Bioinformatics Advance Access published March 20, 2008

Original Papers

Bayesian Inference of the Sites of Perturbations in Metabolic Pathways via Markov Chain Monte Carlo

Bayu Jayawardhana[1,2,*], Douglas B. Kell[1,2,] and Magnus Rattray[3,]

[1]Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St., Manchester M1 7DN, UK.

[2]School of Chemistry, The University of Manchester, Manchester M13 9PL, UK.

[3]School of Computer Science, The University of Manchester, Manchester M13 9PL, UK.

Associate Editor: Prof. 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: bayujw@ieee.org, dbk@manchester.ac.uk, magnus.rattray@manchester.ac.uk

Supplementary information: Supplementary material is available from the journal website.

1 INTRODUCTION

Drug discovery is now recognised as a problem of integrative systems biology requiring genome-wide analyses (see, e.g., Butcher, 2005; Dobson & 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; Parsons *et al*., 2006). For instance, chemical-genetic profiling (Parsons *et al*., 2004; Parsons *et al*., 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 haplo-insufficient 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 Marton *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 paper, 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 *S. 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 nonlinear 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 moiety 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 centred around a local maximum of the posterior. Their approach assumes that the problem of interest has a unimodal and localized posterior distribution which may not be the case in complex systems such as metabolic pathways, especially when data are scarce. Tamaddoni-Nezhad *et al*. (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 recognize 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 in a state space form:

$$\begin{aligned} \dot{x} &= f(x, y_{ext}, \theta, t), \qquad x(0) = x_0, \\ y_{int} &= h(x,t), \end{aligned} \tag{1}$$

subject to M algebraic constraints:

$$\begin{aligned} &\sum_{i \in \mathfrak{A}_j} x_i(t) = C_j, \qquad j = 1, 2, ..., M, \forall t \\ &\mathfrak{A}_j \cap \mathfrak{A}_k = \varnothing \quad \text{whenever} \quad j \neq k \end{aligned} \tag{2}$$

where x denotes the metabolite concentrations, θ 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, $\mathfrak{A}_\mathrm{j}$ 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 moiety 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 θ, y_{ext} and x_0, the state equations (1), (2) converge to a steady-state. Let us denote the steady-state value by x_{ss}. The dependence of x_{ss} on θ, y_{ext} and x_0 implies that the corresponding steady-state observation y_{ss} can be described as a mapping from (θ, y_{ext}, x_0). In other words, $y_{ss} = g(\theta, y_{ext}, x_0)$ 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 θ are also uncertain and therefore probabilistic quantities. Let the prior distribution of x_0 be denoted by $\mathrm{p}_{\mathrm{x}0}(.)$ and the prior distribution of θ be denoted by $\mathrm{p}_\theta(.)$. The prior distributions represent our best knowledge or estimate of the distribution of x_0 and θ before making any observations of the system. Lognormal, 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 $\mathfrak{D}_\mathrm{int}$ and $\mathfrak{D}_\mathrm{ext}$ with distribution functions $\mathrm{p}_{int}(.)$ and $\mathrm{p}_{ext}(.)$ that capture the statistics of a sufficiently large number of steady-state observations y_{ss} 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 standard deviations. 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 kinetic model) (Davey & Kell, 1996). Our model is a simplification in which we assume a single ODE with a measurement distribution capturing all sources of experimental variation.

The Bayesian inference problem (see, e.g., Gelman, 2004) asks whether we can compute the posterior distribution $\mathrm{p}(\theta|\mathfrak{D}_\mathrm{int}, \mathfrak{D}_\mathrm{ext})$. Define the conditional distribution of θ given ($\mathfrak{D}_\mathrm{int}$, $\mathfrak{D}_\mathrm{ext}$) by

$$\mathrm{p}(\theta \,|\, \mathfrak{D}_{int}, \mathfrak{D}_{ext}) = \frac{\mathrm{p}(\theta, \mathfrak{D}_{int}, \mathfrak{D}_{ext})}{\mathrm{p}(\mathfrak{D}_{int}, \mathfrak{D}_{ext})}$$

where $\mathrm{p}(\theta, \mathfrak{D}_{int}, \mathfrak{D}_{ext}) = \int\int \mathrm{p}_{int}\left(g(\theta, y_{ext}, x_0) \,|\, \theta, y_{ext}, x_0\right)$
$\mathrm{p}(\theta, x_0)\mathrm{p}_{ext}(y_{ext})\,\mathrm{d}x_0\mathrm{d}y_{\mathrm{ext}}.$

