Identification of metabolic network models from incomplete high-throughput datasets

Sara Berthoumieux, Matteo Brilli, Hidde de Jong, Kahn Daniel, Eugenio Cinquemani

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Identification of metabolic network models from incomplete high-throughput datasets

Sara Berthoumieux — Matteo Brilli — Hidde de Jong — Daniel Kahn — Eugenio Cinquemani

N° 7524

Janvier 2011

Computational Biology and Bioinformatics

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7524--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE – RHÔNE-ALPES

Identification of metabolic network models from incomplete high-throughput datasets

Sara Berthoumieux*, Matteo Brilli , Hidde de Jong , Daniel Kahn , Eugenio Cinquemani*

Theme : Computational Biology and Bioinformatics
Computational Sciences for Biology, Medicine and the Environment
Équipes-Projets Ibis and Bamboo

Rapport de recherche n° 7524 — Janvier 2011 — 32 pages

Abstract: High-throughput measurement techniques for metabolism and gene expression provide a wealth of information for the identification of metabolic network models. Yet, missing observations scattered over the dataset restrict the number of effectively available datapoints and make classical regression techniques inaccurate or inapplicable. Thorough exploitation of the data by identification techniques that explicitly cope with missing observations is therefore of major importance.

We develop a maximum-likelihood approach for the estimation of unknown parameters of metabolic network models that relies on the integration of statistical priors to compensate for the missing data. In the context of the linlog metabolic modeling framework, we implement the identification method by an Expectation Maximization (EM) algorithm and by a simpler direct numerical optimization method. We evaluate performance of our methods by comparison to existing approaches, and show that our EM method provides the best results over a variety of simulated scenarios. We then apply the EM algorithm to a real problem, the identification of a model for the *Escherichia coli* central carbon metabolism, based on challenging experimental data from the literature. This leads to promising results and allows us to highlight critical identification issues.

Key-words: metabolic network, parameter estimation, system identification, Expectation Maximization algorithm, incomplete high-throughput datasets, proteomics, metabolomics, kinetic modeling, linlog models, central carbon metabolism of *Escherichia coli*

* To whom correspondance should be adressed

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Identification de modèles de réseaux métaboliques à partir de jeux incomplets de données à haut-débit

Résumé : Les techniques actuelles de mesures à haut-débit pour le métabolisme et l'expression génique fournissent de très nombreuses données pour l'identification de modèles de réseaux métaboliques. Cependant, l'existence de données manquantes tout au long du jeu de données restreint le nombre effectif de données disponibles et rend les techniques classiques de régression imprécises ou inapplicables. Il est donc primordial d'utiliser des techniques d'identification qui tiennent compte explicitement de ces observations manquantes.

Nous développons une approche basée sur le maximum de vraisemblance pour l'estimation de paramètres de modèles de réseaux métaboliques. Elle repose sur l'intégration de distributions *a priori* pour compenser les données manquantes. Nous implémentons cette méthode d'identification dans le cadre de la modélisation métabolisme par le formalisme linlog à l'aide d'un algorithme EM (Expectation-Maximization) et d'une méthode plus directe d'optimisation numérique. Nous évaluons la performance de nos méthodes en les comparant à des approches existantes et nous montrons que notre méthode EM produit les meilleurs résultats sur différents scénarios simulés. Nous appliquons ensuite l'algorithme EM à un problème réel, l'identification d'un modèle du métabolisme central du carbone chez *Escherichia coli*, basée sur un important jeu de données expérimentales de la littérature. Les résultats obtenus sont prometteurs et nous permettent de mettre en évidence certains aspects critiques des jeux de données pour l'identification.

Mots-clés : réseaux métaboliques, estimation de paramètres, identification de systèmes, algorithme Expectation Maximization, jeu incomplet de données haut-débits, protéomique, métabolomique, modélisation cinétique, modèle linlog, métabolisme central du carbone chez *Escherichia coli*

1 Introduction

To further our understanding of the cellular processes shaping the response of microbial cells to changes in their environment requires the study of the interactions between gene expression and metabolism. In recent years high-throughput datasets comprising simultaneous measurements of metabolism (fluxes, metabolite concentrations) and gene expression (protein and mRNA concentrations) have become available [Hardiman et al., 2007, Ishii et al., 2007]. These datasets provide a rich store of information for modeling the dynamics of the biochemical reaction systems underlying cellular processes. In particular, they promise to relieve what is currently a bottleneck for modeling in systems biology, obtaining reliable estimates of parameter values in kinetic models [Ashyraliyev et al., 2009, Crampin, 2006].

Notwithstanding these experimental advances, parameter estimation remains a particularly challenging problem, among other things due to incomplete knowledge of the molecular mechanisms, noisy and partial observations, heterogeneous experimental methods and conditions, and the large size of networks [Marucci et al., 2011]. As a consequence, the models may not be identifiable, may not generalize to new situations due to overfitting, and nonlinear rate functions may make them cumbersome to analyze. This has led to the proposal of simplified kinetic modeling frameworks, including linlog kinetics [Visser and Heijnen, 2003], loglin kinetics [Hatzimanikatis and Bailey, 1997], power-law kinetics [Savageau, 1976], and more recently, convenience kinetics [Liebermeister and Klipp, 2006].

Linlog models are a particularly interesting choice for modeling metabolism [Heijnen, 2005, Visser and Heijnen, 2003]. Simulation studies on the level of both individual enzymatic reactions [Heijnen, 2005] and metabolic networks [Costa et al., 2010, Hadlich et al., 2009, Visser et al., 2004] have shown that they provide reasonable approximations of classical enzymatic rate laws. Moreover, a recent genome-scale linlog model of yeast metabolism, parametrized using previously-published kinetic models, has been shown able to identify key steps in the network, that is, reactions exerting most control over glucose transport and biomass production [Smallbone et al., 2010].

A major advantage of linlog models is that, when measurements of fluxes, enzyme concentrations, and metabolite concentrations are available, the parameter estimation problem reduces to multiple linear regression. However, the performance of regression approaches quickly degrades in the presence of missing data, as is often the case in high-throughput datasets due to experimental limitations or instrument failures.

In order to deal with this problem, we propose in this paper a maximum-likelihood method for the identification of linlog models of metabolism from incomplete datasets. The specific contributions of the paper are twofold. On the theoretical side, we develop a method for the optimization of the likelihood based on Expectation Maximization (EM) [Dempster et al., 1977]. The method is constructed for linlog models, but is more generally applicable to other regression problems. In particular, we derive analytical expressions for the expectation step that are well-suited for numerical maximization. This guarantees the applicability of the approach even when modeling large networks. We show by means of simulation experiments on synthetic data that our approach outperforms both regression and a reference method from statistical literature for dealing with incomplete data, multiple imputation [Rubin, 1976, 1996]. In comparison with earlier work on treating incomplete high-throughput datasets [Oba et al., 2003, Scholz et al., 2005], our aim is not to estimate the missing values, but rather to improve the estimation of the model parameters from the incomplete datasets. This is a different problem that necessitates the development of novel methods.

On the biological side, we apply the method to a linlog model of central metabolism in *E. coli*, consisting of some 16 variables. We estimate the 100 parameters of this model from a high-throughput dataset published in the literature [Ishii et al., 2007]. The data consists of measurements of metabolic fluxes and metabolite and enzyme levels in glucose-limited chemostat under 29 different conditions such a wild-type strain and single-gene mutant strains or different dilution rates. Standard linear regression is difficult to apply in this case due to missing data, which disqualifies for 7 reactions too many datapoints, leaving a dataset of size inferior to the number of parameters to estimate. Application of our approach allows one to compute reasonable estimates for most of the identifiable model parameters even when regression is inapplicable.

2 Parameter estimation in linlog models

The dynamics of metabolic networks are described by kinetic models having the form of systems of ordinary differential equations (ODEs) [Heinrich and Schuster, 1996]:

$$\dot{x} = N \cdot v(x, u, e) \tag{1}$$

where $x \in \mathbb{R}_+^n$ denotes the vector of (nonnegative) internal metabolite concentrations, $u \in \mathbb{R}_+^p$ the vector of external metabolite concentrations, $e \in \mathbb{R}_+^m$ the vector of enzyme concentrations, and $v : \mathbb{R}_+^{n+p+m} \to \mathbb{R}^m$ the vector of reaction rate functions. $N \in \mathbb{Z}^{n\times m}$ is a stoichiometry matrix.

The reaction rates v are nonlinear and generally complex functions of x, u, and e, with many kinetic parameters that are difficult to reliably estimate from the data. This has motivated the use of approximate rate functions, like the linear-logarithmic (linlog) functions considered in this paper [Heijnen, 2005, Visser and Heijnen, 2003]. The linlog approximation expresses the reaction rates as proportional to the enzyme concentrations and to a linear function of the logarithms of internal and external metabolite concentrations. This leads to the rate equation

$$v(x, u, e) = \mathrm{diag}(e) \cdot \big(a + B^x \cdot \ln(x) + B^u \cdot \ln(u)\big) \tag{2}$$

where the logarithm of a vector means the vector of logarithms of its elements. For conciseness, in the sequel we shall drop the dependence of v on (x, u, e) from the notation. An in-depth discussion of linlog models and comparison with other approximative rate functions can be found in the review by [Heijnen, 2005].

We are interested in the estimation of the (generally unknown) parameters $a \in \mathbb{R}^m$, $B^x \in \mathbb{R}^{m\times n}$ and $B^u \in \mathbb{R}^{m\times p}$ from q experimental datapoints $(v^{(k)}, x^{(k)}, u^{(k)}, e^{(k)})$, $k = 1, \ldots, q$. That is, the data used for parameter estimation are parallel measurements of enzyme and metabolite levels as well as metabolic fluxes. Notice that in practice reaction rates are measured at (quasi-)steady state. The datapoints $(v^{(k)}, x^{(k)}, u^{(k)}, e^{(k)})$ are obtained under different experimental conditions, for instance different dilution rates in continuous cultures or different mutant strains. For the purpose of parameter estimation, it is convenient to rewrite (2) in the form of a regression model:

$$\left(\frac{v}{e}\right)^T = [1\ \ln(x)^T\ \ln(u)^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix} \tag{3}$$

where the ratio of two vectors (here v/e) denotes elementwise division. Let us use an upperbar to denote the mean of a quantity over its q experimental observations, for instance: $\overline{v/e} = (1/q)\sum_{k=1}^{q} v^{(k)}/e^{(k)}$. By the linearity of (3), it holds that

$$\overline{\left(\frac{v}{e}\right)} = [1\ \overline{\ln(x)}^T\ \overline{\ln(u)}^T] \cdot \begin{bmatrix} a^T \\ (B^x)^T \\ (B^u)^T \end{bmatrix} \tag{4}$$

This allows (3) to be reformulated as a mean-removed model

$$\left(\frac{v}{e} - \overline{\left(\frac{v}{e}\right)}\right)^T = \begin{bmatrix} \ln(x) - \overline{\ln(x)} \\ \ln(u) - \overline{\ln(u)} \end{bmatrix}^T \cdot \begin{bmatrix} (B^x)^T \\ (B^u)^T \end{bmatrix} \tag{5}$$

and we obtain the following parameter estimation problem:

Problem 1 *Given the data matrices*

$$\underbrace{\begin{bmatrix} \left(\frac{v^{(1)}}{e^{(1)}} - \overline{\left(\frac{v}{e}\right)}\right)^T \\ \vdots \\ \left(\frac{v^{(q)}}{e^{(q)}} - \overline{\left(\frac{v}{e}\right)}\right)^T \end{bmatrix}}_{\triangleq W}, \quad \underbrace{\begin{bmatrix} \left(\ln(x^{(1)}) - \overline{\ln(x)}\right)^T & \left(\ln(u^{(1)}) - \overline{\ln(u)}\right)^T \\ \vdots & \vdots \\ \left(\ln(x^{(q)}) - \overline{\ln(x)}\right)^T & \left(\ln(u^{(q)}) - \overline{\ln(u)}\right)^T \end{bmatrix}}_{\triangleq Y}$$

find parameters $C \triangleq \begin{bmatrix} B^x & B^u \end{bmatrix}^T$ *solving the regression problem*

$$W = Y \cdot C + \varepsilon \tag{6}$$

where $\varepsilon \in \mathbb{R}^{q\times m}$ *is measurement noise on* W.

