MERRIN: MEtabolic Regulation Rule INference from time series data

Kerian Thuillier, Caroline Baroukh, Alexander Bockmayr, Ludovic Cottret, Loïc Paulevé, Anne Siegel

MERRIN: MEtabolic Regulation Rule INference from time series data

Kerian Thuillier [1*], Caroline Baroukh [2], Alexander Bockmayr [3], Ludovic Cottret [2], Loïc Paulevé [4], Anne Siegel [1*]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[2]LIPME, INRAE, CNRS, Université de Toulouse, Castanet-Tolosan, France
[3]Freie Universität Berlin, Institute of Mathematics, D-14195 Berlin, Germany
[4]Univ. 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
Supplementary information: See supplementary PDF and https://doi.org/10.5281/zenodo.6670165

1 Introduction

The regulation of metabolic gene expression is essential for an organism to respond appropriately to changes in its environment. For three decades now, methods have been developed to model, simulate and infer gene regulatory networks (de Jong, 2002; Bernot *et al.*, 2004; Chaves *et al.*, 2010). Even with the advances of next generation -omics, such networks remain largely incomplete and unable to accurately predict complex responses of organisms submitted to changes in diverse environments.

The methods developed so far to infer Boolean dynamics of regulatory and signaling networks only rely on information on the regulatory layer of the cell, mainly transcriptomics, proteomics and phosphoproteomics (Saez-Rodriguez *et al.*, 2009; Videla *et al.*, 2017; Razzaq *et al.*, 2018; Tsiantis *et al.*, 2018; Chevalier *et al.*, 2019). However, studying the metabolic layer could help to better infer the regulatory rules. Catabolic repression is a good illustration of how metabolism can highlight regulations inside the cell. This happens when the cell first consumes one substrate (e.g. hexose) until it is exhausted before starting to consume other substrates present in the environment (Monod, 1942). Looking only at the metabolites in the environment, we can infer that a regulation takes place inside the cell, probably on transporters.

Up to now, very few approaches exploited the metabolic layer of the organism to obtain regulatory information. In (Tournier *et al.*, 2017), Resource Balance Analysis (RBA) (Goelzer *et al.*, 2015) is used to infer logical rules governing the activation of metabolic fluxes in response to diverse extracellular media. However, the authors assume that no feedback from metabolism to regulation occurs, which does not correspond to the biological functioning of the cell in most cases.

The fact that metabolic and regulatory layers are of different nature, and thus formalized differently, makes the inference of regulations challenging. The metabolic layer is usually modeled by a metabolic network consisting of a weighted hypergraph with metabolites as nodes, reactions as hyperarcs, and stoichiometry as weights. The (dynamic) response of the metabolism to the environment is usually modeled by Flux Balance Analysis (FBA) (Orth

et al., 2010) resp. dynamic FBA (dFBA) (Mahadevan *et al.*, 2002). This approach assumes that the metabolism of the cell is at quasi steady-state and that the cellular behavior is optimal with respect to some objective (usually growth). FBA and dFBA require solving linear programming problems; the output is the prediction of metabolic fluxes and the concentrations of environmental metabolites and biomass, which are all continuous quantitative data. On the contrary, the dynamics of the regulatory layer is often modeled by Boolean networks (BNs). Combining both layers to infer regulations of the cell and taking into account feedbacks between them thus requires to use a hybrid discrete-continuous modeling and inference framework, such as Satisfiability Modulo Theories (SMT), which was used in Frioux *et al.* (2019) to solve a metabolic network completion problem.

In this study, we present a hybrid discrete-continuous approach to infer metabolic regulations, which combines linear programming for metabolism with answer set programming for regulations. The input consists of a metabolic network, a prior knowledge regulatory network with potential regulations, and time series data. These can be metabolomics data (kinetics of environmental metabolites/biomass and/or fluxomics) and/or expression data from proteomics or transcriptomics. The output is a set of Boolean regulatory networks that best explain the available data. We tested our method on data generated from a dynamic regulatory FBA (d-rFBA) model of a core regulated metabolic network (Covert *et al.*, 2001; Marmiesse *et al.*, 2015), by simulating both the regulatory and the metabolic layer in five environments. In order to assess its robustness, the method was also evaluated with noisy and partial data, *e.g.* transcriptomics and kinetics of environmental metabolites only.

2 Methods and implementation

2.1 d-rFBA: coupling metabolic and regulatory networks

2.1.1 Regulated metabolic networks (RMN), influence graph

A *regulated metabolic network* (RMN) consists of (i) a metabolic layer characterized by linear constraints on metabolic fluxes and (ii) a regulatory layer specified by a Boolean network (BN) which models the interplay between metabolic fluxes, input metabolites, and regulatory proteins.

Formally, a RMN is a quadruple $(\mathcal{N}, \mathrm{Inp}, \mathcal{P}, f)$ composed of (i) a metabolic network $\mathcal{N} = (\mathrm{Int}, \mathrm{Ext}, \mathcal{R}, S)$ with a set of internal metabolites Int, a set of external metabolites Ext, a set of irreversible reactions $\mathcal{R}$ and a stoichiometric matrix $S \in \mathbb{R}^{(|\mathrm{Int}|+|\mathrm{Ext}|)\times|\mathcal{R}|}$. Each reaction $r \in \mathcal{R}$ is associated with flux bounds $l_r, u_r \in \mathbb{R}, 0 \leq l_r \leq u_r$; (ii) a set of input metabolites $\mathrm{Inp} \subseteq \mathrm{Ext}$; (iii) a set of regulatory proteins $\mathcal{P}$; (iv) a BN $f : \mathbb{B}^n \to \mathbb{B}^n, \mathbb{B} = \{0, 1\}$, of dimension $n = |\mathrm{Inp}| + |\mathcal{R}| + |\mathcal{P}|$. We call $f_i : \mathbb{B}^n \to \mathbb{B}$ the *local function* of component i.

