BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1892–1898
doi:10.1093/bioinformatics/btu069

Systems biology

Advance Access publication February 26, 2014

Derivative processes for modelling metabolic fluxes

Justina Žurauskienė*, Paul Kirk, Thomas Thorne, John Pinney and Michael Stumpf*

Theoretical Systems Biology Group, Centre for Bioinformatics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: One of the challenging questions in modelling biological systems is to characterize the functional forms of the processes that control and orchestrate molecular and cellular phenotypes. Recently proposed methods for the analysis of metabolic pathways, for example, dynamic flux estimation, can only provide estimates of the underlying fluxes at discrete time points but fail to capture the complete temporal behaviour. To describe the dynamic variation of the fluxes, we additionally require the assumption of specific functional forms that can capture the temporal behaviour. However, it also remains unclear how to address the noise which might be present in experimentally measured metabolite concentrations.

Results: Here we propose a novel approach to modelling metabolic fluxes: derivative processes that are based on multiple-output Gaussian processes (MGPs), which are a flexible non-parametric Bayesian modelling technique. The main advantages that follow from MGPs approach include the natural non-parametric representation of the fluxes and ability to impute the missing data in between the measurements. Our derivative process approach allows us to model changes in metabolite derivative concentrations and to characterize the temporal behaviour of metabolic fluxes from time course data. Because the derivative of a Gaussian process is itself a Gaussian process, we can readily link metabolite concentrations to metabolic fluxes and vice versa. Here we discuss how this can be implemented in an MGP framework and illustrate its application to simple models, including nitrogen metabolism in *Escherichia coli*.

Availability and implementation: R code is available from the authors upon request.

Contact: j.norkunaite@imperial.ac.uk; m.stumpf@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 20, 2013; revised on December 28, 2013; accepted on January 26, 2014

1 INTRODUCTION

It is generally impossible to simultaneously measure the abundance of all the molecular entities making up biological systems. In gene expression assays, for example, we typically measure messenger RNA expression, but not the activity of transcription factors and/or the occupancy of transcription-factor binding sites. Similarly, in metabolomic analyses (Chou and Voit, 2012; Voit, 2013), key metabolites can be measured using, e.g. mass spectrometry or nuclear magnetic resonance quantification, but it is rarely possible to comprehensively quantify the metabolites even within a single pathway. Typically, more interesting than metabolite and enzyme abundance are the fluxes through biochemical reactions and metabolic networks (Orth *et al.*, 2010; Schuster *et al.*, 1999). Fluxes, $v = (v_1, \ldots, v_m)^T$, correspond to the rates at which molecules, $\mathbf{x} = (x_1, \ldots, x_n)^T$, are turned over by the m reactions; regulation of fluxes in light of changes in environmental and physiological conditions is also intimately linked to cellular physiology.

Although the fluxes are of central concern, they are hard to measure directly. Estimates for intracellular fluxes can be obtained by tracking products from isotope-labeled (^{13}C and ^{15}N metabolic flux analysis) metabolites through the metabolic network (Blank and Ebert, 2012; Zamboni, 2011). However, such an approach is restricted to a metabolically steady-state analysis and is not appropriate for capturing dynamical flux variations. Instead, theoretical analysis has often progressed by assuming stationarity of the metabolic processes, which in turn allows for characterizing the sets of steady-state fluxes under a set of suitable assumptions (Klamt and Stelling, 2003; Schwartz and Kanehisa, 2006; Voit and Almeida, 2004). Flux-balance analysis is the most popular example of this strategy, but it becomes questionable once the steady-state assumption can no longer be upheld. Furthermore, as more data on enzyme abundance become available, we should attempt to include such information and the impact on metabolic processes (Colijn *et al*, 2009; Rossell *et al.*, 2013).

