

MERRIN: MEtabolic Regulation Rule INference from time series data

Kerian Thuillier^{1*}, Caroline Baroukh², Alexander Bockmayr³, Ludovic Cottret²,
Loïc Paulevé⁴, Anne Siegel^{1*}

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

²LIPME, INRAE, CNRS, Université de Toulouse, Castanet-Tolosan, France

³Freie Universität Berlin, Institute of Mathematics, D-14195 Berlin, Germany

⁴Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, F-33400 Talence, France

Abstract

Motivation: Many techniques have been developed to infer Boolean regulations from a prior knowledge network and experimental data. Existing methods are able to reverse-engineer Boolean regulations for transcriptional and signaling networks, but they fail to infer regulations that control metabolic networks.

Results: We present a novel approach to infer Boolean rules for metabolic regulation from time series data and a prior knowledge network. Our method is based on a combination of answer set programming and linear programming. By solving both combinatorial and linear arithmetic constraints we generate candidate Boolean regulations that can reproduce the given data when coupled to the metabolic network. We evaluate our approach on a core regulated metabolic network and show how the quality of the predictions depends on the available kinetic, fluxomics or transcriptomics time series data.

Availability: Software available at <https://github.com/bioasp/merrin>

Contact: anne.siegel@irisa.fr

Availability: Software available at <https://github.com/bioasp/merrin>

1 Notations

$|X|$. The cardinality of a finite set X is denoted by $|X|$.

x_I . Given a vector $x \in D^n$ and a set of indices $I \subseteq \{1, \dots, n\}$, x_I denotes the vector of dimension $|I|$ equal to $(x_i)_{i \in I}$.

\mathbb{B} . The Boolean domain is denoted by $\mathbb{B} = \{0, 1\}$.

$\beta(s)$. Given a non-negative real vector $s \in \mathbb{R}_{\geq 0}^n$, we denote by $\beta(s) \in \mathbb{B}^n$ its *binarization*, i.e. $\forall i \in \{1, \dots, n\}, \beta(s)_i = 1$, if $s_i > 0$, and $\beta(s)_i = 0$, if $s_i = 0$.

2 Boolean over-approximation of RMSS

This section is a complement to Sect. 2.1.3 and details the Boolean relaxation of the Eq.(2.c).

Boolean over-approximation of regulatory-metabolic steady state (\mathbb{B} -RMSS) of a RMN $(\mathcal{N}, \text{Inp}, \mathcal{P}, f)$ can be defined as a triplet $(\bar{v}, \bar{c}, \bar{x}) \in \mathbb{B}^{|\mathcal{R}|} \times \mathbb{B}^{|\text{Ext}|} \times \mathbb{B}^{|\text{Inp}|+|\mathcal{P}|+|\mathcal{R}|}$ (Thuillier *et al.*, 2021) associating binary reaction states \bar{v} , external metabolites availability \bar{c} , and a regulatory state \bar{x} . The binary reaction states \bar{v} must satisfy a relaxed form of Eqs.(1):

$$(1.a_{\text{relaxed}}) \quad \forall m \in \text{Int}, \quad \bigvee_{r \in \mathcal{R}, S_{mr} > 0} \bar{v}_r \iff \bigvee_{r \in \mathcal{R}, S_{r,m} < 0} \bar{v}_r$$

$$(1.b_{\text{relaxed}}) \quad \forall r \in \mathcal{R}, \bar{x}_r = 0 \implies \bar{v}_r = 0$$

$$(1.c_{\text{relaxed}}) \quad \forall m \in \text{Inp}, \forall r \in \mathcal{R}, S_{mr} < 0 \implies \bar{v}_r \leq \bar{c}_m$$

Let us denote by $\bar{\mathbb{S}}$ the set of all the \mathbb{B} -RMSS of the RMN $(\mathcal{N}, \text{Inp}, \mathcal{P}, f)$ (satisfying the relaxed equations Eqs.(1_{relaxed})). Eq.(2.c) can be relaxed by considering \mathbb{B} -RMSS instead of RMSS, thus:

$$(2.c_{\text{relaxed}}) \quad (v', c', x') \in \bar{\mathbb{S}}$$

It must be noted that the set of binarised RMSS is included in $\bar{\mathbb{S}}$, i.e. $\forall (v, c, x) \in \mathbb{S}, (\beta(v), \beta(c), x) \in \bar{\mathbb{S}}$. The converse is not true.