The *influence graph* $G(f)$ summarizes the regulatory dependencies. It is a signed directed graph with node set $\mathrm{Inp} \cup \mathcal{R} \cup \mathcal{P}$ and a positive (resp. negative) edge from j to i if there exists $x \in \mathbb{B}^n$ such that an increase of x_j leads to an increase (resp. decrease) of $f_i(x)$. We assume that f is *locally monotone*, *i.e.*, there exists at most one edge from j to i, but our method does not rely on this assumption. In RMNs, the regulation of reactions has to be mediated by regulatory proteins $\mathcal{P}$. Therefore, there is no edge from j to i in $G(f)$ where both $i, j \in \mathrm{Inp} \cup \mathcal{R}$. Edges between regulatory proteins $i, j \in \mathcal{P}$, however, are possible.

2.1.2 Regulatory-metabolic steady states (RMSSs)

Dynamic regulatory Flux Balance Analysis (d-rFBA) (Covert *et al.*, 2001) extends FBA to derive a discrete time series of steady states optimal for a linear objective. In d-rFBA, a *regulatory-metabolic steady state* (RMSS) of a RMN $(\mathcal{N}, \mathrm{Inp}, \mathcal{P}, f)$ is a triple (v, c, x) associating reaction fluxes v at steady state, concentrations c of external metabolites, and the state x of the Boolean network, which comprises the Boolean regulatory state of reactions and regulatory proteins, and the binarization of the concentration of input metabolites. The reaction fluxes v are constrained by both the regulatory variables x, which can force reaction fluxes to be zero, and by the concentration of external metabolites c, which set upper bounds on uptake fluxes. Formally, a RMSS is a triple $(v, c, x) \in \mathbb{R}^{|\mathcal{R}|} \times \mathbb{R}^{|\mathrm{Ext}|} \times \mathbb{B}^{|\mathrm{Inp}|+|\mathcal{R}|+|\mathcal{P}|}$ such that

$(1.a)\ S_{\mathrm{Int},\mathcal{R}} \cdot v = 0,$

$(1.b)\ \forall r \in \mathcal{R}, l_r \cdot x_r \leq v_r \leq u_r \cdot x_r$

$(1.c)\ \forall m \in \mathrm{Inp}, r \in \mathcal{R}, S_{mr} < 0 \Rightarrow v_r \leq \mathrm{uptake\_bound}(c_m),$

where $S_{\mathrm{Int},\mathcal{R}}$ is the submatrix of S whose rows correspond to internal metabolites and $\mathrm{uptake\_bound}(c_m)$ is the maximum flux through uptake reaction r for input metabolite concentration c_m (Varma and Palsson, 1994).

2.1.3 Dynamics of RMNs and admissible time series

The d-rFBA models are executed at two time scales: the metabolic network, considered as a fast system, depending on the activity of input metabolites and regulatory proteins, rapidly converges to a steady state; the regulatory network, considered as a slow system, gets updated once the metabolic network is in steady state. The overall dynamics is guided by the objective of maximizing the flux through reaction *Growth*, assumed to reflect the growth of the cell (Feist and Palsson, 2010).

Let $\beta : \mathbb{R}^n_{\geq 0} \to \mathbb{B}^n$ be a binarization function such that $\forall s \in \mathbb{R}^n_{\geq 0}, \forall i \in \{1, \ldots, n\}$, $\beta(s)_i = 1$ if and only if $s_i > 0$, else $\beta(s)_i = 0$. Given a RMSS (v^k, c^k, x^k) at time t^k, a successor RMSS $(v^{k+1}, c^{k+1}, x^{k+1})$ at time t^{k+1} is computed as follows:

1. The external metabolite concentrations c^{k+1} are computed from the previous concentrations c^k by considering constant uptake/secretion fluxes v^k for the whole time period $[t^k, t^{k+1}]$.

2. The Boolean state x^{k+1} is computed by applying the regulatory function f to the binarized input metabolites concentrations $x'_{\text{Inp}} = \beta(c^{k+1}_{\text{Inp}})$ at time t^{k+1}, together with the binarized reaction fluxes $x'_{\mathcal{R}} = \beta(v^k)$ and the Boolean values $x'_{\mathcal{P}} = x^k_{\mathcal{P}}$ of the regulatory proteins at time t^k, *i.e.*, $x^{k+1} = f(x')$.

3. $(v^{k+1}, c^{k+1}, x^{k+1})$ is a RMSS maximizing the flux through the *Growth* reaction, *i.e.*, there is no RMSS (v', c^{k+1}, x^{k+1}) such that $v'_{Growth} > v^{k+1}_{Growth}$.

Such simulations can be computed with the FlexFlux implementation of d-rFBA (Marmiesse *et al.*, 2015), which considers a fixed time step τ between successive RMSS, see Thuillier *et al.* (2021) for details.

Let $\mathbb{S}$ be the set of all RMSSs of the RMN $(\mathcal{N}, \text{Inp}, \mathcal{P}, f)$. For input metabolite concentrations $c_0 \in \mathbb{R}^{|\text{Ext}|}$ and the regulatory state $x_0 \in \mathbb{B}^{|\text{Inp}|+|\mathcal{P}|+|\mathcal{R}|}$, we denote by $\max_{Growth} \text{rMSS}(c_0, x_0) = \max\{v_{Growth} \mid (v, c_0, x_0) \in \mathbb{S}\}$ the maximum growth flux given c_0 and x_0. Given reaction fluxes $v, v' \in \mathbb{R}^{|\mathcal{R}|}$, external metabolite concentrations $c, c' \in \mathbb{R}^{|\text{Ext}|}$, and regulatory states $x, x' \in \mathbb{B}^{|\text{Inp}|+|\mathcal{R}|+|\mathcal{P}|}$, d-rFBA enables a transition from (v, c, x) to (v', c', x') if and only if the following constraints are satisfied:
(2.*a*) $c' = \text{update}(c, v)$,
(2.*b*) $x' = f(\beta(c'_{\text{Inp}}), \beta(v), x_{\mathcal{P}})$,
(2.*c*) $(v', c', x') \in \mathbb{S}$,
(2.*d*) $v'_{Growth} = \max_{Growth} \text{rMSS}(c', x')$,
where $\text{update}(c, v)$ updates the external metabolite concentrations c according to reaction fluxes, stoichiometry, and cell volume changes. Eq.(2.c) encompasses Eqs.(1.a-c). As shown in Thuillier *et al.* (2021), one can derive a necessary Boolean condition for these constraints (see Suppl. Sect. 2), which we denote by Eq.($2.\text{c}_{\text{relaxed}}$).