Here we provide a new framework that allows us to model metabolic fluxes and their dynamics, and which deals with the missing data problem in metabolic analysis in a flexible and consistent manner. Gaussian processes (GP) belong to the armoury of non–parametric Bayesian methods and have been widely used to describe dynamical processes (Kirk and Stumpf, 2009) and to infer hidden states, e.g. transcription-factor activities (Honkela *et al.*, 2010). In applications to metabolic modelling, parametric approaches can offer potentially incorrect representations of the underlying fluxes (Voit, 2013). The strengths of GP models arise from their non-parametric nature, which enables us to put priors directly on a function rather than on the parameters of a parametric function. With a multiple-output GPs (MGPs), single GP framework can be extended to handle many outputs, enabling us to learn the unknown relationships between metabolic species. In turn, MGPs can be used to infill the sparsely sampled data (Boyle and Frean, 2004). This means that by using MGPs, it is possible to impute the missing data in between the metabolic measurements more efficiently.

Here we develop a more general framework that uses so-called derivative GPs (Solak *et al.*, 2003), which allow us to link

*To whom correspondence should be addressed.

metabolite abundance, $\mathbf{x}$ (or concentrations) and fluxes v. This in turn enables us to also treat time course data on metabolites and monitor the changes that occur in fluxes, e.g. over the course of physiological responses, such as to changes in the environment (Bryant *et al.*, 2013).

2 METHODS

2.1 GP regression

Gaussian process regression (GPR) can be applied to recover an underlying dynamical process from noisy observations. A GP defines a prior distribution over all possible functions, and to specify a GP, we need expressions for the mean and covariance function that describe the behaviour of the system output over time (Haykin and Moher, 2010). Below we review the standard GPR methodology.

In a typical regression problem, we connect inputs $\mathbf{x}$ and outputs $\mathbf{z}$ via functions, $\mathbf{z} = f(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$ are continuous n-dimensional real-valued vectors. The observed values of the dependent variable, $\mathbf{z}$, can be related to the independent variables, $f(\mathbf{x})$ through,

$$y_i = f(x_i) + \epsilon, \quad i = 1, \ldots, n,$$

where ϵ is a noise term, which is here assumed to be independent and identically distributed according to a Gaussian distribution, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. In GPR, we place a GP (Haykin and Moher, 2010; McKay, 1998) prior over the functions $f(\mathbf{x})$, $f \sim \mathcal{GP}$, meaning that at any finite number of input points $x_1, \ldots, x_n$ the values $f(x_i)$ have a multivariate Gaussian distribution with zero mean and covariance function, K,

$$[f(x_1), \ldots, f(x_n)]^T \sim \mathcal{N}(0, K(\mathbf{x}, \mathbf{x}')).$$

Different functional forms can be chosen for the covariance function (Rasmussen and Williams, 2006), either to simplify computations or to reflect constrains imposed by the data. A flexible and generic choice is to set the covariance function to

$$K(\mathbf{x}_p, \mathbf{x}_q) = \sigma_g^2 \exp\left(-\frac{1}{2l}|\mathbf{x}_p - \mathbf{x}_q|^2\right),$$

where $\theta = \left(\sigma_g^2, l\right)$ represent a set of unknown hyper-parameters, and $\mathbf{x}_p$ and $\mathbf{x}_q$ are inputs. Thus, $\mathbf{y} = (y_1, \ldots, y_n)^T$ has a multivariate normal distribution with zero mean and covariance matrix $C(\theta) = K + \sigma_\epsilon^2 \mathbf{I}$, with $\mathbf{I}$ the identity matrix. The unknown set of hyper-parameters, θ, can be estimated from the data by evaluating the following log-likelihood function,

$$\mathcal{L}(\theta) = -\frac{1}{2}\log|C(\theta)| - \frac{1}{2}\mathbf{y}^T C(\theta)^{-1}\mathbf{y} - \frac{n}{2}\log 2\pi, \tag{1}$$

using either a maximum likelihood approach or by sampling from the posterior distribution with Markov chain Monte Carlo methods (Neal, 1997).

