Identification of metabolic network models from incomplete high-throughput datasets
Sara Berthoumieux, Matteo Brilli, Hidde De Jong, Daniel Kahn, Eugenio Cinquemani

To cite this version:
Sara Berthoumieux, Matteo Brilli, Hidde De Jong, Daniel Kahn, Eugenio Cinquemani. Identification of metabolic network models from incomplete high-throughput datasets. Bioinformatics, 2011, 27 (13), pp.i186-i195. 10.1093/bioinformatics/btr225. hal-00793039

HAL Id: hal-00793039
https://inria.hal.science/hal-00793039
Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License
Identification of metabolic network models from incomplete high-throughput datasets

Sara Berthoumieux1,*, Matteo Brilli1,2, Hidde de Jong1, Daniel Kahn2 and Eugenio Cinquemani1,*

1INRIA Grenoble - Rhône-Alpes, Montbonnot and 2Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR 5558, Université Lyon 1, INRA, Villeurbanne, France

ABSTRACT

Motivation: 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.

Results: 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.

Contact: sara.berthoumieux@inria.fr; eugenio.cinquemani@inria.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

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 systems biology, obtaining reliable estimates of parameter values in kinetic models (Ashyraliyev et al., 2009; Crampin, 2006).

*To whom correspondence should be addressed.

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, 2000).

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, with the help of a recent genome-scale linlog model of yeast metabolism, parametrized using previously-published kinetic models, it has been possible 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 (Nikerel et al., 2006). Power-law models, up to a logarithmic transformation, and loglin models also have this convenient property. 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 article a maximum-likelihood method for the identification of linlog models of metabolism from incomplete datasets. The specific contributions of the paper are 2-fold. 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 approximate kinetic models whose identification can be formulated as a regression problem. 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 (Oha 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 Escherichia coli, consisting of some 23 variables. We estimate the 100 parameters of this model from a high-throughput dataset published in the literature (Ishi et al., 2007). The data consists of measurements of metabolic fluxes and metabolite and enzyme concentrations in glucose-limited chemostat under 29 different conditions such as 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 (Heijnen models and comparison with other approximative rate expressions of its elements. For conciseness, in the sequel we shall drop the notation. An in-depth discussion of linlog models and comparison with other approximate rate functions can be found in the review by Heijnen (2005).

We are interested in the estimation of the (generally unknown) parameters $a \in \mathbb{R}^{m}$, $b \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^{m \times q}$ from $q$ experimental datapoints $(v^{(k)}X^{(k)}Y^{(k)})$, $k=1, ..., q$. That is, the data used for parameter estimation are parallel measurements of enzyme and metabolite concentrations as well as metabolic fluxes. The datapoints $(v^{(k)}X^{(k)}Y^{(k)})$ are obtained under different experimental conditions, for instance different dilution rates in continuous cultures or different mutant strains. Notice that in practice reaction rates are most of the time measured at (quasi-)steady state (see also Section 6). That is, on the time-scale of interest the derivatives of metabolite concentrations vanish and Equation (1) can be rewritten as $N \cdot v = 0$.

For the purpose of parameter estimation, it is convenient to rewrite (2) in the form of a regression model:

$$\begin{align*}
\begin{bmatrix}
\alpha\vphantom{\|}
\end{bmatrix}^T &= \begin{bmatrix} [\ln(x)]^T & [\ln(u)]^T & \vphantom{\|} \end{bmatrix} \begin{bmatrix} a^T & (b^T)^T \end{bmatrix}^T \\
\gamma &\phantom{=}
\end{align*}
$$

(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: $\overline{v_k} = (1/q)\sum_{i=1}^q v_k^{(i)}$. By the linearity of (3), it holds that

$$\begin{align*}
\begin{bmatrix}
\alpha\vphantom{\|}
\end{bmatrix}^T &= \begin{bmatrix} [\ln(x)]^T & [\ln(u)]^T \end{bmatrix} \begin{bmatrix} a^T & (b^T)^T \end{bmatrix}^T
\end{align*}
$$

(4)

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

$$\begin{align*}
\begin{bmatrix}
\overline{v} - \overline{e}\vphantom{\|}
\end{bmatrix}^T &= \begin{bmatrix} \ln(x) & \ln(u) \end{bmatrix} \begin{bmatrix} a^T & (b^T)^T \end{bmatrix}^T
\end{align*}
$$

(5)

and we obtain the following parameter estimation problem:

**Problem 1.** Given the data matrices

$$\begin{align*}
\begin{bmatrix}
\overline{v} - \overline{e}\vphantom{\|}
\end{bmatrix}^T &= \begin{bmatrix} \ln(x) & \ln(u) \end{bmatrix} \begin{bmatrix} a^T & (b^T)^T \end{bmatrix}^T
\end{align*}
$$

find parameters $C \triangleq [B^T \ b^T]^T$ solving the regression problem

$$W = Y \cdot C + \varepsilon$$

(6)

where $x \in \mathbb{R}^{m \times n}$ 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 $C = [B^T \ b^T]^T$ by way of Equation (4).