2.2 The inference problem for regulatory rules

Next we address the compatibility between the d-rFBA dynamics of a RMN and given time series data for reaction fluxes, regulatory protein states and input metabolite concentrations.

Observed time series. An *observation* is a triple $o = (v_{Growth}, c, x_{\mathcal{P}})$, where (i) $v_{Growth} \in \mathbb{R}$ denotes a *Growth* flux, (ii) $c \in \mathbb{R}^{|\text{Inp}|}$ the input metabolite concentrations, (iii) $x_{\mathcal{P}} \in (\mathbb{B} \cup \{\perp\})^{|\mathcal{P}|}$ represents regulatory protein states, which can be either Boolean values or undefined ("$\perp$"). An *observed time series* is a sequence of observations $T_O = (o_0, \cdots, o_m), m \geq 0$.

Compatibility between an observed time series and a RMN. A RMN and an observed time series $T_O = (o_0, \cdots, o_m)$, with $o_i = (v_{Growth_i}, c_i, x_{\mathcal{P}_i}), 0 \leq i \leq m$, are said to be *compatible with maximum distance* $K \in \mathbb{N}$ *and noise rate* $0 \leq \epsilon < 1$ if there exists a d-rFBA simulation $T_S = (\hat{s}_0, \ldots, \hat{s}_l), l \geq m$, of the RMN, with RMSS $\hat{s}_j = (\hat{v}_j, \hat{c}_j, \hat{x}_j), 0 \leq j \leq l$, and a function $g : \{0, \ldots, m\} \to \{0, \ldots, l\}$ associating each observation with a RMSS, such that the following conditions are satisfied for $0 \leq i \leq m$:
(3.*a*) $0 < g(i+1) - g(i) \leq K$,
(3.*b*) $\hat{x}_{g(i)_{\text{Inp}}} = \beta(c_i)$,
(3.*c*) $\forall p \in \mathcal{P}, x_{i_p} \neq \perp \implies \hat{x}_{g(i)_p} = x_{i_p}$,
(3.*d*) $\frac{v_{Growth_i}}{1+\epsilon} \leq \max_{Growth} \text{rMSS}(c_i, \hat{x}_{g(i)}) \leq \frac{v_{Growth_i}}{1-\epsilon}$.
Eq.(3.a) states that consecutive observations are separated by at most K d-rFBA simulation steps. Eq.(3.b) ensures the complete match between the discretized values of the d-rFBA simulation and the observed inputs. Eq.(3.c) constrains the Boolean states of proteins in the d-rFBA simulation to be equal to the observed ones, when available. Eq.(3.d) states that the simulated growth is close (up to the allowed noise) to the observed growth.

Inference problem. Eqs.(2) in Sect. 2.1.3 characterize the admissible sequences of RMSSs w.r.t. a given RMN and Eqs.(3) the compatibility between

a RMN and an observed time series. The problem of inferring regulatory rules compatible with a set of observed time series is:

Problem statement tackled by MERRIN: Inferring regulatory rules from observed time series

Input:

1: a set of observed time series $\{T^1, \cdots, T^q\}, q \geq 1$;
2: a metabolic network $\mathcal{N} = (\text{Int}, \text{Ext}, \mathcal{R}, S)$;
3: a set of regulatory proteins $\mathcal{P}$;
4: a prior knowledge network (PKN) $\mathcal{G}$ whose nodes belong to $\text{Inp} \cup \mathcal{P} \cup \mathcal{R}$ and such that there is no $i \xrightarrow{s} j \in \mathcal{G}$ with $i, j \in \text{Inp} \cup \mathcal{R}$;
5: a noise parameter $\epsilon \in [0, 1[$;
6: a maximum distance $K \in \mathbb{N}$ between observations.

Output: All BNs $f \in \mathbb{B}^{|\text{Inp}|+|\mathcal{R}|+|\mathcal{P}|}$ such that:

1: f is locally monotone;
2: $G(f) \subseteq \mathcal{G}$;
3: for each T^i the associated RMN $(\mathcal{N}, \text{Inp}, \mathcal{P}, f)$ has a d-rFBA simulation T_S compatible with T^i (satisfying Eqs.(3));
4: there is no BN $f' \in \mathbb{F}$ smaller than f considering the local functions in disjunctive normal form (subset minimality ordering).

In practice, we focus on the *smallest* (subset-minimal) compatible BNs by considering a partial ordering between BNs based on the disjunctive normal form (DNF) of the local functions (Chevalier *et al.*, 2019). However, our approach can be used to enumerate all compatible BNs, not only the subset-minimal ones.

2.3 Resolution using hybrid Answer Set Programming

The inference problem relies on hybrid optimization as it requires exploring the combinatorial domain of putative regulatory BNs constrained by the PKN, and checking both combinatorial constraints linking consecutive states of regulatory proteins according to a given observed time series (Eq.(2.b) and Eqs.(3.b-c)) and linear arithmetic constraints related to the characterization of RMSSs and v_{Growth} optimization (Eqs.(1), Eqs.(2.c-d), Eq.(3.d)). To solve this problem, we used SMT (Satisfiability Modulo Theory) solving (Barrett and Tinelli, 2018; Janhunen *et al.*, 2017), by implementing a resolution framework relying on constraint propagation: whenever a solution satisfying the combinatorial part is found, the linear part is checked. If the linear check succeeds then the solution is accepted.

