^K, \rightarrow^K) where $S^K = \prod_{v \in V} \rho(v)$ is a set of states and $\rightarrow^K \subseteq S^K \times S^K$ is a

transition relation obtained from a parametrization $K \in \mathcal{K}^G$ s.t. $s \rightarrow^K s'$ if and only if one of the following two, mutually exclusive conditions holds:

$$\begin{aligned} & \forall v \in V : s'_v = s_v \wedge K_v(s) = s_v, \\ & \exists u \in V, \forall v \in V \setminus \{u\} : s'_v = s_v \wedge s_u \neq s'_u = K_u(s). \end{aligned}$$

To examine a behaviour of a TS we use the model checking procedure, for details please refer to [14]. This method allows to query whether the system satisfies a certain formally described property. In case of our system we verify whether a path of a certain kind exists in a TS, more specifically, whether a path that matches a sequence of *measurements* exists. Formally a sequence of measurements is described via a vector $M^K = (M_1, \dots, M_m)$ for some $m \in \mathbb{N}$ where for any $i \in [1, m]$ it holds that $M_i \subseteq S^K$. Understandably, a state $s \in S^K$ matches a measurement M_i iff also $s \in M_i$. Then a path $w = (s_1, \dots, s_n) \in (\rightarrow^K)^{n-1}$ satisfies M^K iff there is a vector $I = (i_1, \dots, i_m)$ of indices such that for any $k \in [1, m]$ it holds that $s_{i_k} \in M_k$ (measurements are matched) and for each pair $k, l \in [1, m]$ we have that if $k < l$ then also $i_k < i_l$ (ordering is preserved). The path w is then called the *witness* of satisfaction of M^K by (S^K, \rightarrow^K) , written $w \models M^K$.

3 Labels

Having a model, one is usually interested in what its properties are, e.g. which of the regulations are effective, how it behaves dynamically etc. We call functions that provide such information *labels*. We recall some previously introduced labels [5, 16] one can assign to a parametrization, now updated to fit the workflow, and some new ones. As the domain of a label usually depends on the respective regulatory graph, we use the symbol l for a label in general, and l^K to denote that the label depends on the graph encoded by K and is evaluated under K .

All the labels are illustrated in Fig. 3 on the toy example in Fig. 2.

3.1 Sign

This label is based on the usual interpretation of an effect of a regulation:

$$\begin{aligned} (u, t, v) \in E \text{ is activating} & \iff \exists \omega \in \Omega_v : K_v(\omega_{u \leftarrow [t, t_+)}) > K_v(\omega_{u \leftarrow [t_-, t)}), \\ (u, t, v) \in E \text{ is inhibiting} & \iff \exists \omega \in \Omega_v : K_v(\omega_{u \leftarrow [t, t_+)}) < K_v(\omega_{u \leftarrow [t_-, t)}), \end{aligned}$$

where $\omega_{u \leftarrow [t_-, t)}$ denotes that the regulatory interval ω_u is substituted by $[t_-, t)$. From this definition we derive the $\text{Sign}^K : E^K \rightarrow \{0, +, -, 1\}$ where:

$$\text{Sign}^K(e) = \begin{cases} 0 & \text{iff } e \text{ is not activating and not inhibiting,} \\ + & \text{iff } e \text{ is activating and not inhibiting,} \\ - & \text{iff } e \text{ is not activating and inhibiting,} \\ 1 & \text{iff } e \text{ is activating and inhibiting.} \end{cases}$$

a)	Structural Labels	
	$Sign^K(A, 1, B)$	+
	$Sign^K(B, 1, A)$	-
	$Sign^K(B, 2, B)$	-
	$Indegree^K(A)$	1
	$Indegree^K(B)$	2
	$Indegree^K(SUM)$	3
	$Bias^K(A)$	1
	$Bias^K(B)$	2
	$Impact^K(A, 1, B)$	0.905
	$Impact^K(B, 1, A)$	-1
	$Impact^K(B, 2, B)$	-0.302

b)	Regulatory Functions	
	F_A^K	$1 \& B\{0\}$
	F_B^K	$1 \& A\{1\} \& B\{2\} + 2 \& A\{1\} \& B\{0,1\}$

c)	Property Labels	
	$Cost^K(M^K)$	4
	$Robustness^K(M^K)$	0.25

Fig. 3. Illustrative labels for the toy example from Fig. 2. **a)** All the possible structural labels. For clarity we use the symbol SUM to denote the sum of the INDEGREE values. **b)** The REGULATORY FUNCTION labels corresponding to the given parametrization. **c)** The dynamic labels for the measurement series M^K from Fig. 2f.

Note that the 0 value means that the regulation has no effect on its target and could be removed without affecting the dynamics, which is utilized in the following label INDEGREE. The value 1 describes the situation where a regulation has ambiguous semantics, not meeting the so-called *Snoussi condition* [10], which is usually contrary to the expectation of the modeller about the system.