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

$$w_i = Y \cdot c_i + \varepsilon_i$$

(7)

where $w_i$ and $c_i$ are the $i$th columns of $W$ and $C$, respectively.

The values of the parameter matrices $B^T$ and $b^T$ admit an interesting biological interpretation. Notice that one can immediately find values $s_i \in \mathbb{R}$, $s_0 \in \mathbb{R}^+$, $c_i \in \mathbb{R}^+$, and $v_i \in \mathbb{R}^n$ such that $v_i/c_i \equiv r_i$, $\ln(s_0) = \ln(x_0)$, and $\ln(s_0) = \ln(x_0)$. As a consequence, Equation (5) can be rearranged into the common relative formulation of linlog models,

$$\begin{align*}
\begin{bmatrix}
\overline{v} - \overline{e}\vphantom{\|}
\end{bmatrix}^T &= \begin{bmatrix} \ln(x) & \ln(u) \end{bmatrix} \begin{bmatrix} a^T & (b^T)^T \end{bmatrix}^T
\end{align*}
$$

(8)
where \( \mathbf{1}_m \) is an \( m \times 1 \) vector of ones, \((v_0, x_0, \theta_0, \mathbf{e}_0)\) is a so-called reference state (Heijnen, 2005) and \( B_{0}^i, B_{10}^i \) are matrices of elasticity constants, where
\[
B_{0}^i = \text{diag} \left( \frac{v_0}{x_0} \right) B^i, \quad B_{10}^i = \text{diag} \left( \frac{v_0}{x_0} \right) B^i.
\]

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 \( B^i \) and \( B^k \) found by the solution of Problem 1, and the equality \( c_0/v_0 = 1/(\theta/e) \).

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

(i) Since the measurements are carried out at (quasi-)steady state, we have \( N \cdot \mathbf{f} \equiv 0 \). This introduces dependencies among the data and thus reduces the information content of the data matrix \( \mathbf{Y} \), in the sense that \( \mathbf{Y} \) becomes rank deficient. Like in earlier work (Nikerel et al., 2009), we use standard approaches to solve this problem. We notably rely on principal component analysis (PCA) (Jolliffe, 1986; Nikerel et al., 2009) applied to the data matrix \( \mathbf{Y} \) to reduce the model order, i.e. the number of independent parameters, and ensure well-posedness of the regression problem (see Supplementary Section S1 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 \( \mathbf{W} \) and \( \mathbf{Y} \) are zero-mean, an important requirement for the correctness of the outlined analysis.

(ii) 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 article 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 \( \mathbf{c}_i \) of the model given in (7) in the case where some entries of \( \bar{\mathbf{Y}} \) are unknown. We address the estimation problem by a maximum-likelihood approach, which is known to yield optimal (unbiased and minimum variance) estimates for our problem setting in the case where \( \mathbf{Y} \) is fully known. As the problem is identical for all reactions \( i \), in the remainder of the section we will drop for simplicity index \( i \) from the notation.

Let \( \mathcal{F} \) be the set of indices (row, column) corresponding to the known entries of \( \mathbf{Y} \), i.e. \( (j,k) \in \mathcal{F} \) if and only if \( \bar{Y}_{j,k} \) is available. It is convenient to introduce the decomposition \( \bar{\mathbf{Y}} = \bar{\mathbf{Y}} + \bar{\mathbf{F}} \), where
\[
\bar{Y}_{j,k} = \begin{cases} \bar{Y}_{j,k} & \text{if } (j,k) \in \mathcal{F}, \\ 0 & \text{otherwise} \end{cases}, \quad \bar{F}_{j,k} \begin{cases} 0 & \text{if } (j,k) \in \mathcal{F}, \\ \bar{Y}_{j,k} & \text{otherwise}. \end{cases}
\]

Matrix \( \bar{\mathbf{Y}} \) is fully determined: Once measurements \( \bar{Y} \) of \( \bar{\mathbf{Y}} \) are collected, we treat \( \bar{\mathbf{Y}} = \bar{\mathbf{Y}} \) as fixed parameters of the regression problem. Matrix \( \bar{\mathbf{Y}} \) collects the unknown entries of \( \mathbf{Y} \). We model these missing data as unobserved independent random variables, whose prior distributions encode our generic knowledge about them. Assuming that the \textit{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 \( (j,k) \notin \mathcal{F} \) and \( \bar{F}_{j,k} \), we let
\[
\bar{Y}_{j,k} \sim \mathcal{N}(\mu_{j,k}, \sigma_{j,k}^2),
\]
where \( \mu_{j,k} = \text{mean}(\bar{Y}_{j,k}) \), \( \sigma_{j,k} = \text{std}(\bar{Y}_{j,k}) \).