By prior independence of θ and x_0, we can write

$$\begin{aligned} \mathrm{p}(\theta, \mathfrak{D}_{int}, \mathfrak{D}_{ext}) = \mathrm{p}_\theta(\theta) \int\int \mathrm{p}_{int}\left(g(\theta, y_{ext}, x_0) \,|\, \theta, y_{ext}, x_0\right) \\ \mathrm{p}_{\mathrm{x}0}(x_0)\mathrm{p}_{ext}(y_{ext})\,\mathrm{d}x_0\mathrm{d}y_{\mathrm{ext}} \end{aligned} \tag{3}.$$

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

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

$$p(\mathfrak{D}_{int}, \mathfrak{D}_{ext} \mid \theta) = \int\int p\left(\mathfrak{D}_{int} \mid \theta, y_{ext}, x_0\right) p_{x0}(x_0) p_{ext}(y_{ext})\, dx_0 dy_{ext}. \quad (4)$$

Markov Chain Monte Carlo (MCMC) can be used to draw samples from the posterior distribution $p(\theta|\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})$ and use it to approximate various statistical properties related to $p(\theta|\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})$, 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 nonlinear 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 $p(\mathfrak{D}_{\mathrm{int}} \mid \theta, y_{ext}, x_0) p_{x0}(x_0) p_{ext}(y_{ext})$ over x_0 and y_{ext}. Therefore, drawing samples from the posterior distribution $p(\theta|\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})$ is equivalent to getting samples from

$$p\left(\mathfrak{D}_{int} \mid \theta, y_{ext}, x_0\right) p_{x0}(x_0)\, p_{ext}(y_{ext}) p_{\theta}(\theta) \quad (5)$$

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

In this paper, 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 $\mathfrak{D}_{\mathrm{int}}$ and $\mathfrak{D}_{\mathrm{ext}}$. This leads to the next problem where one has two measurement data sets that we may summarize as $(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{normal}}$ and $(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{perturbed}}$ 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 $p(\theta|(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{normal}})$ and the posterior distribution for the drug-treated (mutant) organism $p(\theta|(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{perturbed}})$. 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 θ. Then we can infer whether the enzymatic reaction with parameter θ_i has been perturbed by computing

$$\begin{aligned} & p(\theta_{i,\mathrm{perturbed}} > \theta_{i,\mathrm{normal}} \mid (\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{normal}}, (\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{perturbed}}) := \\ & \quad \int\int p(\theta_{i,\mathrm{perturbed}} > \theta_{i,\mathrm{normal}} \mid (\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{perturbed}}) p(\theta_{i,\mathrm{normal}} \mid (\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}})_{\mathrm{normal}}) d\theta_{i,\mathrm{normal}} d\theta_{i,\mathrm{perturbed}} \end{aligned} \quad (6)$$

where $p(\theta_i|(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}}))$ is the marginalization of $p(\theta|(\mathfrak{D}_{\mathrm{int}}, \mathfrak{D}_{\mathrm{ext}}))$ over the complement of θ_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 $p(\theta_{i,\mathrm{perturbed}} > \theta_{i,\mathrm{normal}})$ to represent the left-hand term of (6).