Notice that the parameter vector a no longer appears in the regression problem, but an estimate of it can be recovered from estimates of $C = \begin{bmatrix} B^x & B^u \end{bmatrix}^T$ by way of Eq. (4).

In the remainder of the paper, we make the assumption that each column $\varepsilon_{\cdot i}$ of ε follows a Gaussian distribution, indicated by $\varepsilon_{\cdot i} \sim \mathcal{N}(0, \Sigma_{\varepsilon_i})$, where Σ_{ε_i} is diagonal, *i.e.* the measurement errors in different experiments are mutually uncorrelated. We further assume that $\varepsilon_{\cdot i}$ is independent of $\varepsilon_{\cdot j}$ for $i \neq j$. Then, Problem 1 can be subdivided into m independent subproblems, one for each reaction i:

$$w_{\cdot i} = Y \cdot c_{\cdot i} + \varepsilon_{\cdot i}, \tag{7}$$

where $w_{\cdot i}$ and $c_{\cdot i}$ are the i-th columns of W and C, respectively.

The values of the parameter matrices B^x and B^u admit an interesting biological interpretation. Notice that one can immediately find values $x_0 \in \mathbb{R}^n_+$,

$u_0 \in \mathbb{R}^p_+$, $e_0 \in \mathbb{R}^m_+$ and $v_0 \in \mathbb{R}^m$ such that $v_0/e_0 = \overline{v/e}$, $\ln x_0 = \overline{\ln(x)}$, and $\ln u_0 = \overline{\ln(u)}$. As a consequence, Eq. (5) can be rearranged into the common relative formulation of linlog models,

$$\frac{v}{e} = \text{diag}\left(\frac{v_0}{e_0}\right)\left[1 + B_0^x \ln\left(\frac{x}{x_0}\right) + B_0^u \ln\left(\frac{u}{u_0}\right)\right] \tag{8}$$

where 1 is an $m \times 1$ vector of ones, (v_0, x_0, u_0, e_0) is a so-called reference state [Heijnen, 2005] and B_0^x, B_0^u are matrices of elasticity constants, where

$$B_0^x = \text{diag}\left(\frac{e_0}{v_0}\right) \cdot B^x, \quad B_0^u = \text{diag}\left(\frac{e_0}{v_0}\right) \cdot B^u. \tag{9}$$

The elasticities, introduced in the context of Metabolic Control Analysis (MCA) [Heinrich and Schuster, 1996], describe the normalized local response of the reaction rates to changes in metabolite concentrations. The interest is that they can thus be immediately computed from the values of B^x and B^u found by the solution of Problem 1, and the equality $e_0/v_0 = 1/(\overline{v/e})$.

Although straightforward in theory, solving the regression problem (6) encounters two complications in practice.

1. Since the measurements are carried out at (quasi-)steady state, we have $N \cdot v(x, u, e) = 0$. This introduces dependencies among the data and thus reduces the information content of the data matrix Y, in the sense that Y becomes rank deficient. Like in earlier work [Nikerel et al., 2006], we use standard approaches to solve this problem. We notably rely on Principal Component Analysis (PCA) [Jolliffe, 1986, Nikerel et al., 2006] applied to the data matrix Y to reduce the model order, *i.e.*, the number of independent parameters, and ensure well-posedness of the regression problem (see Appendix A for technical details). In summary, we use Singular Value Decomposition (SVD), a technique decomposing the data matrix into dominant and marginal components according to a variance criterion. For the purpose of linear regression, this corresponds to decomposing the parameter vector into a reduced number of components that can be determined with certainty based on the data, while the remaining components are poorly determined, *i.e.*, they are 'nonidentifiable', and are discarded with negligible effect on the fit. We note in passing that the columns of W and Y are zero-mean, an important requirement for the correctness of the outlined analysis.

2. The high-throughput datasets contain a substantial amount of missing values, due to experimental limitations or instrument failures. If, for any given reaction, we only used the datapoints in which all relevant metabolite concentrations, enzyme concentrations, and metabolic fluxes playing a role in that reaction are available, then a large amount of data would have to be thrown away. In practice, we would run the risk that the parameters cannot be reliably identified. The development of a method that is capable of maximally exploiting the information contained in incomplete datasets for solving Problem 1 is the main subject of the paper and will be fully developed in the later sections.

3 Likelihood-based identification of linlog models from missing data

For every reaction i, we are concerned with the problem of estimating the unknown parameters $c_{\cdot i}$ of the model given in (7) in the case where some entries of Y are unknown. We address the estimation problem by a likelihood-maximization approach, which is known to yield optimal (unbiased and minimum variance) estimates for our problem setting in the case where Y is fully known. As the problem is identical for all reactions i, in the remainder of the section we will drop for simplicity index $\cdot i$ from the notation.

Let $\mathscr{I}$ be the set of indices (row, column) corresponding to the known entries of Y, *i.e.*, $(j,k) \in \mathscr{I}$ if and only if $Y_{j,k}$ is available. It is convenient to introduce the decomposition $Y = \breve{Y} + \widetilde{Y}$, where

$$\breve{Y}_{j,k} = \begin{cases} Y_{j,k}, & \text{if } (j,k) \in \mathscr{I}, \\ 0, & \text{otherwise;} \end{cases} \qquad \widetilde{Y}_{j,k} = \begin{cases} 0, & \text{if } (j,k) \in \mathscr{I}, \\ Y_{j,k}, & \text{otherwise.} \end{cases}$$

Matrix $\breve{Y}$ is fully determined: Once measurements $\breve{y}$ of $\breve{Y}$ are collected, we treat $\breve{Y} = \breve{y}$ as fixed parameters of the regression problem. Matrix $\widetilde{Y}$ collects the unknown entries of Y. We model these missing data as unobserved independent random variables, whose prior distributions encode our generic knowledge about them. Assuming that the *a-priori* distributions are not known (worst case), we define a Gaussian prior for each quantity that is missing in an experiment based on the measurements of the same quantity available from other experiments. For every $(j,k) \notin \mathscr{I}$ and $\mathscr{Y}_{j,k} = \{Y_{j',k} : \ (j',k) \in \mathscr{I}\}$ (assumed nonempty), we let

$$\begin{cases} \widetilde{Y}_{j,k} \sim \mathscr{N}(\mu_{j,k}, \sigma_{j,k}^2), \\ \mu_{j,k} = \mathrm{mean}(\mathscr{Y}_{j,k}), \\ \sigma_{j,k} = \mathrm{std}(\mathscr{Y}_{j,k}). \end{cases} \tag{10}$$

We can now formulate the estimation problem.

Problem 2 *Given measurements $W = w$ and $\breve{Y} = \breve{y}$, compute the estimate $\hat{c} = \arg\max_c \log \mathscr{L}(c)$, with $\mathscr{L}(c) = f_{W|\breve{y},c}(w)$, where, for any c, $f_{W|\breve{y},c}(\cdot)$ is the probability density function of W given $\breve{Y} = \breve{y}$ corresponding to model* (7)–(10).

Note that $\mathscr{L}(c)$ is a likelihood function for a linear model with missing data, in the sense that it is defined with respect to available data $\breve{Y}$ only. One can express $\mathscr{L}(c)$ by marginalization,

$$\log \mathscr{L}(c) = \log \int f_{W|\breve{y},\tilde{y},c}(w) f_{\widetilde{Y}|\breve{y},c}(\tilde{y}) d\tilde{y}, \tag{11}$$

where $f_{W|\breve{y},\tilde{y},c}(\cdot)$ is the standard likelihood function for model (7) given $\breve{Y} = \breve{y}$ and $\tilde{y}$, with $\tilde{y}$ varying over all possible values of $\widetilde{Y}$, and $f_{\widetilde{Y}|\breve{y},c} = f_{\widetilde{Y}|\breve{y}}$ is determined by the prior (10). The explicit solution to the integral is reported in Appendix B. A direct approach to solving Problem 2 is to maximize (11) by numerical optimization. However, the function is not convex in c, whence its direct optimization is prone to end up in local minima and requires the use of global optimization strategies.

Alternatively, we propose to tackle Problem 2 by an Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. EM provides a general methodology for the optimization of a likelihood function with missing information. It is based on an iterative two-step procedure that, for the problem at hand, we implement as follows. Let us define the random variable $Z = \widetilde{Y} \cdot c$, so that model (7) becomes $W = \breve{Y} \cdot c + Z + \varepsilon$. Note that $Z \sim \mathcal{N}(\mu_{\breve{y},c}, \Sigma_{\breve{y},c})$, where for any given c, mean and variance can be derived from (10). Let $\hat{c}^0$ be an initial guess of c. At every iteration $\ell = 1, 2, 3, \ldots$, compute an updated estimate $\hat{c}^\ell$ from the estimate $\hat{c}^{\ell-1}$ at the previous iteration by performing the following EM steps:

Expectation: Compute

$$\begin{aligned} Q(c|\hat{c}^{\ell-1}) &= \mathbb{E}[\log f_{Z,W|\breve{y},c}(Z,w)|\breve{y}, \hat{c}^{\ell-1}, w] \\ &= \int \log f_{Z,W|\breve{y},c}(z,w) f_{Z|\breve{y},\hat{c}^{\ell-1},w}(z) dz; \end{aligned} \tag{12}$$

Maximization: Solve

$$\hat{c}^\ell = \arg\max_c Q(c|\hat{c}^{\ell-1}). \tag{13}$$

In (12), $f_{Z,W|\breve{y},c}$ is the joint probability density function of Z and W given $\breve{Y} = \breve{y}$ and c, while $f_{Z|\breve{y},\hat{c}^{\ell-1},w}$ is the probability density function of Z given $\breve{Y} = \breve{y}$, $W = w$ and $\hat{c}^{\ell-1}$. In fact, this quantity is independent of w and can be computed by our definition (10).

It can be proven that, at every iteration ℓ, the EM algorithm increases the value of $\mathscr{L}(\hat{c}^\ell)$, and eventually converges to a maximum of $\mathscr{L}$ [Little and Rubin, 2002]. While this is not necessarily a global maximum, EM has proven effective in many applications [Graham, 2009, Horton and Kleinman, 2007]. A key property is that convergence to a maximum is achieved even if (13) is not solved exactly: It suffices that $\hat{c}^\ell$ is such that $Q(\hat{c}^\ell|\hat{c}^{\ell-1}) \geq Q(\hat{c}^{\ell-1}|\hat{c}^{\ell-1})$, which is easily achieved even by a local optimization algorithm. In practice, we can use the explicit expression of $\mathscr{L}$ in Problem 2 for stopping the iterations, *e.g.*, when the relative improvement on $\mathscr{L}$ falls below a specified threshold $\tau > 0$:

$$|\mathscr{L}(\hat{c}^\ell) - \mathscr{L}(\hat{c}^{\ell-1})|/|\mathscr{L}(\hat{c}^\ell)| \leq \tau.$$

To complete the implementation of the algorithm, one must express $Q(c|\hat{c}^{\ell-1})$ in a form convenient for maximization. As explained in Appendix B, we were able to express (12) as an explicit function of c for any given $\hat{c}^{\ell-1}$. In compact form:

$$Q(c|\hat{c}^{\ell-1}) \propto -KL(f_c||f_{\hat{c}^{\ell-1}}) - H(f_{\hat{c}^{\ell-1}}) + \log(\kappa_{f_c}), \tag{14}$$

where f_c stands for a Gaussian distribution with variance $\Sigma_{f_c} = [\Sigma_\varepsilon^{-1} + \Sigma_{\breve{y},c}^{-1}]^{-1}$ and mean $\mu_{f_c} = \Sigma_{f_c} \cdot (\Sigma_\varepsilon^{-1} \cdot (w - \breve{y} \cdot c) + \Sigma_{\breve{y},c}^{-1} \cdot \mu_{\breve{y},c})$, κ_{f_c} is a function depending on c via μ_{f_c} and Σ_{f_c}, and the proportionality factor that we dropped (indicated by the presence of $\propto$ in place of $=$) depends on $\hat{c}^{\ell-1}$ but not on c. Finally, $KL(\cdot||\cdot)$ and $H(\cdot)$ are the Kullback-Leibler distance between distributions and the entropy of a distribution, respectively, for which, in the Gaussian case at hand, explicit formulas are available [Cover and Thomas, 2006, Stoorvogel and van Schuppen, 1996]. A slight technical complicacy is needed in case $\Sigma_{\breve{y},c}$ is singular (see Appendix B for all the mathematical details).