We can now formulate the estimation problem.

PROBLEM 2. Given measurements \( \mathbf{W} = \mathbf{w} \) and \( \bar{\mathbf{Y}} = \bar{\mathbf{Y}} \), compute the estimate \( \hat{\mathbf{c}} = \arg\max_{\mathbf{c}} \log \mathcal{L}(\mathbf{c}) \), with \( \mathcal{L}(\mathbf{c}) = \mathcal{L}(\mathbf{c})(\mathbf{w}) \), where, for any \( c, \bar{f}(\mathbf{w}, \mathbf{c}) \) is the probability density function of \( \mathbf{W} \) given \( \bar{\mathbf{Y}} = \bar{\mathbf{Y}} \) corresponding to model (7)-(10).

Note that \( \mathcal{L}(\mathbf{c}) \) is a likelihood function for a linear model with missing data, in the sense that it is defined with respect to available data \( \mathbf{Y} \) only. One can express \( \mathcal{L}(\mathbf{c}) \) by marginalization,
\[
\log \mathcal{L}(\mathbf{c}) = \log \int \bar{f}(\mathbf{w}, \mathbf{c}) \mathcal{L}(\mathbf{c})(\mathbf{w}) \mathbf{d}\bar{\mathbf{Y}}
\]
where \( \bar{f}(\mathbf{w}, \mathbf{c}) \) is the standard likelihood function for model (7) given \( \bar{\mathbf{Y}} = \bar{\mathbf{Y}} \) and \( \bar{\mathbf{Y}} \), with \( \bar{\mathbf{Y}} \) varying over all possible values of \( \bar{\mathbf{Y}} \), and \( \bar{f}(\mathbf{w}, \mathbf{c}) \) is determined by the prior (10). The explicit solution to the integral is reported in Supplementary Section S2. A direct approach to solving Problem 2 is to maximize (11) by numerical optimization. However, the function is not convex in \( \mathbf{c} \), whence its direct optimization is prone to end up in local minima and the use of global optimization strategies is required.

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 \( Z = \bar{\mathbf{Y}} + \mathbf{c} \), so that model (7) becomes \( \mathbf{W} = \bar{\mathbf{Y}} + Z = \bar{\mathbf{Y}} + c \). Note that \( Z = \bar{\mathbf{Y}} + c \mathbf{B}^i \), where for any given \( c \), mean and variance can be derived from (10). Let \( \mathbf{E}^i \) be an initial guess of \( c \). At every iteration \( t = 1, 2, 3, \ldots \), compute an updated
To complete the implementation of the algorithm, one must express complicacy is needed in case entropy of a distribution, respectively, for which, in the Gaussian function of $Z$ given $L$ maximized $L$ stands for a Gaussian distribution with variance $Q_{\ell} = \arg\max_{c} Q(c|z^{\ell-1})$. Finally, $KLL$ is the $\ell$th component of the model equations. We generate data matrices $F$ 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 Section 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 Section 3 with the direct maximization of the log-likelihood (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 dataset $\tilde{F}(\beta_{k}, \delta_{k}, \chi_{k}, \nu_{k})$ 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 $\tilde{F}$, 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 Section 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, where $F$ stands for Full datasets). 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 (Section 2). 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

\begin{equation}
Q(c|z^{\ell-1}) = \mathbb{E}[\log f_{Z}^{\ell-1}(Z|w, w)|z^{\ell-1}, w] = \int \log f_{Z}^{\ell-1}(z|w, w)f_{Z}^{\ell-1}(w, w)c |c| dc.
\end{equation}
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 was performed in place of the identification of the original model. $Y^* \in \mathbb{R}^{n \times p}$, with $r \leq n \leq p$, 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 Supplementary Section S3).

We implemented the different parameter estimation algorithms in MATLAB, using the lsqnonlin function for the regression-based methods and finsearch 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. Figure 1 is dedicated to the scenario with 40% missing data and 10% noise, whereas Figure 2 reports on 75% 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.

### Analysis

- **Figure 2** 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. Figure 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.

### Discussion

- **Multiple imputation (MI)** gives the worst results (largest bias) in 3 out of the 4 reactions shown in Figure 1, and in 5 out of 6 reactions in Figure 2. In reactions 11 of Figure 1 and 22 of Figure 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 Figure 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. Figure 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.

