Bayesian inference of the sites of perturbations in metabolic pathways via Markov chain Monte Carlo

Bayu Jayawardhana1,2,*, Douglas B. Kell1,2 and Magnus Rattray3

1Manchester Interdisciplinary Biocentre, The University of Manchester, Manchester M1 7DN, 2School of Chemistry, The University of Manchester, Manchester M13 9PL and 3School of Computer Science, The University of Manchester, Manchester M13 9PL, UK

Received on January 29, 2008; revised on February 22, 2008; accepted on March 15, 2008
Advance Access publication March 20, 2008
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Genetic modifications or pharmaceutical interventions can influence multiple sites in metabolic pathways, and often these are ‘distant’ from the primary effect. In this regard, the ability to identify target and off-target effects of a specific compound or gene therapy is both a major challenge and critical in drug discovery.

Results: We applied Markov Chain Monte Carlo (MCMC) for parameter estimation and perturbation identification in the kinetic modeling of metabolic pathways. Variability in the steady-state measurements in cells taken from a population can be caused by differences in initial conditions within the population, by variation of parameters among individuals and by possible measurement noise. MCMC-based parameter estimation is proposed as a method to help in inferring parameter distributions, taking into account uncertainties in the initial conditions and in the measurement data. The inferred parameter distributions are then used to predict changes in the network via a simple classification method. The proposed technique is applied to analyze changes in the pathways of pyruvate metabolism of mutants of Lactococcus lactis, based on previously published experimental data.

Availability: MATLAB code used in the simulations is available from ftp://anonymous@dbkweb.mib.man.ac.uk/pub/Bioinformatics_BJ.zip

Contact: bayu@ieee.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Drug discovery is now recognized as a problem of integrative systems biology requiring genome-wide analyses (see, e.g. Butcher, 2005; Dobson and Kell, 2008; Williams, 2005). In particular, genomics, transcriptomics, proteomics and metabolomics are being exploited to identify the mode-of-action and toxicity of possible compounds.

A number of articles describe genome-wide studies for detecting mode-of-action (Baetz et al., 2004; Clarke et al., 2001; Giaever et al., 1999; Marton et al., 1998; Parsons et al., 2004, 2006) and genomic screening via induced haploinsufficiency (Baetz et al., 2004; Giaever et al., 1999) have been proposed to detect the sites of interaction of a compound in biochemical pathways. These techniques compare a large number of mutant strains or induced haploinsufficient organisms for hypersensitivity to a set of compounds. The strains that show greater sensitivity to a compound are used to identify the implicated pathways. This approach relies on the availability of large numbers of mutants or haploinsufficient strains.

A related approach is to use DNA microarray analysis for validating drug targets and off-targets as pursued by Betts et al. (2002) and Marion et al. (1998). Gene-expression profiles from the untreated and the drug-treated cells are analyzed and clustered based on the levels of expression. The functionality of genes which have large changes in their expression level is taken to indicate the chemical pathway(s) affected by the compound. Based on a similar principle, the use of proteomics in mode-of-action studies involves the identification of proteins that are significantly altered in the drug-treated organism (Chapal et al., 2004).

Metabolomics studies have been performed from a similar perspective (Aranibar et al., 2001; Ott et al., 2003) and knowledge of metabolite transactions and metabolomics are also important in drug discovery (Kell, 2006). The metabolites represent the product of biochemical pathways in an organism and, potentially, can be used to infer the changes in the activities of enzymes in specific chemical pathways consequent upon pharmacological or genetic perturbations.

In metabolomics, one typically observes the metabolite concentrations or fluxes in a quasi-steady state. Because of the rapid turnover of cellular pools, it is fairly difficult to extract the metabolites reliably during transients, as these can occur on a time scale smaller than the ability to inhibit metabolism. Therefore, only steady-state data will be considered in this article, although the technique is sufficiently general to deal with time-series data where they exist.

Given measurement data statistics and a kinetic model of the metabolic pathways, we will use Bayesian inference to reflect the degree of uncertainty in the model parameters due to uncertainties in the measurements and in the initial conditions. We show that by comparing the inferred parameter distribution of the normal and drug-induced metabolic pathways,
significant parameter changes can be identified and such perturbations can be taken to reflect the mode-of-action of a drug.

