Deriving Gene Associations for Metabolic Models

Razanne Issa, David James Sherman

Deriving Gene Associations for Metabolic Models

Razanne Issa, David James Sherman

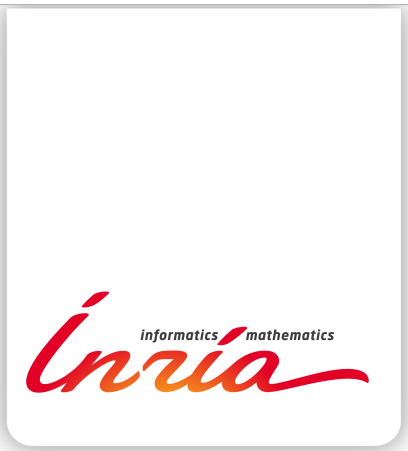
RESEARCH REPORT

N° 9246

December 2018

Project-Team Pleiade

ISSN 0249-6399 ISRN INRIA/RR--9246--FR+ENG



Deriving Gene Associations for Metabolic Models

Razanne Issa, David James Sherman

Project-Team Pleiade

Research Report n° 9246 — December 2018 — 25 pages

Abstract: We define a formal procedure for inferring gene-protein-reaction (GPR) relations from complete metabolic models, using a logical representation of knowledge and a small set of inference rules. We show that different use cases of metabolic models requires difference GPR relations. Three examples from the yeast *Saccharomyces cerevisae* illustrate the procedure and demonstrate its usefulness.

Key-words: metabolic networks, comparative genomics, inference

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Dériver des associations de gène pour des modèles métaboliques

Résumé : Nous définissons une procédure formelle pour l'inférence de relations gène-protéine-réaction (GPR) à partir de modèles métaboliques complets, par moyen d'une représentation logique des connaissances et un petit ensemble de règles d'inférence. Nous montrons que les différents cas d'utilisation requièrent des différentes relations GPR. Trois exemples tirés de la levure *Saccharomyces cerevisae* illustrent la procédure et démontrent son utilité.

Mots-clés : réseaux métaboliques, génomique comparée, inférence

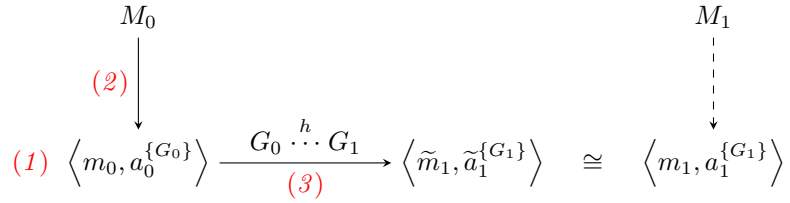


Figure 1: Incremental modeling: we (*1*) define a tuple $\langle m, a^{\{G_0\}} \rangle$ that annotates a minimal enzymatic model m with annotations a_0 defined over a set of variables G, and (*2*) project the tuple from a complete model M. The objective is that (*3*) the inferred model $\langle \widetilde{m}_1, \widetilde{a}_1^{\{G_1\}} \rangle$ is congruent to the ideal model $\langle m_1, a_1^{\{G_1\}} \rangle$ that we would have obtained had M_1 been available.

1 Introduction

Gene Associations or Gene-Protein-Reaction relations (GPR) [RVSP03, NvEF+06] in metabolic models record knowledge about the direct and indirect relations between genes and biochemical reactions. Whether written as boolean formulas or as collections of ⟨gene, protein, reaction)⟩ triplets, gene associations define boolean functions that assert whether a reaction may take place. One widespread use of gene associations is in metabolic model inference for new target organisms. For such organisms there is typically a genome but very little systematic experimental data. The genes in the genome are used to evaluate the gene associations for every reaction in a reaction database; a reaction is included in the inferred model it its gene association is satisfied.

Gene associations play an essential role in incremental model inference (figure 1) as defined in [Iss15]. The incremental method is designed for cases when the target organism is closely related to an organism for which a minimal model $\langle m, a^{\{G_0\}} \rangle$, derived (or derivable) from a complete model M_0, already exists. The gene associations $a^{\{G_0\}}$ annotate the metabolic model m, and are essential in inferring the target model $\langle \widetilde{m}_1, \widetilde{a}_1^{\{G_1\}} \rangle$.

In the general case, however, gene associations are an annotation determined by a human curator on the basis of existing knowledge about the biochemistry and gene regulation of the target organism. While expert creation is the of course the ultimate criterion for correctness, it would be useful to also have a formal criterion defined in terms of the knowledge base.

In this report we describe how a generalized form of gene associations can be used to annotate a minimal enzymatic model, in order to infer specific metabolic models for related organisms based on genomic comparisons. The gene associations will implicitly represent knowledge derived from a hypothetical complete model. We will moreover consider how to rewrite the gene associations during model inference.

Different kinds of models, with different degrees of sophistication, can be defined for the same organism. For this work we have selected four complementary use cases for model inference using gene associations. We will show that each requires a different kind of assertion on the reaction, and consequently a different kind of gene association. Consequently, the step (*1*) of figure 1, which projects a complete model to a minimal model annotated by gene associations, depends on the type of models we wish to infer.

- *Model inference* uses an annotated reference model and homology between the genes in that reference and the genes in a target organism, to infer a model for the target organism [cite]. The assertion that must be tested for each reaction is whether the genome of the target contains the required genes, for example those coding for the enzymes that catalyze

the reaction.

- *Flux balance analysis* (FBA) calculates the fluxes through the reactions that maximize an objective function under a steady-state assumption [cite]. FBA can be used to simulate the effect of gene knockouts on metabolic fluxes. The assertion that must be tested is whether a reaction should be removed from the system because its requisite genes are knocked out or deactivated.

- *Hierarchical modeling* decomposes the complex cellular system into a hierarchy of independently-modeled stochastic transition systems, that communicate through flows and can synchronously perform state transitions through firing of global events [cite]. The assertion that must be tested is whether the *guard* on a transition, defined in terms of state variables such as protein expression or metabolite concentration, permits it to be executed.

- *Dynamic simulation* converts the metabolic model into a system of differential equations that define the evolution of molecule concentrations over time. Gene associations in dynamic simulation assert dependencies between variables that ultimately might be any molecular species involved in the model.

For each of these uses we define the appropriate gene associations and show how to derive them from a complete metabolic model. This derivation will provide a different gene association correctness criterion for each use case.

To illustrate the differences between the four use cases and the corresponding gene associations, we use three examples from *Saccharomyces cerevisiae*:

- Ribonuclease reductase, reaction *RNR*;

- Ferrocytochrome-C oxidoreductase, reaction *CYC*;

- 3-methyl-2-oxopentanoate decarboxylase, reaction *PDC*.

Each of these examples presents a different challenge: the RNR system must take the cell cycle into account, the CYC system must take compartments into account, and the PDR system must take gene regulation into account.

By showing how the appropriate gene associations can be derived from a complete model, we provide the means to define a gold standard for annotation of a metabolic model. However, in practical applications we rarely have a complete model for which gene associations could be derived; rather, gene association annotations are defined at the same time as the definition or the inference of the biochemical reactions in the model. The interesting question in these applications is how to improve the quality of the reactions and their annotations. We show how the use-specific definition of gene associations provides a consistency check on the model being defined.

2 Definitions and Notation

Gene associations or Gene-Protein-Reaction (GPR) relations are boolean formulas constructed using the operators "*AND*" and "*OR*" over two levels: genes to proteins and proteins to reactions. In the first level, the formulas express how proteins are coded by their genes. In the second level, the formulas express the dependencies of reactions on proteins (Figure 2). Identifying GPR relations is essential for the reconstruction of metabolic networks[FP08, FST05, FHT$^+$09] and a critical step in that reconstruction.

The gene associations in a metabolic network can be identified through manual curation, a task that requires extensive experimental data [FGAT09] (e.g UniProtDB [C+11]), or by automatic methods. The latter are founded on knowledge engineering, and are principally based on sequence similarity[KDWV13, WLT05, AJW+08]. Because extensive experimental data are rarely available, efficient automatic methods are necessary to decipher correct annotations [RCO+13].

Gene associations are stored in many public data banks such as KEGG [KG00] and MetaCyc [CFF+06].