The availability of the closed-form expression (14) allows us to implement EM efficiently, *i.e.*, with an explicit maximization problem that is solved numerically at all iterations. Once the parameter estimates are obtained, several methods from the literature can be used to assess the accuracy of the results by inferring confidence intervals. Examples are randomized methods such as bootstrapping [Manly, 1997] and the profile likelihood method by [Raue et al., 2009]. This method derives confidence intervals using a threshold on a function called the profile likelihood. In our application, this is obtained separately for each parameter c_j by re-maximization of (11) with respect to all parameters $c_{k\neq j}$, for all values c_j in a neighborhood of $\hat{c}_j$.

4 Validation on synthetic data

Before applying the EM algorithm to actual biological identification problems, we test the performance of the method on simulated data. For this purpose, a synthetic model has been developed, a simplified variant of the linlog model of *E. coli* central metabolism studied in Sec. 5 below. The model, in the form (2), contains 17 variables, representing internal and external metabolites involved in 25 reactions, and 78 parameters (see Appendix C for the model equations). We generate data matrices Y from this model by means of simulation, for different percentages of missing data and experimental noise. Using the model structure and the simulated data, we solve Problem 1 for each reaction independently, as described in Sec. 3.

In order to assess the added value of our specific implementation of likelihood optimization, we first compare the performance of the EM algorithm of Sec. 3 with the direct maximization of the loglikelihood (11) implemented with a general-purpose MATLAB optimization routine. This method will be referred to as MaxLL in the sequel.

Second, we compare the likelihood-based identification approaches with standard methods, notably linear regression (referred to as Rg) and the commonly-used multiple imputation (MI) method [Rubin, 1976, 1996]. Regression is performed based on full datasets only, *i.e.*, it does not consider an experimentally-determined datapoint $(v^{(k)}, x^{(k)}, u^{(k)}, e^{(k)})$ when at least one of the measurements is missing. MI is based on imputation of missing data by random draws of the missing values, *i.e.*, non-zero elements of $\tilde{Y}$, from the *a-priori* distribution defined in (10). Both methods thus exploit only part of the information contained in an incomplete dataset and provide a lower limit for quantifying the performance of the methods proposed in Sec. 3.

Third, we compare the results of EM with the least-squares identification of the model on complete datasets (a method referred to as RgF). Though inapplicable to real data with missing measurements, the method is statistically optimal. Hence, it provides us an upper performance bound that can be used to assess the role of missing data in performance degradation, separately from the role of noise.

Most of the high-throughput datasets available in the literature have been obtained when metabolism is at (quasi-)steady-state. In order to mimic available experimental data as closely as possible, simulated data obtained from the synthetic model should therefore be steady-state data. We generated steady states of (1)-(2), and recorded the corresponding metabolite concentrations and metabolic flux values for 30 different conditions, each consisting of a random change in the enzyme concentration with respect to a reference value.

We compared performance of the five methods described above (EM, MaxLL, MI, Rg, RgF) on datasets with different amounts of missing data (40% and 75%) for the metabolite concentrations and noise levels (10% and 20%) for w. The only difference with the dataset used for the reference method RgF is that the latter has no missing data. A noise level of 10% means that the distribution used to generate the noise has a standard deviation equal to 10% of the values in w. The percentages of missing data in the simulation study are comparable to those observed in practice (Sec. 5 and [Ishii et al., 2007]). For every different combination of missing data percentage and noise level, a dataset was generated by homogeneously distributing missing data among columns of Y, the indices for each column being chosen at random. For every simulated scenario, randomly generated noise was added to w in the dataset.

For all the above scenarios, identification of each reaction was addressed separately, in accordance with the discussion of Sec. 2. For every reaction, we first tested the identifiability of the synthetic linlog model by PCA of the full data matrix Y. In our simulation, 9 reactions out of the 25 composing the model were detected as having nonidentifiable parameters. For those reactions, identification of a reduced order model

$$w = Y^* \cdot c^* + \varepsilon \tag{15}$$

was performed in place of the identification of the original model. $Y^* \in \mathbb{R}^{q \times r}$, with $r \leq n+p$, is a reduced-order data matrix obtained by linear transformation of Y and $c^* \in \mathbb{R}^r$ is a parameter vector, smaller than c, that is 'identifiable', in the sense that it is well determined by the data (see Appendix C).

We implemented the different parameter estimation algorithms in Matlab, using the `lscov` function for the regression-based methods and `fminsearch` for global optimization in MaxLL and the maximization step in EM. Both EM and MaxLL require an initial guess of the parameters to be specified. We proposed 10 different initial parameter vectors, including the estimation obtained with the baseline method Rg where available. In order to draw statistics for the estimation performance, each of the five algorithms was applied on 100 Monte Carlo repetitions of the identification problem. The complete performance test over all methods, conditions and 100 repetitions took about 7 h 40 min in Matlab 7.4.0 on a Linux PC workstation (1862 MhZ, 2 Gb RAM).

The most informative results from all identification methods are summarized by boxplots of the ratio of the estimated parameter values c over the reference parameter values c_{ref} used to simulate the data, the closer the ratio to 1, the better the estimates. Ensemble statistics are drawn for all parameters corresponding to the same reaction. Fig. 1 is dedicated to the scenario with 40% missing data and 10% noise, whereas Fig. 2 reports on 75% missing data and 20% noise. Complete results for all reactions under all conditions can be found in Appendix C.

Since the individual reactions of the model involve only a small subset of metabolites, each of the m identification subproblems consists of the estimation of a limited number of parameters, mostly 2 or 3. For the case with 40% missing

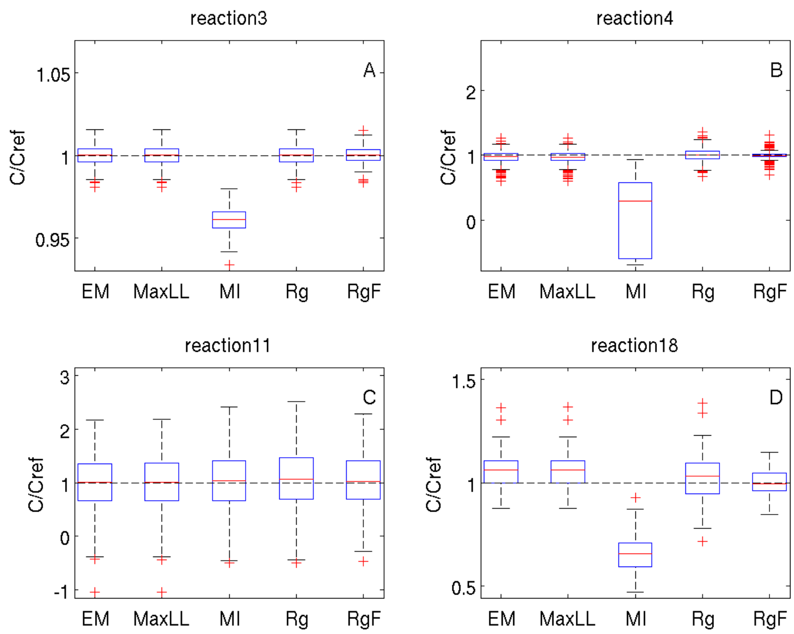

Figure 1: Statistics of estimated parameter values for datasets with 40% of missing data and 10% noise. The results are shown as boxplots of the ratio of the estimated parameter values c and reference parameter values c_{ref}. Statistics have been computed for each of the 5 methods from 100 datasets. For each method, the red line displays the median and the lower and upper blue lines represent the lower and upper quartile values, respectively. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the ends of the box and outliers are shown with red crosses. The tested algorithms are Expectation Maximization (EM), direct optimization of loglikelihood (MaxLL), multiple imputation (MI), regression on incomplete datasets (Rg) and regression on complete datasets (RgF). (A–D) Boxplots for reactions 3, 4, 11 and 18 of the network, respectively.

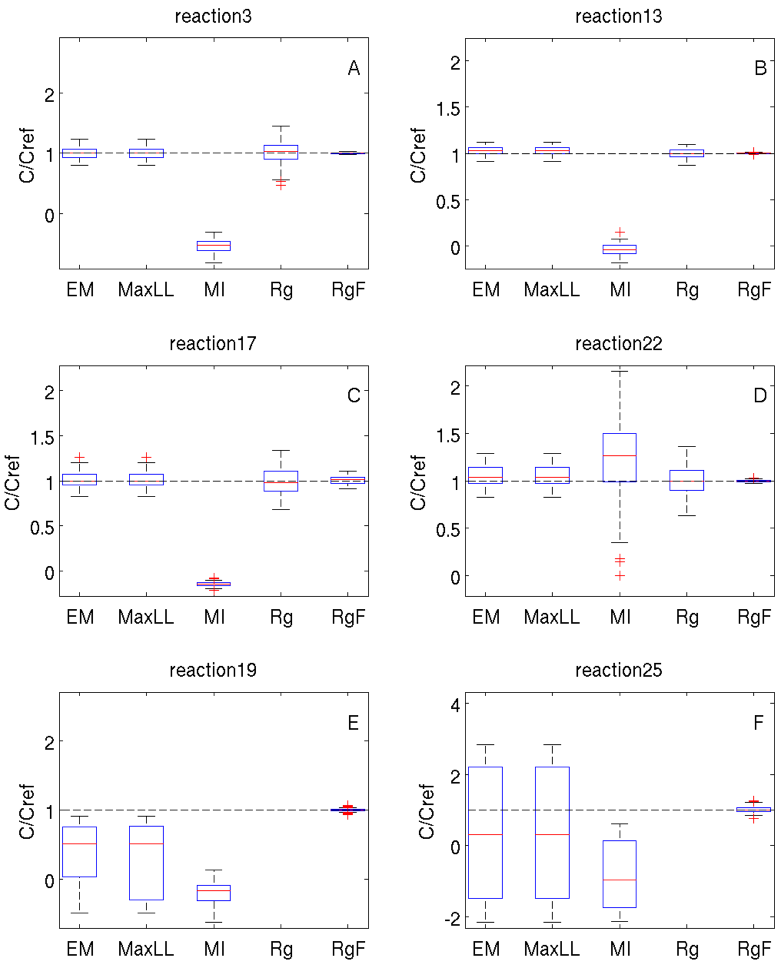

Figure 2: Statistics of estimated parameter values for datasets with 75% of missing data and 20% noise. The graphical notations are the same as for Fig. 1. (A–F) Boxplots for reactions 3, 13, 17, 22, 19 and 22 of the network, respectively.

data, Rg can therefore be performed in all runs for every reaction of the model. On the contrary, with 75% missing data, regression cannot be applied to 6 reactions which is apparent from the absence of the Rg statistics for 2 reactions in Fig. 2.

In comparison with the other methods, multiple imputation (MI) gives the worst results (largest bias) in 3 out of the 4 reactions shown in Fig. 1, and in 5 out of 6 reactions in Fig. 2. In reactions 11 of Fig. 1 and 22 of Fig. 2, the relatively small biases are accompanied by an estimation uncertainty wider than for EM and MaxLL. This could be explained by a restricted use of information contained in the distribution of missing data. Indeed, MI only considers random draws from the distribution while EM and MaxLL are based on all possible values taken by missing data through integration of the distribution.