3.2 Indegree

This self-explanatory label counts the number of effective incoming regulations. Formally we denote $Indegree^K : V^K \rightarrow \mathbb{N}$ the number of non-zero incoming regulations, defined as:

$$Indegree^K(v) = |\{(u, t, v) \in E \mid u \in V, Sign(u, t, v) \neq 0\}|.$$

Additionally, the function is extended to capture the sum of the INDEGREE values of all the components, such that $Indegree^K(V) = \sum_{v \in V^K} Indegree^K(v)$. The sum of INDEGREE values is of a special interest, as quite often one is interested in structures that are minimal w.r.t. number of regulations.

3.3 Cost

The COST [14] of a measurement series is equal to the number of states of its shortest witness. The value is of interest under the assumption that a shorter witness in general means a lower number of qualitative changes and in turn a

slower energy consumption by the system. Even in the cases where the energy assumption is not realistic (e.g. due to different time scales) the COST value still reflects on how functionally complex the system is. Thus, one is usually interested in minimizing it.

Denote \mathcal{M}^K the set of possible measurement series for S , then the label has the form $Cost^K : \mathcal{M}^K \rightarrow \mathbb{N}_0$ s.t. if there is no witness for M^K then $Cost^K(M^K) = 0$, otherwise $Cost^K(M^K) = \min\{m \mid \exists w \in (\rightarrow^K)^{m-1} : w \models M^K\}$.

3.4 Robustness

The ROBUSTNESS [14] label $Robustness^K : \mathcal{P}^K \rightarrow [0, 1]$ is closely related to the COST label. In general terms it denotes the probability that M such that $Cost^K(M) = m$ will be satisfied by a random walk of length m that starts from M_1 . Formally:

$$Robustness^K(M) = \frac{|\{w \in (\rightarrow^K)^{m-1} \mid w_1 \in M_1, w \models M\}|}{|\{w \in (\rightarrow^K)^{m-1} \mid w_1 \in M_1\}|},$$

where w_1 denotes the first state on the path w . Understandably, if a measurement is not satisfiable, then there are no witnesses and the dividend and therefore also the ROBUSTNESS is equal to 0.

This particular notion of ROBUSTNESS reflects on the ability of the model to keep the requested behaviour even though uncertainty is introduced to the model through the modelling framework. The non-determinism of the simulation arises in states where the qualitative behaviour in reality depends on quantitative nuances indistinguishable by our abstraction. The higher the ROBUSTNESS of the model w.r.t. a measurement series, the less sensitive the model is to these quantitative nuances, respectively to perturbation in these.

3.5 Impact

The IMPACT label represents the relation between a regulator and its target via the function $Impact^K : E^K \rightarrow [-1, 1]$. We have introduced this value in [16] and here we present a definition that uses regulatory contexts as its domain. For a regulation $(u, t, v) \in E$ we obtain the IMPACT of u on v by computing the correlation of the activity level of the regulator and the respective parameter value. As we are interested only in parameters that are directly affected by this regulation, we take a subset of regulatory contexts on the border of the threshold value t . These we list as an arbitrarily ordered vector $\Omega_v^t = (\omega \in \Omega_v \mid \omega_u \in \{[t_-, t), [t, t_+])\})$. To indicate presence or absence of the said regulation, we use the function $Pres_u : \Omega_u \rightarrow [0, \rho(u)]$ that projects the activity interval of u on its lower boundary, i.e. if $\omega_u = [t_-, t)$ then $Pres_u(\omega) = t_-$. The IMPACT of (u, t, v) is then equal to the Pearson correlation coefficient between the image of Ω_v^t under $Pres_u$ and K_v :

$$Impact^K(u, t, v) = \frac{cov(Pres_u(\omega)_{\omega \in \Omega_v^t}, K_v(\omega)_{\omega \in \Omega_v^t})}{std(Pres_u(\omega)_{\omega \in \Omega_v^t}) \cdot std(K_v(\omega)_{\omega \in \Omega_v^t})},$$

where *cov* is the covariance and *std* is the standard deviation. This value is quite helpful when one is searching for the key regulators of a certain component. The further the value is from 0, the more prominent the regulation is.

3.6 Bias

By the term BIAS we here mean the general tendency of a parametrization to push a component towards higher or lower activity levels. The BIAS label $Bias^K : V \rightarrow [0, 1]$ is obtained simply as $Bias^K(v) = \sum_{\omega \in \Omega_v} K_v(\omega) \cdot |\Omega_v|^{-1} \cdot \rho(v)^{-1}$. For a Boolean component this coincides with the notion as defined by other authors, e.g. [9].

As a component has in general more effect on the other components at higher activity levels, the BIAS label allows to distinguish the components whose presence seems to be crucial for the activity of the network.

3.7 Regulatory Function

While not being a label *per se* we also assign a logical REGULATORY FUNCTION, providing a more human-readable description of a parametrization. In particular, we describe each partial parametrization as a Post Algebra [7] expression in a disjunctive normal form (DNF) of cardinality $\max\{\rho(v) \mid v \in V^K\}$.