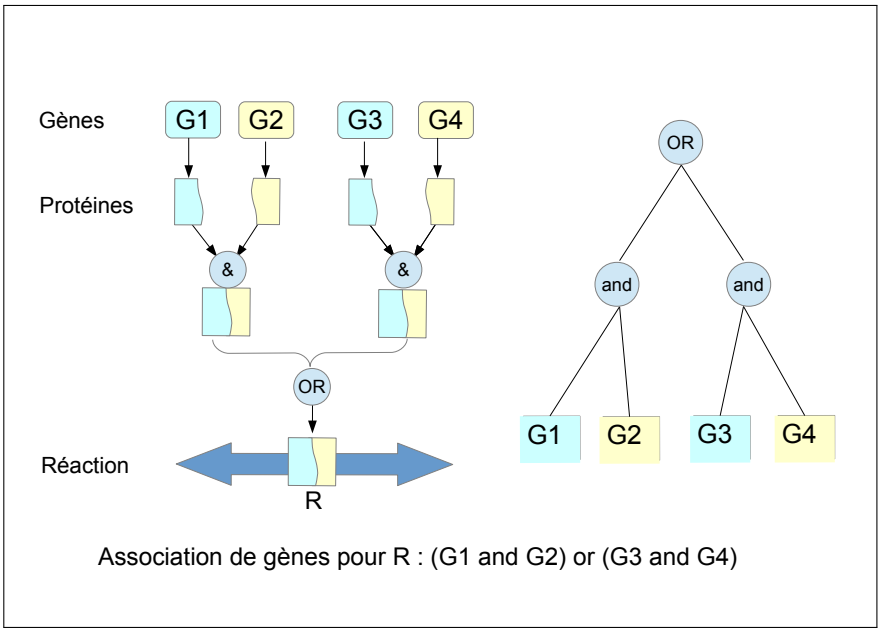


Figure 2: (left) The gene association for a reaction **R** catalyzed by a protein complex comprised either of the aggregation of products of genes **G1** and **G2**, or of the aggregation of the products of **G3** and **G4**. (right) The formula for the gene association, written as a binary tree.

The knowledge represented in the model by gene associations is indispensible for predicting cell phenotypes under genetic modifications such as gene knockouts in flux balance analysis (FBA) [JLP11], and for taking into account the effects of gene regulation in dynamic simulation [KSM12].

Let us commence by defining the logical representation we will use to document a state of knowledge about a set of reactions in a metabolic model and a set of gene-protein-reaction relations associated with those reactions. In this representation, genes and proteins are atomic formulas (or *atoms*), and the relations between these objects and the state of the system are *clauses*. For example, the clause **"A:B ⊢ A, B"** means that, to the complex **A:B** can only be formed if the proteins **A** and **B** are already present. We define inference rules that combine and rewrite clauses in order to derive the formulas that are required for each use.

The following notation is used for the assertions employed by the logical representation:

- $(\mathbf{A}, \mathbf{c})$ - the molecule **A** is present in compartment **c**
- **A:B** - a macromolecular complex is formed from **A** and **B**
- $\textsc{degrad}(\mathbf{A})$ - the molecule **A** is degraded by the cell

- **DIS(A:B)** - the complex **A:B** is dissociated by the cell
- **OX(A)** - the molecule **A** is oxidized
- **RED(A)** - the molecule **A** is reduced
- **OX-RED(A)** - the redox cycle of the molecule **A** is present

This representation permits one to write clauses of the kind **C, D** $\vdash$ **A, B**, where **A** and **B** imply **C** and **D**, for example as a result of a reaction **A** + **B** $\rightarrow$ **C** + **D**. We will see, in what follows, that the formula required for each use can be obtained by rewriting clauses, up to a use-specific halting criterion.

BioRica [AGS+11] is a high-level modeling framework that integrates discrete, continuous, stochastic, nondeterministic, and timed behaviors. It provides a means to hierarchically compose models in a nonambiguous way. BioRica extends the ideas of AltaRica [ABCC94] [APGR99] to biological systems, allowing specifications that use of differential equations as well as automata with stochastic and nondeterministic behaviors. BioRica is designed for modeling complex systems through the combination of existing models of subsystems, in particular those imported from SBML models.

The basic building block of BioRica is the *node*, which has discrete or continuous behavior. The specification of a node has eight fields: *state, flow, sub, event, trans, assert, init* and *extern*. The field *state* declares state variables and their value domains. Hierarchical connections to other BioRica nodea are specified using the field *flow*, which defines input-output connections, and the field *sub*, which lists nodes that the current node uses hierarchically. The field *event* declares the name of the discrete events that are possible, and the field *trans* defines the state transitions that are caused by these events. The field *assert* imposes constraints on input-output flows and on states. The field *init* declares the initial values of state variables. Finally, the field *extern* is used to specify event delay distributions and priorities between events.

3 Strategies for Deriving Gene Associations

In what follows we propose strategies for deriving gene associations for each of the use cases presented in 1. All of the strategies are based on a common method: substitution and simplification rules are applied top-down until a halting criterion is reached. The choice of halting criterion to use depends on the nature of the use case being studied.

In the case of **model inference**, to determine whether a reaction can occur in an organism, it suffices to know whether the enzymes that catalyze it are present or not. Consequently, gene association rewriting must halt as soon as the genes that code for the enzymes catalyzing the reaction are observed.

Steady-state **flux balance analysis** supposes that a reaction can carry any flux that is physiologically possible, as long as the necessary genes are expressed. Assuming steady-state, gene association reqriting must halt before any substitution, such as gene expression, for example, that would correspond to a change of state.

In the case of **hierarchical modeling**, the nodes in the BioRica specification correspond to a model that defines a natural limit of knowledge. Consequently, gene association rewriting is limited to the confines of each node. These limited can be identified syntactically in the model's specification.

Finally, in **dynamic simulation**, since any reaction may contribute to the simulation, gene association rewriting must continue as long as clauses can be rewritten, and can in principle concern the entire model. In the following, we will not illustrate this strategy.

The three examples that follow have been chosen to illustrate the process of deriving gene associations in metabolic networks. Each example starts from one central reaction of interest, but may include up to fifteen associated reactions, required for understanding the prerequisite conditions of the central reaction.

- RNR, the ribonucleotide reductase system modifying **RNR** by **Trx1p**, specifically the reaction catalyzed by the bisulfite **Trx1p** (r_8 dans Yeastnet).
- CYC, the complex **Cyc7p:Ccp1p** that catalyzes the transformation of ferricytochrome-C to ferrocytochrome-C (r_{0469} dans Yeastnet).
- PDC, the transcription factor complex **Thi3p:Pdc2p** or **Thi3p:Thi2p** that induces the four paralogs of **Pcd1p**, who catalyze the transformation of 3-methyl-2-oxopentanoate en 2-methyl butanol (r_{0067} dans Yeastnet)

4 Ribonucleotide reductase system (RNR)

Ribonucleotide reductase (RNR) converts ribonucleotides to deoxiribonucleotides during the **S** phase of the cell cycle. Recall that the cell cycle corresponds to the set of steps constituting and delimiting the life of a cell, and is composed of several growth stages: the **Mitose/Méiose** phase, where the cell divides, the **Gap1** phase, where the cell grows until it reaches a critical size that signals transition to the **S** phase, where DNA is replicated, before transitioning to the **Gap2** phase.

During the **Gap1**, **Gap2** and **Mitose/Méiose** phases, when ribonucleotide reduction is not necessary, the heterodimer complex **Rnr2p:Rnr4p** is imported into the nucleus through the action first of the addressing protein **Dif1p**, which brings the complex **Rnr2p :Rnr4p** in proximity to the nucleus, and second of the protein **Kap122p** that imports the complex into the nucleus. The **Rnr2p:Rnr4p** complex is retained in the nucleus by the anchoring protein **Wtm1p**.

During the **S** phase of the cell cycle, the **Rnr2p:Rnr4p** complex is moved from the nucleus to the cytoplasm by two cooperating mechanisms:

- The degradation **Dif1p**, which prevents entry of new **Rnr2p:Rnr4p** complexes into the nucleus.
- The action of the export protein **Crm1p**, which transports the complex **Rnr2p:Rnr4p** from the nucleus to the cytoplasm.

The **Rnr2p:Rnr4p** complex associates with the homodimère complex **Rnr1p:Rnr1p** to form the new complex **Rnr1p:Rnr1p:Rnr2p:Rnr4p**, which is then reduced by one of the other of the isozymes od thioredoxine **Trx1p** or **Trx2p**. These proteins are coded respectively by the **TRX1** and **TRX2**. The reduced complex **Rnr1p:Rnr1p:Rnr2p:Rnr4p** in turn catalyzes the reduction of diphosphate ribonucleoside **(NDP)** into diphosphate deoxyribonucleoside **(dNDP)** and is thus oxidized.

A graphic representation of the model used here is presented in figure 3. Starting from this representation, we determined a logicial representation and a BioRica specification corresponding to the reaction $\mathrm{NDP} \wedge \mathrm{O_2} \wedge 2\mathrm{H_2} \rightarrow \mathrm{dNDP} \wedge 2\mathrm{H_2O}$.