3 Results

3.1 Experiment conditions of Covert *et al.* (2001)

This section is a complement to Sect. 3.2 and details the initial states of the 5 experiment conditions described in Covert *et al.* (2001).

The 5 experiment conditions of Covert *et al.* (2001), used to generate our experimental time series, are shown in Tab. S1. Each experiment is based on a different set of initial input metabolite concentrations c and the regulatory state x is initialized such that: (i) $\forall r \in \mathcal{R}, x_r = 0$, (ii) $\forall i \in \text{Inp}, x_i = \beta(c_i)$, (iii) for each regulatory protein we apply the associated regulatory rule: $x_{\text{RPc1}} = \beta(c_{\text{Carbon1}})$, $x_{\text{RPo2}} = \beta(c_{\text{Oxygen}})$, $x_{\text{RPb}} = 0$ and $x_{\text{RPb}} = \beta(c_{\text{Hex}})$.

Experiment	Input metabolite concentration (mmol.L ⁻¹)					Regulatory protein state			
	cCarbon1	cCarbon2	cOxygen	cFest	cHex	xRPc1	xRPo2	xRPb	xRPb
1	10	10	100	0	0	1	0	0	0
2	0	10	5	0	0	0	0	0	0
3	0	10	100	0	2	0	0	0	1
4	0	5	100	0	10	0	0	0	1
5	1	10	100	0.1	5	1	0	0	1

Table S1: Experiment conditions used to generate the 5 simulations of (Covert *et al.*, 2001).

3.2 Inferring from non-complete noisy time series

This section is a complement to Sect. 3.4.

3.2.1 Comparing the d-rFBA simulations

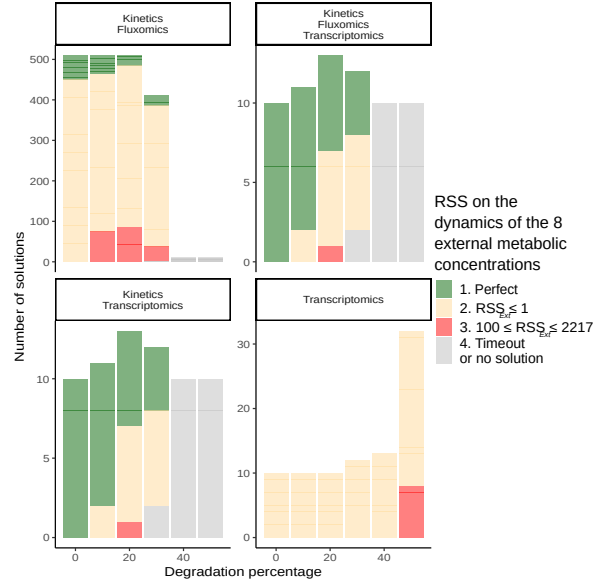
RSS scores (RSS_{Ext} and $RSS_{\mathcal{P}}$) of the regulatory inferred on the 240 instances.

For each of regulatory BN inferred for the 240 instances, we compared the associated d-rFBA time series of external metabolites and regulatory proteins to the ones of the gold standard model using the RSS_{Ext} score (Fig. S1(a)) and the $RSS_{\mathcal{P}}$ score (Fig. S1(b)).