A Post Algebra expression P in DNF of cardinality n is in our case described using the grammar:

$$\begin{aligned} P &\rightarrow M|M \mid M \\ M &\rightarrow V\&A \mid V \\ V &\rightarrow 0 \mid \dots \mid n \\ A &\rightarrow A\&A \mid v\{L\} \\ L &\rightarrow LL \mid V \end{aligned}$$

where $v \in V, |, \&, \{, \}, 0, \dots, n$ are terminals, and P, M, V, A, L are non-terminals. The semantics are such that an atom, i.e. an expression of the form $v\{L\}$, evaluates to n if the variable v is at a level listed in L and to 0 otherwise. The binary operator $\&$ evaluates to the smaller of its operands and the binary operator $|$ evaluates to the bigger of its operands. E.g. consider the function in Fig. 3b and an interpretation $A = 1, B = 1$. Then we can do the following valuation:

$$1\&A\{1\}|2\&B\{0, 2\} \mapsto 1\&2|2\&B\{0, 2\} \mapsto 1\&2|2\&0 \mapsto 1|2\&0 \mapsto 1|0 \mapsto 1.$$

Note that in the Boolean case the operator $\&$ corresponds to the logical conjunction, $|$ to the disjunction, $v\{1\}$ to the simple v , and $v\{0\}$ to $\neg v$.

We obtain the REGULATORY FUNCTION label by enumerating all the prime implicants and joining them via a disjunction.

4 Parametrization Sets Analysis and Comparison

While the individual parametrizations can be at least partially ordered by the values of their labels, it is only seldom that a single parametrization would appear as an optimal one. Moreover, even if one aims to find a parametrization that scores the best in all the metrics, i.e. minimum COST, maximum ROBUSTNESS, minimum INDEGREE etc., usually there are multiple parametrizations with the best score or those that are pairwise incomparable. We therefore focus on so-called *selections*, i.e. sets of parametrizations that fit certain criteria on the labels and analyse the whole selection.