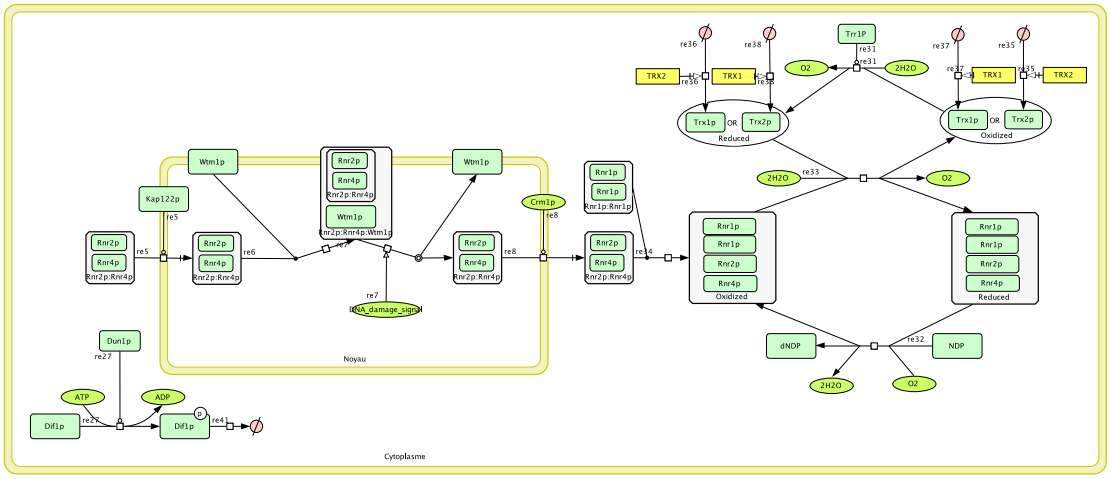


Figure 3: Explicit model for the ribonucleotide reductase reaction (RNR)

4.1 Logical Description

$$\mathrm{NDP} \wedge \mathrm{O_2} \wedge 2\mathrm{H_2} \rightarrow \mathrm{dNDP} \wedge 2\mathrm{H_2O} \vdash \textsc{ox-red}(\mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p}) \quad (1)$$

$$\mathrm{dNDP} \vdash \mathrm{NDP} \wedge \textsc{ox-red}(\mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p}) \quad (2)$$

$$\textsc{ox-red}(\mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p}) \vdash \mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p} \wedge (\textsc{ox-red}(\mathrm{Trx1p}) \vee \textsc{ox-red}(\mathrm{Trx2p})) \quad (3)$$

$$\textsc{ox-red}(\mathrm{Trx1p}) \vee \textsc{ox-red}(\mathrm{Trx2p}) \vdash (\mathrm{Trx1p} \vee \mathrm{Trx2p}) \wedge \mathrm{Trr1p} \quad (4)$$

$$\mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p} \vdash \mathrm{Rnr1p{:}Rnr1p} \wedge \mathrm{Rnr2p{:}Rnr4p} \quad (5)$$

$$\mathrm{Rnr1p{:}Rnr1p} \vdash \mathrm{Rnr1p} \wedge \mathrm{Rnr1p} \quad (6)$$

$$\mathrm{Rnr2p{:}Rnr4p} \vdash \mathrm{Rnr2p} \wedge \mathrm{Rnr4p} \quad (7)$$

$$(\mathrm{Rnr1p{:}Rnr1p{:}Rnr2p{:}Rnr4p, c}) \vdash (\mathrm{Rnr1p{:}Rnr1p, c}) \wedge (\mathrm{Rnr2p{:}Rnr4p, c}) \quad (8)$$

$$(\mathrm{Rnr1p{:}Rnr1p, c}) \vdash (\mathrm{Rnr1p{:}Rnr1p, n}) \wedge \mathrm{Crm1p} \wedge \mathrm{Phase(S)} \quad (9)$$

$$(\mathrm{Rnr1p{:}Rnr1p, n}) \vdash (\mathrm{Rnr1p{:}Rnr1p, c}) \wedge \mathrm{Kap122p} \wedge \mathrm{Dif1p} \wedge \neg\mathrm{Phase(S)} \quad (10)$$

$$\mathrm{Dif1p} \vdash \neg\mathrm{Dun1p} \wedge \neg\mathrm{Phase(S)} \quad (11)$$

$$\mathrm{Rnr2p{:}Rnr4p{:}Wrm1p} \vdash \mathrm{Rnr2p} \wedge \mathrm{Rnr4p} \wedge \mathrm{Wrm1p} \quad (12)$$

$$\mathrm{Crm1p} \vdash \mathrm{CRM1} \quad (13)$$

$$(14)$$

$$\text{Dif1p} \vdash \text{DIF1} \quad (15)$$
$$\text{Dun1p} \vdash \text{DUN1} \quad (16)$$
$$\text{Kap122p} \vdash \text{KAP122} \quad (17)$$
$$\text{Rnr1p} \vdash \text{RNR1} \quad (18)$$
$$\text{Rnr2p} \vdash \text{RNR2} \quad (19)$$
$$\text{Rnr4p} \vdash \text{RNR4} \quad (20)$$
$$\text{Trr1p} \vdash \text{TRR1} \quad (21)$$
$$\text{Trx1p} \vdash \text{TRX1} \quad (22)$$
$$\text{Trx2p} \vdash \text{TRX2} \quad (23)$$
$$\text{Wrm1p} \vdash \text{WRM1} \quad (24)$$

4.2 BioRica Description

The hierarchical model of the RNR system is comprised of three BioRica nodes:

- **Subcomplex**, which describes the formation of the **Rnr2p:Rnr4p** heterodimer and its importation into the nucleus. We suppose that the formation of the complex is the limiting step in the reaction; the kinetics of the addressing and importing reactions are assumed to be much more rapid, and are not explicitly modeled.
- **Complex**, which describes the formation of the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** heterotetramer in the cytoplasm during **S** phase. We suppose that this formation is only limited by the presence of **Rnr2p:Rnr4p** in the cytoplasm, which is in turn controlled by the transition to **S** phase. The transporter **Crm1** is not explicitly modeled.
- **Reaction**, which described the system of reactions neighboring ribonucleotide reductase. This includes the cycles of reduction-oxidation of the reductase heterotetramer and both of the thioredoxines **Trx1** and **Trx2**.

The **Main** node links these three sub-nodes by boolean variables that indicate the presence in the cytoplasm of the heterodimer **Rnr2p:Rnr4p** and the heterotetramer **Rnr1p:Rnr1p:Rnr2p:Rnr4p**. It also synchronizes the transitions to the **S** phase by the sub-nodes **Subcomplex** and **Complex**. The complete BioRica specification is as follows.

node Subcomplex

state
$\mathrm{T}_c, \mathrm{T}_n$: $\{0,1\}$;
ε : $\{0,1\}$;
flow
(Rnr2p:Rnr4p, c) : FLOAT;
init
(Rnr2p:Rnr4p, c_i) ;
diff
$d(\text{Rnr2p:Rnr4p, n}) = \mathrm{T}_n.(\text{cell}.V1.(1 + 1.\ln(\frac{(\text{Rnr2p:Rnr4p,c})}{ic(\text{Rnr2p:Rnr4p,c})}) - \ln(\frac{(\text{Rnr2p:Rnr4p,n})}{ic((\text{Rnr2p:Rnr4p,n})})))$
$+\varepsilon(\text{Rnr2p:Rnr4p}, c_i)$;
event
phaseS;
sortieS ;

trans

true $\vdash$ phaseS $\rightarrow$ $T_n := 0$, $T_c := 1$, $\varepsilon := 1$;
true $\vdash$ sortieS $\rightarrow$ $T_n := 1$, $T_c := 0$, $\varepsilon := 0$;

edon

node Complex

state

(Rnr1p:Rnr1p, c) = BOOL;

flow

(Rnr2p:Rnr4p, c) = BOOL;
(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c) = BOOL;

event

phaseS;
associate;

trans

true $\vdash$ phaseS $\rightarrow$ (Rnr2p:Rnr4p, c) := true;
(Rnr2p:Rnr4p, c), (Rnr1p:Rnr1p, c) $\vdash$ associate $\rightarrow$ (Rnr1p:Rnr1p:Rnr2p:Rnr4p, c):= true;

edon

node Reaction

state

TRX1 : BOOL;
TRX2 : BOOL;
Trx1p : BOOL;
Trx2p : BOOL;
Trr1p : BOOL;
ox(Trx1p or Trx2p) = BOOL ;
D : $\{0, 1\}$;
NDP : FLOAT;
dNDP : FLOAT;
dNDP : FLOAT;
O_2 : FLOAT;
H_2O : FLOAT;
ox(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c) : BOOL;

flow

(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c) : BOOL;
ox(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c) : BOOL;

event

ox(TRX);
ox(RNR);
G-expr(TRX1);
G-expr(TRX2);
active;

constants

iC(Ndp);
iC(dNdp);
iC(O_2) ;
iC(H_2O);

diff

$$d(dNDP) = D.(cell\ .V0.(1 + 1.\ln(\tfrac{C(NDP)}{ic(NDP)}) + \ln(\tfrac{C(O_2)}{ic(O_2)}) - 2\ln(\tfrac{C(H_2O)}{ic(H_2O)}) - \ln(\tfrac{C(dNDP)}{ic(dNDP)})));$$

trans

$$\begin{array}{rrl}
\text{TRX1} \vdash & \text{G-expr(TRX1)} \rightarrow & \text{Trx1p} := \text{true;} \\
\text{TRX2} \vdash & \text{G-expr(TRX2)} \rightarrow & \text{Trx2p} := \text{true;} \\
\text{Trr1p} \vdash & \text{ox(TRX)} \rightarrow & \text{ox(Trx1p | Trx2p)} := \text{true;} \\
\text{(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c),ox(Trx1p | Trx2p)} \vdash & \text{ox(RNR)} \rightarrow & \\
\text{ox(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c)} := \text{true;} & & \\
\text{ox(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c)} \vdash & \text{active} \rightarrow & \text{D} = 1 \text{ ;}
\end{array}$$

edon

node Main

sub

N : Subcomplex;
C : Complex;
R : Reaction;

assert

C.(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c) = R.(Rnr1p:Rnr1p:Rnr2p:Rnr4p, c);
N.(Rnr2p:Rnr4p, c) = C.(Rnr2p:Rnr4p, c);

sync

<N.phaseS, C.phaseS>;

edon

4.3 Gene Associations for Model Inference

We can conclude that the ribonucleotide reductase reaction can occur in a target organism if there is sufficient evidence to rewrite its gene association. The reaction is catalyzed by the oxidation cycle of the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** complex, so in order in instantiate the reaction we must be assured that the genes coding for the constituant proteins of the complex are present in the target organism (figure 4).