Analysis of Fig. 1 reveals that, for 40% missing data and 10% noise, the performance of EM and MaxLL is almost identical and similar to that of regression (Rg and RgF), with limited improvements on Rg, *i.e.*, slightly smaller variability. In some cases, such as for reactions 11 and 18, their performance approaches the optimal, unattainable bound provided by RgF, *i.e.*, they have similar bias and variability.

Performance improvements of likelihood-based methods over Rg become more significant when identification is performed on the dataset with higher percentage of missing data and larger noise. Fig. 2A-D show results for reactions where Rg was applicable. Both EM and MaxLL substantially reduce estimation variability in reactions 3, 17 and 22. At the same time, due to the larger amount of missing data, performance loss with respect to RgF is more significant. Turned another way, this shows the accuracy that could be recovered were all datasets complete.

Fig. 2E-F show the results when Rg fails to produce estimates and cannot be used to initialize EM and MaxLL optimization. Still, EM provides estimates of the right order of magnitude and, for the case of Fig. 2E, of the right sign in at least 75% of the runs (box entirely above 0), while the median has the right sign and is reasonably close to 1. The estimation of the sign provided by MaxLL is less reliable (box crossing 0).

Overall, we conclude that the EM-based approach provides the most accurate estimates under all simulated conditions. We will therefore apply this method to the identification of the linlog model of an actual metabolic network from a published high-throughput dataset.

5 Application to central metabolism in *E. coli*

The network of central carbon metabolism in *Escherichia coli* has been studied for a long time from different perspectives, which makes it an ideal model system for our purpose. A rather precise idea of the structure of the network exists, several kinetic models of the network dynamics are available ([Bettenbrock et al., 2005, Kotte et al., 2010] and references therein), and recently a high-throughput dataset containing the required information for solving Problem 1 has been published [Ishii et al., 2007]. The network we consider here gathers enzymes, metabolites and reactions that make up the bulk of *E. coli* central carbon metabolism, including glycolysis, the pentose-phosphate pathway, the tricarboxylic acid cycle and anaplerotic reactions such as glyoxylate shunt and PEP-carboxylase (Fig. 3).

The dataset used for identification of this network was obtained by experiments with 24 single-gene disruptants that were grown at a fixed dilution rate of 0.2 h^{-1} in glucose-limited chemostat and on wild-type cells at 5 different dilution rates [Ishii et al., 2007]. The authors collected data using multiple high-throughput techniques, in particular DNA microarray analysis and two-dimensional differential gel electrophoresis (2D-DIGE) for genes and proteins, capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS) for metabolites, and metabolic flux analysis. They thus obtained a dataset consisting of metabolite concentrations, mRNA and protein concentrations for the enzymes, and metabolic fluxes under 29 different experimental conditions. A large number of different metabolites were measured in the experiments, with missing data in varying amounts, from 0 to 80% of the observations, 28% on average for the metabolites considered below.

From the reactions listed in [Ishii et al., 2007], we have constructed a linlog model of the form (2), with $n = 16$ internal metabolites, $q = 7$ external metabolites and measured cofactors, and $m = 31$ reactions (see Appendix D). Each of the reactions is catalyzed by a single enzyme, which may actually stand for several enzymes in the case of isoenzymes, enzyme complexes, or lumped reactions. Reactions have been simplified or lumped together when a shared metabolite has not been measured, which precludes estimation of the corresponding elements in the parameter matrices B^x and B^u. In comparison with an earlier linlog model of *E. coli* central carbon metabolism [Visser et al., 2004], we extended the scope to include the tricarboxylic acid cycle and the glyoxylate shunt, but due to the above-mentioned simplifications our model is more coarse-grained.

An identifiability analysis was performed by several rounds of missing data imputation using the *a-priori* distribution defined in Eq. (10) and PCA, which led in each case to the same result: 7 out of 31 reactions were detected as having nonidentifiable parameters. For those reactions, the model has been reduced as described in Eq. (15) using a data matrix Y completed by the means $\mu_{j,k}$ of the *a-priori* distributions. For every individual reaction, the reduced model has a parameter vector c^* that is now entirely identifiable.

Apart from the distribution of the *a-priori* missing data, given by Eq. (10), application of the EM requires information about the distribution of ε, the error on the ratios of fluxes and enzyme concentrations. The Ishii dataset provides several replica measurements for a reference experimental condition: wild-type cells grown on glucose-limited chemostat with a dilution rate of 0.2 h^{-1}. These data were used for the computation of the variance of ε. Running the EM method on the model and the data took about 220 s using the implementation of Sec. 4. In order to assess the accuracy of the estimated B^x and B^u, we computed for each parameter a 95% confidence interval, by means of the profile likelihood method outlined in Sec. 3. The computation of the confidence intervals for all parameters required about 23 min.

Contrary to the simulation studies reported in Sec. 4, a reference or 'real' model for the evaluation of the results does not exist in this case. However, *a-priori* biochemical knowledge on the signs of the elasticities is available, *i.e.*, elasticities are positive for substrates and negative for products. This information can be compared with the estimated signs of the elasticities, and their confidence intervals, computed from the parameter matrices using the relations

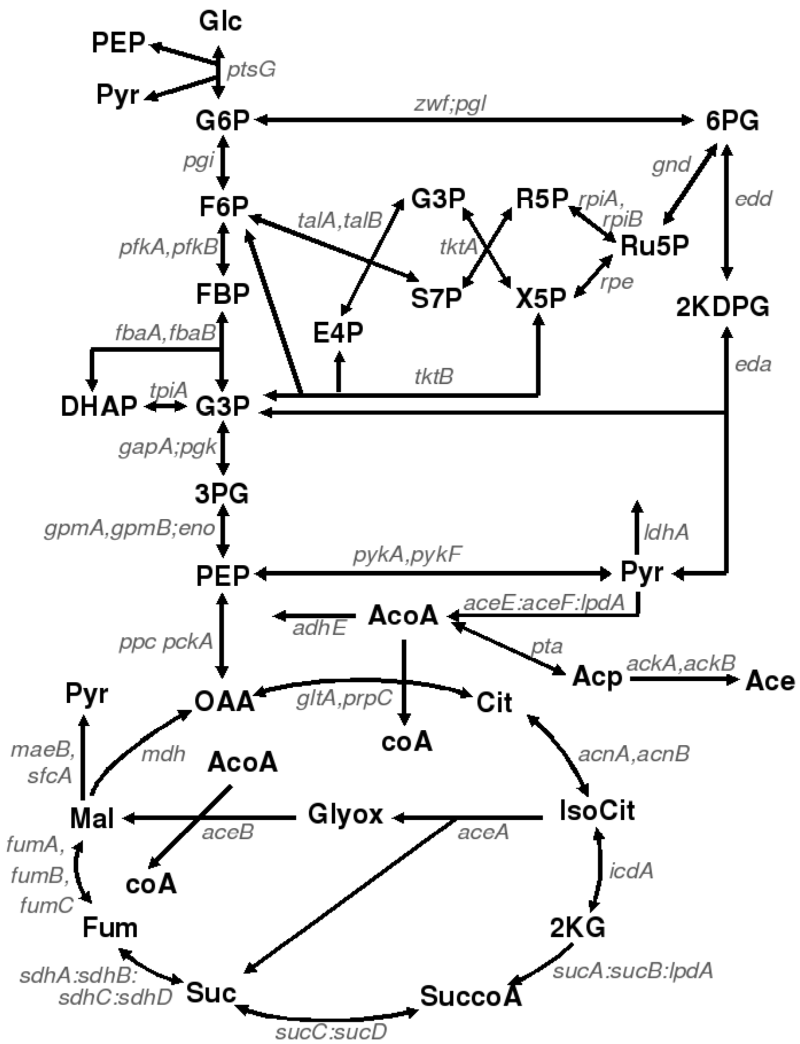

Figure 3: Scheme of *Escherichia coli* central carbon metabolism. This map, showing metabolites (bold fonts) and genes (italic) is adapted from [Ishii et al., 2007]. Abbreviations of metabolites are glucose (Glc), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), fructose 1-6-biphosphate (FBP), dihydroxyacetone phosphate (DHAP), glyceraldehyde 3-phosphate (G3P), 3-phosphoglycerate (3PG), phosphoenolpyruvate (PEP), pyruvate (Pyr), 6-phosphogluconate (6PG), 2-keto-3-deoxy-6-phospho-gluconate (2KDPG), ribulose 5-phosphate (Ru5P), ribose 5-phosphate (R5P), xylulose 5-phosphate (X5P), sedoheptulose 7-phosphate (S7P), erythrose 4-phosphate (E4P), oxaloacetate (OAA), citrate (Cit), isocitrate (IsoCit), 2-keto-glutarate (2KG), succinate-CoA (SuccoA), succinate (Suc), fumarate (Fum), malate (Mal), glyoxylate (Glyox), acetyl-CoA (AcoA), acetylphosphate (Acp) and acetate (Ace). Cofactors impacting the reactions are not shown. The gene names are separated by a comma in the case of isoenzymes, by a colon for enzyme complexes, and by a semicolon when the enzymes catalyze reactions that have been lumped together in the model.

in Eq. (9). The results are shown in Table 1. Similar unshown results are obtained by means of the MaxLL method.

We observe that the EM method obtains estimates for all reactions, including the 7 cases where the insufficient amount of data made regression not applicable. However, 27 of the 100 non-zero elasticities of the model are not identifiable from this dataset. Moreover, out of the remaining 73 elasticity estimates, half of them have signs that are not statistically significant, in the sense that the 95% confidence interval straddles 0. This is most likely due to the magnitude of noise in metabolite concentrations being comparable to the magnitude of relevant information, as for example for PEP where the standard deviation over all experimental conditions equals the standard deviation of the replicates in a single condition (0.06 mM vs 0.05 mM). This precludes the estimation of an unambiguous sign.

Of the elasticities with statistically significant signs, 20 out of 37 are correct, in the sense that they have the expected positive or negative sign. The remaining elasticities, distributed over 11 reactions, are incorrectly estimated. Let us now discuss what we believe are potential sources of these errors, an information that could be used to single out erroneous estimates *a-priori*.

We first note that for 5 of these 11 reactions (GapA;Pgk, GltA,PrpC, Mdh, Edd;Eda and Pta;AckA,AckB, see Table 1), only very few complete datapoints are available (between 1 and 5) and regression mostly fails in these cases. In addition, all of these reactions involve at least one metabolite missing in more than 70% of the experimental conditions. The combination of very few complete datapoints and a high percentage of missing metabolite measurements obviously makes model identification extremely difficult and it is fair to say that here we reach the limit of the applicability of our method, or of any method for that matter, due to the lack of data.

Second, 4 reactions are known to operate close to equilibrium: Pgi, FbaA,FbaB, TpiA and GpmA,GpmB;Eno [Visser et al., 2004]. Theoretically, these reactions are not identifiable, as their elasticities are not independent [Visser et al., 2004], but PCA did not detect this. Most likely, this is due to the above-mentioned noise in metabolite concentrations, which decreases their correlations. A cautious, preemptive strategy would be to reduce the model for any reaction known to be close to equilibrium and eliminate the corresponding dependent variables.

The errors in the signs of some elasticities in the remaining 2 reactions (PtsG and PfkA,PfkB) are less straightforward to explain. It is unlikely that they can be attributed to the EM method, given that regression is applicable here with a relatively large number of complete datapoints available (14 and 18, respectively) and gives the same results. Alternatively, they may be explained by a modeling error or a hidden variable, for instance an unknown cofactor, biasing the estimation results. It is also possible that the approximations of the linlog model are not suitable for these reactions, for instance because there are large variations in metabolite concentrations between conditions, driving the system far from the reference state.

In summary, EM gives reasonable results for a fairly complicated model on a challenging dataset. Even though some puzzling issues remain, we believe that these can be safely attributed to the inherent difficulty of the identification problem.