Simulation results using the glycolytic pathway of *Saccharomyces cerevisiae* show the effectiveness of the proposed method to predict perturbations with reasonable accuracy and high sensitivity. The technique is then applied to analyze changes in the pathways of pyruvate metabolism of mutants of *Lactococcus lactis*, based on experimental data taken from Hoefnagel et al. (2002).

Coleman and Block (2006) and Battogtokh et al. (2002) showed how the Markov Chain Monte Carlo (MCMC) strategy can be used to estimate the posterior distribution of parameters in non-linear systems described by differential equations. However, due to the nature of the problems considered in those papers, our method uses several features that are not available there. First, we focus on steady-state data without any information about the sampling time. Secondly, our technique deals with observable external variables that are not components of the state vector but that are parts of the state equations. Thirdly, metabolic pathways have stoichiometric constraints called moieties conservations and therefore parameter distributions and MCMC proposals should respect these constraints.

Liebermeister and Klipp (2006) have introduced a Bayesian inference method for parameter estimation in systems biology models. They approximate the posterior distribution by a Gaussian distribution centered around a local maximum of the likelihood function. They construct logic-based reasonings based on the effects in metabolic pathways using inductive logic programming (2006) provide an alternative way to infer drug inhibitory effects in metabolic pathways using inductive logic programming. They construct logic-based reasonings based on the relative changes in metabolite concentrations and assume unidirectional reactions. This over-simplification can fail to capture the effect of feedback mechanisms, cofactors and reversible reactions on the behavior of metabolites. A recent related work applying MCMC to systems biology, due to Vyshemirsky and Girolami (2007), implements MCMC for model selection whereas here we focus on parameter estimation and identifying significant parameter changes.

2 METHODS

2.1 Problem formulation and parameter estimation

Leaving aside stochastic systems in which the number of molecules is insufficiently large to be approximated by a continuous quantity, a kinetic model of metabolic pathways is conveniently cast in terms of coupled ordinary differential equations (ODE) in a state space form:

\[
\begin{align*}
\dot{x} &= f(x, y_{ext}, \theta, t), \\
y_{int} &= h(x, t), \\
x(0) &= x_0, \\
y_{ext} &= y_{int},
\end{align*}
\]

subject to \( M \) algebraic constraints:

\[
\sum_{i \in A_k} x_i(t) = C_j, \quad j = 1, 2, \ldots, M, \forall t \]

\[
A_j \cap A_k = \emptyset \quad \text{whenever} \ j \neq k
\]

where \( x \) denotes the metabolite concentrations, \( \theta \) are the system parameters, \( y_{ext} \) are the observable external metabolite fluxes and concentrations, \( y_{int} \) are the observation variables and \( x_0 \) is the initial state of the model. For the \( j \)th algebraic constraint, \( A_j \) is the index set of components of \( x \) and \( C_j \) is a positive constant. In the above formulation, the function \( f \) represents the enzyme kinetics laws governing the reactions while the function \( h \) represents the observation data that are normally full or partial information on the metabolite concentrations and fluxes.

In kinetic modeling of metabolic pathways, the \( M \) algebraic constraints in (2) are called moieties conservations. These are groups of variables whose overall quantities are assumed to be constant throughout the time course (Hofmeyr et al., 1986). The conservation of adenine and pyridine nucleotide moieties are examples.

Suppose that for some \( \theta, y_{ext} \) and \( x_0 \), the state Equations (1), (2) converge to a steady state. Let us denote the steady-state value by \( x_r \). The dependence of \( x_r \) on \( \theta, y_{ext} \) and \( x_0 \) implies that the corresponding steady-state observation \( y_{int} \) can be described as a mapping from \( (\theta, y_{ext}, x_0) \). In other words, \( y_{int} = g(\theta, y_{ext}, x_0) \) for some function \( g \).

Note that it is generally difficult to derive the function \( g \) analytically. Hence, one is forced to do numerical simulations of the kinetic model to obtain the steady-state values of the observables.

It is also assumed that the initial state \( x_0 \) and the system parameters \( \theta \) are also uncertain and therefore probabilistic quantities. Let the prior distribution of \( x_0 \) be denoted by \( p_{x_0}(\cdot) \) and the prior distribution of \( \theta \) be denoted by \( p_{\theta}(\cdot) \). The prior distributions represent our best knowledge or estimate of the distribution of \( x_0 \) and \( \theta \) before making any observations of the system. Log-normal, gamma or Gaussian distributions expanded around parameters obtained from *in vitro* experiments are some possible choices for these priors. For initial states that are subject to a linear constraint, a Dirichlet distribution can be used along with a scale parameter. For the vaguest parameters and initial states we use uniform distributions with estimated lower and upper bounds.