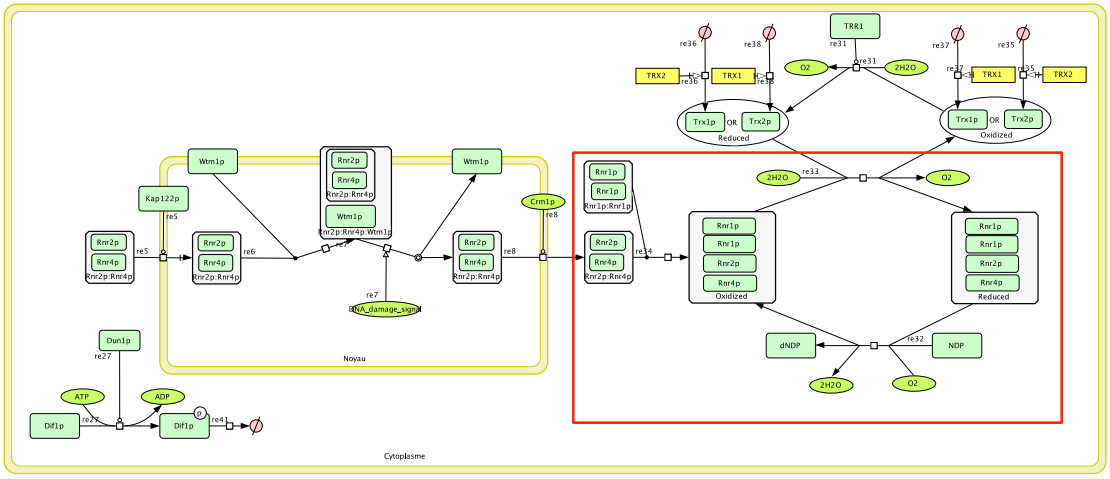


Figure 4: Explicit model of the ribonucleotide reductase reaction, outlining in red the part of the model considered when performing model inference

The gene association of the ribonucleotide reductase reaction is given by RNR1 ∧ RNR2 ∧ RNR4. This is derived formally from the presence of the oxidation cycle of the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** complex:

OX-RED(Rnr1p:Rnr1p:Rnr2p:Rnr4p)

$\vdash$ Rnr1p:Rnr1p:Rnr2p:Rnr4p $\wedge$ (OX-RED(Trx1p) $\vee$ OX-RED(Trx2p)) eq(3)
$\vdash$ Rnr1p:Rnr1p $\wedge$ Rnr2p:Rnr4p eq(5)
$\vdash$ Rnr1p $\wedge$ Rnr2p $\wedge$ Rnr4p eqs(6) (7)
$\vdash$ RNR1 $\wedge$ RNR2 $\wedge$ RNR4 eqs(18) (19) (20)

4.4 Gene associations for Flux Balance Analysis

Flux Balance Analysis takes into account fluxes through reactions, under the hypothesis that no state change occurs. In consequence, we may not apply substitution or simplification rules that lead to gene expression (e.g. eq(22)) or to formation or dissociation of protein complexes (e.g. eq(5)). In this precise case, we suppose that the enzymes and enzymatic complexes are present, and we do not attempt to take into account the cycles of reduction reactions (Figure 5).

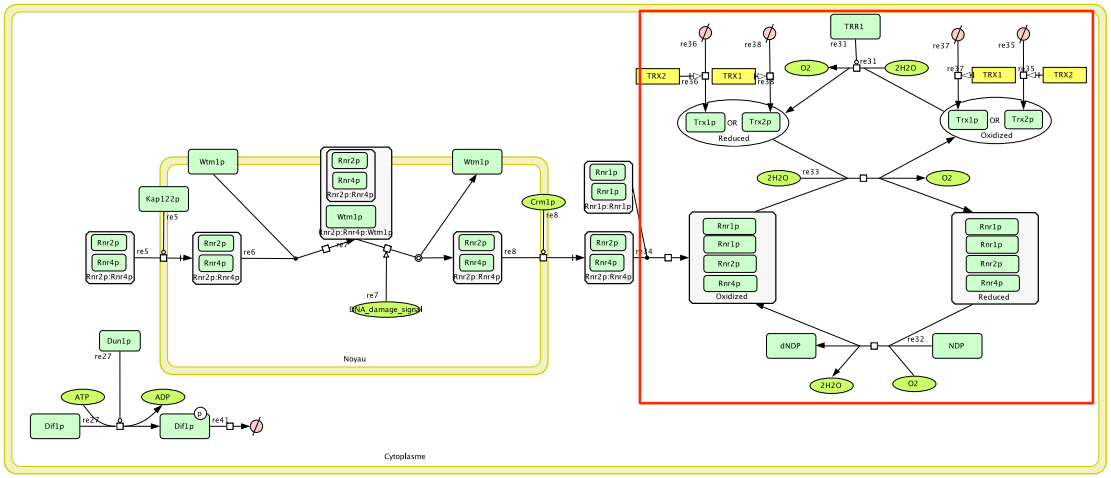


Figure 5: Explicit model of the ribonucleotide reductase reaction, outlining in red the part of the model considered when performing flux balance analysis

The gene association of the ribonucleotide reductase reaction is given by (Trx1p $\vee$ Trx2p) $\wedge$ Trr1p. This is derived formally from the presence of the oxidation cycle of the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** complex using rules that do not compise gene expression or protein complex formation:

OX-RED(Rnr1p:Rnr1p:Rnr2p:Rnr4p)
$\vdash$ Rnr1p:Rnr1p:Rnr2p:Rnr4p $\wedge$ OX-RED(Trx1p) $\vee$ OX-RED(Trx2p) eq(3)
$\vdash$ Rnr1p:Rnr1p:Rnr2p:Rnr4p $\wedge$ (Trx1p $\vee$ Trx2p) $\wedge$ Trr1p eq(4)

4.5 Gene Associations for Hierarchical Modeling

The reduction of ribonucleotide to deoxyribonucleotide is performed by the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** heretotetramer complex, which is formed by the association of the **Rnr1p:Rnr1p** homodimer with the **Rnr2p:Rnr4p** heterodimer released from the nucleus. The reductase complex is activated by a **Trx1p** or **Trx2p** thioredoxine, which is in turn activated by the **Trr1p** thioredoxine reductase. These reactions occur at the same time: each ribonucleotide reduction exidizes the reductase complex, which is reduced by the thioredoxine, itself reduced by the

thioredoxine reductase. Thus, to describe these dependencies, we require the genes coding the **Rnr1p:Rnr1p:Rnr2p:Rnr4p** heterotetramer reductase complex, the thioredoxine that performs the reduction, as well as the thioredoxine reductase (Figure 6).

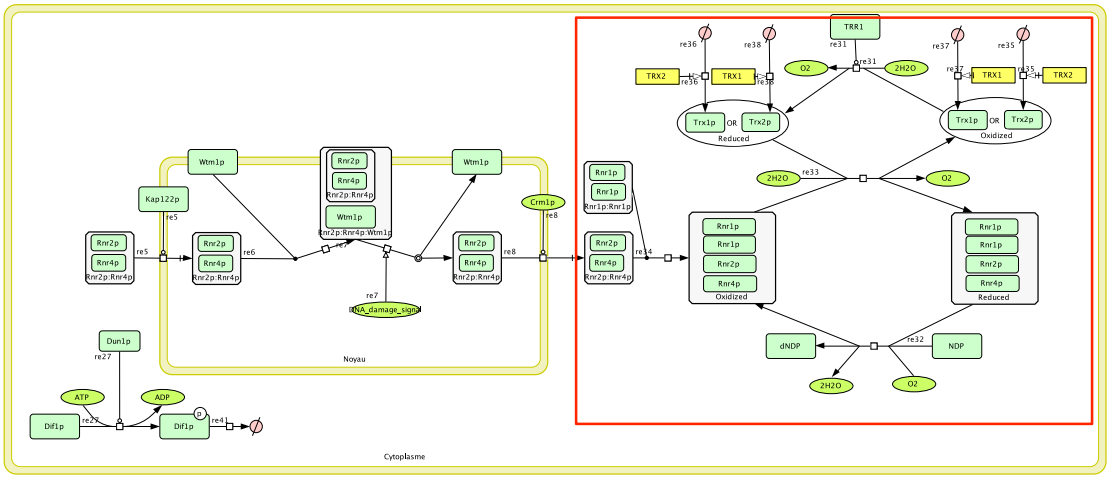


Figure 6: Explicit model of the ribonucleotide reductase reaction, outlining in red the part of the model considered when performing hierarchical modeling

$$
\begin{aligned}
&\text{Rnr1p:Rnr1p:Rnr2p:Rnr4p} \ \wedge (\text{Trx1p} \vee \text{Trx2p}) \wedge \text{Trr1p} \\
&\vdash \quad \text{Rnr1p:Rnr1p} \wedge \text{Rnr2p:Rnr4p} \wedge (\text{Trx1p} \vee \text{Trx2p}) \wedge \text{Trr1p} && \text{eqs(4)(5)} \\
&\vdash \quad \text{Rnr1p} \wedge \text{Rnr2p} \wedge \text{Rnr4p} \wedge (\text{Trx1p} \vee \text{Trx2p}) \wedge \text{Trr1p} && \text{eqs(6)(7)} \\
&\vdash \quad \text{RNR1} \wedge \text{RNR2} \wedge \text{RNR4} \wedge (\text{TRX1} \vee \text{TRX2}) \wedge \text{TRR1} && \text{eqs(18) (19)} \\
& && \text{(20) (22)} \\
& && \text{(23) (21)}
\end{aligned}
$$