Table 1: Elasticity matrix $[B_0^x \; B_0^u]$ estimated by EM from the data of Ishii et al. [2007] for the linlog model of *E. coli* central carbon metabolism (the columns of the matrix have been permuted for readability). Unidentifiable elasticities are shown in grey, uncertain elasticities, *i.e.*, having a sign that is not significant with 95% confidence, in yellow, and correctly/incorrectly identified elasticities, *i.e.*, having a sign that is significant with 95% confidence, in green/red. Abbreviations are as in Fig. 3. Some of the cofactors are modeled as ratios of metabolite concentrations, *e.g.* ATP/ADP. Reaction 27, labeled μ, is a phenomenological reaction for biomass production. The last row indicates the percentage of missing data per metabolite and the right-most column displays the amount of complete datapoints available for each reaction. For reactions labeled with *, regression was not able to produce any result.

	Enzyme \ Metabolite	Glc	PEP	G6P	Pyr	F6P	FBP	DHAP	3PG	AcoA/coA	6PG	Ru5P	R5P	S7P	2KG	Suc	Fum	Mal	ATP/ADP	Cit	NADPH/NADP	NADH/NAD	FAD	Ace	# complete datapoints
1	PtsG	0.29	-0.88	0.79	1.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
2	Pgi	0	0	-0.33	0	0.23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27
3	PfkA,PfkB	0	-0.13	0	0	-0.11	-0.23	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	18
4	FbaA,FbaB	0	0	0	0	0	-0.29	0.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
5	TpiA	0	0	0	0	0	0	-0.07	0.22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
6	GapA;Pgk	0	0	0	0	0	-0.18	0	-0.05	0	0	0	0	0	0	0	0	0	0.32	0	0	-0.17	0	0	5
7	GpmA,GpmB;Eno	0	0.26	0	0	0	0	0	-0.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24
8	PykA,PykF	0	0.19	0	0.43	0	-0.14	0	0	0	0	0	0	0	0	0	0	0	0.11	0	0	0	0	0	11
9	AceE:AceF:LpdA	0	0	0	0.05	0	0	0	0	-0.05	0	0	0	0	0	0	0	0	0	0	0	0.27	0	0	6
10	Zwf;Pgl	0	0	-0.1	0	0	0	0	0	0	-0.21	0	0	0	0	0	0	0	0	0	-1.38	0	0	0	2
11	Gnd	0	0	0	0	0	0	0	0	0	0.33	-0.08	0	0	0	0	0	0	0	0	-0.76	0	0	0	2*
12	Rpe	0	0	0	0	0	0	0	0	0	0	0.46	0	-0.39	0	0	0	0	0	0	0	0	0	0	28
13	RpiA,RpiB	0	0	-0.64	0	0	0	0	0	0	0	0.26	-0.1	0	0	0	0	0	0	0	0	0	0	0	18
14	TktA	0	0	0	0	0	0	0	0	0	0	0	0.01	-0.1	0	0	0	0	0	0	0	0	0	0	18
15	TalA,TalB	0	0	0	0	-0.25	0	0	0	0	0	0	0	0.29	0	0	0	0	0	0	0	0	0	0	27
16	TktB	0	0	0	0	-0.58	0	0	0	0	0	0.35	0	0	0	0	0	0	0	0	0	0	0	0	27
17	GltA,PrpC	0	0	0	0	0	0	0	0	0.22	0	0	0	0	0.01	0	0	0	0	0.49	0	-0.02	0	0	1*
18	AcnA,AcnB	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0.55	0	0	0	0	5
19	IcdA	0	0	0	0	0	0	0	0	0.07	0	0	0	0	0.002	0	0	0	0	0	0.001	0	0	0	4
20	SucA:SucB:LpdA;SucC:SucD	0	0	0	0	0	0	0	0	0	0	0	0	0	1.26	0.3	0	0	0	0	0	-0.48	0	0	2*
21	SdhA:SdhB:SdhC:SdhD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.08	-0.44	0	0	0	0	0	0.4	0	21
22	FumA,FumB,FumC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.46	0.1	0	0	0	0	0	0	25
23	Mdh	0	0.15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.01	0	0.31	0	-0.12	0	0	3*
24	Ppc;PckA	0	0.29	0	0	0	-0.13	0	0	0	0	0	0	0	0	0	0	0.1	-1.15	0.21	0	0	0	0	9
25	MaeB,SfcA	0	0	0	0.08	0	0	0	0	-0.2	0	0	0	0	0	0	0	0.26	0	0	-0.003	0.04	0	0	2*
26	AceA;AceB	0	0	0	0	0	0	0	0	-0.11	0	0	0	0	0	-0.01	0	-0.02	0	0	0	0	0	0	25
27	μ	0	0.16	-0.11	0.06	-0.04	0	0	0.11	0.1	0	0	0.09	0	-0.06	0	0	0	-0.44	0	0.05	-0.01	0	0	0*
28	Edd;Eda	0	0	0	2.47	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	5
29	Pta;AckA,AckB	0	0	0	0.0015	0	0	0	0	-0.33	0	0	0	0	0	0	0	0	-1.42	0	-0.26	-0.83	0	2.1	2*
30	LdhA	0	0	0	2.67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.11	0	0	6
31	AdhE	0	0	0	0	0	0	0	0	0.91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28
	% Missing Data	3	17	0	48	7	34	59	10	3	72	3	38	3	59	3	14	14	0	62	79	79	17	17	

6 Discussion

In this work we have addressed the problem of estimating parameters of approximate models of metabolic networks from incomplete datasets. Even with the largest datasets available at present, such as in [Ishii et al., 2007], the absence or corruption of a large number of measurements may reduce the effective number of datapoints to a handful of experimental conditions, thus making simple regression techniques ineffective or even inapplicable. Making full use of all the available data is therefore essential to render identification well-posed and improve the quality of the estimated models.

To this aim, we have proposed a maximum-likelihood method for the identification of linlog metabolic network models that compensates for the missing data by the use of statistical priors. We developed an algorithm that attains maximization of the likelihood based on Expectation Maximization, a well accepted paradigm for the numerical optimization of likehood functions in the presence of unobserved variables. A simpler implementation based on direct likelihood maximization via general-purpose numerical optimization algorithms was also considered and found slightly less powerful. The performance of EM was compared to that of an existing method of reference, namely multiple imputation, and to worst-case and best-case scenarios given by least-squares regression on the sole complete datapoints and on complete datasets, respectively. We showed that EM outperforms multiple imputation by a wide margin. In comparison with worst-case regression, it reduces the estimation variability and is able to produce reasonable estimation results even when regression on incomplete datasets is inapplicable. It also approaches the ideal performance of regression on complete datasets for low rates of missing data, regardless of noise.

Based on these findings, we applied EM to the identification of a linlog model for the central carbon metabolism in the bacterium *E. coli* from the experimental data presented by Ishii et al. [2007]. Even with the large amount of incomplete datapoints, due to the difficulty of experimentally measuring metabolite concentrations, EM was able to estimate many of the model parameters (elasticies) in agreement with the current understanding of the system. This is even true for reactions where the reduced number of complete datapoints impairs the applicability of least squares regression. On the other hand, the challenging quality of the data sheds light on the performance limits of the method, which tends to fail when large measurement noise makes the estimation of small parameters statistically unreliable, when the same variable cannot be measured in most conditions, or when reactions operate near equilibrium.

Overall, results from the simulations and the application on real data showed that our EM approach is able to make the most of incomplete, noisy high-throughput datasets for the estimation of parameters in approximate kinetic models. In the future, we expect to improve performance by developing a number of technical points, including approximate analytic/dedicated numerical solutions for the EM maximization steps, the refinement of the identifiability analysis via SVD of incomplete data matrices [Brand, 2002], and a more detailed modeling of measurement noise. It is worth noting that, while the method has been developed on linlog models, it is more generally applicable to any other metabolic network model that can be put in a form linear in the parameters by straightforward manipulations, such as generated mass action models that provide advantages when some metabolite concentrations approach 0 [Savageau, 1976, del Rosario et al., 2008]. In addition, estimated parameters of approximate metabolic models, such as elasticities of linlog models, provide useful hints for the identification of more detailed nonlinear kinetic models.

From a broader perspective, the application of the EM method to a unique multi-omics dataset for *E. coli* carbon metabolism allowed us to isolate issues that are critical for the appropriate exploitation of the data for parameter estimation, and that may need to be taken into account during the design of the experiments. One such issue is that a high percentage of missing data for some of the individual variables, even at a relatively low average percentage over the entire dataset, was found to be detrimental to the identification results. This may influence sampling strategies, expecially for metabolites that are difficult to measure. Another issue is the identifiability problems caused by steady-state measurements, that cannot always be resolved by genetic mutation or by varying physiological conditions. From this perspective time-resolved observations of the network dynamics carry great promise [Hardiman et al., 2007], although much more demanding experimentally.

Appendices

A Identifiability analysis

First of all, we note that in model (7), for each reaction i, only a subset of metabolites is involved. That is, only a subset of the entries of vector $c_{.i}$ need to be identified, while the values of the remaining entries are fixed to 0. Let us call n_{c_i} with $0 < n_{c_i} < (n+p)$ the effective number of parameters to identify. A straightforward reformulation of the regression model is the following:

$$w_{.i} = Y^{'} \cdot c' + \varepsilon_{.i} \tag{16}$$

with $c' \in \mathbb{R}^{n_{c_i}}$ a vector collecting the nonzero values of $c_{.i}$ and $Y^{'} \in \mathbb{R}^{q \times n_{c_i}}$ a matrix composed of the corresponding columns of Y,*i.e.,* the metabolites involved in reaction i. Bearing this in mind, for simplicity, we will drop index i in the sequel writing n_c in place of n_{c_i} sticking to the usual notation $w = Y \cdot c + \varepsilon$.

To deal with identifiability issues, we use Principal Component Analysis (PCA) on the model (16) [Jolliffe, 1986, Nikerel et al., 2006]. To detect non-identifiable parameters, we decompose the data matrix Y using Singular Value Decomposition (SVD):

$$Y = U \cdot diag(s_1, s_2, ..., s_{n_c}) \cdot V^T \tag{17}$$

with $U \in \mathbb{R}^{q \times q}$ and $V \in \mathbb{R}^{n_c \times n_c}$ orthonormal matrices and $s_1 \geq \cdots \geq s_{n_c} \geq 0$ the singular values of Y. Note that the sum of squared singular values is equal to the square of the Frobenius norm of Y, $\|Y\|^2 = \sum_k \sum_j Y_{j,k}^2$, that is, in an equivalent statistical interpretation, to the sum of the variances of all metabolite concentrations over q experiments (recall that each column of Y has zero mean by definition).

In presence of dependencies among data, there exists an index r with $1 \leq r < n_c$ such that $s_{r+1} = \cdots = s_{n_c} = 0$. Then Y is of rank r. As a consequence, for any two vectors w and c such that $w = Y \cdot c$, there exists an $(n_c - r)$-dimensional vector space $K_Y \subseteq \mathbb{R}^{n-r}$ (the kernel of Y) such that $w = Y \cdot (c + k_Y)$ also holds for any $k_Y \in K_Y$. For the purpose of identification, this implies that c cannot be uniquely reconstructed from the data. (The case where $(s_{r+1}, \cdots, s_{n_c})$ are only approximately 0 will be discussed later in this section.)

We rely on the observation that $K_Y = \text{range}(V_{r+1:n_c})$, the vector space generated by the last $n_c - r$ columns of V. In order to formulate a regression problem with a well-defined solution, we rewrite model (16) in terms of a reduced data matrix $Y^* \in \mathbb{R}^{q \times r}$ and a reduced parameter vector $c^* \in \mathbb{R}^r$ as follows:

$$\begin{cases} w = Y^* \cdot c^* + \varepsilon \\ Y^* = Y \cdot V_{1:r} \end{cases} \tag{18}$$

where $V_{1:r} \in \mathbb{R}^{n_c \times r}$ is the matrix obtained by extracting the first r columns of V. Since Y^* is full-column rank, $Y^* \cdot c^*$ is a linear combination in c^* of independent data vectors. This ensures that the solution to the regression (18) is unique, hence we call c^* 'identifiable'.