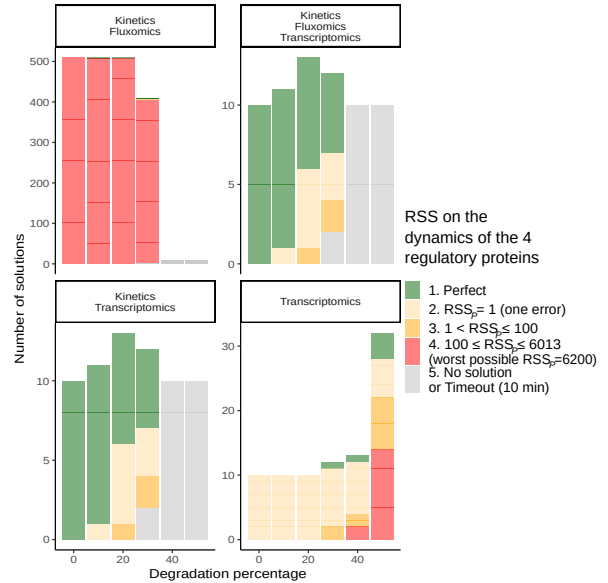
Kinetic-fluxomics instances. Kinetic-fluxomics (KF) instances do not contain any information on the four regulatory proteins states. As expected, MERRIN is not able to determine the regulatory proteins states (Fig. S1(b)) from this datatype leading to the enumeration of a huge number of compatible BNs. Here, we have restricted the number of solutions to enumerate to 51, this limit was reached for each (KF) instance that admit a solution to the inference problem. Thus, we do not recommend to use MERRIN on KF instances.

Impact of the degradation rate. Our results show that MERRIN could not infer any regulatory BNs

on the datatype KFT, KF and KT with a degradation rate strictly greater than 30%. For T instance, the number of inferred BNs increased significantly at 40% and 50% of degradation. Moreover, at high



(a) RSS_{Ext} on the dynamics of the 8 external metabolite concentrations



(b) $RSS_{\mathcal{P}}$ on the dynamics of the 4 regulatory proteins

Fig. S1: **RSS depending on datatype and degradation level.** Each vertical bar corresponds to the results of MERRIN on the 10 instances associated with a considered datatype (KF, KFT, KT, T) and degradation level (0%, 10%, 20%, 30%, 40%, 50%). The different colors represents the score range ((a) RSS_{Ext} and (b) $RSS_{\mathcal{P}}$) of the solution (see legend).

degradation level, both $RSS_{\mathcal{P}}$ and RSS_{Ext} decrease drastically: a huge part of the BNs inferred for T instances have $RSS_{\mathcal{P}}$ and a RSS_{Ext} greater than 100. Thus, we do not recommend to use MERRIN on instances having a degradation level higher than 30%.

3.2.2 Inferred regulatory BNs

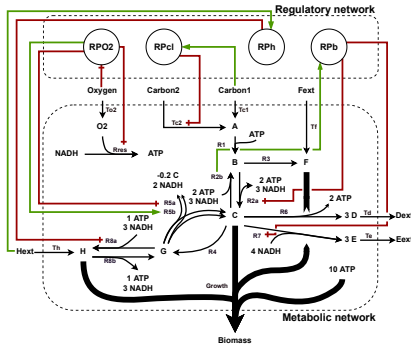
Enumerations of the 15 different regulatory BNs inferred from 120 instances.

In this section, we focus on the results obtained on 120 different time-series instances: complete (KFT), kinetic-transcriptomics (KT) and transcriptomics (T) instances with a noise ranging from 0% to 30% (Sect. 3.4).

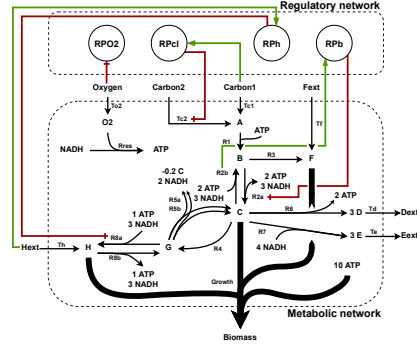
Let us consider the metabolic network \mathcal{N} , the set of inputs metabolites Inp and the set of regulatory proteins \mathcal{P} given as input to MERRIN. There are 15 different regulatory BNs that have been inferred on the 120 instances, for each inferred regulatory BN f , the RMN $(\mathcal{N}, Inp, \mathcal{P}, f)$ is shown in Fig. S2 with their respective scores: precision, recall, $RSS_{\mathcal{P}}$, RSS_{Ext} .

Best result. The regulatory BN of Fig. S2(b) is the one inferred on the complete (KFT) datatype with 0% of degradation. It allows exactly reproducing the 5 d-rFBA simulations of Covert *et al.* (2001) used to generate the input time series. This BN has been inferred on 58 of the 120 instances and only on the datatype KFT and KT.