Suppose that the MCMC samples for both cases are given by $w_{\mathrm{normal}}(n)$ and $w_{\mathrm{perturbed}}(n)$ where $n = 1,2,\ldots,N$ and N is the total number of samples. Then can be approximated by

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

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

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

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

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

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

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

In the next section, an optimal ε is computed based on the glycolysis pathway model of sixteen strains of *in-silico* organism. A total of 240 ordered pairs of datasets are obtained from the permutation of sixteen sets of MCMC samples. By considering the classification of these pairs of datasets for being up-regulated and for being not up-regulated, an ROC curve (Broadhurst & Kell, 2006) can be drawn by varying ε. 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 $acc = c_1\, TPR + c_2\, TNR$ 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 ε 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_{max} and the concentrations of external glucose, F26BP, glycerol and ethanol. The prior distributions for V_{max} are lognormal 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[1]. 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.

[1] They are set such that the true value lies within the interval $(\exp(-m\sqrt{\sigma})$, $\exp(m\sqrt{\sigma}))$ where σ is the log-variance and m is the log-mean. Note that $\exp(\sqrt{0.4805}) \approx 2$ and $\exp(\sqrt{7.6872}) \approx 16$.

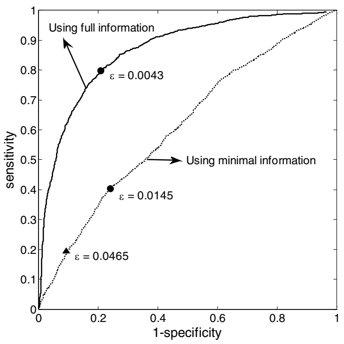

Figure 1. ROC curve of MCMC-based perturbations identification using full information (solid line) and using minimal information (dashed line). Circles – maximum accuracy using the weighting ratio (c_1:c_2) of 1:1; Triangle – maximum accuracy using the weighting ratio (c_1:c_2) of 2:3. The area under the ROC curve for full information is 0.8357 and for minimal information is 0.6114.

Following the method described in Section 2.1, MCMC is used to draw samples from the target distribution $p_{target}(w)$ in (5). This implies that the proposed move is defined in the space of parameters θ, external observable metabolite concentrations and fluxes y_{ext}, and initial metabolite concentrations x_0. We use lognormal 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_{ext} *viz.* the measured glycogen and trehalose fluxes.

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

Parameters	Case A	Case B	Case C	Case D	Case E
$V^1_{\max}$	101.3	81.3	121.3	110.75	120.81
$V^2_{\max}$	670.5	570.5	770.5	340.83	690.77
$V^3_{\max}$	1933	1633	2433	200.29	1270.64
$V^4_{\max}$	121.5	101.5	181.5	552.06	284.36
$V^5_{\max}$	101	81	161	161.25	384.52
$V^{7,f}_{\max}$	2336	1336	3336	2074.87	1712.2
$V^{7,r}_{\max}$	3298	2298	4298	3333.92	4936.26
$V^8_{\max}$	2291	1291	2991	958.51	819.58
$V^9_{\max}$	2423	1423	3223	2278.63	1494.45
$V^{10}_{\max}$	240.4	180.4	290.4	238.36	297.29
$V^{11}_{\max}$	700.5	650.5	790.5	790.02	816.08
$V^{12}_{\max}$	869.9	809.9	969.9	1448.25	5939.07
$V^{13}_{\max}$	50.2	40.2	70.2	113.76	781.56
$V^{14}_{\max}$	47.2	37.2	57.2	53.91	94.63

The unit for $V_{\max}$ is mmol(l-internal vol)$^{-1}$min^{-1}. The limiting step constant $V^1_{\max}$ corresponds to glucose transport, $V^2_{\max}$ corresponds to hexokinase, $V^3_{\max}$ corresponds to phosphogluco isomerase, $V^4_{\max}$ corresponds to phosphofructokinase, $V^5_{\max}$ corresponds to fructose-1,6-biphosphate aldolase, $V^{7,f}_{\max}$ and $V^{7,r}_{\max}$ correspond to the forward and reverse reaction of D-glyceraldehyde-3-phosphate dehydrogenase, $V^8_{\max}$ corresponds to phosphoglycerate kinase, $V^9_{\max}$ corresponds to phosphoglycerate mutase, $V^{10}_{\max}$ corresponds to phosphopyruvate hydratase, $V^{11}_{\max}$ corresponds to pyruvate kinase, $V^{12}_{\max}$ corresponds to pyruvate decarboxylase, $V^{13}_{\max}$ corresponds to the reverse reaction of alcohol dehydrogenase and $V^{14}_{\max}$ corresponds to glycerol 3-phosphate dehydrogenase.