The measurement data are summarized by statistical models \( D_{int} \) and \( D_{ext} \) with distribution functions \( p_{int}(\cdot) \) and \( p_{ext}(\cdot) \) that capture the statistics of a sufficiently large number of steady-state observations \( y_{int} \) and external observations \( y_{ext} \), respectively. We use distribution functions, instead of a set of data points, because steady-state measurement data published in the literature are often summarized by means and SDs. These are naturally mapped onto a Gaussian distribution. Note that the distributions represent the metabolism of an aggregation of cells, where each cell has a different steady state (the ODEs give the unicellular state Equations (1), (2)) for some function \( g \).

Define the conditional distribution of \( x_0 \) given \( (D_{int}, D_{ext}) \) by

\[
p(\theta, D_{int}, D_{ext}) = \frac{p(\theta, D_{int}, D_{ext})}{p(D_{int}, D_{ext})}
\]

where

\[
p(\theta, D_{int}, D_{ext}) = \int p(\theta, y_{ext}, x_0) p_{x_0}(x_0) p_{y_{ext}}(y_{ext})dy_{ext} dy_{int}.
\]

By prior independence of \( \theta \) and \( x_0 \), we can write

\[
p(\theta, D_{int}, D_{ext}) = p_{\theta}(\theta) \int p_{x_0}(x_0) p_{y_{ext}}(y_{ext})dy_{ext} dy_{int}.
\]

Since \( p(D_{int}, D_{ext}) \) is a normalizing constant, the posterior distribution \( p(\theta | D_{int}, D_{ext}) \) is proportional to the right-hand side of (3).

The integral on the right hand side of (3) is the likelihood function \( p(D_{int}, D_{ext} | \theta) \). It depends on the existence of steady-state values in the
dynamical equations and with parameter $\theta$, for all sampled initial conditions $x_0$. This condition creates a problem since most non-linear systems have unstable regions (including oscillatory behavior). For a given $\theta, y_{\text{ext}}, x_0$, let us define $p(D_{\text{int}} | \theta, y_{\text{ext}}, x_0) = 0$ if no steady state is reached and $p(D_{\text{int}}(\theta, y_{\text{ext}}, x_0) = \text{final}(\theta, y_{\text{ext}}, x_0)$ if a steady state is obtained. Then this problem can be avoided with the likelihood function given by

$$p(D_{\text{int}}, D_{\text{ext}} | \theta) = \int p(D_{\text{int}} | \theta, y_{\text{ext}}, x_0) p(x_0) p(y_{\text{ext}}(x_0) d\theta d
u_{\text{ext}}.$$  (4)

MCMC can be used to draw samples from the posterior distribution $p(\theta | D_{\text{int}}, D_{\text{ext}})$ and use it to approximate various statistical properties related to $p(\theta | D_{\text{int}}, D_{\text{ext}})$, for example, approximating the confidence interval, median and expected value. Coleman and Block (2006) and Battogtokh et al. (2002) give an overview of the MCMC technique applied to non-linear systems. Various techniques for MCMC can be found in Gelman (2004) and Spiegelhalter et al. (1996). In this work we use a standard Metropolis–Hastings algorithm.

In order to draw samples from the posterior distribution using MCMC, we would like to avoid dealing explicitly with (4) since the integral that appears on the right-hand side is computationally intractable. Notice that this integral corresponds to the marginalization of $p(D_{\text{int}} | \theta, y_{\text{ext}}, x_0)p(x_0)p(y_{\text{ext}}(x_0))$ over $x_0$ and $y_{\text{ext}}$. Therefore, drawing samples from the posterior distribution $p(\theta | D_{\text{int}}, D_{\text{ext}})$ is equivalent to getting samples from

$$p(D_{\text{int}} | \theta, y_{\text{ext}}, x_0) p(x_0) p(y_{\text{ext}}(x_0) p(\theta)$$  (5)

and marginalizing the samples over the initial state $x_0$ and over the external variables $y_{\text{ext}}$. We will refer to (5) as the target distribution $p_{\text{target}}(w)$ where $w = (\theta, y_{\text{ext}}, x_0)$.