all parametrizations					—					Cost(p) = 4, Robustness(p) = 1, Sign(B,2,B) = 0					
a	Label	#	Elements			Label	#	Elements			Label	#	Elements		
	Cost(M)	2	0:66.67, 4:33.33,			Cost(M)	2	0:66.67, 4:66.67,			Cost(M)	1	4:100,		
	F _A	4	0:25, 1:25, B(0):25, B(12):25,			F _A	4	0:35, 1:25, B(0):15, B(12):25,			F _A	2	0:60, B(0):40,		
	F _B	81	0:1.23, 1:1.23, 1&1A:1.23, 1&1A&B(0):1.23,			F _B	81	0:1.23, 1:1.23, 1&1A:1.23, 1&1A&B(0):1.23,			F _B	3	1&1A 2&1A:40, 2:40, 2&1A:20,		
	Indegree(A)	2	0:50, 1:50,			Indegree(A)	2	0:10, 1:10,			Indegree(A)	2	0:60, 1:40,		
	Indegree(B)	3	0:3.7, 1:14.81, 2:81.48,			Indegree(B)	3	0:36.3, 1:45.19, 2:81.48,			Indegree(B)	2	0:40, 1:60,		
	Indegree(SUM)	4	0:1.85, 1:9.26, 2:48.15, 3:40.74,			Indegree(SUM)	4	0:18.15, 1:50.74, 2:28.15, 3:40.74,			Indegree(SUM)	3	0:20, 1:60, 2:20,		
	K _A (B(0))	2	0:50, 1:50,			K _A (B(0))	2	0:10, 1:10,			K _A (B(0))	2	0:60, 1:40,		
	K _A (B(1))	2	0:50, 1:50,			K _A (B(1))	2	0:50, 1:50,			K _A (B(1))	1	0:100,		
	K _B (A(0),B(0,1))	3	0:33.33, 1:33.33, 2:33.33,			K _B (A(0),B(0,1))	3	0:33.33, 1:33.33, 2:66.67,			K _B (A(0),B(0,1))	1	2:100,		
	K _B (A(0),B(2))	3	0:33.33, 1:33.33, 2:33.33,			K _B (A(0),B(2))	3	0:33.33, 1:33.33, 2:66.67,			K _B (A(0),B(2))	1	2:100,		
	K _B (A(1,2),B(0,1))	3	0:33.33, 1:33.33, 2:33.33,			K _B (A(1,2),B(0,1))	3	0:13.33, 1:6.67, 2:6.67,			K _B (A(1,2),B(0,1))	3	0:20, 1:40, 2:40,		
K _B (A(1,2),B(2))	3	0:33.33, 1:33.33, 2:33.33,			K _B (A(1,2),B(2))	3	0:13.33, 1:6.67, 2:6.67,			K _B (A(1,2),B(2))	3	0:20, 1:40, 2:40,			
Sign(A,1,B)	4	+33.33, -33.33, 0:11.11, 1:22.22,			Sign(A,1,B)	4	+33.33, -26.67, 0:-28.89, 1:22.22,			Sign(A,1,B)	2	-60, 0:40,			
Sign(B,1,A)	3	+25, -25, 0:50,			Sign(B,1,A)	3	+25, -15, 0:-10,			Sign(B,1,A)	2	-40, 0:60,			
Sign(B,2,B)	4	+33.33, -33.33, 0:11.11, 1:22.22,			Sign(B,2,B)	4	+33.33, -33.33, 0:-88.89, 1:22.22,			Sign(B,2,B)	1	0:100,			
b	Label	Count	Min	Max	Mean	Label	Count	Min	Max	Mean	Label	Count	Min	Max	Mean
	K _A (B(0))	162	0	1	0.5	K _A (B(0))	160	0	0	0.099...	K _A (B(0))	2	0	1	0.4
	K _A (B(1))	162	0	1	0.5	K _A (B(1))	162	0	1	0.5	K _A (B(1))	0	0	0	0
	K _B (A(0),B(0,1))	216	0	2	1	K _B (A(0),B(0,1))	211	-2	0	-1	K _B (A(0),B(0,1))	5	2	2	2
	K _B (A(0),B(2))	216	0	2	1	K _B (A(0),B(2))	211	-2	0	-1	K _B (A(0),B(2))	5	2	2	2
	K _B (A(1,2),B(0,1))	216	0	2	1	K _B (A(1,2),B(0,1))	212	0	0	-0.19...	K _B (A(1,2),B(0,1))	4	0	2	1.2
	K _B (A(1,2),B(2))	216	0	2	1	K _B (A(1,2),B(2))	212	0	0	-0.19...	K _B (A(1,2),B(2))	4	0	2	1.2
	Indegree(A)	162	0	1	0.5	Indegree(A)	160	0	0	0.099...	Indegree(A)	2	0	1	0.4
	Indegree(B)	312	0	2	1.777...	Indegree(B)	309	0	1	1.177...	Indegree(B)	3	0	1	0.6
	Indegree(SUM)	318	0	3	2.277...	Indegree(SUM)	314	0	1	1.277...	Indegree(SUM)	4	0	2	1
	Bias(A)	243	0	1	0.5	Bias(A)	241	0	0.5	0.3	Bias(A)	2	0	0.5	0.2
	Bias(B)	320	0	1	0.5	Bias(B)	315	-0.5	0	-0.30...	Bias(B)	5	0.5	1	0.8
c	Impact(B,1,A)	162	-1	1	0	Impact(B,1,A)	160	0	1	0.4	Impact(B,1,A)	2	-1	0	-0.4
	Impact(A,1,B)	248	-1	1	4.386...	Impact(A,1,B)	245	0	1	0.6	Impact(A,1,B)	3	-1	0	-0.6
	Impact(B,2,B)	248	-1	1	1.370...	Impact(B,2,B)	248	-1	1	1.370...	Impact(B,2,B)	0	0	0	0
	Cost(M)	108	0	4	1.333...	Cost(M)	103	-4	0	-2.66...	Cost(M)	5	4	4	4
	Robustness(M)	108	0	1	0.1875	Robustness(M)	103	-1	0	-0.8125	Robustness(M)	5	1	1	1
	d	Label	Count	Min	Max	Mean	Label	Count	Min	Max	Mean	Label	Count	Min	Max
K _A (B(0))		162	0	1	0.5	K _A (B(0))	160	0	0	0.099...	K _A (B(0))	2	0	1	0.4
K _A (B(1))		162	0	1	0.5	K _A (B(1))	162	0	1	0.5	K _A (B(1))	0	0	0	0
K _B (A(0),B(0,1))		216	0	2	1	K _B (A(0),B(0,1))	211	-2	0	-1	K _B (A(0),B(0,1))	5	2	2	2
K _B (A(0),B(2))		216	0	2	1	K _B (A(0),B(2))	211	-2	0	-1	K _B (A(0),B(2))	5	2	2	2
K _B (A(1,2),B(0,1))		216	0	2	1	K _B (A(1,2),B(0,1))	212	0	0	-0.19...	K _B (A(1,2),B(0,1))	4	0	2	1.2
K _B (A(1,2),B(2))		216	0	2	1	K _B (A(1,2),B(2))	212	0	0	-0.19...	K _B (A(1,2),B(2))	4	0	2	1.2
Indegree(A)		162	0	1	0.5	Indegree(A)	160	0	0	0.099...	Indegree(A)	2	0	1	0.4
Indegree(B)		312	0	2	1.777...	Indegree(B)	309	0	1	1.177...	Indegree(B)	3	0	1	0.6
Indegree(SUM)		318	0	3	2.277...	Indegree(SUM)	314	0	1	1.277...	Indegree(SUM)	4	0	2	1
Bias(A)		243	0	1	0.5	Bias(A)	241	0	0.5	0.3	Bias(A)	2	0	0.5	0.2
Bias(B)		320	0	1	0.5	Bias(B)	315	-0.5	0	-0.30...	Bias(B)	5	0.5	1	0.8

Fig. 4. Reports produced by TREMPPI for the graph in Fig. 2. For the interactive version please see [12]. **Left:** Reports for \mathcal{K}^G . **Right:** Reports for $\mathcal{K}^{G,\Psi}$ with $\Psi \equiv \text{Cost}(p) = 4 \wedge \text{Robustness}(p) = 1 \wedge \text{Sign}(B, 2, B) = 0$. **Middle:** A comparison left - right. **a)** A QUALITATIVE report. The label F_B is not fully listed. **b)** A QUANTITATIVE report. **c)** A REGULATION graph. **d)** A CORRELATION graph.