Algorithm 1 Hybrid Resolution: $T = \{T^1, \cdots, T^q\}, \mathcal{N}, P, \mathcal{G}, \epsilon, K$

1: $\text{Inp} \leftarrow \{m \mid m \in \text{Ext}, \exists r \in \mathcal{R}, S_{mr} > 0\}$
2: $n \leftarrow |\text{Inp}| + |\mathcal{R}| + |\mathcal{P}|$
3: $\mathbb{F} \leftarrow \{f \mid f \in \mathbb{B}^n \rightarrow \mathbb{B}^n, G(f) \subseteq \mathcal{G} \wedge f \text{ is locally monotone}\}$

[ASP solving]

4: select $\hat{f} \in \mathbb{F}$ verifying (2.a), (2.b) and $(2.c_{\text{relaxed}})$
5: $\mathcal{RMN} \leftarrow (\mathcal{N}, \text{Inp}, \mathcal{P}, \hat{f})$
6: **for all** $T^i \in T$ **do**
7: select a family of RMSS $\{\hat{s}^i_0, \cdots, \hat{s}^i_{l_i}\}$ of the $\mathcal{RMN}$ satisfying constraints (3.a), (3.b) and (3.c)
8: **end for**

[Linear solving]

9: check with linear programming whether (2.c) and (3.d) hold
10: **if** (2.c) and (3.d) hold **then**
11: $\hat{f}$ is a solution
12: **else**
13: **for all** o^i_j and its associated RMSS $\hat{s}^i_k$ **do**
14: $o^i_j = (v^i_{Growth_j}, c^i_j, x^i_j)$ and $\hat{s}^i_k = (\hat{v}^i_k, \hat{c}^i_k, \hat{x}^i_k)$
15: **if** $\hat{v}^i_{Growth_k} > (v^i_{Growth_j})/(1-\epsilon)$ **then**
16: add Eq.(4) with $x = \hat{x}^i_k$
exclude any RMSS associated with o^i_j that do not verify Eq.(4).
17: **else if** $\hat{v}^i_{Growth_k} < (v^i_{Growth_j})/(1+\epsilon)$ **then**
18: add Eq.(5) with $x = \hat{x}^i_k$
exclude any RMSS associated with o^i_j that do not verify Eq.(5)
19: **end if**
20: **end for**
21: return to step 4
22: **end if**

If it fails then the solution is rejected and new constraints are added to the combinatorial part to avoid alternative solutions which would for sure fail the linear check as well.

The inference from purely combinatorial constraints was formulated using Answer Set Programming (ASP) (Baral, 2003; Gebser *et al.*, 2012), a logic programming framework for expressing symbolic satisfiability problems. Modern solvers like Clingo (Gebser *et al.*, 2017) support various reasoning modes, including subset-minimal enumeration. The linear arithmetic constraints were formulated in linear programming.

The constraint propagation exploits a monotonicity property of the objective v_{Growth} of RMSSs: for fixed input metabolite concentrations,

inhibiting (*resp.* releasing an inhibition of) a reaction cannot increase (*resp.* decrease) the maximum value of v_{Growth}. Thus, given input metabolite concentrations $c_0 \in \mathbb{R}^{|\text{Inp}|}$ and an optimal RMSS (v, c_0, x), we can characterize optimal RMSS (v', c_0, x') for which $v'_{Growth} \leq v_{Growth}$ (Eq.(4)) resp. $v'_{Growth} \geq v_{Growth}$ (Eq.(5)) by requiring

(4) $\forall r \in \mathcal{R}, x'_r \leq x_r$ resp. (5) $\forall r \in \mathcal{R}, x'_r \geq x_r$.

This allows performing constraint propagation during the combinatorial resolution and further reducing the number of linear programming checks.

Algorithm and implementation. The hybrid resolution of the inference problem is detailed in Algorithm 1. For the sake of simplicity, we explain the global solving scheme on the full time series T, although the software implementation extends this algorithm to incomplete time series. In practice, Algorithm 1 is implemented by extending the Clingo solver, using its Python API, with a linear constraint propagator, implemented with the python PuLP library, and the solver COIN (Forrest *et al.*, 2022). Each problem instance was executed on Fedora 34 with an 8 core processor i7-1165G7@2.80GHz and 16GB of RAM.

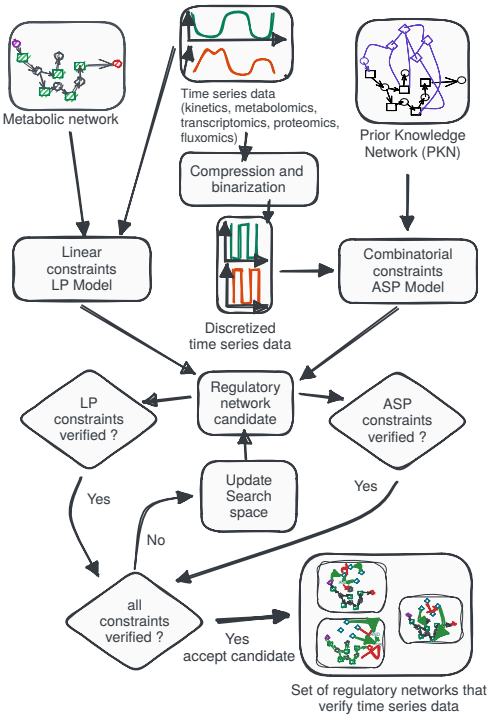
3 Results

3.1 MERRIN workflow

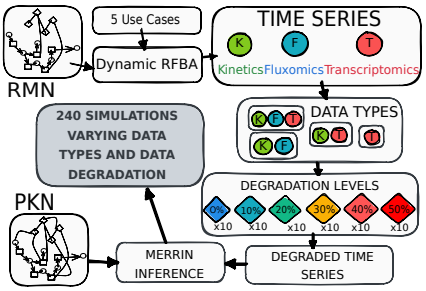
The MEtabolic Regulation Rule Inference (MERRIN) software implements the workflow in Fig. 1(a) to infer regulatory rules of a RMN from possibly incomplete and noisy observed time series (Sect. 2.2 and 3.2) using Algorithm 1.

MERRIN takes as *input* (i) a metabolic network $\mathcal{N} = (\text{Int}, \text{Ext}, \mathcal{R}, S)$ in SBML format, (ii) a set of regulatory proteins $\mathcal{P}$ (ii) a set of observed time series $T = \{T^1, \ldots, T^q\}$ with their type (complete, kinetic-fluxomic, kinetic-transcriptomic, transcriptomic) in CSV format, and (iii) a prior knowledge network (PKN) $\mathcal{G}$ in text format. To allow for incomplete and noisy time series, two parameters can be set: (i) $K \in \mathbb{N}$ the maximum number of intermediate unobserved RMSSs for each time series; (ii) $\epsilon \in [0, 1[$ the estimated noise rate. For the rest of the paper, we will consider $\epsilon = 0.3$ and $K = 10$.