In this article, we use the Metropolis–Hastings algorithm to generate several parallel Markov chains and the convergence of the parallel chains to a target distribution is monitored using the measure proposed by Gelman (2004). Details about the prior and proposal distributions (including those for conserved variables), the Metropolis–Hastings algorithm, and the convergence measure are given in the Supplementary Material.

### 2.2 Prediction of perturbations

The method described so far enables us to approximate the distribution of the system parameters given measurement data summarized by $D_{\text{int}}$ and $D_{\text{ext}}$. This leads to the next problem where one has two measurement datasets that we may summarize as $(D_{\text{int}}, D_{\text{ext}})_{\text{normal}}$ and $(D_{\text{int}}, D_{\text{ext}})_{\text{perturbed}}$ corresponding to two different conditions that one may take as a reference state and a perturbed state, respectively.

By using MCMC, we can estimate the posterior distribution for the normal (wild-type) organism $p(\theta | D_{\text{int}}, D_{\text{ext}})_{\text{normal}}$ and the posterior distribution for the drug-treated (mutant) organism $p(\theta | D_{\text{int}}, D_{\text{ext}})_{\text{perturbed}}$. Given these two posterior distributions, one can compare both distributions to infer the effect of the perturbation.

Let the subscript $i$ denote the index of element in $\theta$. Then we can infer whether the enzymatic reaction with parameter $\theta_i$ has been perturbed by computing

$$p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}} | D_{\text{int}}, D_{\text{ext}})_{\text{normal}} := \int \int p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}} | D_{\text{int}}, D_{\text{ext}})_{\text{perturbed}}$$

where $p(\theta_i | D_{\text{int}}, D_{\text{ext}})$ is the marginalization of $p(\theta | D_{\text{int}}, D_{\text{ext}})$ over the complement of $\theta_i$. The above method was used by Liu et al. (2006) to determine the significance of differential gene expression in samples exposed to a treatment compared with those from a control.

For brevity, we use the notation $p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}})$ to represent the left-hand term of (6). Suppose that the MCMC samples for both cases are given by $w_{\text{normal}}(n)$ and $w_{\text{perturbed}}(n)$ where $n = 1,2, \ldots, N$ and $N$ is the total number of samples. Then (6) can be approximately by

$$p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}}) \approx \frac{1}{N} \sum_{i=1}^N (w_{i, \text{perturbed}}(n) - w_{i, \text{normal}}(n))$$

where $w_i$ is the $i$th component of $w$ and $\chi$ is an indicator function given by $\chi(s) = 1$ for all $s \geq 0$ and $\chi(s) = 0$ elsewhere.

Since $p(\theta_{i, \text{normal}} > \theta_{i, \text{perturbed}}) = 1 - p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}})$, the MCMC samples can be classified into three classes based on the value of $p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}})$ with a cut-off $0 < \epsilon < 0.5$:

(A1) If $p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}}) > 0.5 + \epsilon$ then $\theta_{i, \text{perturbed}}$ is up-regulated;

(A2) If $p(\theta_{i, \text{perturbed}} > \theta_{i, \text{normal}}) < 0.5 - \epsilon$ then $\theta_{i, \text{perturbed}}$ is down-regulated;

(A3) Otherwise $\theta_{i, \text{perturbed}}$ is unchanged.

This classification can be used to predict whether the enzymatic reaction which corresponds to the parameter $\theta_i$ is up-regulated, relatively unchanged or down-regulated in the perturbed case compared to the normal one.

In the next section, an optimal $\epsilon$ is computed based on the glycolysis pathway model of 16 strains of *in-silico* organism. A total of 240 ordered pairs of datasets are obtained from the permutation of 16 sets of MCMC samples. By considering the classification of these pairs of datasets for being up-regulated and for being not up-regulated, an receiver operating characteristics (ROC) curve (Broadhurst and Kell, 2006) can be drawn by varying $\epsilon$. Note that if one computes an ROC curve that compares the case of being down-regulated and being not down-regulated, the symmetry of the classification algorithm and the symmetry of the permuted ordered pair of datasets ensure that the same curve is obtained. If the accuracy acc is defined by $acc = c_1 + c_2$ TNR where TPR is the true positive rate (sensitivity), TNR is the true negative rate (specificity), and $c_1, c_2$ are the weighting constants, the optimal $\epsilon$ that maximizes acc can be found from the ROC curve.