Have a parametrization space \mathcal{K} and a sequence of predicates $\Phi = \Phi_1, \dots, \Phi_n$ where $\Phi_i : \mathcal{K} \rightarrow \mathbb{B}$ for each $i \in [1, n]$. A *selection by Φ* we call the set of parametrizations denoted \mathcal{K}^Φ s.t. for each $K \in \mathcal{K}$ we have that $K \in \mathcal{K}^\Phi$ if and only if $\bigwedge_{i=1}^n \Phi_i(K)$ holds true. As the selections may contain millions or more parametrizations in size, approaches that allow to evaluate the whole selection at once are necessary to gain understanding of the nature of the selection. We present four different methods, each used to depict some of the labels in a manner that generalizes the values of the labels from members of the selection to the whole selection. A visual representation of such data is then called a *report*. Additionally, each of the reports features an individual method of *comparison*—having two different selections $\mathcal{K}^\Phi, \mathcal{K}^\Psi$ we create a third report which illustrates the difference between the two selections. This we denote using the minus ($-$) symbol, illustrating the fact that it is a non-commutative difference operation. All the reports are illustrated in Fig. 4 on the example network from Fig. 2. Each report provides a comparison between the set of all 324 parametrizations, i.e. a selection by $\Phi \equiv \text{true}$ and a selection where the M^K from Fig. 3 has minimal COST and maximal ROBUSTNESS and where the self-regulation of the component B is not present, i.e. a selection by $\Psi \equiv (\text{Cost}(\text{series}) = 4 \wedge \text{Robustness}(\text{series}) = 1 \wedge \text{Sign}(B, 2, B) = 0)$.

4.1 Explicit Qualitative Report

The first tool we employ is a QUALITATIVE summary, which describes an image of a label in the selection, i.e. all the distinct label values that appear in the selection and their frequency in percent. Have a label $l : X \rightarrow Y$, where X, Y are some sets and a selection \mathcal{K}^Φ . For example in the case $l = \text{Sign}^K$, we have $X = E$ and $Y = \{0, +, -, 1\}$. For each value $x \in X$ we then set:

$$\begin{aligned} \text{qual}(\mathcal{K}^\Phi, l, x) &= (\text{size}(\mathcal{K}^\Phi, l, x), \text{elems}(\mathcal{K}^\Phi, l, x)), \\ \text{size}(\mathcal{K}^\Phi, l, x) &= |\text{elems}(\mathcal{K}^\Phi, l, x)|, \\ \text{elems}(\mathcal{K}^\Phi, l, x) &= \{(y, q) \mid q = |\{K \in \mathcal{K}^\Phi \mid l^K(x) = y\}| \cdot 100 \cdot |\mathcal{K}^\Phi|^{-1}\}. \end{aligned}$$

A comparison of two selections $\mathcal{K}^\Phi, \mathcal{K}^\Psi$, denoted $\text{qual}(\mathcal{K}^\Phi, l, x) - \text{qual}(\mathcal{K}^\Psi, l, x)$, is obtained by subtracting the two pairs, where $\text{elems}(\mathcal{K}^\Phi, l, x) - \text{elems}(\mathcal{K}^\Psi, l, x)$ is computed as:

$$\{(y, q^\Phi - q^\Psi) \mid (y, q^\Phi) \in \text{elems}(\mathcal{K}^\Phi, l, x), (y, q^\Psi) \in \text{elems}(\mathcal{K}^\Psi, l, x)\}.$$

Since the set of parametrizations is finite, all values have finite domain and are thus suitable to this form of presentation. However, in the case of labels that project to rational numbers, i.e. ROBUSTNESS, BIAS, and IMPACT values, the size of the image quite often threatens to be almost as big as the selection itself, therefore we chose not to include them in the QUALITATIVE report.