Given a unique solution c^* to the regression (18), the space of undistinguishable solutions for the original parameter vector c in (16) can then be defined as follows:

$$c \in \{V_{1:r} \cdot c^* + k_Y \ , \ k_Y \in K_Y\} \tag{19}$$

Depending on the structure of the orthonormal matrix V, we may be able to isolate some entries of c that can be uniquely determined from the reduced model, that is, from the estimates of c^*. This happens when all elements of at least one row of $V_{r+1:n_c}$ are equal to 0. Indeed, this is the criterion we used in Sec. 5 to isolate identifiable parameters in nonidentifiable reactions, such as reactions 17 and 19 in Table 1.

In practice, singular values are rarely exactly 0, even in presence of data dependencies. This can be due to several causes, including measurement noise, numerical roundoffs, etc. Still, for some $r < n_c$, values of $(s_{r+1}, \cdots, s_{n_c})$ sufficiently close to 0 can make the estimates of c solving regression poorly determined. Thus a criterion to discard those singular values needs to be defined.

In this paper, we approximate by zero all singular values whose total contribution to the variance of Y, *i.e.,* to the 'informativity' of the metabolite data, is under some suitable threshold λ, that is, we define

$$r = \min \left\{ t \in [1..(n_c - 1)] \ , \ \frac{\sum_{k=1}^{t} s_k^2}{\sum_{k=1}^{n_c} s_k^2} \geq \lambda \right\} \tag{20}$$

and set $s_{r+1} = \cdots = s_{n_c} = 0$. By this approximation, from (17) we obtain a data matrix Y of rank $r < n_c$. The PCA method described before then applies. The results discussed in Sec. 4 were obtained with $\lambda = 0.99$. In the application to the biological model of Sec. 5 , the choice of λ was driven by the level of measurement noise corrupting the data matrix Y. This was quantified based on the variance of the data over the available replicas in a fixed experimental condition [Ishii et al., 2007]. Depending on the reaction, the value of λ ranged from 0.94 to 0.999.

B Likelihood-based identification of linlog models

We rely on the notation of Sec. 3, *i.e.,* we focus on a single reaction and drop index i from the notation. The loglikelihood of the model is:

$$\log \mathscr{L}(c) = \log \int f_{W|\breve{y},\tilde{y},c}(w) f_{\widetilde{Y}|\breve{y},c}(\tilde{y}) d\tilde{y} \tag{21}$$

For convenience, we rewrite (21) in terms of the random variable $Z = \widetilde{Y} \cdot c$ introduced in Sec. 3 so that it becomes:

$$\log \mathscr{L}(c) = \log \int f_{W|\breve{y},z,c}(w) f_{Z|\breve{y},c}(z) dz. \tag{22}$$

Here $f_{W|\breve{y},z,c}(\cdot)$ is the Gaussian likelihood function of model (7), equivalently rewritten as $W = \breve{Y} \cdot c + Z + \varepsilon$, given $\breve{Y} = \breve{y}$ and $Z = z$, with z varying over all possible values of Z, and $f_{Z|\breve{y},c}$ is the Gaussian prior of $Z = \widetilde{Y} \cdot c$ following from (10). The expressions of $f_{W|\breve{y},z,c}$ and $f_{Z|\breve{y},c}$ are thus

$$\begin{cases} f_{W|\breve{y},z,c}(w) = \frac{1}{\sqrt{\det(2\pi\Sigma_\varepsilon)}} \exp(-\frac{1}{2}[w - \breve{Y} \cdot c - z]^T \Sigma_\varepsilon^{-1} [w - \breve{Y} \cdot c - z]), \\ f_{Z|\breve{y},c}(z) = \frac{1}{\sqrt{\det(2\pi\Sigma_{\breve{y},c})}} \exp(-\frac{1}{2}[z - \mu_{\breve{y},c}]^T \Sigma_{\breve{y},c}^{-1} [z - \mu_{\breve{y},c}]), \end{cases} \tag{23}$$

with $\mu_{\breve{y},c} = M \cdot c$, where the entry $M_{j,k}$ of matrix M is the mean $\mu_{j,k}$ of the distribution of $\widetilde{Y}_{j,k}$, and $\Sigma_{\breve{y},c}$ is the variance matrix of the random variable Z. By the independence assumptions on $\widetilde{Y}$, it turns out that

$$\Sigma_{\breve{y},c} = diag\left(\sum_{k=1}^{n_c} c_k^2 \cdot [\sigma_{1,k}^2 \cdots \sigma_{q,k}^2]^T\right). \tag{24}$$

where $\sigma_{j,k}$ is defined in (10).

Assume for the moment that $\Sigma_{\breve{y},c}$ is invertible. Defining

$$f_{p_c}(z) = f_{W|\breve{y},z,c}(w) \cdot f_{Z|\breve{y},c}(z), \tag{25}$$

after simple but tedious calculations, we obtain

$$f_{p_c}(z) = \kappa_{f_c} \cdot f_c(z) \tag{26}$$

with f_c the density function of a Gaussian distribution $\mathscr{N}(\mu_{f_c}, \Sigma_{f_c})$ and

$$\begin{cases} \Sigma_{f_c} = [\Sigma_\varepsilon^{-1} + \Sigma_{\breve{y},c}^{-1}]^{-1} \\ \mu_{f_c} = \Sigma_{f_c} \cdot [\Sigma_\varepsilon^{-1} \cdot (w - \breve{Y} \cdot c) + \Sigma_{\breve{y},c}^{-1} \cdot \mu_{\breve{y},c}] \\ \kappa_{f_c} = \frac{\exp(-\frac{1}{2}[w - \breve{Y} \cdot c - \mu_{\breve{y},c}]^T \cdot [\Sigma_\varepsilon + \Sigma_{\breve{y},c}]^{-1} \cdot [w - \breve{Y} \cdot c - \mu_{\breve{y},c}])}{\sqrt{\det(2\pi[\Sigma_\varepsilon + \Sigma_{\breve{y},c}])}}. \end{cases} \tag{27}$$

The proportionality factor κ_{f_c} does not depend on the integration variable z, so it can be taken out of the integral and (22) can be rewritten as follows:

$$\log \mathscr{L}(c) = \log(\kappa_{f_c}) + \log\left(\int f_c(z) dz\right). \tag{28}$$

The integral of a normalized Gaussian density function being 1, we finally have an analytical expression for the loglikelihood: $\log \mathscr{L}(c) = \log(\kappa_{f_c})$.

The above results are used in the expectation step of the EM algorithm. Recall the definition

$$Q(c|\hat{c}^{\ell-1}) = \int \log(f_{Z,W|\breve{y},c}(z,w)) f_{Z|\breve{y},\hat{c}^{\ell-1},w}(z)dz. \tag{29}$$

The Bayes theorem allows us to rewrite (29) as follows:

$$Q(c|\hat{c}^{\ell-1}) = \int \log(f_{W|\breve{y},z,c}(w) f_{Z|\breve{y},c}(z)) \frac{f_{W|\breve{y},z,\hat{c}^{\ell-1}}(w) f_{Z|\breve{y},\hat{c}^{\ell-1}}(z)}{f_{W|\breve{y},\hat{c}^{\ell-1}}(w)} dz. \tag{30}$$

Function $f_{W|\breve{y},\hat{c}^{\ell-1}}(w)$ does not depend on z so it can be taken out of the integral. Moreover, this function does not depend on c so it will have no impact on the maximization step of EM. Thus, we can ignore this function from the computation of the expectation function above.

Using definitions (25) and (26), we can rewrite (30) in the following way:

$$\begin{aligned} Q(c|\hat{c}^{\ell-1}) &\propto \int \kappa_{f_{\hat{c}^{\ell-1}}} f_{\hat{c}^{\ell-1}}(z) \log(\kappa_{f_c} f_c(z)) dz \\ &\propto \int f_{\hat{c}^{\ell-1}}(z) \log(\kappa_{f_c} f_c(z)) dz. \end{aligned} \tag{31}$$

We have dropped the constant factor $\kappa_{f_{\hat{c}^{\ell-1}}}$ as it does not depend on c and thus does not influence the maximization step of EM. By replacing $\log(\kappa_{f_c} f_c(z))$ by $\log(\kappa_{f_c} f_c(z) f_{\hat{c}^{\ell-1}}(z)/f_{\hat{c}^{\ell-1}}(z))$ and separating the integrand in a sum of terms, we can rewrite (31) as

$$-Q(c|\hat{c}^{\ell-1}) \propto \int f_{\hat{c}^{\ell-1}}(z) \log\left(\frac{f_{\hat{c}^{\ell-1}}(z)}{f_c(z)}\right) dz - \int f_{\hat{c}^{\ell-1}}(z) \log(f_{\hat{c}^{\ell-1}}(z)) dz - \log(\kappa_{f_c}). \tag{32}$$

We recognize in the first term the definition of the Kullback-Leibler divergence $KL(f_c||f_{\hat{c}^{\ell-1}})$ between the two probability distributions f_c and $f_{\hat{c}^{\ell-1}}$ and in the second term the entropy $H(f_{\hat{c}^{\ell-1}})$ of $f_{\hat{c}^{\ell-1}}$ [Cover and Thomas, 2006, Stoorvogel and van Schuppen, 1996]. For Gaussian distributions, these can be written explicitely as

$$\begin{aligned} KL(f_c||f_{\hat{c}^{\ell-1}}) = \frac{1}{2}(\log\left(\frac{\det(\Sigma_{f_c})}{\det(\Sigma_{f_{\hat{c}^{\ell-1}}})}\right) + Tr(\Sigma_{f_c}^{-1}\Sigma_{f_{\hat{c}^{\ell-1}}}) \\ + [\mu_{f_c} - \mu_{f_{\hat{c}^{\ell-1}}}]^T \Sigma_{f_c}^{-1} [\mu_{f_c} - \mu_{f_{\hat{c}^{\ell-1}}}]), \end{aligned} \tag{33}$$

where $Tr(\ldots)$ stands for trace and

$$H(f_{\hat{c}^{\ell-1}}) = \log\left(\sqrt{\det(2\pi e \Sigma_{f_{\hat{c}^{\ell-1}}})}\right). \tag{34}$$

To summarize, together with (27), this gives us the explicit formula

$$Q(c|\hat{c}^{\ell-1}) \propto -KL(f_c||f_{\hat{c}^{\ell-1}}) - H(f_{\hat{c}^{\ell-1}}) + \log(\kappa_{f_c}), \tag{35}$$

which we employ in our implementation of EM.

In more generality, for some values of c, $\Sigma_{\breve{y},c}$ may be singular or poorly conditioned. To avoid this circumstance, we can adapt our procedure as follows. We consider a decomposition

$$W = \breve{Y} \cdot c + Z + (\varepsilon' + \varepsilon'') = \breve{Y} \cdot c + (Z + \varepsilon') + \varepsilon'' \tag{36}$$

where ε' and ε'' are independent zero-mean Gaussian random vectors such that $\Sigma_{\varepsilon'} \triangleq Var(\varepsilon') = \alpha\Sigma_\varepsilon$ and $\Sigma_{\varepsilon''} \triangleq Var(\varepsilon'') = (1-\alpha)\Sigma_\varepsilon$, with $\alpha \in (0,1)$ a tunable parameter. Since $\Sigma_\varepsilon > 0$ by assumption, it follows that $\Sigma_{\varepsilon'} > 0$ and $\Sigma_{\varepsilon''} > 0$. Moreover, $\Sigma_\varepsilon = \Sigma_{\varepsilon'} + \Sigma_{\varepsilon''}$, *i.e.,* the statistics of ε and of $\varepsilon' + \varepsilon''$ are identical. Since $Var(Z + \varepsilon') = \Sigma_{\breve{y},c} + \Sigma_{\varepsilon'} > 0$, if we interpret $Z + \varepsilon'$ as the unknown observations (in place of Z) and ε'' as the model noise (in place of ε), we ensure that the variance of the 'missing data' is invertible. Thus, in practice, we apply all formulas developed above with $\Sigma_{\breve{y},c} + \Sigma_{\varepsilon'}$ in place of $\Sigma_{\breve{y},c}$ and $\Sigma_{\varepsilon''}$ in place of Σ_ε.