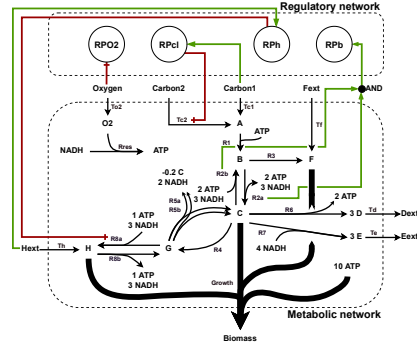
Worst result. Among the 15 regulatory BN, the regulatory BN of Fig. S2(p) has the worst RSSs scores: $RSS_{\mathcal{P}} = 89$ and $RSS_{Ext} = 1194.07$. These scores are due to the absence of regulation controlling the reaction *Tc2* which inhibit the consumption of *Carbon2* if some *Carbon1* is available.



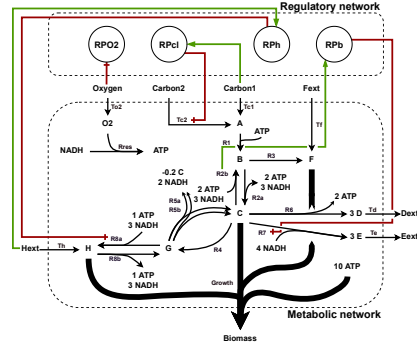
Precision: 1; Recall: 1; $RSS_{\mathcal{P}}$: 0; RSS_{Ext} : 0
(a) Gold standard RMN



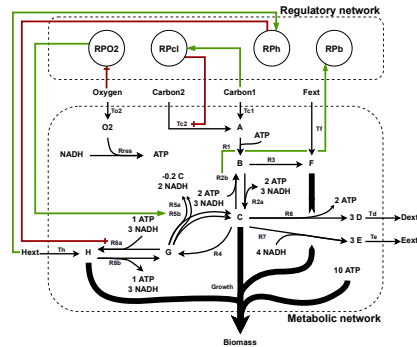
Precision: 1; Recall: 0.64; $RSS_{\mathcal{P}}$: 0; RSS_{Ext} : 0
(b)



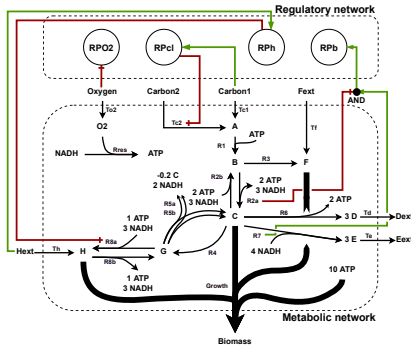
Precision: 0.86; Recall: 0.55; $RSS_{\mathcal{P}}$: 68; RSS_{Ext} : 0.18
(c)



Precision: 1; Recall: 0.64; $RSS_{\mathcal{P}}$: 1; RSS_{Ext} : 0.12
(d)

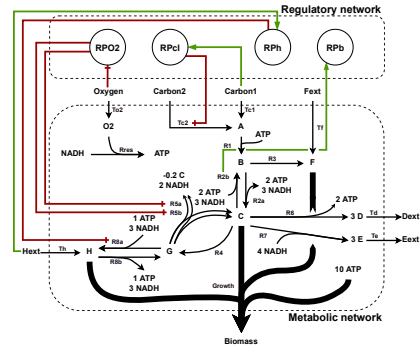


Precision: 1; Recall: 0.64; $RSS_{\mathcal{P}}$: 1; RSS_{Ext} : 0.18
(e)



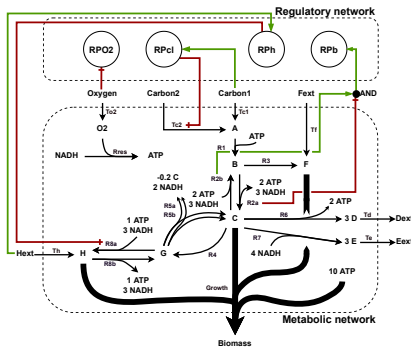
Precision: 0.71; Recall: 0.45; RSS_P : 68; RSS_{Ext} : 0.18