The *search space* $\mathbb{F}$ consists of all Boolean networks (BNs) f of dimension $n = |\text{Inp}| + |\mathcal{R}| + |\mathcal{P}|$ whose influence graph $G(f)$ is a subgraph of the PKN $\mathcal{G}$. The size of $\mathbb{F}$ is doubly exponential in n. MERRIN returns as *output* all subset-minimal locally monotone regulatory BNs $f \in \mathbb{F}$ such that the associated RMN $(\mathcal{N}, \text{Inp}, \mathcal{P}, f)$ is compatible with the observed time series $T = \{T^1, \ldots, T^q\}$.



(a) MERRIN software

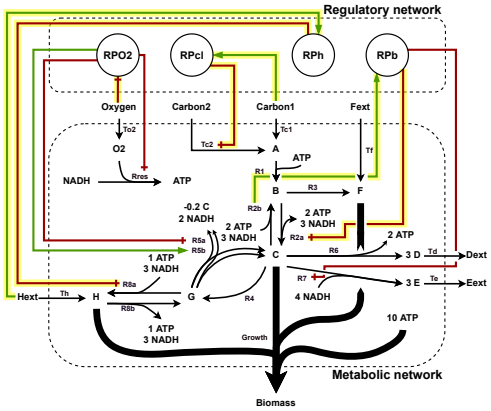


(b) Data generation procedure

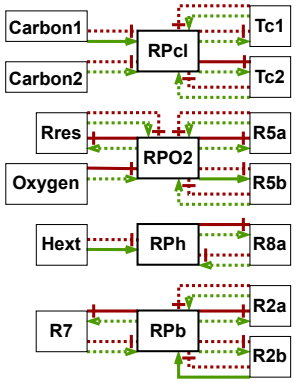
Fig. 1: **(a)** Workflow of the **MERRIN software** for metabolic regulation rule inference. **(b) Degraded time series generation procedure**: generation of 240 time series for the RMN of (Covert *et al.*, 2001), with different levels of incompleteness and noise.

3.2 Application to a core regulated metabolic model

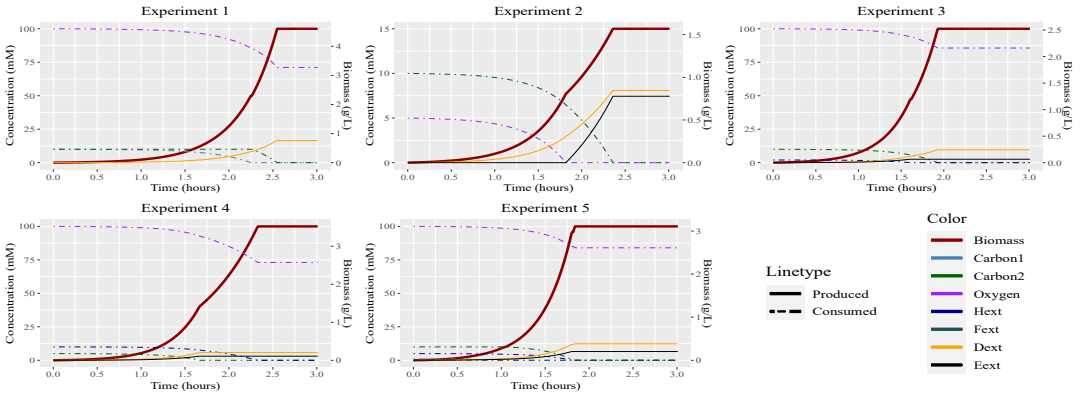
Problem instance. To validate our approach, we applied MERRIN to synthetic data generated for a core regulated metabolic network originally proposed in (Covert *et al.*, 2001), which we refer to as the *gold standard*. (i) The *metabolic layer* of the gold standard (see Fig. 2(a)), also serving as input for MERRIN, contains 20 reactions and 8 exter-



(a) Regulated metabolic network (RMN)



(b) Prior knowledge network (PKN)



(c) Inferred RMN simulations

Fig. 2: **(a)** Regulated metabolic network from (Covert *et al.*, 2001). Lower part is the metabolic network. The nodes are metabolites and the black hyperedges are reactions. Upper part is the regulatory network. The nodes are regulatory proteins. Edges represent the Boolean functions: *green edges* denote activation, *red edges* inhibition. Yellow highlighted edges are the inferred regulation from the complete noise-free time series. **(b)** Set of permitted interactions use for the inference. Red edges, solid and dot, are inhibitions. Green edges, solid and dot, are activations. The set of solid edges describes the influence graph of the regulatory network of (a). **(c)** FlexFlux simulations of the inferred RMN (yellow highlighted regulations in (a)) using the experimental conditions of (Covert *et al.*, 2001). These simulations are identical to the simulations of the reference RMN.

nal metabolites, among them the 5 inputs Carbon1, Carbon2, Oxygen, Fext, Hext. (ii) The *regulatory layer* of the gold standard involves the four regulatory proteins RPcl, RPO2, RPb, RPh. (iii) In order to explore alternative regulatory rules that could explain the observed time series data, we consider the *PKN* in Fig. 2(b), which includes for each edge in the influence graph of the gold standard all possible combinations of signs and directions. Moreover, two edges from Carbon2 to RPcl, and four edges between RPcl and Tc1 were added as possible alternative regulations to be explored. It follows that the search space to be explored by MERRIN contains $\approx 1.8 \times 10^{15}$ locally monotone BNs, including the gold standard.