The effect of the specific choice of α is under investigation. In this work, we took $\alpha = 0.2$, a value that leads to good results in practice.

C Validation on synthetic data

The model used for comparing peformance of the identification algorithms is a reduced synthetic linlog model of the *E. coli* central carbon metabolism network (Fig. 4). This network contains 17 variables, describing internal and external metabolites, and 25 reactions, summarized in Table 2 and Table 3, respectively. The linlog model has the form of Eq. (1)-(2).

Index	Name	Symbol
1	Pyruvate	PYR
2	Phosphoenol-pyruvate	PEP
3	Glyceraldehyde-3-phosphate	G3P
4	Fructose-6-phosphate	F6P
5	Glucose-6-phosphate	G6P
6	3-phosphoglycerate	3PG
7	Dihydroxyacetonephosphate	DHAP
8	Ribulose-5-phosphate	Ru5P
9	Ribose-5-phosphate	R5P
10	6-phosphogluconate	6PG
11	Erythrose-4-phosphate	E4P
12	Xylulose-5-phosphate	X5P
13	2-phosphoglycerate	2PG
14	1,3-diphosphophoglycerate	1,3DP
15	Fructose-1,6-bisphosphate	F1,6P
16	2-keto-3-deoxy-6-phosphogluconate	KDPG
17	Sedoheptulose-7-phosphate	S7P

Table 2: Metabolites included in the synthetic linlog model.

Index	Name
1	Phosphotransferase system
2	Glucose-6-phosphate isomerase
3	Glucose-6-phosphate dehydrogenase
4	Phosphofructokinase
5	Transaldolase
6	Transketolase a
7	Transketolase b
8	Aldolase
9	Glyceraldehyde-3-phosphate dehydrogenase
10	Triosephosphate isomerase
11	Glycerol-3-phosphate dehydrogenase
12	Phosphoglycerate kinase
13	Serine synthesis
14	Phosphoglycerate mutase
15	Enolase
16	Pyruvate kinase
17	PEP carboxylase
18	Pyruvate synthesis
19	6-Phosphogluconate dehydrogenase
20	Ribose-phosphate isomerase
21	Ribulose-phosphate epimerase
22	Ribose-phosphate pyrophosphokinase
23	Phosphogluconate dehydratase
24	KDPG aldolase
25	Fructose bisphosphatase

Table 3: Reactions included in the synthetic linlog model.

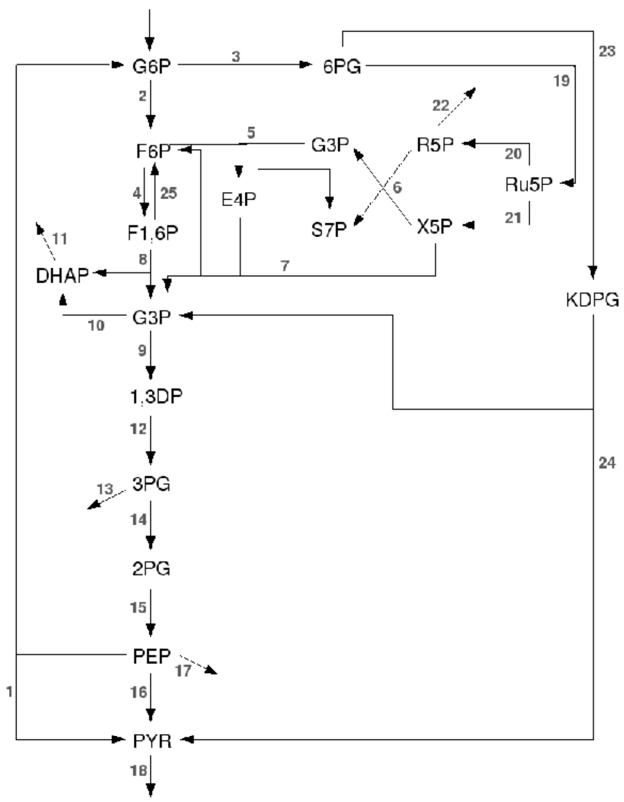

Figure 4: Network for the synthetic model, a reduced version of the *E. coli* central carbon metabolism network.

A dataset was generated from the synthetic linlog model by setting all enzyme concentrations to 1 and choosing plausible values for the parameter vector a and matrices B^x, B^u, that is, values consistent with existing kinetic models of carbon metabolism in *E. coli* [Bettenbrock et al., 2005]. Then $q = 30$ different experimental conditions were simulated by randomly changing enzyme concentrations. For each condition $j \in [1..q]$, vectors $\ln(x^{(j)})$, $\ln(u^{(j)})$ and $v^{(j)}$ were determined by the equations resulting from the formulation of the linlog model and the (quasi-)steady-state equation $N \cdot v = 0$:

$$\left\{ \begin{array}{l} \begin{bmatrix} \ln(x^{(j)}) \\ \ln(u^{(j)}) \end{bmatrix} = -\left[N \cdot \mathrm{diag}(e^{(j)}) \cdot [B^x \, B^u]\right]^{-1} N \cdot \mathrm{diag}(e^{(j)}) \cdot a \\ v^{(j)} = \mathrm{diag}(e^{(j)}) \cdot \left(a + B^x \cdot \ln(x^{(j)}) + B^u \cdot \ln(u^{(j)})\right) \end{array} \right. \tag{37}$$

For this dataset, four scenarios were considered, corresponding to more or less favorable conditions for identification: 40 % and 75 % missing entries and 10% and 20% noise. For each column of Y, *i.e.*, each metabolite of the model, the 40% or 75% missing data were distributed randomly over the q measurements.

Randomly generated noise was added to the same incomplete dataset in each of 100 Monte Carlo repetitions.

Identifiability analysis was performed following the approach described in Appendix A, with $\lambda = 0.99$. 10 reactions were found to be nonidentifiable (reactions 2, 5, 6, 7, 8, 12, 14, 15, 20 and 21). Among these reactions only 3 identifiable parameters could be isolated (one in reaction 2, one in reaction 7 and one in reaction 12).

Results from all identification methods on identifiable reactions are summarized in Fig. 5 for the most favorable scenario with 40% missing data and 10% noise, and in Fig. 6 for the least favorable scenario with 75% missing data and 20% error. The results for the other scenarios fall between those shown in Fig. 5 and Fig. 6, and are not shown here.

D Application to central metabolism in *E. coli*

Fig. 3 shows a (simplified) representation of central carbon metabolism in *E. coli*. The network could not be directly transformed into a linlog model of the form (1)-(2), since metabolites G3P, E4P, X5P, 2KDPG, OAA, IsoCit, SuccoA, Acp and Glyox were not measured by Ishii et al. [2007]. This prevents the estimation of elasticities for the above metabolites and their inclusion in the model. We overcome this limitation by lumping reactions not measured by Ishii et al. [2007].

In addition to the above model simplification imposed by the available dataset, we added a phenomenological reaction μ to model biomass production. The reaction involves 11 metabolites, the reaction flux is equal to the dilution rate under the experimental conditions of Ishii et al. [2007] and the enzyme concentration is set to 1.

The linlog model thus obtained contains 16 internal metabolites and 7 external metabolites or cofactors, listed in Table 4, as well as 31 reactions, listed in Table 5.

Internal metabolites			
Index	Symbol	Index	Symbol
1	PEP	9	Ru5P
2	G6P	10	R5P
3	Pyr	11	S7P
4	F6P	12	2KG
5	FBP	13	Suc
6	DHAP	14	Fum
7	3PG	15	Mal
8	6PG	16	Cit

Index	External metabolites or cofactors
17	Glc
18	AcoA/coA
19	ATP/ADP
20	NADPH/NADP
21	NADH/NAD
22	FAD
23	Ace

Table 4: Internal and external metabolites and cofactors of the linlog model of carbon metabolism in *E. coli*. Some of the cofactors are modeled as ratios of metabolite concentrations, *e.g.*, ATP/ADP.

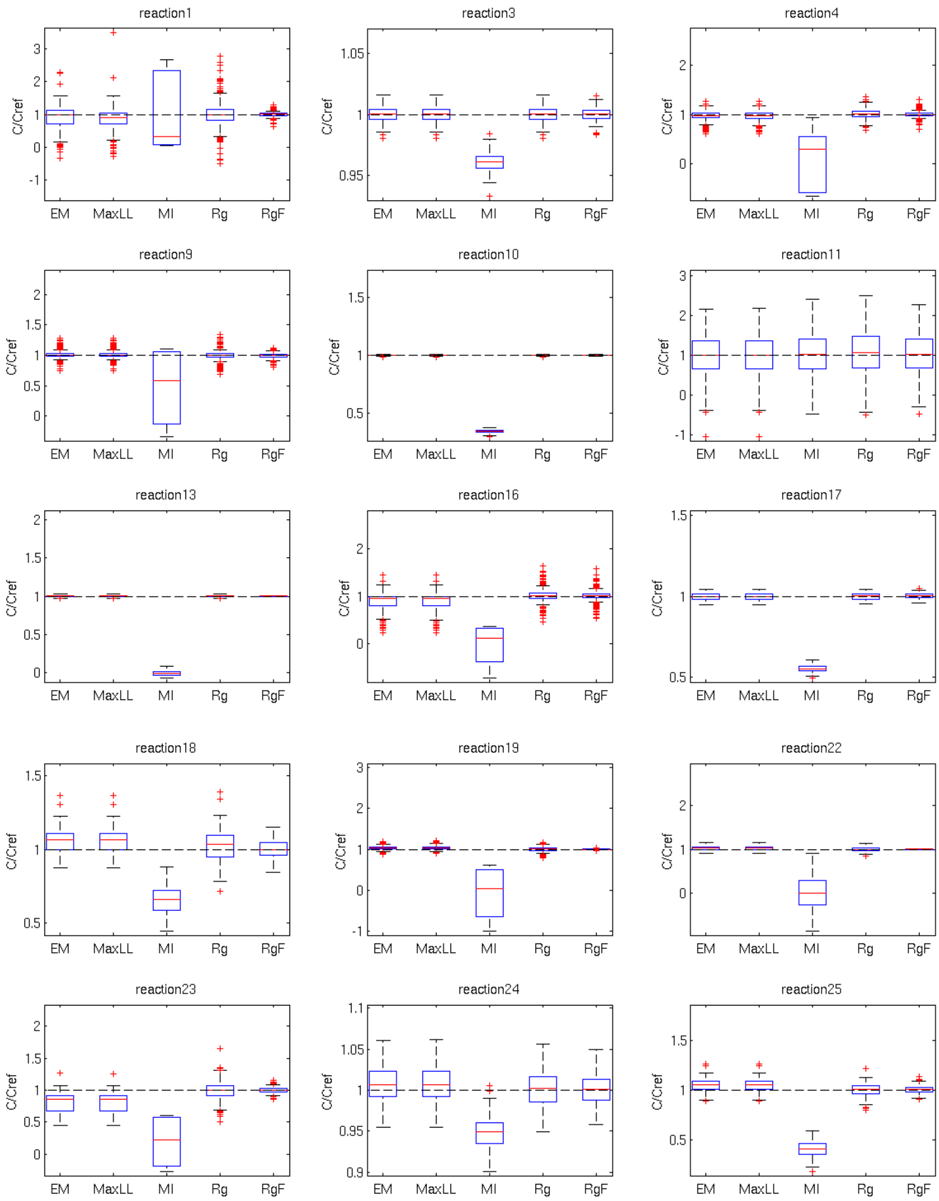

Figure 5: Statistics of estimated parameter values in identifiable reactions for datasets with 40% of missing data and 10% noise. The graphical notations are the same as for Fig. 1.