- **Figure 2E and 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 Figure 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).
Identification of metabolic network models from incomplete high-throughput datasets

Fig. 2. Statistics of estimated parameter values for datasets with 75% of missing data and 20% noise. The graphical notations are the same as for Figure 1. (A–F) Boxplots for reactions 3, 13, 17, 22, 19 and 25 of the network, respectively.

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 E.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 authors collected data using multiple high-throughput techniques, in particular DNA microarray analysis and 2D 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, \( p = 7 \) external metabolites and measured cofactors, and \( m = 31 \) reactions (see Supplementary Section S4). 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 \( B^r \) and \( B^p \). 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 \textit{a priori} distribution defined in Equation (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 Equation (15) using a data matrix \( Y \) completed by the means \( \mu_{Y,k} \) of the \textit{a priori} distributions. For every individual reaction, the reduced model has a parameter vector \( \epsilon^* \) that is now entirely identifiable.

Apart from the distribution of the \textit{a priori} missing data, given by Equation (10), application of EM requires information about the distribution of \( \epsilon \), 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 in a glucose-limited chemostat with a dilution rate of \( 0.2 \) h\(^{-1}\). These data were used for the computation of the variance of \( \epsilon \). In order to assess the accuracy of the estimated \( B^r \) and \( B^p \), we computed for each parameter a 95\% confidence interval, by means of the profile likelihood method outlined in Section 3. Running the EM method on the model and the data took about 220 s using the implementation of Section 4. The computation of the confidence intervals for all parameters required about 23 min.

Contrary to the simulation studies reported in Section 4, a reference or ‘real’ model for the evaluation of the results does not exist in this case. However, \textit{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 in Equation (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, 26 of the 100 non-zero elasticities of the model are not identifiable from this dataset. Moreover, out of the remaining 74 elasticity estimates, more than 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 fact that the magnitude of noise in metabolite concentrations is comparable to the magnitude of relevant information. For example, for PEP the standard deviation over the experimental conditions equals the standard deviation of the replicates in a single condition (0.06 mM versus 0.05 mM). This precludes the estimation of an unambiguous sign.

Of the elasticities with statistically significant signs, 20 out of 34 are correct, in the sense that they have the expected positive or negative sign. The remaining elasticities, distributed over 9 reactions, are incorrectly estimated. Let us now discuss what we believe are potential sources of these errors, giving information that could be used to single out erroneous estimates \textit{a priori}.

We first note that for 3 of these 9 reactions (GapA, Pgk, Mdh and Edd,Eda, see Table 1), only very few complete datapoints are available (between 3 and 5) and regression mostly fails in these cases. In addition, all of these reactions involve at least one metabolite missing in >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 PykA,PykF) 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 11, 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.

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 those reported 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 inappplicable. 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 optimisation of likelihood functions in the presence of unobserved variables. A simpler implementation based on direct likelihood maximization via general-purpose numerical
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>GIC</th>
<th>PEP</th>
<th>G6P</th>
<th>Pyr</th>
<th>F6P</th>
<th>FBP</th>
<th>DHAP</th>
<th>3PG</th>
<th>AcoA</th>
<th>Ace</th>
<th>completed datapoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.1</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

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 assign that is significant with 95% confidence) in green. Abbreviations are as in Figure 3. Some of the cofactors are modeled as ratios of metabolite concentrations, e.g., ATP/ADP. Reaction 27, labeled \(\mu\), 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. *Regression was not able to produce any result.
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 Escherichia 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 metabolic concentrations, EM was able to estimate many of the model parameters (elasticities) 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 analytical/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 for 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 generalized mass action models that provide advantages when some metabolite concentrations approach 0 (del Rosario et al., 2008; Savageau, 1976). 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 Escherichia coli carbon metabolism allowed us to isolate issues that are critical for the appropriate exploitation of the data for parameter estimation. These issues 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 much detrimental to the identification results. This may influence sampling strategies, especially for metabolites that are difficult to measure. Another issue is the identifiability problems caused by steady-state measurements, which cannot always be resolved by genetic mutation or by varying physiological conditions. From this perspective time-resolved observations of the network dynamics, although much more demanding experimentally, carry great promise (Hardiman et al., 2007).

**ACKNOWLEDGEMENTS**

The authors would like to thank Delphine Ropers for useful discussions.

**Funding** Agence Nationale de la Recherche under project MetaGenoReg (ANR-06-BYOS-0003).

**Conflict of Interest** none declared.

**REFERENCES**


Identification of metabolic network models from incomplete high-throughput datasets