Degraded time series generation. We used the workflow in Fig. 1(b) to generate a benchmark of 240 time series sets. First FlexFlux (Marmiesse *et al.*, 2015) was used to generate complete *kinetic-fluxomic-transcriptomic* (KFT) d-rFBA simulation data for the five environmental conditions of the core RMN (see Suppl. Sect. 3.1), each yielding 301 RMSS (initial biomass $= 0.1 g.L^{-1}$, steps $= 300$, intervals $= 0.01h$). Then, for each complete KFT time series, we generated (i) a *kinetic-fluxomic* (KF) time series by removing the values of the regulated proteins, (ii) a *kinetic-transcriptomic* (KT) time series by discretizing all fluxes to binary values (iii) a *transcriptomic* (T) time series by discretizing all fluxes and metabolite concentrations to binary

values. The resulting time series were further compressed by removing redundant time points to emulate biological experiments where only a few selected measurements are made. Finally, for each of the five environmental conditions and each type of data (KFT, KF, KT, T), we generated 60 random time series at different noise rates (0%, 10%, 20%, 30%, 40% and 50%), by randomly deleting time points and increasing or decreasing quantitative values. Altogether we obtained 240 sets of 5 incomplete and/or noisy time series, each including 6 to 18 time points after the compression step.

Inference scores. The quality of MERRIN predictions was evaluated on two different levels. First, we measured the distance between the observed time series, on which the inference was based, and the time series obtained by simulating the inferred model. The distance between two RMSS time series $S = \{s^0, \cdots, s^m\}$ and $\hat{S} = \{\hat{s}^0, \cdots, \hat{s}^m\}$ w.r.t. a set of components A was computed as the *residual sum of squares* (RSS): $RSS_A = \sum_{i=0}^{m} \sum_{a \in A} (s_a^i - \hat{s}_a^i)^2$. We used $\text{RSS}_{\mathcal{P}}$ to measure the accuracy of the prediction of the time series for the four regulatory proteins (RPcl, RPO2, RPh, RPb) and RSS_{Ext} to measure the accuracy of the prediction of the time series of the eight external metabolites (Carbon1, Carbon2, Oxygen, Hext, Fext, Dext, Eext, Biomass).

Second, we measured the ability of MERRIN to infer the expected regulations using the recall and precision of the inferred BN. Given BNs f and $\hat{f}$, the *recall* of $G(\hat{f})$ w.r.t. $G(f)$ is the fraction of edges of $G(f)$ in $G(\hat{f})$, *i.e.*, recall $= |G(f) \cap G(\hat{f})|/|G(f)|$, where $|G(f)|$ denotes the number of edges. The *precision* of $G(\hat{f})$ w.r.t. $G(f)$ is the fraction of edges of $G(\hat{f})$ in $G(f)$, *i.e.*, precision $= |G(f) \cap G(\hat{f})|/|G(\hat{f})|$.

3.3 Performance of MERRIN on complete data

MERRIN was first applied to the complete noise-free kinetic-fluxomics-transcriptomics (KFT) time series corresponding to the five different environmental conditions. On this input, MERRIN inferred exactly one smallest regulatory BN in 6.95s. The inferred regulatory rules are shown with yellow highlighted edges in Fig. 2(a). The BN contains seven regulatory rules (for RPO2, RPcl, RPh, RPb, Tc2, R2a and R8a) of the gold standard, three of which regulate reaction activity. It has a *precision* of 1, meaning that all seven regulatory rules are in the gold standard; and a *recall* of 0.64, because four of the regulatory rules of the gold standard have not been retrieved (rules for R5a, R5b, R7 and Rres). Both *RSS*s are equal to 0: although the recall is not 1, the d-rFBA simulations of the five experiments with the inferred regulatory BN (Fig. 2(c)) match exactly the complete noise-free time series. The unrecovered regulatory rules of the gold standard are not necessary to explain the observed time series.

This is consistent with the discussion in (Covert *et al.*, 2001) that the regulation of *Rres* is not necessary for the optimal solution. Biologically, this regulation is only present to ensure that unnecessary respiratory enzymes decay in an anaerobic environment. However, since enzyme amounts are not explicitly represented in the d-rFBA framework, the time series do not reflect this biological behavior, hampering the inference of the regulation. Similarly, R5a and R5b were introduced in the RMN to model that aerobic and anaerobic carbon synthesis is catalyzed by different enzymes. However, these enzymes are not included in the model and both reactions are strictly equivalent. It is therefore not surprising that MERRIN cannot infer the regulation stating which of the two reactions should be selected. Finally, the missing regulation of R7 in the inferred RMN is explained by the fact that R7 cannot be activated in d-rFBA simulations optimizing growth because its activation would consume carbon and energy, leading to a decrease in biomass synthesis. Therefore, regulating *R7* is not necessary to explain its activity in the simulations.

3.4 Impact of data incompleteness and noise

Range of application of MERRIN. When considering higher degradation rates (40% and 50%), 9 of the 60 test instances reached the time limit of 600s (see Suppl. Sec. 3.2.1). The number of BNs also increased drastically at 50% degradation, as well as the RSS scores, suggesting that the degradation rate of 30% is the limit for the MERRIN approach. As shown in Suppl. Sect. 3.2.2, we also tested the case of kinetic-fluxomics instances. Such instances do not contain any information on the four regulatory protein states, making it difficult to infer regulatory rules between proteins and reactions. As expected, MERRIN is not able to correctly determine the regulatory rules controlling them. This leads to time-consuming enumeration of a very large number of BNs, all compatible with the observed time series, but considering all the possible regulatory protein states. Based on these results, we suggest to use MERRIN only on kinetics and transcriptomics real data sets. According to the design of MERRIN, proteomics data can be viewed as alternative to transcriptomics data if they are available. There-