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.

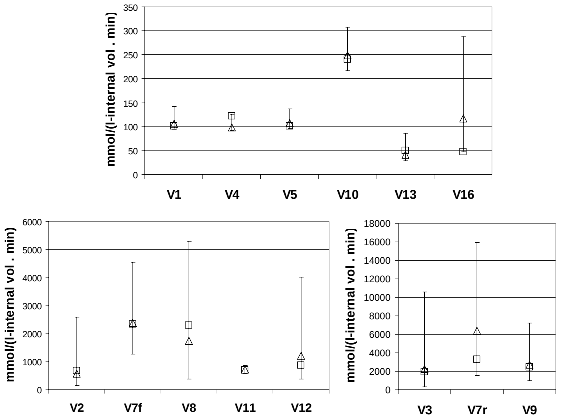

Figure 2. Parameter estimation and credible interval in Case A using MCMC. (•) - the target $V_{\max}$ in Case A; (△) - the estimated parameter using the mean of MCMC samples. The credible interval approximation uses the 2.5 and 97.5 percentile sample.

A permutation of sixteen 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 ε 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 sixteen observation datasets and use the capital letter to indicate each case. The first data set (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 $V_{\max}$ values for all five cases are summarized in Table 1. Details of all sixteen 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 $V^i_{\max}$ 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 thirty measurements) with the posterior predictive distribution (i.e. simulated output samples y_{int} using the parameters from MCMC samples).

The classification algorithm presented in the previous section is used to show whether $V_{\max}$ in the mutant case is higher, relatively constant or lower than that in wild-type case. Using the optimal ε = 0.0043, the resulting classifications are shown in Figure 4.

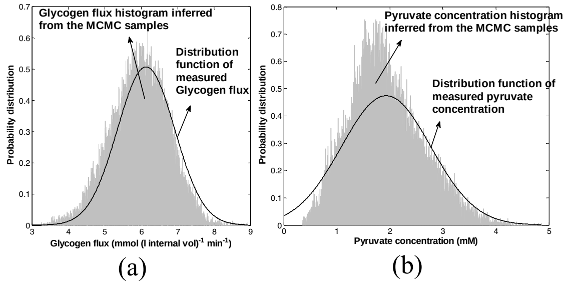

Figure 3. Plot of the inferred metabolite fluxes and concentrations based on the MCMC samples with full measurement data. (a). Glycogen flux; (b). Pyruvate concentration.

Figure 4(a) shows that thirteen out of fourteen perturbations are correctly predicted by the classification method. When all $V_{\max}$ are up-regulated (Case C), the classification method is able to correctly identify ten parameter changes with four false negatives. In Cases D and E, where some $V_{\max}$ are increased and some are decreased, the technique yields two and no false positives, respectively. These results show the efficacy of MCMC for identifying mode-of-action given steady-state data of realistic quality.