5 Ferrocytochrome c System (CYC)

The **Cyc1p** and **Cyc7p** isoforms of cytochrome c are electron transport proteins taking part in the electron transport chain in mitochondria. In this system, the **Cyc1p** or **Cyc7p** proteins form a complex with cytochrome c peroxidase **Ccp1p**. In this complex, **Ccp1p** c catalyzes the oxidation of 2-ferrocytochrome c, which is the reduced form of **Cyc1p** or **Cyc7p** (figure 7).

Les cytochrome c isoformes **Cyc1p** et **Cyc7p** sont des protéines de transport d'électrons dans la chaîne de transport d'électrons mitochondriale. Dans ce système, la protéine **Cyc1p** ou **Cyc7p** forme d'abord un complexe avec cytochrome c peroxidase **Ccp1p**, et dans ce complexe, **Ccp1p** catalyse l'oxydation du 2-ferrocytochrome c, qui est la forme réduite de **Cyc1p** or **Cyc7p** (Figure 7).

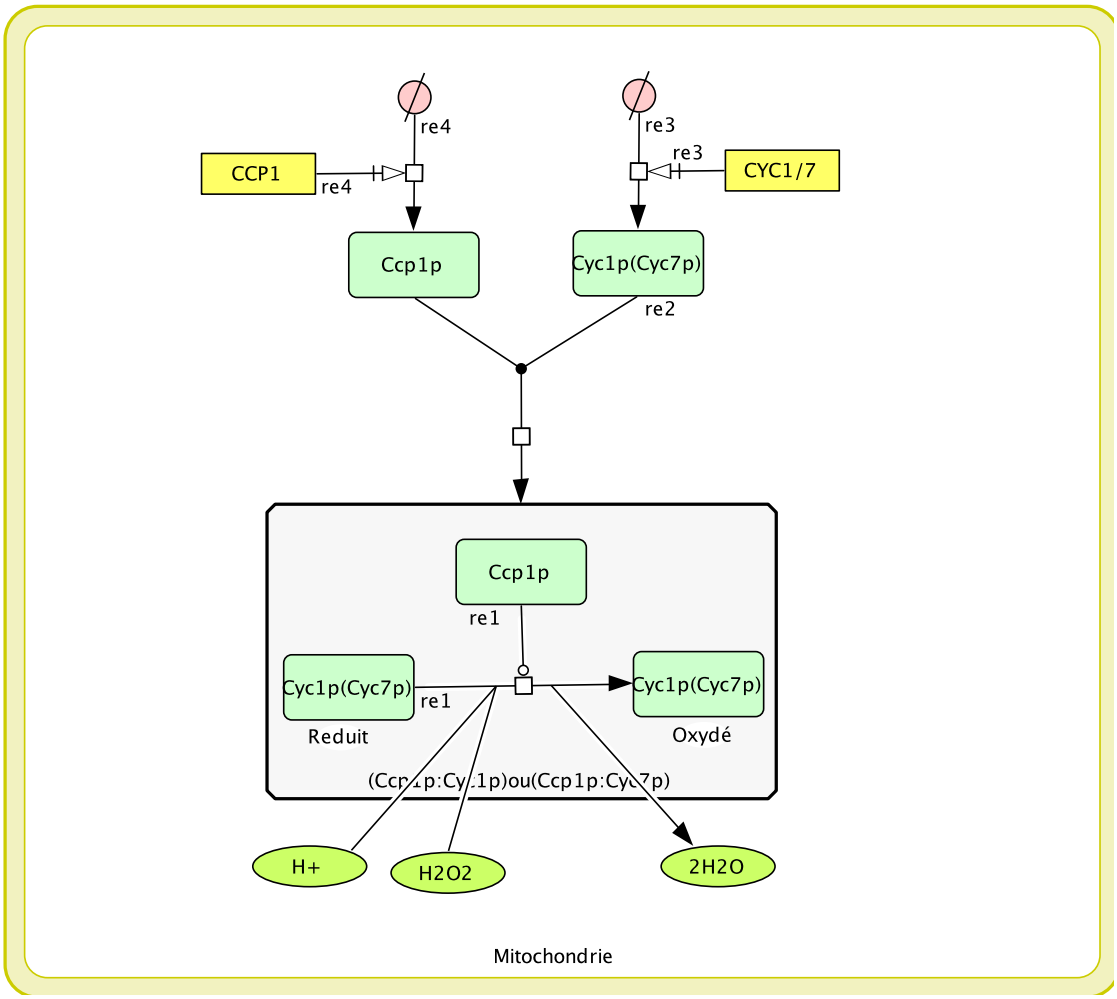


Figure 7: Explicit model of the ferrocytochrome c reaction

5.1 Logical Description

$$R = \text{RED}(\text{Cyc1p}) \wedge \text{H}^+ \wedge \text{H}_2\text{O}_2 \rightarrow \text{OX}(\text{Cyc1p}) \wedge 2\text{H}_2\text{O}$$

$$\begin{aligned} \text{R} &\vdash \text{Ccp1p} \wedge \text{Cyc1p:Ccp1p} && (25)\\ \text{Cyc1p:Ccp1p} &\vdash \text{Cyc1p} \wedge \text{Ccp1p} && (26)\\ \text{Cyc1p} &\vdash \text{CYC1} && (27)\\ \text{Ccp1p} &\vdash \text{CCP1} && (28) \end{aligned}$$

$$R' = \text{RED}(\text{Cyc7p}) \wedge \text{H}^+ \wedge \text{H}_2\text{O}_2 \rightarrow \text{OX}(\text{Cyc7p}) \wedge 2\text{H}_2\text{O}$$

$$\begin{aligned} \text{R'} &\vdash \text{Ccp7p} \wedge \text{Cyc7p:Ccp1p} && (29)\\ \text{Cyc7p:Ccp1p} &\vdash \text{Cyc7p} \wedge \text{Ccp1p} && (30)\\ \text{Cyc7p} &\vdash \text{CYC7} && (31)\\ \text{Ccp1p} &\vdash \text{CCP1} && (32) \end{aligned}$$

5.2 BioRica Description

The hierarchical model of the **CYC** system contains a single hybrid node **Main**, composed of discrete transitions that represent gene expression and the formation of the **Ccp1p:cytochrome c** complex, and one continuous equation describing the oxidation of cytochrome c by **Ccp1**. The oxidation of **Ccp1** may only occur once the **Ccp1p:cytochrome c** complex is formed.

node Main

state

CCP1	:	BOOL;
CYC1	:	BOOL;
CYC7	:	BOOL;
Ccp1p	:	BOOL;
Cyc1p	:	BOOL;
Cyc7p	:	BOOL;
Ccp1p:(Cyc1p|Cyc7p)	:	BOOL;
R_0	=	$\{0,1\}$;
$(\text{Cyc1p}\|\text{Cyc7p})_O$	:	FLOAT;
$(\text{Cyc1p}\|\text{Cyc7p})_R$	:	FLOAT;

diff

$$d((Cyc1p|Cyc7p)_O) = R_0(\text{cell}.V_0.(1 + 1.\ln(\tfrac{(Cyc1p|Cyc7p)_R}{ic(Cyc1p|Cyc7p)_R}) + \ln(\tfrac{H^+}{ic(H^+)}) + \ln(\tfrac{H_2O_2}{ic(H_2O_2)}) - \ln(\tfrac{(Cyc1p|Cyc7p)_O}{ic(Cyc1p|Cyc7p)_O}) - \ln(\tfrac{H_2O}{ic(H_2O)})));$$

event

G-expr(CCP1); G-expr(CYC1); G-expr(CYC7);
associateCOM; Oxydate;

trans

CCP1	⊢	G-expr(CCP1)	→	Ccp1p := true;
CYC1	⊢	G-expr(CYC1)	→	Cyc1p := true;
CYC7	⊢	G-expr(CYC7)	→	Cyc7p := true;
Ccp1p, (Cyc1p | Cyc7p)	⊢	associateCOM	→	Ccp1p:(Cyc1p | Cyc7p) := true;
Ccp1:(Cyc1p | Cycp7)	⊢	Oxydate	→	R_0 := 1;

edon

5.3 Gene Associations for Model Inference

The genes needed so that this reaction may occur in the target organism are the genes coding for the catalyst **Cc1p** and the complex **Cyc1p|Cyc7p:Ccp1p** (figure 8).