4.2 Explicit Quantitative Report

Similarly to the previous, we summarize the overall nature of quantitative labels, i.e. those whose image is a subset of rational numbers, using the quadruple:

$$\begin{aligned} \text{quan}(\mathcal{K}^\Phi, l, x) &= (\text{count}(\mathcal{K}^\Phi, l, x), \min(\mathcal{K}^\Phi, l, x), \max(\mathcal{K}^\Phi, l, x), \text{mean}(\mathcal{K}^\Phi, l, x)), \\ \text{count}(\mathcal{K}^\Phi, l, x) &= |\{K \in \mathcal{K}^\Phi \mid l^K(x) \neq 0\}|, \\ \min(\mathcal{K}^\Phi, l, x) &= \min\{l^K(x) \mid K \in \mathcal{K}^\Phi\}, \\ \max(\mathcal{K}^\Phi, l, x) &= \max\{l^K(x) \mid K \in \mathcal{K}^\Phi\}, \\ \text{mean}(\mathcal{K}^\Phi, l, x) &= \sum_{K \in \mathcal{K}^\Phi} l^K(x) \cdot |\mathcal{K}^\Phi|^{-1}. \end{aligned}$$

The difference between the QUANTITATIVE reports of two selections $\mathcal{K}^\Phi, \mathcal{K}^\Psi$ is then set simply as the subtraction of the two quadruples.

Note that the *count* has a somewhat special meaning, as the 0 value is of particular interest for some of the labels. In the case of COST for example, it denotes that the respective measurement series is not satisfiable or for SIGN it states that the edge is absent.

4.3 Inferred Regulation Graph

Based on IMPACT and SIGN, we can summarize the average effect of regulations of a sample. The IMPACT can be easily extended from a parametrization to a sample as $\text{Impact}^{\mathcal{K}^\Phi}(e) = \sum_{K \in \mathcal{K}^\Phi} \text{Impact}^K(e) \cdot |\mathcal{K}^\Phi|^{-1}$ for each $e \in E$. For the SIGN we take a supremum under the partial ordering $0 < - < 1, 0 < + < 1$, i.e. for any $e \in E$ we set $\text{Sign}^{\mathcal{K}^\Phi}(e) = \sup\{\text{Sign}^K(e) \mid K \in \mathcal{K}^\Phi\}$. Lastly we depict the FREQUENCY of a regulation, which states how often a regulation is active at all, i.e. has a non-zero SIGN, formally $\text{Frequency}^{\mathcal{K}^\Phi}(e) = |\{K \in \mathcal{K}^\Phi \mid \text{Sign}^K(e) \neq 0\}|$.

Visually, the IMPACT value is mapped to a color gradient of the regulation edge with the color red representing the value -1 , yellow representing 0 , and green representing 1 . The FREQUENCY is mapped to the width of an edge. When the FREQUENCY is equal to 0 , the edge is then displayed as dotted. Lastly, the SIGN is reflected in the shape of the head of the edge. The $+$ SIGN is mapped to a pointed arrow shape, the $-$ to a rectangle shape (also known as *blunt arrow*), the 1 to a combination of both and the 0 to a circle.

To create a comparison, the IMPACT and FREQUENCY values are directly subtracted. The SIGN can not be clearly interpreted in the comparison and for simplicity it is kept from the minuend. Note that the subtraction means that the result lies behind the original boundaries of a value. The color gradient is therefore stretched to the range $[-2, 2]$ and a negative FREQUENCY value is depicted by a dashed edge.

4.4 Correlation Graph

Similarly to the REGULATION graph we also create a correlation graph, based on the BIAS label. The label extended similarly to the IMPACT label, i.e. $Bias^{\mathcal{K}^\Phi}(v) = \sum_{K \in \mathcal{K}^\Phi} Bias^K(v) \cdot |\mathcal{K}^\Phi|^{-1}$ for $v \in V$. Additionally, one is usually interested in whether there is a relation between activities of multiple components, e.g. if one component seems to be taking over if another is missing. This is obtained as the correlation between the BIAS of individual components in a sample, i.e.:

$$Correlation^{\mathcal{K}^\Phi}(v, u) = \frac{cov(Bias^K(v)_{K \in \mathcal{K}^\Phi}, Bias^K(u)_{K \in \mathcal{K}^\Phi})}{std(Bias^K(v)_{K \in \mathcal{K}^\Phi}) \cdot std(Bias^K(u)_{K \in \mathcal{K}^\Phi})}.$$

The CORRELATION value is mapped to a color gradient in the same manner as the IMPACT value in the REGULATIONS graph. The BIAS value is mapped to the width of the border of the respective component in a manner similar to the edge width in the case of the FREQUENCY value.

To create a comparison both values are simply subtracted.

5 Case Study

To provide a practical demonstration of our methodology, we have utilized the data provided by D’Allesandro *et al.* in their study of hepatocyte growth factor (HGF) signalling [2].

In the original article the authors constructed a core network, illustrated in Fig. 5, with a set of regulations that are with high certainty present. Afterwards, a qualitative method is used to find an optimal structure that combines the core network with a subset of possible edges. To this end the authors obtained a rich set of experimental data, which they later discretized to meet the needs of their qualitative framework. The discretized data features measurements of 6 components in 6 different experimental set-ups. In each of the experiments one or two of the components of the network are inhibited and later the HGF stimuli is added. Additionally the authors provide a control measurement where no inhibition is present.