### 3 RESULTS

#### 3.1 Glycolytic pathways in *Saccharomyces cerevisiae*

In this subsection, the perturbation analysis is evaluated using simulated glycolysis data. The glycolysis model is taken from Pritchard and Kell (2002) and Teusink et al. (2000).

The parameters are the limiting step constants $V_{\text{max}}$ and the concentrations of external glucose, F26BP, glycerol and ethanol. The prior distributions for $V_{\text{max}}$ are log-normal distributions with the log-mean values taken from the *in vitro* measurements presented by Teusink et al. (2000) and the log-variance values set between 0.4805 and 7.6872. The prior distribution for the log-variance values set between 0.4805 and 7.6872. The prior distribution for the concentrations of external glucose, F26BP, glycerol and ethanol are uniform distributions with intervals: (0.01,1000) for external glucose, (0.001,10) for F26BP, (0.01,100) for glycerol and (0.01,10) for ethanol. The prior distributions for initial metabolite concentrations are uniform distributions.

Following the method described in Section 2.1, MCMC is used to draw samples from the target distribution $p_{\text{target}}(w)$ in Equation (5). This implies that the proposed move is defined in the space of parameters $\theta$, external observable metabolite
concentrations and fluxes $y_{\text{ext}}$ and initial metabolite concentrations $x_0$. We use log-normal distributions as the proposal distributions for parameters, external observable variables and the initial concentrations of non-conserved metabolites. The conserved variables, which are a consequence of the conservation of adenine and pyridine nucleotides, have initial concentrations parametrized according to the method described in the Supplementary Material.

The full measurement data are the concentrations of internal glucose, ATP, G6P, ADP, F6P, F16BP, AMP, DHAP, GAP, NAD, BPG, NADH, P3G, P2G, PEP, PYR, acetaldehyde and the fluxes of glucose, glycerol, succinate, pyruvate, glycogen, trehalose. In this model, there are external observable metabolite fluxes $y_{\text{ext}}$ namely the measured glycogen and trehalose fluxes.

Sixteen observation datasets are generated using randomly selected parameter values (except in the specific cases discussed below) with the rest of the constants (e.g. equilibrium constants $K_{eq}$, Michaelis–Menten constants, Hill coefficients) set to the same values as used in Teusink et al. (2000). For each case, 30 samples are generated with uniformly distributed initial conditions and with normally distributed glycogen flux and trehalose flux. After the steady-state samples are obtained, additive Gaussian noise is added and the resulting dataset is summarized as a Gaussian distribution.

Using the methodology described in the previous section, MCMC samples are generated to estimate the posterior distribution for each of the datasets. Three parallel chains were used for each case and simulations were run until the sequences converged.

A permutation of 16 sets of resulting MCMC samples with the permutation size of two is used to evaluate the classification method described in Section 2.2. The permutation gives 240 ordered pairs of datasets, from which, the ROC curve that compares the case of being up-regulated and not being up-regulated can be drawn by varying $\varepsilon$ from $-0.5$ to $0.5$. Figure 1 shows the ROC curve with the approximate ROC area of 0.836 (the area is estimated based on the trapezoidal area under the curve).

Let us evaluate five out of 16 observation datasets and use the capital letter to indicate each case. The first dataset (Case A) can be regarded as the wild-type yeast, and the rest are in silico mutants. All limiting step constants in Case B are lower than those in Case A, while limiting step constants in Case C are higher than those in Case A. The $V_{\text{max}}$ values for all five cases are summarized in Table 1. Details of all 16 cases are provided in the Supplementary Material.

Figure 2 shows the result of MCMC for parameter estimation in Case A. The estimation uses the mean of the samples as the estimated parameter for each $V_{\text{max}}$ and the credible interval is approximated by the 2.5th and 97.5th percentile of the marginalized samples. Figure 2 shows that MCMC is indeed able to produce credible intervals where the true values lie.

The capability of MCMC to produce parameter distribution taking into account uncertainties in the measurement data and in the initial conditions is illustrated in Figure 3 which shows the MCMC samples from Case A. It compares the statistical model of measurement data (which are Gaussian distributions calculated from 30 measurements) with the posterior predictive distribution (i.e. simulated output samples $y_{\text{int}}$ using the parameters from MCMC samples).