(f)



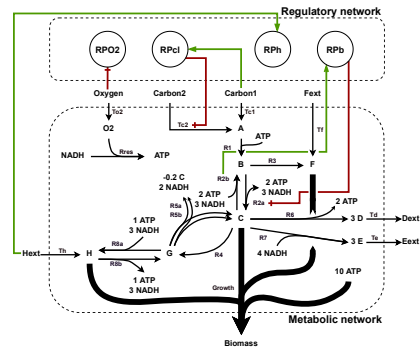
Precision: 0.88; Recall: 0.64; RSS_P : 1; RSS_{Ext} : 0.18

(j)



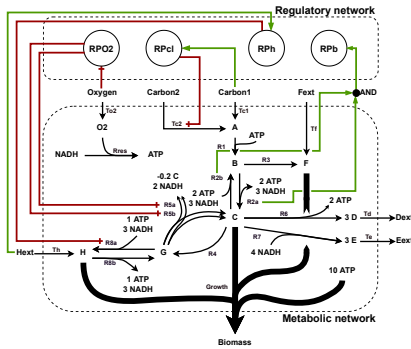
Precision: 0.86; Recall: 0.55; RSS_P : 1; RSS_{Ext} : 0.18

(g)



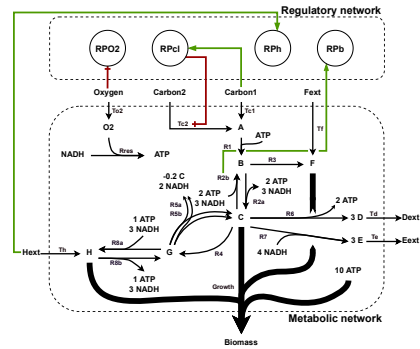
Precision: 1; Recall: 0.55; RSS_P : 0; RSS_{Ext} : 0.41

(k)



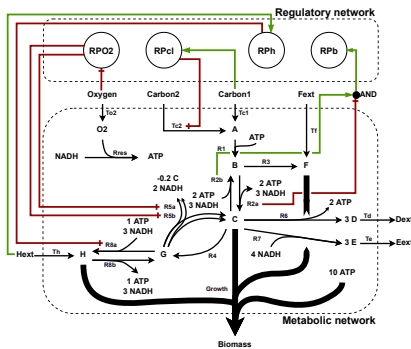
Precision: 0.78; Recall: 0.64; RSS_P : 68; RSS_{Ext} : 0.18

(h)



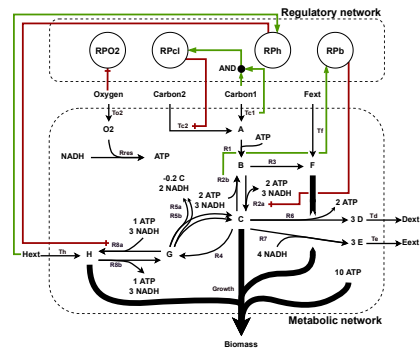
Precision: 1; Recall: 0.45; RSS_P : 1; RSS_{Ext} : 0.5

(l)



Precision: 0.78; Recall: 0.64; RSS_P : 1; RSS_{Ext} : 0.18

(i)



Precision: 0.75; Recall: 0.55; RSS_P : 6; RSS_{Ext} : 0.94

(m)