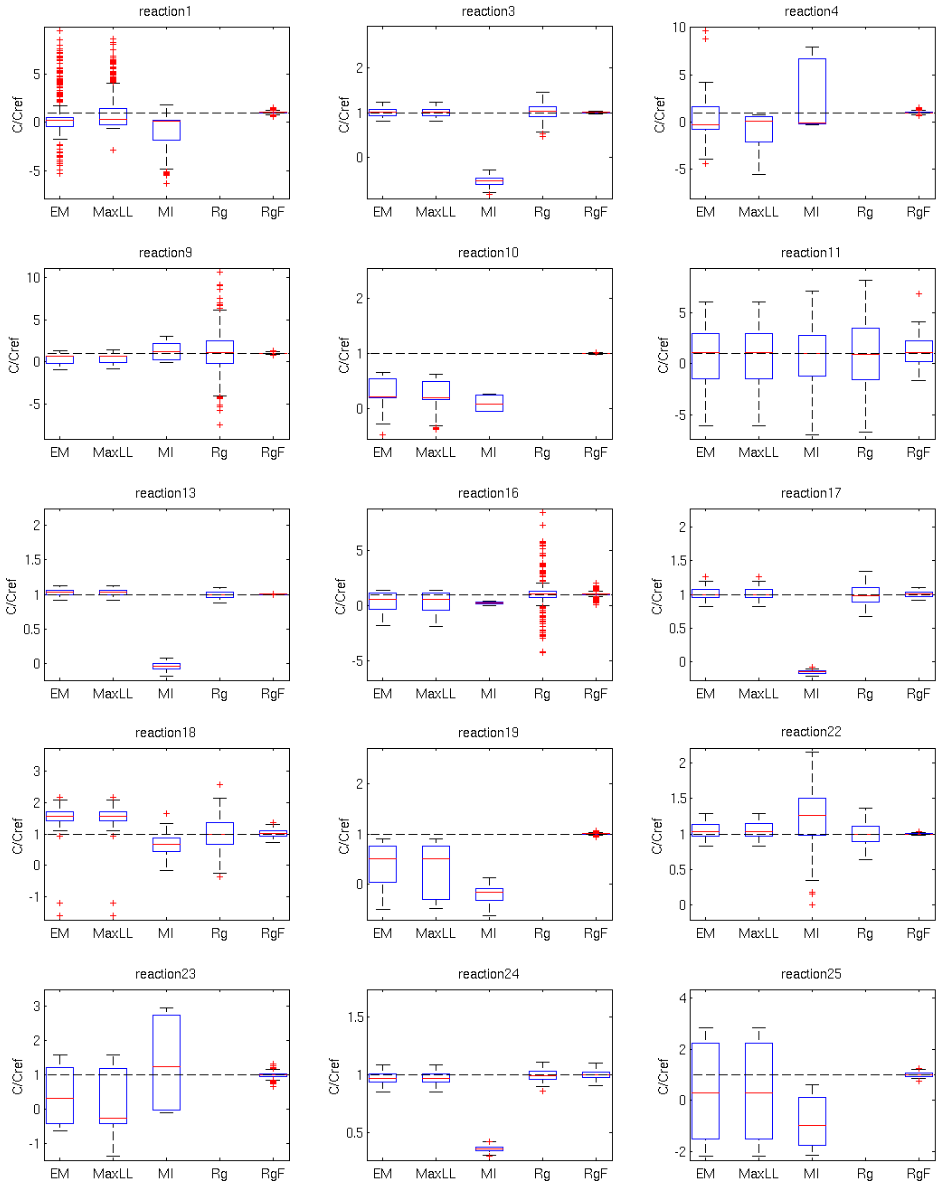

Figure 6: Statistics of estimated parameter values in identifiable reactions for datasets with 75% of missing data and 20% noise. The graphical notations are the same as for Fig. 1.

Index	Reaction
1	Glc + PEP $\xleftrightarrow{ptsG}$ Pyr+G6P
2	G6P $\xleftrightarrow{pgi}$ F6P
3	F6P + ATP/ADP $\xleftrightarrow{pfkA,pfkB}$ FBP [PEP]$_{in}$
4	FBP $\xleftrightarrow{fbaA,fbaB}$ DHAP
5	DHAP $\xleftrightarrow{tpiA}$ 3PG
6	FBP + ATP/ADP $\xleftrightarrow{gapA;pgk}$ 3PG + NADH/NAD
7	3PG $\xleftrightarrow{gpmA,gpmB;eno}$ PEP
8	PEP + ATP/ADP $\xleftrightarrow{pykA,pykF}$ Pyr [FBP]$_{act}$
9	Pyr $\xleftrightarrow{aceE:aceF:lpdA}$ AcoA/coA + NADH/NAD
10	G6P $\xleftrightarrow{zwf;pgl}$ 6PG + NADPH/NADP
11	6PG $\xleftrightarrow{gnd}$ Ru5P + NADPH/NADP
12	Ru5P $\xleftrightarrow{rpe}$ S7P
13	Ru5P $\xleftrightarrow{rpiA,rpiB}$ R5P [G6P]$_{in}$
14	R5P $\xleftrightarrow{tktA}$ S7P
15	S7P $\xleftrightarrow{talA,talB}$ F6P
16	Ru5P $\xleftrightarrow{tktB}$ F6P
17	AcoA/coA $\xleftrightarrow{gltA,prpC}$ Cit [2KG]$_{in}$ [NADH/NAD]$_{act}$
18	Cit $\xleftrightarrow{acnA,acnB}$ 2KG
19	AcoA/coA $\xleftrightarrow{icdA}$ 2KG + NADPH/NADP
20	2KG $\xleftrightarrow{sucA:sucB:lpdA;sucC:sucD}$ Suc + NADH/NAD
21	Suc + FAD $\xleftrightarrow{sdhA:sdhB:sdhC:sdhD}$ Fum
22	Fum $\xleftrightarrow{fumA,fumB,fumC}$ Mal
23	Mal + PEP $\xleftrightarrow{mdh}$ Cit +NADH/NAD
24	PEP $\xleftrightarrow{ppc;pckA}$ Mal + Cit + ATP/ADP [FBP]$_{act}$
25	Mal $\xleftrightarrow{maeB,sfcA}$ Pyr + NADPH/NADP [AcoA/coA]$_{in}$ [NADH/NAD]$_{act}$
26	AcoA/coA $\xleftrightarrow{aceA;aceB}$ Suc + Mal
27	PEP+G6P+Pyr+F6P+3PG+AcoA/coA+R5P+2KG+ATP/ADP $\xrightarrow{\mu}$NADPH/NADP+NADH/NAD
28	6PG $\xleftrightarrow{edd;eda}$ Pyr
29	AcoA/coA $\xleftrightarrow{pta;ackA,ackB}$ Ace+ATP/ADP [Pyr]$_{act}$ [NADPH/NADP]$_{in}$ [NADH/NAD]$_{in}$
30	Pyr + NADH/NAD $\xrightarrow{ldhA}$
31	AcoA/coA $\xrightarrow{adhE}$

Table 5: Reactions of the linlog model of carbon metabolism in *E. coli.* Activators and inhibitors of the reaction are shown with $[\cdot]_{act}$ and $[\cdot]_{in}$, respectively. Reaction 27, labeled μ, is a phenomenological reaction for biomass production. The enzyme names are separated by a comma in the case of isoenzymes, by a colon for enzyme complexes, and by a semicolon when the enzymes catalyze reactions that have been lumped together in the model. Reactions 20, 26, 28 and 29 result from the merging of reactions due to the absence of measurements of SuccoA, Glyox, 2KDPG and Acp, respectively, in the dataset of Ishii et al. [2007].

Contents

References

M. Ashyraliyev, Y. Fomekong Nanfack, J.A. Kaandorp, and J.G. Blom. Systems biology: Parameter estimation for biochemical models. *FEBS J.*, 276(4):886–902, 2009.

K. Bettenbrock, S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E.D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *J. Biol. Chem.*, 281(5):2578–84, 2005.

M. Brand. Incremental singular value decomposition of uncertain data with missing values. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. 7th Eur. Conf. Comput. Vision (ECCV 2002)*, volume 2350 of *Lecture Notes in Computer Science*, pages 707–20. Springer Verlag, 2002.

R.S. Costa, D. Machado, I. Rocha, and E.C. Ferreira. Hybrid dynamic modeling of *Escherichia coli* central metabolic network combining Michaelis-Menten and approximate kinetic equations. *Biosystems*, 100(2):150–8, 2010.

T.M. Cover and J.A. Thomas. *Elements of Information Theory (2nd edition)*. Wiley, 2006.

E.J. Crampin. System identification challenges from systems biology. In *Proc. 14th IFAC Symp. Syst. Identif. (SYSID 2006)*, pages 81–93, Newcastle, Australia, 2006.

R.C.H. del Rosario, E. Mendoza, and E.O. Voit. Challenges in lin-log modelling of glycolysis in *Lactococcus lactis*. *IET Syst. Biol.*, 2(3):136–49, 2008.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B*, 39(1):1–38, 1977.

J.W. Graham. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.*, 60:549–76, 2009.

F. Hadlich, S. Noack, and W. Wiechert. Translating biochemical network models between different kinetic formats. *Metab. Eng.*, 11(2):87–100, 2009.

T. Hardiman, K. Lemuth, M.A. Kellerand, M. Reuss, and M. Siemann-Herzberg. Topology of the global regulatory network of carbon limitation in *Escherichia coli*. *J. Biotechnol.*, 132(4):359–74, 2007.

V. Hatzimanikatis and J.E. Bailey. Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models. *Biotechnol. Bioeng.*, 54(2):91–104, 1997.

J.J. Heijnen. Approximative kinetic formats used in metabolic network modeling. *Biotechnol. Bioeng.*, 91(5):534–45, 2005.

R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman & Hall, 1996.

N.J. Horton and K.P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.*, 61(1):79–90, 2007.

N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316(5824):593–7, 2007.

I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

O. Kotte, J.B. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Biol.*, 6:355, 2010.

W. Liebermeister and E. Klipp. Bringing metabolic networks to life: Convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Model.*, 3:41, 2006.

R.J.A Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.

B.F.J. Manly. *Randomization, Bootstrap and Monte-Carlo Methods in Biology*. Chapman and Hall, 1997.

L. Marucci, S. Santini, M. di Bernardo, and D. di Bernardo. Derivation, identification and validation of a computational model of a novel synthetic regulatory network in yeast. *J. Math. Biol.*, 2011. In press.

I. Nikerel, W. van Winden, W. van Gulik, and J. Heijnen. A method for estimation of elasticities in metabolic networks using steady state and dynamic metabolomics data and linlog kinetics. *BMC Bioinform.*, 7:540, 2006.

S. Oba, M.A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–96, 2003.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–29, 2009.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–90, 1976.

D.B. Rubin. Multiple imputation after 18+ years. *J. Am. Stat. A.*, 81:473–89, 1996.

M.A. Savageau. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology.* Addison-Wesley, 1976.

M. Scholz, F. Kaplan, C.L. Guy, J. Kopka, and J. Selbig. Non-linear PCA: A missing data approach. *Bioinformatics*, 21(20):3887–95, 2005.

K. Smallbone, E. Simeonidis, N. Swainston, and P. Mendes. Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.*, 4:6, 2010.

A. Stoorvogel and J.H. van Schuppen. System identification with information theoretic criteria. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, volume 153 of *NATO ASI*, pages 289–338. Springer Verlag, 1996.

D. Visser and J.J. Heijnen. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab. Eng.*, 5(3):164–76, 2003.

D. Visser, J.W. Schmid, K. Mauch, M. Reuss, and J.J. Heijnen. Optimal re-design of primary metabolism in *Escherichia coli* using linlog kinetics. *Metab. Eng.*, 6(4):378–90, 2004.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399