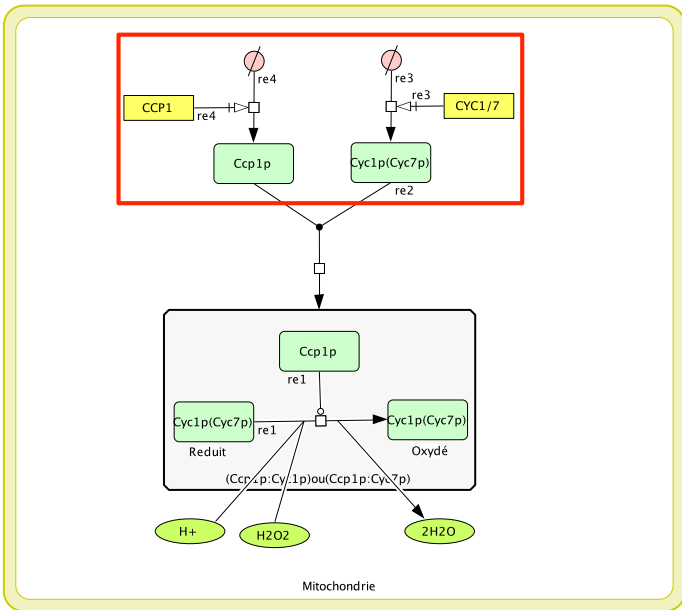


Figure 8: Explicit model for the Ferrocytochrome c system, outlining in red the part of the model considered when performing model inference

R	$\vdash$	Ccp1p $\wedge$ Cyc1p:Ccp1p	eq(25)
	$\vdash$	Ccp1p $\wedge$ Cyc1p	eq(26)
	$\vdash$	CCP1 $\wedge$ CYC1	eqs(28) (27)
R'	$\vdash$	Ccp1p $\wedge$ Cyc7p:Ccp1p	eq(29)
	$\vdash$	Ccp1p $\wedge$ Cyc7p	eq(30)
	$\vdash$	CCP1 $\wedge$ CYC7	eqs(32) (31)

5.4 Gene Associations for Flux Balance Analysis

In this example no state changes occur, so the reaction may be instantiated if the enzyme and the complex are already present. (figure 9).

R	$\vdash$	Ccp1p $\wedge$ Cyc1p:Ccp1p	eq(25)
R'	$\vdash$	Ccp1p $\wedge$ Cyc7p:Ccp1p	eq(29)

5.5 Gene Associations for Hierarchical Modeling

The oxidation of cytochrome c (**Cyc1p** ou **Cyc7p**) by **Ccp1p** takes place within the **Cyc1p:Ccp1p** or the **Cyc7p:Ccp1p**) complex, which associates cytochrome c with cytochrome c peroxidase (figure 10).

R	$\vdash$	Ccp1p $\wedge$ Cyc1p:Ccp1p	eq(25)
	$\vdash$	Ccp1p $\wedge$ Cyc1p	eq(26)
	$\vdash$	CCP1 $\wedge$ CYC1	eqs(28) (27)

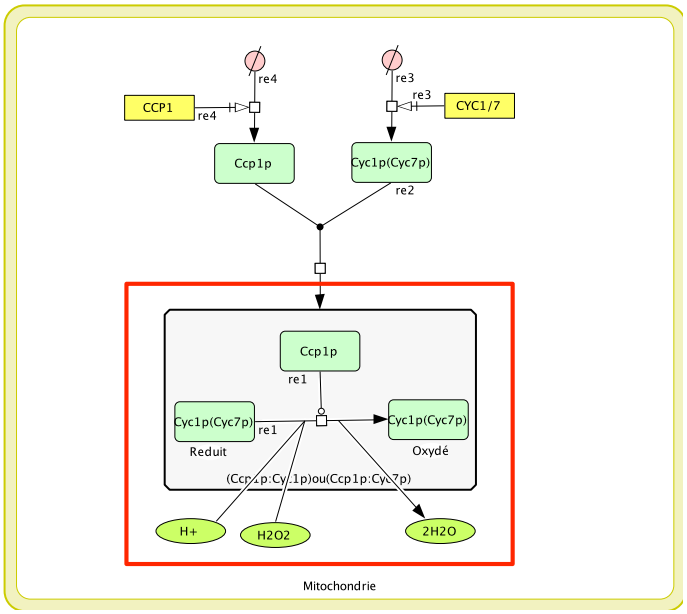


Figure 9: Explicit model of the Ferrocytochrome c system, outlining in red the part of the model considered when performing flux balance analysis

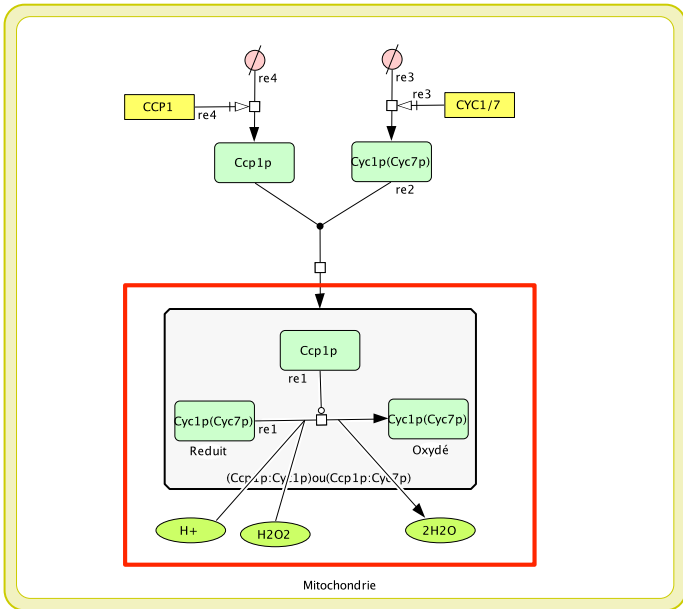


Figure 10: Explicit model of the Ferrocytochrome c system, outlining in red the part of the model considered when performing hierarchical modeling

$$
\begin{array}{lll}
\text{R'} & \vdash & \text{Ccp1p} \wedge \text{Cyc7p:Ccp1p} \qquad \text{eq(29)} \\
 & \vdash & \text{Ccp1p} \wedge \text{Cyc1p} \qquad \text{eq(30)} \\
 & \vdash & \text{CCP1} \wedge \text{CYC7} \qquad \text{eqs(32) (31)}
\end{array}
$$

6 3-methyl-2-oxopentanoate Decarboxylase System (PDC)

The enzymes of 3-methyl-2-oxopentanoate decarboxylase, **Pdc1p**, **Pdc5p**, **Pdc6p**, et **ARO10**, react to 2-oxo-acids like 3-methyl-2-oxopentanoate, which is an intermediate in the synthesis of isoleucine, (figure 11).

Thi3p is a regulatory protein of this reaction. It binds to the transcription factors **Thi2p** and **Pdc2p**. In the presence of thiamine, the **Pdc2p:Thi3p** and **Thi2P:Thi3p** complexes dissociate, and thus deactivate the transcription factors. The genes **PDC1** and **PDC5**, which code respectively for the proteins **Pdc1p** and **Pdc5p**, are regulated by the **Pdc2p:Thia3p** complex or by the **Thi2p:Thi3p** complex, and by the absence of the protein **Pdc1p**. However, the genes **PDC6** and **ARO10**, which code respectively for the proteins **Pdc6p** and **Aro10p**, are not regulated by these mechanisms.