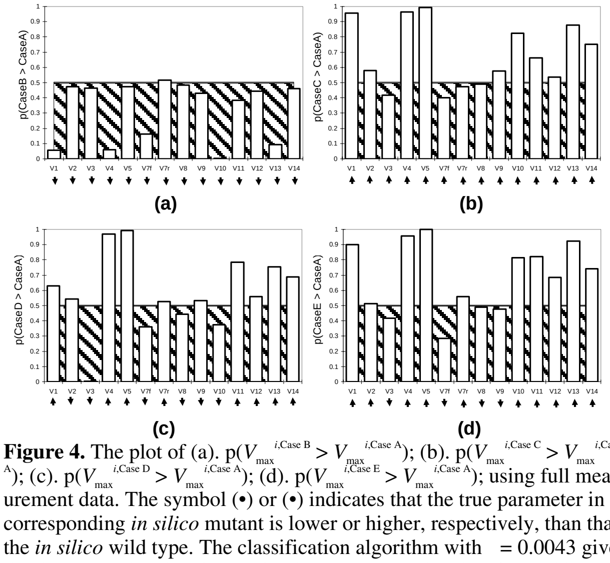

Figure 4. The plot of (a). p($V_{\max}^{i,\text{Case B}} > V_{\max}^{i,\text{Case A}}$); (b). p($V_{\max}^{i,\text{Case C}} > V_{\max}^{i,\text{Case A}}$); (c). p($V_{\max}^{i,\text{Case D}} > V_{\max}^{i,\text{Case A}}$); (d). p($V_{\max}^{i,\text{Case E}} > V_{\max}^{i,\text{Case A}}$); using full measurement data. The symbol (•) or (•) indicates that the true parameter in the corresponding *in silico* mutant is lower or higher, respectively, than that in the *in silico* wild type. The classification algorithm with = 0.0043 gives the down-regulated region (striped area), up-regulated region (white area) and unchanged region (thick line). The label *Vi* in the figure corresponds to the limiting-step constant $V_{\max}^{i}$ (see also Table 1).

We now repeat the whole experiment but with limited amounts of measurement data. Instead of having full measurement of metabolite concentrations and fluxes, we seek to infer the parameters based only on the measurement data of the fluxes of glucose, glycerol, succinate and pyruvate.

It is found that the credible intervals obtained from the experiments using minimal informations is typically larger than that obtained from full measurements (data are available 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 5(a). 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.

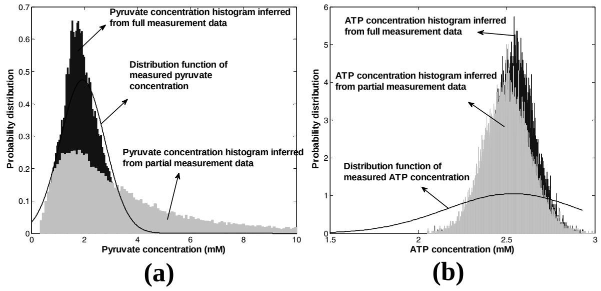

Figure 5. Plots of the inferred measurement based on the MCMC samples using partial and full information. (a). Pyruvate concentration; (b). ATP concentration.

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 ε is computed using the ratio c_1:c_2 of 2:3 which puts higher weight on specificity. Based on this ratio, the optimal ε 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 *Lactococcus 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 $v_{14} = V^{14}_{\max}$ [*pyruvate*]/(K_m+[*pyruvate*]) where K_m is 0.31 mM (Sueda *et al*., 2004).

The parameters are the fourteen limiting step constants $V_{\max}$ corresponding to fourteen reactions in the model and the external concentrations of glucose, lactate, acetoin, O_2, phosphate, ethanol and butanediol. The prior distributions for $V_{\max}$ is uniform distributions defined on (0,20000). The prior distribution 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 lognormal 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.