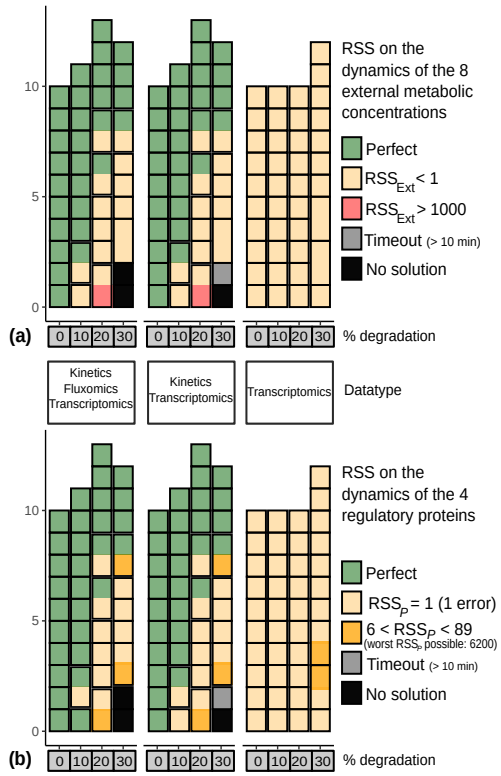


Fig. 3: **RSS depending on data type and degradation level.** Each vertical bar corresponds to the results of MERRIN on the 10 instances associated with a considered data type (KFT, FT, T) and degradation type (0%, 10%, 20% and 30%). Each square corresponds to one solution and its color to RSS ranges (see legend). A black edge separates the MERRIN results on the different instances.

fore, in the following, we focus only on the data types KFT (kinetic-fluxomics-transcriptomics), KT (kinetics-transcriptomics) and T (transcriptomics) with a degradation rate between 0 and 30%, which represents 120 instances.

Number of models inferred by MERRIN. Fig. 3 shows the number of subset-minimal models inferred by MERRIN in the given time limit for the 120 tested instances . When a solution was reached in the time limit, MERRIN inferred at most two subset-minimal models. In total, 134 BNs were inferred from the 120 instances. Among these 134 BNs, there were only 15 different BNs, see Fig. 2 in the Suppl. Sect. 3.2.2. For each of these models, we computed the precision and recall (see Sect.3.2) with respect to the gold standard (see Suppl. Sect. 3.2.3, Fig. 3). For 110 instances out of 120, the precision is equal to 1, meaning that all the regulatory rules inferred in these BNs are present in the gold standard. The maximum recall is equal to 0.64, while the minimum recall is 0.55.

Performance. Among the 120 instances of our benchmark, only one has reached the time limit (grey square in Fig. 3). For this instance, we do not have any information whether or not there is a solution. In 3 out of the 120 instances (Fig. 3), MERRIN reported that no BN satisfied the constraints. This happens only at 30% noise rate. For the 116 other instances, the average inference time was 25.975s.

Simulation scores. For each of the 134 BNs inferred, we compared the associated d-rFBA time series of external metabolites and regulatory proteins to the ones of the gold standard using the $\mathrm{RSS}_{\mathrm{Ext}}$ score (Fig. 3(a)) and the $\mathrm{RSS}_{\mathcal{P}}$ score (Fig. 3(b)). In Fig. 3, green squares correspond to cases where MERRIN inferred a unique BN whose associated RMN has exactly the same r-dFBA simulations as the gold standard ($\mathrm{RSS}_{\mathrm{Ext}} = 0$ (Fig. 3(a)) and $\mathrm{RSS}_{\mathcal{P}} = 0$ (Fig. 3(b))). Interestingly, the same BN was inferred for each green square, and this BN is the same as the one obtained on complete data (Fig. 2(a)) Yellow squares of Fig. 3 stand for BNs reproducing the gold standard RMN simulations with a very small error. These errors are due to missing regulatory rules. For example, all the BNs with $\mathrm{RSS}_{\mathrm{Ext}}$ ¡ 1 and $\mathrm{RSS}_{\mathcal{P}} = 1$ are BNs for which the regulatory rule of reaction R2a has not been inferred. Red squares correspond to the worst possible $\mathrm{RSS}_{\mathrm{Ext}}$ ($> 1\,000$), equivalent to cases in which no regulatory rules were inferred. This happens twice among the 120 experiments.

Impact of degradation rate. A vertical bar of 10 green squares in Fig. 3 means that MERRIN inferred, for each of the 10 test instances, a unique BN that perfectly matches the gold standard. This occurred only for KT and KFT instances with no degradation in the input time series. $\mathrm{RSS}_{\mathrm{Ext}}$ and $\mathrm{RSS}_{\mathcal{P}}$ increased with the degradation rate, as one should expect. However, most of the RSS scores are very small, emphasizing that the inferred BNs can almost perfectly reproduce the gold standard when the degradation rates is less than 30%.

Impact of the type of data. The results are identical for the complete (KFT) and the kinetic-transcriptomics (KT) instances (except one KP at 30%, which reached the time limit of 600s). This could be expected since MERRIN reasons over binarized fluxomics data, which once binarized are

identical to the qualitative information provided by transcriptomics data. In addition, the inferred BNs from the KFT and KT time series reproduce the gold standard with good precision most of the time, except in two cases (red squares).

For transcriptomics (T) time series instances, our results show that no inferred BN was able to perfectly reproduce the gold standard. However, for each inferred BN both $\mathrm{RSS}_{\mathrm{Ext}}$ and $\mathrm{RSS}_{\mathcal{P}}$ are small: $\mathrm{RSS}_{\mathcal{P}} \leq 1$ for all, except for two instances, and $\mathrm{RSS}_{\mathrm{Ext}} < 1$. This suggests that without information on external metabolite concentrations, it is harder for MERRIN to explain if the observed RMSS is due to some regulations or to a specific combination of external metabolite concentrations. In this case, regulatory rules, such as the rule controlling the reaction R2a, are missed.

4 Discussion and conclusion

We introduced MERRIN, a novel approach to infer rules for metabolic regulation in changing environments. MERRIN is based on the d-rFBA framework, which combines discrete simulations of Boolean networks, modeling the activity of regulatory proteins, with the prediction of metabolic response, based on linear programming.

Advantages of using constraint propagators. A characteristic of the inference problem is that the set of BNs verifying both combinatorial and linear constraints is small compared to the set of BNs verifying only the combinatorial constraints. To address this issue, our resolution implements a Satisfiability Modulo Theory (SMT) approach with a dedicated algorithm for combining Boolean satisfiability with linear programming: we designed a constraint propagation strategy on top of the Answer Set Programming solver Clingo by exploiting a monotonicity property of the optimization objective in RMNs. This strategy reduced substantially the number of candidate solutions to be validated, by generalizing counterexamples satisfying the combinatorial constraints but not the linear ones encountered during the search.