### Table 1. Limiting step values used in five different simulation setups

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Case A</th>
<th>Case B</th>
<th>Case C</th>
<th>Case D</th>
<th>Case E</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{\text{max}}$</td>
<td>101.3</td>
<td>81.3</td>
<td>121.3</td>
<td>110.75</td>
<td>120.81</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>670.5</td>
<td>570.5</td>
<td>770.5</td>
<td>340.83</td>
<td>690.77</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>1933</td>
<td>1633</td>
<td>2433</td>
<td>200.29</td>
<td>1270.64</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>121.5</td>
<td>101.5</td>
<td>181.5</td>
<td>552.06</td>
<td>284.36</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>101</td>
<td>81</td>
<td>161</td>
<td>161.25</td>
<td>384.52</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>2336</td>
<td>1336</td>
<td>3336</td>
<td>2074.87</td>
<td>1712.2</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>3928</td>
<td>2298</td>
<td>4298</td>
<td>3537.84</td>
<td>4936.26</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>2291</td>
<td>1291</td>
<td>2991</td>
<td>958.51</td>
<td>819.58</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>2423</td>
<td>1423</td>
<td>3223</td>
<td>2273.63</td>
<td>1494.45</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>240.4</td>
<td>180.4</td>
<td>290.4</td>
<td>238.36</td>
<td>297.29</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>700.5</td>
<td>650.5</td>
<td>790.5</td>
<td>790.02</td>
<td>816.08</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>869.9</td>
<td>809.9</td>
<td>969.9</td>
<td>1448.25</td>
<td>5939.07</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>50.2</td>
<td>40.2</td>
<td>70.2</td>
<td>113.76</td>
<td>781.56</td>
</tr>
<tr>
<td>$g_{\text{max}}$</td>
<td>47.2</td>
<td>37.2</td>
<td>57.2</td>
<td>53.91</td>
<td>94.63</td>
</tr>
</tbody>
</table>
in the Supplementary Material). This shows that the extent of uncertainties in the parameter estimation increases as the availability of information decreases.

It is interesting to observe that, based on minimal information, we are still able to infer the distribution of some metabolite concentrations and fluxes that are not available in the measurement data. Figure 5 shows the measurement data for (a) pyruvate and (b) ATP concentrations in Case A along with the inferred pyruvate and ATP concentration obtained from the MCMC simulation using full information and using minimal information. The inference using minimal information is able to estimate the uncertainties reasonably well, although information scarcity produces a long-tail distribution as shown in Figure 5a. The discrepancy of the measured distribution of ATP concentration with the inferred one using full information suggests that measurement noise has a significant contribution to the uncertainties in the measurement data which cannot be explained by parameter uncertainties alone.

The ROC curve for perturbation identification using minimal information is shown in Figure 1. The area under the curve is 0.611 which is considerably lower than that obtained using full information. Due to this relatively poor performance, the optimal $\epsilon$ is computed using the ratio $c_1/c_2$ of 2.3 which puts higher weight on specificity. Based on this ratio, the optimal $\epsilon$ is 0.0465 with a specificity of 0.9075 and a sensitivity of 0.1928. In the following subsection, we use this optimal value for the analysis of perturbations in pyruvate pathways of L.lactis where the experimental data are minimal. In general the correct balance between specificity and sensitivity will depend on the application.

### 3.2 Pyruvate pathways in L.lactis

In this section, we try to identify the perturbation in the lactic acid bacteria represented by the removal of lactate dehydrogenase and the over-expression of NADH oxidase. The experimental data are taken from Hoefnagel et al. (2002) and the corresponding pathways are shown in Figure 6. Comparing our model with that in Hoefnagel et al. (2002), our model contains an additional pyruvate carboxylase reaction which can serve as an alternative branch for the production of phosphoenolpyruvate. The branch is added to explain the missing carbon flux in the experimental data. In addition to the rate equations used in Hoefnagel et al. (2002), the rate equation for pyruvate carboxylase is given by $v_{4} = \frac{V_{\text{max}}}{K_{m} + [\text{pyruvate}]}$ where $K_{m}$ is 0.31 mM (Sueda et al., 2004).
ment data and incomplete kinetic model.