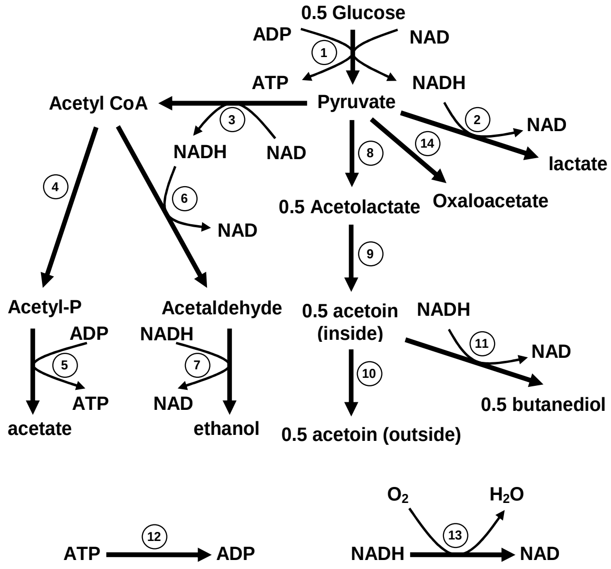

Figure 6. Pyruvate pathways model in *L. lactis* based on the model used in (Hoefnagel, *et al*., 2002) with an additional reaction of pyruvate carboxylase. Numbers in circles represents reactions: 1. 'lumped glycolysis'; 2. Lactate dehydrogenase; 3. pyruvate dehydrogenase; 4. phosphotransacetylase; 5. acetate kinase; 6. acetaldehyde dehydrogenase; 7. alcohol dehydrogenase; 8. Acetolactate synthase; 9. acetolactate decarboxylase and non-enzymic acetolactate decarboxylation; 10. acetoin efflux; 11. acetoin dehydrogenase; 12. ATPase; 13. NADH oxidase; 14. pyruvate carboxylase.

The conserved variables are the conserved moieties of adenine and pyridine nucleotides and the conservation of [Acetyl-CoA] + [CoA]. Their initial concentrations are parametrized 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_{ext}. The data are summarized as Gaussian distributions with the mean values set to those measured in Hoefnagel *et al*. (2002) and the standard deviations set to approximately ten percent of the mean values (following the observation by de Koning and van Dam (1992) that metabolites measurements of this type have standard errors of approximately ten percent). 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.

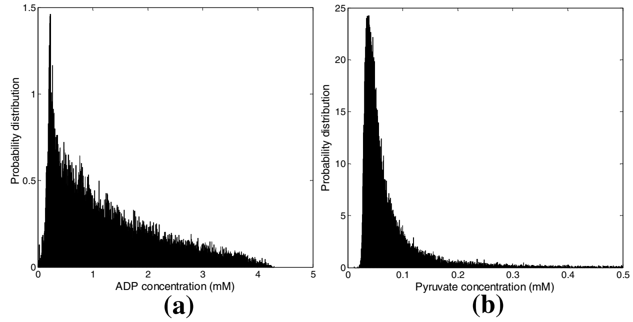

Figure 7. Inferred steady-state concentrations of (a) ADP; and (b) pyruvate; from the simulated output using the resulting MCMC samples in the wild-type case.

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_{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.

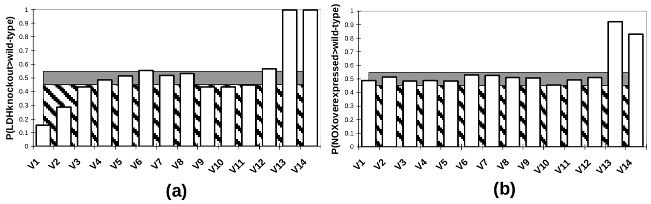

Figure 8. Perturbation identification on the pyruvate pathways in *L. Lactis* using MCMC. (a). The plot of p($V_{\max}^{i,\text{LDHknockout}} > V_{\max}^{i,\text{Normal}}$); (b). The plot of p($V_{\max}^{i,\text{NOXoverexpressed}} > V_{\max}^{i,\text{Normal}}$). The classification algorithm with = 0.0465 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.

The result shown in Figure 8(a) 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 8(b) shows that the overexpression 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.

DISCUSSION

This paper focuses on the application of MCMC to parameter estimation and perturbation analysis in metabolic pathways. Compared to the existing literature, the technique developed in this paper has three distinguishing features. First, the systems are described by first-order nonlinear 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 nonlinear 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 data set 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 (see 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 ε 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 *Lactococcus 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.

FUNDING

The UK Biotechnology and Biological Sciences Research Council (BB/E016065).

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.

REFERENCES

Aranibar, N., Singh, B.K., Stockton, G.W. and Ott, K.H. (2001) Automated mode-of-action detection by metabolic profiling, Biochem Biophys Res Commun, 286, 150-155.