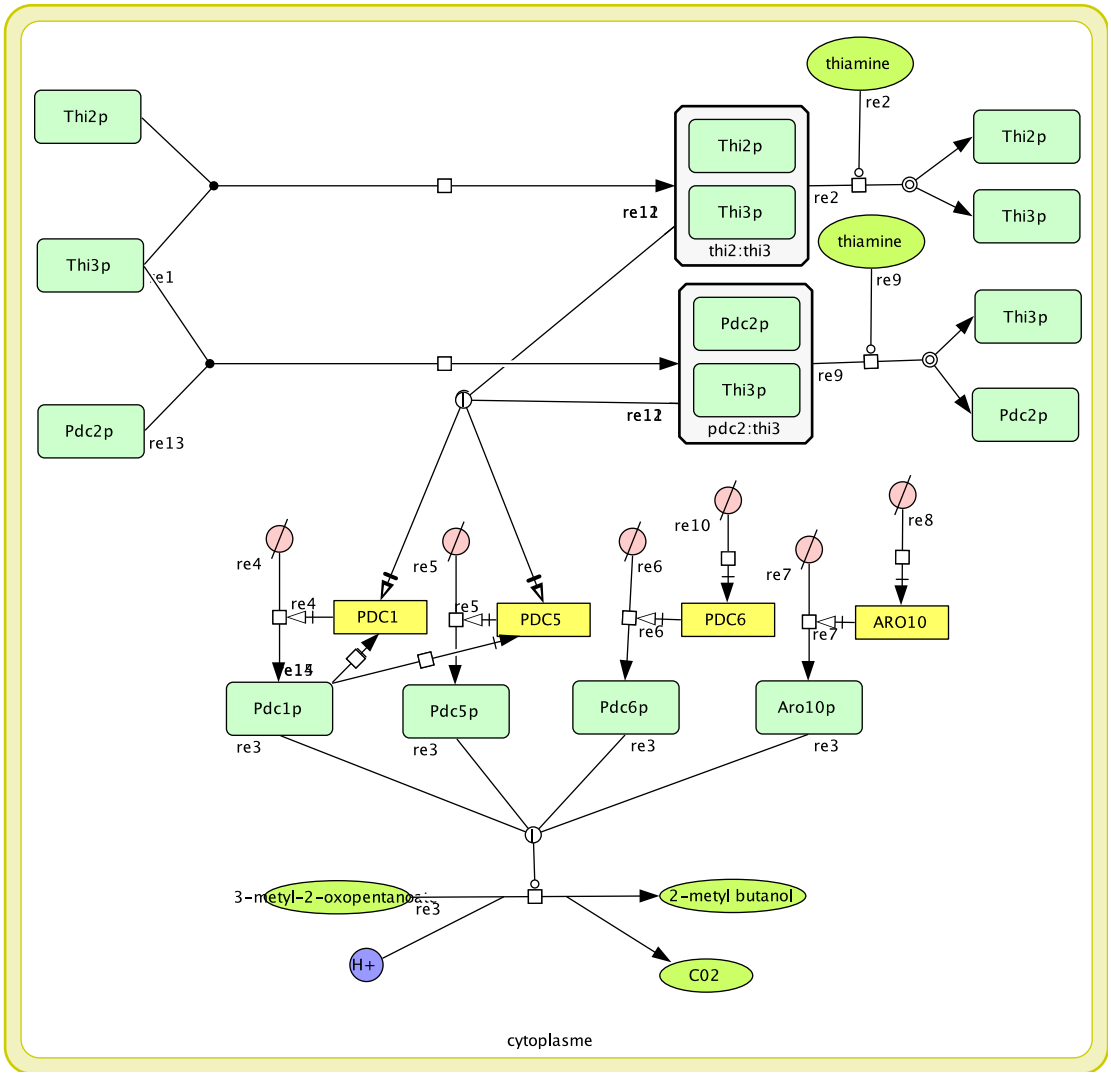


Figure 11: Explicit model of the 3-méthyl-2-oxopentanoate décarboxylase reaction

6.1 Logical Description

$$\text{R(PDC)} = \text{3-metyl-2-oxepentanal} \wedge \text{H}^+ \rightarrow \text{2-metylbutanal} \wedge \text{CO2}$$

$$\begin{aligned}
\text{R(PDC)} &\vdash \text{Pdc1p} \vee \text{Pdc5p} \vee \text{Pdc6p} \vee \text{Aro10} && (33)\\
\text{Pdc1p} &\vdash \text{PDC1} \wedge \text{Thi3p: } (\text{Thi2p} \vee \text{Pdc2p}) \wedge \neg\text{Pdc1p} && (34)\\
\text{Pdc5p} &\vdash \text{PDC5} \wedge \text{Thi3p: } (\text{Thi2p} \vee \text{Pdc2p}) \wedge \neg\text{Pdc1p} && (35)\\
\text{Thi3p:Pdc2p} &\vdash \text{Thi3p} \wedge \text{Pdc2p} \wedge \neg\text{Thiamine} && (36)\\
\text{Thi3p:Thi2p} &\vdash \text{Thi3p} \wedge \text{Thi2p} \wedge \neg\text{Thiamine} && (37)\\
\text{DIS(Thi3p:Pdc2p)} &\vdash \text{Thi3p:Pdc2p} \wedge \text{Thiamine} && (38)\\
\text{DIS(Thi3p:Thi2p)} &\vdash \text{Thi3p:Thi2p} \wedge \text{Thiamine} && (39)\\
\text{Pdc6p} &\vdash \text{PDC6} && (40)\\
\text{Aro10p} &\vdash \text{ARO10} && (41)
\end{aligned}$$

6.2 BioRica Description

The hierarchical model of the **PDC** system is composed of two sub=nodes:

- **Complex**, which describes the association of **Thi3p** with either **Thi2p** or with **Pdc2p**, to form the transcription factor that regulates the expression of the genes **PDC1** and **PDC5**.
- **Factor**, which is a hybrid node composed of discrete transitions that represent the expression of the genes **PDC1**, **PDC5**, **PDC6**, or **ARO10**, and a continuous equation representing the decarboxylase reaction. The expression of genes **PDC1** and **PDC5** depends on the transcription factor.

The **Main** node is connected to **Complex** and **Factor** by boolean variables that indicate the presence of one or the other transcription factor complexes.

node Complex

state
Thi3p : BOOL;
Thi2p : BOOL;
Pdc2p : BOOL;
Thiamin : BOOL;
flow
Thi2p:Thi3p : BOOL;
Pdc2p:Thi3p : BOOL;
event
associate(Thi2p:Thi3p); associate(Pdc2p:Thi3p);
trans
Thi3p , Thi2p , ¬Thiamin ⊢ associate(Thi2p:Thi3p) → Thi2p:Thi3p := true;
Thi3p , Pdc2p, ¬Thiamin ⊢ associate(Pdc2p:Thi3p) → Pdc2p:Thi3p := true;
edon

node Factor

state

PDC1 : BOOL;
PDC5 : BOOL;
PDC6 : BOOL;
ARO10 : BOOL;
Pcd1p : BOOL;
Pcd5p : BOOL;
Pcd6p : BOOL;
Aro10p : BOOL;
3-metyl-2-oxepentanal : FLOAT;
R_0 : $\{0, 1\}$;

flow
Thi2p:Thi3p : BOOL;
Pdc2p:Thi3p : BOOL;

const
ic(3-metyl-2-exepontanal)

diff
$$\text{d(3-metyl-2-oxepentanal)} = \text{R}_0(\text{cell}.V0.(1 + 1.\ln(\tfrac{3-metyl-2-oxepentanal}{ic(3-metyl-2-exepontanal)}) + \ln(\tfrac{H^+}{ic(H^+)}) - \ln(\tfrac{2-metylbutanal}{ic(2-metylbutanal)}) - \ln(\tfrac{CO2}{ic(CO2)})))$$

event
G-expr(PDC1); G-expr(PDC5);
G-expr(PDC6); G-expr(ARO10);
React;

trans
PDC1, (Thi2p:Thi3p | Pdc2p:Thi3p), ¬ Pdc1p ⊢ G-expr(PDC1) → Pdc1p := true ;
PDC5, (Thi2p:Thi3p | Pdc2p:Thi3p), ¬ Pdc5p ⊢ G-expr(PDC5) → Pdc5p := true ;
PDC6 ⊢ G-expr(PDC6) → Pdc6p := true ;
ARO10 ⊢ G-expr(ARO10) → Aro10p := true ;
Pdc1p | Pdc5p | Pdc6p | Aro10p ⊢ React → R_0 :=1 ;

edon

node Main

sub
F : Factor;
C: complex ;

assert
C.Thi2p:Thi3p = F.Thi2p:Thi3p;
C.Pdc2p:Thi3p = F.Pdc2p:Thi3p;

edon

6.3 Gene Associations for Model Inference

The reaction is catalyzed by one of the enzymes **Pdc1p**, **Pdc5p**, **Pdc6p**, and **Aro10p**, which are products of the genes **PDC1**, **PDC5**, **PDC6**, and **ARO10**. Consequently, the reaction may be inferred if at least one of these genes is present (figure 12).

R(PDC) ⊢ Pdc1p ∨ Pdc5p ∨ Pdc6p ∨ Aro10p eq(33)
⊢ PDC1 ∨ PDC5 ∨ PDC6 ∨ ARO10 eqs(34) (35) (40) (41)

6.4 Gene Associations for Flux Balance Analysis

In this specific case, we suppose that the enzymes and enzymatic complexes are present (or not), and we only need to model the cycles of reduction reactions.