The parameters are the 14 limiting step constants $V_{\text{max}}$ corresponding to 14 reactions in the model and the external concentrations of glucose, lactate, acetoin, O$_2$, phosphate, ethanol and butanediol. The prior distributions for $V_{\text{max}}$ are uniform distributions on $(0, 20000)$. The prior distributions for the concentrations of glucose, lactate, acetoin, O$_2$, phosphate, ethanol and butanediol are uniform distributions with intervals: $(0.1, 100)$ for external glucose and phosphate, $(0.01, 10)$ for lactate, acetoin and ethanol, $(0.002, 2)$ for O$_2$ and $(0.0001, 0.1)$ for butanediol. The prior distributions for initial metabolite concentrations are uniform distributions.

We use log-normal distributions as the proposal distributions for parameters, external observable variables and the initial concentrations of non-conserved metabolites. Details on the strategies are available in the Supplementary Material.

The conserved variables are the conserved moieties of adenine and pyridine nucleotides and the conservation of $\text{[Acetyl-CoA]} + [\text{CoA}]$. Their initial concentrations are parameterized according to the method described in the Supplementary Material.

The measurement data are acetate flux, ethanol flux, acetoin flux and lactate flux. In this case, there are no external observable metabolite concentrations or fluxes $y_{\text{ext}}$. The data are summarized as Gaussian distributions with the mean values set to those measured in Hoefnagel et al. (2002) and the SDs set to those measured in Hoefnagel et al. (2002) and the SDs set to $10\%$ of the mean values (following the observation by de Koning and van Dam (1992) that metabolites measurements of this type have standard errors of $\sim 10\%$). While the objective in Hoefnagel et al. (2002) is to maximize the acetoin flux by manipulating the enzyme production, we try to recapture the changes in the pathways based only on the minimal measurement data and incomplete kinetic model.

The conserved variables are the conserved moieties of adenine and pyridine nucleotides and the conservation of $\text{[Acetyl-CoA]} + [\text{CoA}]$. Their initial concentrations are parameterized according to the method described in the Supplementary Material.

The measurement data are acetate flux, ethanol flux, acetoin flux and lactate flux. In this case, there are no external observable metabolite concentrations or fluxes $y_{\text{ext}}$. The data are summarized as Gaussian distributions with the mean values set to those measured in Hoefnagel et al. (2002) and the SDs set to $10\%$ of the mean values (following the observation by de Koning and van Dam (1992) that metabolites measurements of this type have standard errors of $\sim 10\%$). While the objective in Hoefnagel et al. (2002) is to maximize the acetoin flux by manipulating the enzyme production, we try to recapture the changes in the pathways based only on the minimal measurement data and incomplete kinetic model.

The conserved variables are the conserved moieties of adenine and pyridine nucleotides and the conservation of $\text{[Acetyl-CoA]} + [\text{CoA}]$. Their initial concentrations are parameterized according to the method described in the Supplementary Material.

The measurement data are acetate flux, ethanol flux, acetoin flux and lactate flux. In this case, there are no external observable metabolite concentrations or fluxes $y_{\text{ext}}$. The data are summarized as Gaussian distributions with the mean values set to those measured in Hoefnagel et al. (2002) and the SDs set to $10\%$ of the mean values (following the observation by de Koning and van Dam (1992) that metabolites measurements of this type have standard errors of $\sim 10\%$). While the objective in Hoefnagel et al. (2002) is to maximize the acetoin flux by manipulating the enzyme production, we try to recapture the changes in the pathways based only on the minimal measurement data and incomplete kinetic model.

Figure 7 shows the inferred concentration of ADP and pyruvate in the wild-type case that are not measured in Hoefnagel et al. (2002). We can infer them by computing the distribution of simulated output samples $Y_{\text{int}}$ using the parameters from MCMC samples. Based on the figure, the steady state of ADP concentration has the highest posterior probability close to 0.2 mM while the highest posterior distribution of steady-state pyruvate concentration is close to 0.04 mM with the distribution having a long right-hand tail.

The comparison of MCMC samples from the minimal measurement data in three different strains of L. lactis: wild-type, LDH knocked-out and NOX over-expressed, is shown in Figure 8.

The result shown in Figure 8a suggests that knocking out the LDH gene in L. lactis produces pleiotropic effects apart from the down-regulation of the lactate dehydrogenase reaction itself. It is highly likely that there is also an increase in the reactions of NADH oxidase and pyruvate carboxylase following the knocking out of the LDH gene. The reactions of ‘lumped glycolysis’ and lactate dehydrogenase are also likely to be down-regulated in the LDH-mutant of L. lactis. On the other hand, Figure 8b shows that the over-expression of NOX gene results in only a minor pleiotropic effect. Our method shows that the NOX-mutant of L. lactis has significant effects only on increasing the NADH oxidase and pyruvate carboxylase reactions.