In the study, the data are present as fold-change comparisons between some of the experiments. For each component there are 9 time-points measured, however in the discretized form these are divided at the time of 30 min into an early and late response, as it is expected that around that time feedback effects start to play a role in the behaviour of the system. As the fold-change scheme is not suitable for encoding as a time-series, we reinterpreted the data into a measurement scheme, where the fold change translates to a difference between two measurements. This means that from two fold-changes we obtain three measurements. The particular values for the measurements were determined in the following manner:

- In the experiment a Met inhibitor was used that blocks the receptor of the pathway and thereby downregulates all signalling processes even under

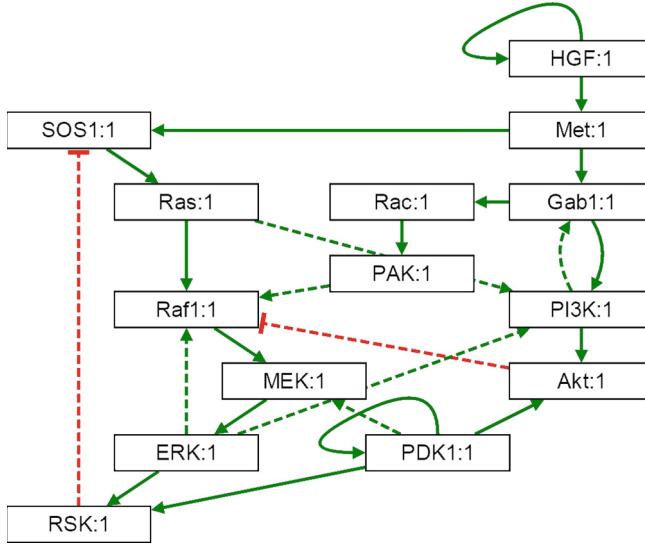


Fig. 5. The structure of the model that was identified as optimal in [2]. The regulations denoted by a full line constitute the core network, whereas those that are dashed are added from the pool of optional regulations. In the enumeration step we place requirements on the edges corresponding to the SIGN label, in particular the full edges with a pointed arrow are required to have the + SIGN, the dashed edges with a pointed arrow to have either + or 0, and the dashed edges with a blunt arrow to have either - or 0.

stimulation. The fold-change comparison to the control shows a significant downregulation in all read-outs therefore we conclude that the control state has active read-outs.

- For other set-ups, if there is a significant decrease [2] in the fold change, the component is expected to be at the level 0 after the change.
- Likewise, if there is a significant increase, the component is expected to be at the level 1.
- If there is no significant change, no requirement is placed on the value.
- We require the full monotonicity in the behaviour, i.e. if a value of a component does not change between two timepoints, we require that it cannot change in the simulation either. If the value differs between two measurements, we require that there is exactly one change of that value. For details please refer to [15].

Altogether we have obtained 5 properties, which are detailed in the supplement [11]. Even though this interpretation of the data is quite strict, we obtained that even the core network is capable of satisfying all the experiments for each of the possible parametrizations. We have therefore focused instead on the structure identified as optimal by the authors to see whether addition of some of the optional edges can disrupt the expected function of the network. This optimal network is depicted in Fig. 5. This is a slightly simplified version of the original,

which features two components for RSK in an activation cascade. We joined this cascade into a single node, which is a preserving operation [8].

For the purposes of the analysis we have created three selections:

1. *ALL* is the set of all 223776 possible parametrizations.
2. *VALID* is the set of 149184 parametrizations that satisfy all the measurement series.
3. *MINCOST* is the set of 135072 parametrizations that have the minimal COST for all the measurement series.

As can be immediately seen, there exist parametrizations over the optional edges that render one or more of the measurement series not satisfiable.

Data of all the reports are in an interactive form available in the supplement [11]. Here we provide some of the possible observations about the data. Firstly we analyse the comparison *VALID* – *ALL*. In the QUANTITATIVE report we see that $K_{\text{MEK}}(\text{PDK1} = 0, \text{Raf} = 1)$ is for the *VALID* set bound to the value 1, meaning it is necessary that Raf1 alone can activate MEK. We also see that three out of the five measurement series are satisfied by all parametrizations, whereas the remaining two are satisfied exactly by those in the *VALID* selection, meaning that both place the same requirement on the behaviour of the network. In the QUALITATIVE report we can see that the function $\text{MEK} = \text{Raf1} \& \text{PDK1}$ is completely missing, in accordance with the quantitative observation, and the functions $\text{MEK} = \text{Raf1}$ and $\text{MEK} = \text{Raf1} | \text{PDK1}$ are now present with the same FREQUENCY, suggesting that the $(\text{PDK1}, 1, \text{MEK})$ regulation is superfluous and should probably be removed. In the REGULATIONS graph we then see that both the FREQUENCY and the IMPACT of PDK1 on MEK decreases and lastly in the CORRELATIONS report we see an increase in the BIAS of MEK.