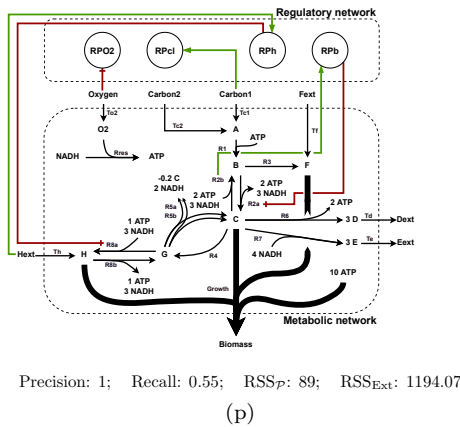
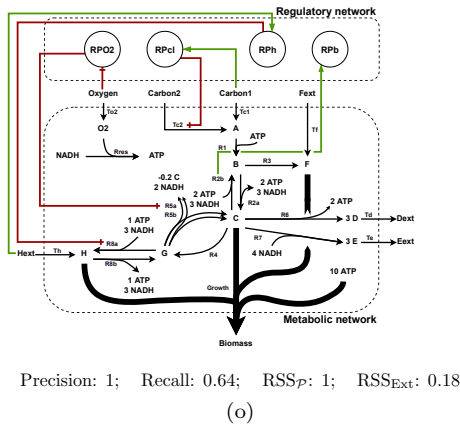
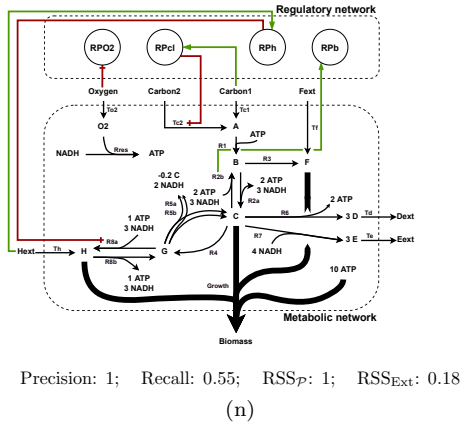


Fig. S2: (a) Regulatory BN of the gold standard model (Covert *et al.*, 2001).

(b)-(p) Set of 15 regulatory BNs inferred from the 120 instances representing kinetic-fluxes-transcriptomics (KFT), kinetic-transcriptomics (KT) and transcriptomics (T) observations with a noise ranging from 0% to 30%.

3.2.3 Comparisons with the gold standard regulatory BN

Recall and precision scores of the inferred BNs.

Let us focus on the 120 instances (datatype KFT, KT and T with a degradation level between 0% and

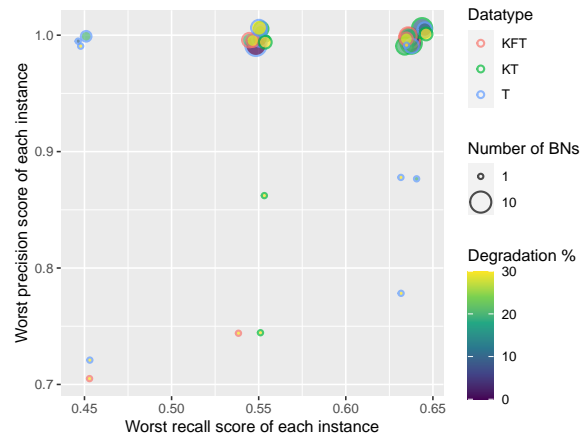


Fig. S3: Worst recall and precision depending of data type and degradation level. For each instance of the considered datatype (KFT, KT, T) and degradation level (0%, 10%, 20%, 30%), only the worst recall and worst precision are considered. Each circle corresponds to a set of instances of identical datatype and degradation level having the same worst recall and worst precision.

30%). For each inferred BN, we computed the recall and the precision according to the gold standard regulatory BN. Fig. S3 represents the worst recall and the worst precision of each one of the 120 instances depending of the datatype and the degradation level. Our results show that, except for 8 instances, MERRIN inferred BNs having a precision of 1 and a recall between 0.45 and 0.64, meaning that at least 50% of the edge of the influence graph of the gold standard are correctly retrieved.

The degradation level seems to have the bigger impact on the precision score: all, except one, instances with a worst precision lower than 1 have a degradation level of 30%. For the recall, it appears that it is the datatype which have the bigger impact: T instances have smaller recall than the other KFT and KT instances. This last result can be easily explained by the fact that T instances do not have any information on the input metabolite concentrations, thus it is harder to define if an observed RMSS is due to a specific concentration of input metabolites or to some regulatory states.

References

Covert, M.W.*et al.* (2001). Regulation of gene expression in flux balance models of metabolism. *J of Theo Biol*, **213**(1), 73–88.

Thuillier, K.*et al.* (2021). Learning boolean controls in regulated metabolic networks: A case-study. In *CMSB*, volume 12881 of *LNCS*, pages 159–180. Springer.