4 DISCUSSION

This article focuses on the application of MCMC to parameter estimation and perturbation analysis in metabolic pathways. Compared to the existing literature, the technique developed

Fig. 7. Inferred steady-state concentrations of (a) ADP and (b) pyruvate; from the simulated output using the resulting MCMC samples in the wild-type case.

Fig. 8. Perturbation identification on the pyruvate pathways in L. lactis using MCMC. (a). The plot of $p(V_4\text{LDHknockout}^\text{max} < V_4\text{Normal}^\text{max})$. (b). The plot of $p(V_6\text{NOXover}^\text{max} > V_6\text{Normal}^\text{max})$. The classification algorithm with $\epsilon = 0.0465$ gives the down-regulated region (striped area), up-regulated region (white area) and unchanged region (grey area). The label $V_i$ in the figure corresponds to the limiting-step constant of the $i$th reaction as shown in Figure 6.

Figure 7 shows the inferred concentration of ADP and pyruvate in the wild-type case that are not measured in Hoefnagel et al. (2002). We can infer them by computing the distribution of simulated output samples $Y_{\text{int}}$ using the parameters from MCMC samples. Based on the figure, the steady state of ADP concentration has the highest posterior probability close to 0.2 mM while the highest posterior distribution of steady-state pyruvate concentration is close to 0.04 mM with the distribution having a long right-hand tail.

The comparison of MCMC samples from the minimal measurement data in three different strains of L. lactis: wild-type, LDH knocked-out and NOX over-expressed, is shown in Figure 8.

The result shown in Figure 8a suggests that knocking out the LDH gene in L. lactis produces pleiotropic effects apart from the down-regulation of the lactate dehydrogenase reaction itself. It is highly likely that there is also an increase in the reactions of NADH oxidase and pyruvate carboxylase following the knocking out of the LDH gene. The reactions of ‘lumped glycolysis’ and lactate dehydrogenase are also likely to be down-regulated in the LDH-mutant of L. lactis. On the other hand, Figure 8b shows that the over-expression of NOX gene results in only a minor pleiotropic effect. Our method shows that the NOX-mutant of L. lactis has significant effects only on increasing the NADH oxidase and pyruvate carboxylase reactions.
in this article has three distinguishing features. First, the systems are described by first-order non-linear ODEs subject to algebraic constraints corresponding to the moiety conservations. Secondly, it focuses on steady-state measurement data without any information about the sampling time. Thirdly, it copes with observable external variables that are not components of the state vector but that do appear in the non-linear ODEs.

The ability to produce a broad credible region in parameter space is a feature distinguishing MCMC-based parameter estimation from most parameter identification techniques. This complements the standard parameter fitting methods, such as genetic or gradient descent-based algorithms, which can be used to obtain a point estimate of the model parameters and for which the credibility of estimated parameters is less easily assessed. Approaches for estimating credible intervals based on expansions around a point estimate, using e.g. the Hessian of the log-likelihood, are essentially asymptotic in the size of dataset and work under the assumption that the posterior distribution can be well-approximated by a Gaussian distribution. In the examples considered here, data were very limited and therefore the credible intervals obtained were broad and asymmetrical, reflecting the non-Gaussian nature of the posterior distribution. The posterior distribution also allows for the investigation of higher order relationships between parameters, e.g. the correlation or mutual information between parameters (Lüdtke et al., 2008). We have not pursued this here, as we have focused only on changes in the marginal distribution of each parameter. In addition, we have confined these studies to two comparatively small systems, and future work will determine the extent to which these methods scale to larger networks.

The perturbation analyses are done via a simple classification algorithm applied to MCMC samples from the two cases. The algorithm relies on a parameter $\theta$ which can be selected based on simulated cases. It can be chosen at an appropriate level depending on the availability of measurement data and in order to balance specificity and sensitivity.

Our analysis of measurement data from three strains of $L. lactis$ gives insights into the possible pleiotropic effects due to genetic modification. It also confirms that the mutants have major alterations in the known target reactions.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Joshua Knowles for many useful discussions and Dr Simon Rogers for the technical advice on the proposal distributions.

Funding: This study was funded by UK Biotechnology and Biological Sciences Research Council (BB/E016065).

Conflicts of Interest: none declared.

REFERENCES