Secondly we analyse the comparison *MINCOST* – *VALID*. From the QUANTITATIVE report we see that there is a slight increase in BIAS of Raf1. In the QUALITATIVE report we can see that 14 regulatory functions for Raf1 disappear completely, however as there are still 134 remaining, this does not provide too much information. A much cleaner picture can be gained from the REGULATIONS graph where we can see a decrease in the IMPACT and the FREQUENCY of ERK on Raf1 in favour of both inhibition by Akt and activation by PAK.

6 Conclusion

We have presented numerous methods for analysis and evaluation of qualitative parametrizations sets. The methods are gathered in two groups: labels, which evaluate individual parametrizations, and reports, which subsequently evaluate whole parametrizations sets. All of the methods are experimentally implemented in the tool TREMPPI [13]. The performance of the methods and the implementation is sufficient for application to realistic problems, as illustrated on a case study of HGF signalling.

While the method was mainly developed due to the specific nature of the problem of parameter identification in Thomas Networks, we believe that

the general approach should be also applicable to more complicated frameworks. The approach should be readily convertible to frameworks which are, in certain ways, only an extension of the Thomas method, e.g. piece-wise affine models [3]. However even for frameworks with infinite parametrization pool, the label-report-compare approach could work if combined with an appropriate sampling method.

As the tools we were using for the purposes of this article are becoming more mature, we would like to make them more available for public use. To this end a public web-service version of TREMPPI is planned.

References

1. Alexopoulos, L.G., Saez-Rodriguez, J., Cosgrove, B.D., Lauffenburger, D.A., Sorger, P.K.: Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Mol. Cell. Proteomics* **9**(9), 1849–1865 (2010)
2. DAlessandro, L.A., Samaga, R., Maiwald, T., Rho, S.-H., Bonefas, S., Raue, A., Iwamoto, N., Kienast, A., Waldow, K., Meyer, R., Schilling, M., Timmer, J., Klamt, S., Klingmüller, U.: Disentangling the complexity of HGF signaling by combining qualitative and quantitative modeling. *PLoS Comput. Biol.* **11**(4), e1004192 (2015)
3. de Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**(1), 67–103 (2002)
4. Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., Saez-Rodriguez, J.: Exhaustively characterizing feasible logic models of a signaling network using answer set programming. *Bioinformatics* **30**, 2320–2326 (2013)
5. Klärner, H., Siebert, H., Bockmayr, A.: Time series dependent analysis of unparametrized Thomas networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **99**, 1338–1351 (2012)
6. Lee, W.-P., Tzou, W.-S.: Computational methods for discovering gene networks from expression data. *Briefings Bioinf.* **10**(4), 408–423 (2009)
7. Miller, D.M., Thornton, M.A.: Multiple Valued Logic: Concepts and Representations, vol. 2. Morgan & Claypool Publishers, San Rafael (2007)
8. Saadatpour, A., Albert, R., Reluga, T.C.: A reduction method for boolean network models proven to conserve attractors. *SIAM J. Appl. Dyn. Syst.* **12**(4), 1997–2011 (2013)
9. Shmulevich, I., Kauffman, S.A.: Activities and sensitivities in boolean network models. *Phys. Rev. Lett.* **93**(4), 048701 (2004)
10. Snoussi, E.H.: Qualitative dynamics of piecewise-linear differential equations: a discrete mapping approach. *Dyn. Stab. Syst.* **4**(3–4), 565–583 (1989)
11. Streck, A.: HGF network analysis (2015). http://dibimath.github.io/HGF_4.8.12/. Accessed 18 June 2015
12. Streck, A.: HSB 2015 example model data (2015). http://dibimath.github.io/HSB_2015/. Accessed 18 June 2015
13. Streck, A.: TREMPPI source repository (2015). <https://github.com/xstreck1/TREMPPI/>. Accessed 18 June 2015
14. Streck, A., Siebert, H.: Extensions for LTL model checking of Thomas networks. In: *Advances in Systems and Synthetic Biology*, vol. 14, pp. 101–114. EDP Sciences (2015)

15. Streck, A., Thobe, K., Siebert, H.: Analysing cell line specific EGFR signalling via optimized automata based model checking. In: Roux, O., Bourdon, J. (eds.) CMSB 2015. LNCS, vol. 9308, pp. 264–276. Springer, Heidelberg (2015)
16. Thobe, K., Streck, A., Klarner, H., Siebert, H.: Model Integration and crosstalk analysis of logical regulatory networks. In: Mendes, P., Dada, J.O., Smallbone, K. (eds.) CMSB 2014. LNCS, vol. 8859, pp. 32–44. Springer, Heidelberg (2014)
17. Thomas, R.: Regulatory networks seen as asynchronous automata: a logical description. *J. Theoret. Biol.* **153**(1), 1–23 (1991)

Hybrid Systems Biology

Fourth International Workshop, HSB 2015, Madrid,
Spain, September 4-5, 2015. Revised Selected Papers

Abate, A.; Šafránek, D. (Eds.)

2015, XIV, 249 p. 78 illus. in color., Softcover

ISBN: 978-3-319-26915-3