Baetz, K. *et al*. (2004) Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action, Proc Natl Acad Sci U S A, 101, 4525-4530.

Battogtokh, D. *et al*. (2002) An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa, Proc Natl Acad Sci U S A, 99, 16904-16909.

Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A. and Duncan, K. (2002) Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling, Mol Microbiol, 43, 717-731.

Broadhurst, D. and Kell, D.B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics, 2, 171-196.

Butcher, E.C. (2005) Can cell systems biology rescue drug discovery?, Nat Rev Drug Discov, 4, 461-467.

Chapal, N. *et al*. (2004) Pharmacoproteomic approach to the study of drug mode of action, toxicity, and resistance: applications in diabetes and cancer, Fundamental & Clinical Pharmacology, 18, 413-422.

Clarke, P.A., Poele, R.T., Wooster, R. and Workman, P. (2001) Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential, Biochemical Pharmacology, 62, 1311-1336.

Coleman, M.C. and Block, D.E. (2006) Bayesian parameter estimation with informative priors for nonlinear systems, Aiche Journal, 52, 651-667.

Davey, H.M. and Kell, D.B. (1996) Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analysis, Microbiol. Rev., 60, 641-696.

de Koning, W. and van Dam, K. (1992) A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH, Anal Biochem, 204, 118-123.

Dobson, P. D. & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? Nat Rev Drug Discov, in press.

Gelman, A. (2004) Bayesian data analysis. Chapman & Hall/CRC, Boca Raton, Fla. ; London.

Giaever, G. *et al*. (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency, Nature Genetics, 21, 278-283.

Hoefnagel, M.H.N. *et al*. (2002) Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis, Microbiology-Sgm, 148, 1003-1013.

Hofmeyr, J.H., Kacser, H. and van der Merwe, K.J. (1986) Metabolic control analysis of moiety-conserved cycles, Eur J Biochem, 155, 631-641.

Kell, D.B. (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development, Drug Discovery Today, 11, 1085-1092.

Liebermeister, W. and Klipp, E. (2006) Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data, Theor Biol Med Mod, 3, 42.

Liu, X.J., Milo, M., Lawrence, N.D. and Rattray, M. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression, Bioinformatics, 22, 2107-2113.

Lüdtke, N. *et al*. (2008) Information-theoretic Sensitivity Analysis: a general method for credit assignment in complex networks J Roy Soc Interface, 5, 223-235.

Marton, M.J. *et al*. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays, Nat Med, 4, 1293-1301.

Ott, K.H., Aranibar, N., Singh, B. and Stockton, G.W. (2003) Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts, Phytochemistry, 62, 971-985.

Parsons, A.B. *et al*. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways, Nature Biotechnology, 22, 62-69.

Parsons, A.B. et.al. (2006) Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast, Cell, 126, 611-625.

Pritchard, L. and Kell, D.B. (2002) Schemes of flux control in a model of Saccharomyces cerevisiae glycolysis, European Journal of Biochemistry, 269, 3894-3904.

Rojas, I. *et al*. (2007) SABIO-RK: a database for biochemical reactions and their kinetics, BMC Systems Biology, 1, S6.

Spiegelhalter, W.R., Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London.

Sueda, S., Islam, M.N. and Kondo, H. (2004) Protein engineering of pyruvate carboxylase: investigation on the function of acetyl-CoA and the quaternary structure, Eur J Biochem, 271, 1391-1400.

Tamaddoni-Nezhad, A. *et al*. (2006) Application of abductive ILP to learning metabolic network inhibition from temporal data, Machine Learning, 64, 209-230.

Teusink, B. *et al*. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry, European Journal of Biochemistry, 267, 5313-5329.

Vyshemirsky, V. and Girolami, M.A. (2007) Bayesian Ranking of Biochemical System Models, Bioinformatics.

Williams, M. (2005) Systems and integrative biology as alternative guises for pharmacology: prime time for an iPharm concept?, Biochem Pharmacol, 70, 1707-1716.