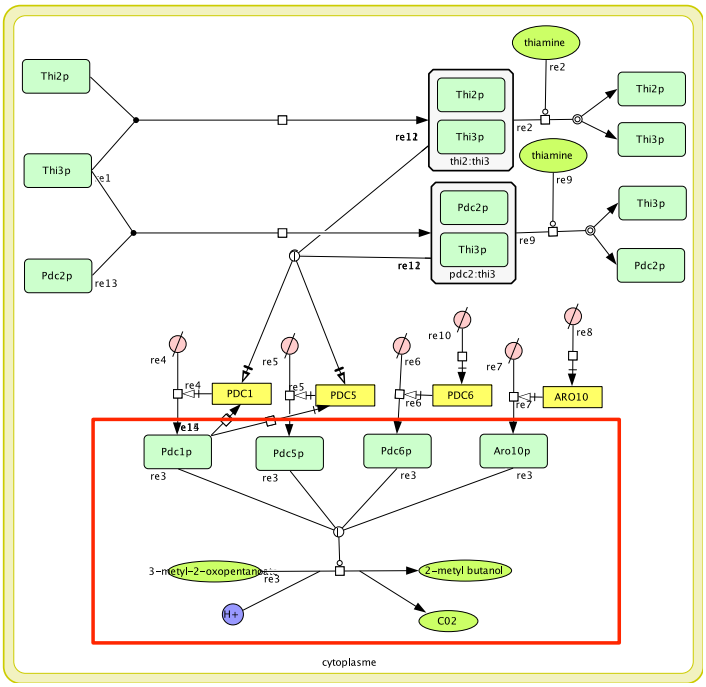


Figure 12: Explicit model for the 3-methyl-2-oxopentanoate decarboxylase system, outlining in red the part of the model considered when performing model inference

$$\text{R(PDC)} \quad \vdash \quad \text{Pdc1p} \vee \text{Pdc5p} \vee \text{Pdc6p} \vee \text{Aro10p} \qquad \text{eq(33)}$$

6.5 Gene Associations for Hierarchical Modeling

The transformation of 3-méthyl- 2-oxopentanoate into 2-méthyl butanol is performed by one of the enzymes **Pdc1p**, **Pdc5p**, **Pdc6p**, or **Aro10p** (figure 14) .

$$\begin{aligned} \text{R(PDC)} \quad &\vdash \quad \text{Pdc1p} \vee \text{Pdc5p} \vee \text{Pdc6p} \vee \text{Aro10p} && \text{eq(33)} \\ &\vdash \quad \text{PDC1} \vee \text{PDC5} \vee \text{PDC6} \vee \text{ARO10} && \text{eqs(34),(35),(40),(41)} \end{aligned}$$

7 Discussion

This report defines a formal criterion for correctness of gene associations: a gene association is correct if it is the one that would have been derived from a complete model. Our goal has been to define this criterion, and not so much to provide a method for deriving the gene association. In most practical cases a complete model is not available.

Starting from the definition of gene associations, the requirements of specific use cases, and the possible influences of a (gene or protein) variable, we defined simplification and substitution rules that produce the most pertinent gene association for a given reaction. These rules may also be used to provide guidelines for curators who build gene associations by hand.

The criterion may be used to define methods for finding gene associations through the analysis of experimental results. It may also be used to evaluate the form of boolean formulas used in gene associations, in order to facilitate knowledge extraction by inference procedures.

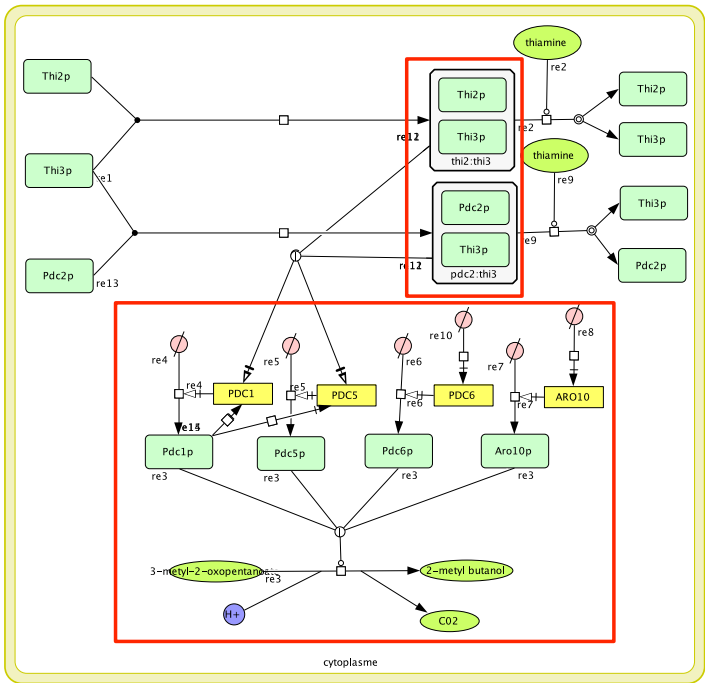


Figure 13: Explicit model of the 3-methyl-2-oxopentanoate decarboxylase system, outlining in red the part of the model considered when performing flux balance analysis

References

[ABCC94] André Arnold, Didier Bégay, Paul Crubillé, and P Crubille. *Construction and analysis of transition systems with MEC*. Number 3 in AMAST series in computing. World Scientific, 1994.

[AGS+11] Rodrigo Assar, Alice Garcia, David James Sherman, et al. Modeling stochastic switched systems with biorica. In *Journées Ouvertes en Biologie, Informatique et Mathématiques JOBIM 2011*, pages 297–304, 2011.

[AJW+08] Michiel E Adriaens, Magali Jaillard, Andra Waagmeester, Susan LM Coort, Alex R Pico, and Chris TA Evelo. The public road to high-quality curated biological pathways. *Drug discovery today*, 13(19):856–862, 2008.

[APGR99] André Arnold, Gérald Point, Alain Griffault, and Antoine Rauzy. The altarica formalism for describing concurrent systems. *Fundam. Inf.*, 40(2,3):109–124, August 1999.

[C+11] UniProt Consortium et al. Ongoing and future developments at the universal protein resource. *Nucleic acids research*, 39(suppl 1):D214–D219, 2011.

[CFF+06] Ron Caspi, Hartmut Foerster, Carol A Fulcher, Rebecca Hopkinson, John Ingraham, Pallavi Kaipa, Markus Krummenacker, Suzanne Paley, John Pick, Seung Y Rhee, et al. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic acids research*, 34(suppl 1):D511–D516, 2006.

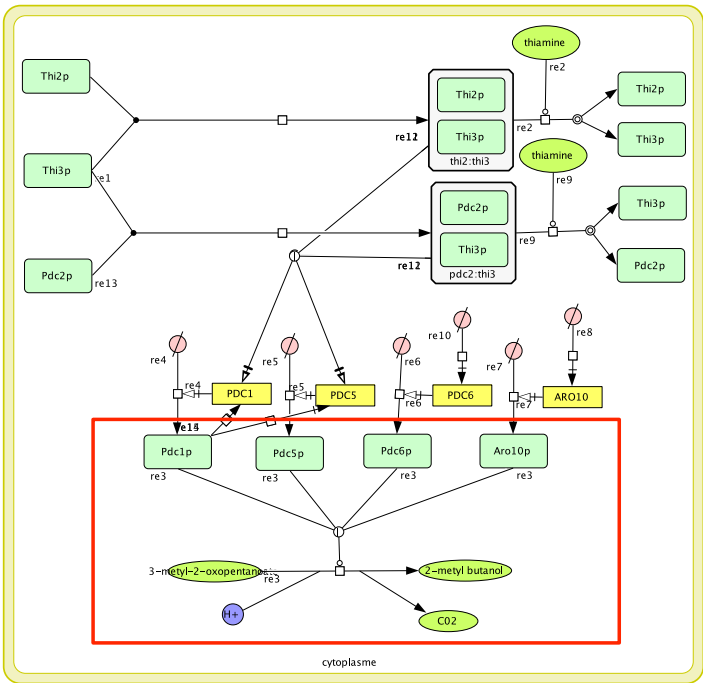


Figure 14: Explicit model of the 3-methyl-2-oxopentanoate decarboxylase system, outlining in red the part of the model considered when performing hierarchical modeling

[FGAT09] Nicholas Furnham, John S Garavelli, Rolf Apweiler, and Janet M Thornton. Missing in action: enzyme functional annotations in biological databases. *Nature chemical biology*, 5(8):521–525, 2009.

[FHT+09] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

[FP08] Adam M Feist and Bernhard Ø Palsson. The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. *Nature biotechnology*, 26(6):659–667, 2008.

[FST05] Christof Francke, Roland J Siezen, and Bas Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology*, 13(11):550–558, 2005.

[Iss15] Razanne Issa. *symbolic analysis and inference of metabolic models*. Theses, Université de Bordeaux, July 2015.

[JLP11] Paul A Jensen, Kyla A Lutz, and Jason A Papin. Tiger: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC systems biology*, 5(1):147, 2011.

[KDWV13] S Krishnakumar, Dilip A Durai, Pramod P Wangikar, and Ganesh A Viswanathan. Sharp: genome-scale identification of gene–protein–reaction associations in cyanobacteria. *Photosynthesis research*, 118(1-2):181–190, 2013.

[KG00] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[KSM12] Akhil Kumar, Patrick F Suthers, and Costas D Maranas. Metrxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics*, 13(1):6, 2012.

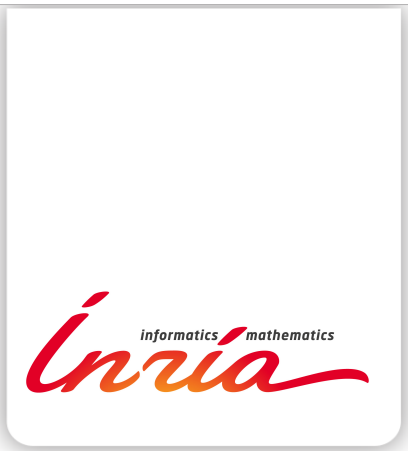
[NvEF+06] R. A. Notebaart, F. H. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, 7:296, Jun 2006.

[RCO+13] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

[RVSP03] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.*, 4(9):R54, 2003.

[WLT05] James D Watson, Roman A Laskowski, and Janet M Thornton. Predicting protein function from sequence and structural data. *Current opinion in structural biology*, 15(3):275–284, 2005.

Contents



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399