Possible strategies to infer all regulatory rules. MERRIN infers regulations only when they improve the fitting between observations and simulations, which depends on the underlying optimality principle (here optimizing growth). Since the presence of some regulations from the gold standard does not affect the fitting, it is not possible for MERRIN to infer them. Inferring more regulations would require to introduce enzyme amounts and their synthesis. Methods such as r-deFBA (Liu and Bockmayr, 2020), should allow solving this issue.

Impact of the synchronous simulation assumption. The d-rFBA framework as defined in (Covert *et al.*, 2001; Marmiesse *et al.*, 2015) uses synchronous simulation of BNs (the state of all regulatory proteins is updated simultaneously). While our implementation allows considering asynchronous simulation, this results in a less constrained model. Indeed, the fact that a regulatory protein has the same state in two consecutive steady states could be explained either with the application of a regulatory rule, or by the absence of an update. Therefore, considering asynchronous updates would probably require considering further time constraints in order to match the experimental observations.

Use of synthetic data to validate network inference. The validation of methods related to the inference of regulatory rules can be misleading since there is no reference multi-layer data set or reference RMN allowing large-scale validations. As discussed in (Covert *et al.*, 2001) and confirmed in (Thuillier *et al.*, 2021), even in the most complete (small-scale) gold standard RMN introduced in (Covert *et al.*, 2001), some regulatory rules introduced according to literature-based knowledge have no impact on the RMN simulation. To address this issue and to test our approach, we used a benchmark strategy consisting in generating several types of data from the simulations of a gold standard. This allowed testing the robustness of the MERRIN approach in different *scenarios* of data types (combinations of kinetics, fluxomics and transcriptomics data) and noise (up to 50% noise introduced in the data). We argue that such a benchmark strategy could be used in a similar way to test the robustness of any other dynamical network inference method when only few reference data are available.

Impact of data types and quality. According to our results, the performance of MERRIN on kinetic and transcriptomics data is similar to complete data (kinetic, fluxomics and transcriptomics). This suggests that inferring regulatory rules of metabolic networks actually would not require fluxomics data, which are most probably the hardest data to obtain experimentally. In this direction, a perspective to extend the MERRIN approach would be to identify the best experimental designs to discriminate the models associated with the PKN. In addition, MERRIN seems to be sensitive to noise only for single fluxomics data. In all other cases, up to 30%

noise in the data has few impact of the MERRIN performance.

Scalability. The computation times in this study are encouraging for inferring regulations in larger networks. Handling linear constraints reduces to FBA, which can be done efficiently on genome-scale networks. However, this has to be done many times during combinatorial search. Thus, for inferring large-scale regulated metabolic networks improved constraint propagation techniques may become necessary to further prune the combinatorial search space.

Funding

Work of LP is supported by the French Agence Nationale pour la Recherche (ANR), grant number ANR-20-CE45-0001. Work of LC and CB is supported by the French Laboratory of Excellence project "TULIP" (grant number ANR-10-LABX-41; ANR-11-IDEX-0002-02).

References

Baral, C. (2003). *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, New York, NY, USA.

Barrett, C.*et al.* (2018). *Satisfiability Modulo Theories*, pages 305–343. Springer International Publishing, Cham.

Bernot, G.*et al.* (2004). Application of formal methods to biological regulatory networks: extending thomas' asynchronous logical approach with temporal logic. *J of Theo Biol*, **229**(3), 339–347.

Chaves, M.*et al.* (2010). Comparing boolean and piecewise affine differential models for genetic networks. *Acta Biotheor*, **58**(2-3), 217–232.

Chevalier, S.*et al.* (2019). Synthesis of boolean networks from biological dynamical constraints using answer-set programming. In *ICTAI*. IEEE.

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

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *J of Comp Biol*, **9**, 67–103.

Feist, A.M.*et al.* (2010). The biomass objective function. *Curr Opin Microbiol*, **13**(3), 344–349.

Forrest, J.*et al.* (2022). coin-or/cbc: Release releases/2.10.7.

Frioux, C.*et al.* (2019). Hybrid metabolic network completion. *Theory and Practice of Logic Programming*, **19**(1), 83–108.

Gebser, M.*et al.* (2012). *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

Gebser, M.*et al.* (2017). Multi-shot ASP solving with clingo. *CoRR*, **abs/1705.09811**.

Goelzer, A.*et al.* (2015). Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic Engineering*, **32**, 232–243.

Janhunen, T.*et al.* (2017). Clingo goes linear constraints over reals and integers. *Theory and Practice of Logic Programming*, **17**(5-6), 872–888.

Liu, L.*et al.* (2020). Regulatory dynamic enzyme-cost flux balance analysis: A unifying framework for constraint-based modeling. *J of Theo Biol*, **501**, 110317.

Mahadevan, R.*et al.* (2002). Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical Journal*, **83**(3), 1331–1340.

Marmiesse, L.*et al.* (2015). FlexFlux: combining metabolic flux and regulatory network analyses. *BMC Systems Biology*, **9**(1).

Monod, J. (1942). Recherches sur la croissance des cultures bacteriennes. *Ann. Inst. Pasteur*, **69**, 179.

Orth, J.D.*et al.* (2010). What is flux balance analysis? *Nat Biotechnol*, **28**(3), 245–248.

Razzaq, M.*et al.* (2018). Computational discovery of dynamic cell line specific boolean networks from multiplex time-course data. *PLOS Comp Biol*, **14**(10), e1006538.

Saez-Rodriguez, J.*et al.* (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol*, **5**(1), 331.

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.

Tournier, L.*et al.* (2017). Optimal resource allocation enables mathematical exploration of microbial metabolic configurations. *J. Math. Biol.*, **75**(6-7), 1349–1380.

Tsiantis, N.*et al.* (2018). Optimality and identification of dynamic models in systems biology: an inverse optimal control framework. *Bioinf*, **34**(14), 2433–2440.

Varma, A.*et al.* (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl Environ Microbiol*, **60**(10), 3724–3731.

Videla, S.*et al.* (2017). caspo: a toolbox for automated reasoning on the response of logical signaling networks families. *Bioinf*, page btw738.