For any finite number of input (test) points, $x_1^\star, \ldots, x_r^\star$, we define the joint prior probability distribution

$$[\mathbf{y}, f(x_1^\star), \ldots, f(x_r^\star)]^T \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\mathbf{x}_p, \mathbf{x}_q) + \sigma_\epsilon^2 \mathbf{I} & K(\mathbf{x}_p, \mathbf{x}_q^\star) \\ K(\mathbf{x}_p^\star, \mathbf{x}_q) & K(\mathbf{x}_p^\star, \mathbf{x}_q^\star) \end{pmatrix}\right).$$

With the GP prior, it is possible to evaluate the posterior distribution over the functions; the values of f evaluated at inputs $\left(x_1^\star, \ldots, x_r^\star\right)$ and conditioned on the observations $\mathbf{y}$ are jointly distributed as (Rasmussen and Williams, 2006),

$$[f(x_1^\star), \ldots, f(x_r^\star)]^T | \mathbf{y} \sim \mathcal{N}(m_p, K_p), \tag{2}$$

where

$$m_p = K(\mathbf{x}_p^\star, \mathbf{x}_q)[K(\mathbf{x}_p, \mathbf{x}_q) + \sigma_\epsilon^2 \mathbf{I}]^{-1}\mathbf{y},$$

and

$$K_p = K(\mathbf{x}_p^\star, \mathbf{x}_q^\star) - K(\mathbf{x}_p^\star, \mathbf{x}_q)[K(\mathbf{x}_p, \mathbf{x}_q) + \sigma_\epsilon^2 \mathbf{I}]K(\mathbf{x}_p, \mathbf{x}_q^\star).$$

Although Equation (2) defines an appropriate GP posterior, which allows us to make predictions about a single variable $\mathbf{y}$, it remains unclear how to deal with several variables simultaneously: if outputs are correlated then the standard GPR framework may fail in providing an adequate description.

2.2 Multiple–output GPs

Boyle and Frean (2004) introduced MGPs, where a set of dependent GPs is constructed via multiple-input multiple-output linear filters. This perspective can capture the dependencies between several variables by solving a convolution integral and specifying a suitable covariance function, which in turn includes the cross and auto correlations among related variables. Our construction of derivative processes below builds on MGPs.

Dealing with linear filters is central to signal processing where such filters describe a physical systems that can generate an output signal in response to a given input signal (Haykin and Moher, 2010; Roberts, 2008). Linear filters are characterized by their kernel function (an impulse response) $h(t)$, and the output $z(t)$ can be expressed via convolution integral,

$$z(t) = h(t) \otimes x(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau,$$

where the symbol '$\otimes$' denotes the convolution operator. To transmit the signal that has the mathematical properties of a GP, the kernel function, $h(t)$ must be absolutely integrable, i.e.

$$\int_{-\infty}^{\infty} |h(t)|dt < \infty,$$

Then if the input $X(t)$ is specified to be a Gaussian white noise process, the output process, $Z(t)$, will also be a GP.

Specifying a stable linear time-invariant filter with M white noise processes as inputs, $X_1(t), \ldots, X_M(t)$, K outputs, $Z_1(t), \ldots, Z_K(t)$ and $M \times K$ impulse responses results in a dependent GP model (Boyle and Frean, 2005). A multiple-input multiple-output filter can thus be defined as

$$Z_k(t) = \sum_{m=1}^{M} \int_{-\infty}^{\infty} h_{mk}(\tau)X_m(t-\tau)d\tau,$$

where $h_{mk}(t)$ are kernel functions and $Z_k(t)$ is the kth output. As discussed previously, the observed variables might differ from expected variables owing to the measurement noise, and we thus consider

$$Y_k(t) = Z_k(t) + W_k(t), \tag{3}$$

where $W_k(t)$ is a Gaussian white noise process with variance σ_k^2.

Multiple-input multiple-output filters are able to capture the relationships between several variables $Y_k(t)$; in the model, these kind of dependencies are build in via shared input noise sources that enable the specification of valid covariance functions. For the sake of simplicity, let the impulse response be a Gaussian kernel, $h_{mk}(t) = v_{mk}\exp\{-\frac{1}{2}(t-\mu_{mk})^2 A_{mk}\}$. Then evaluating the convolution integral leads to the following covariance function,

$$\begin{aligned} \mathcal{C}_{ij}(d) &= \sum_{m=1}^{M} \int_{-\infty}^{\infty} h_{mi}(\tau)h_{mj}(\tau + d)d\tau \\ &= \sum_{m=1}^{M} \frac{(2\pi)^{\frac{1}{2}} v_{mi} v_{mj}}{\sqrt{A_{mi} + A_{mj}}} \exp\left\{-\frac{1}{2}(d - [\mu_{mi} - \mu_{mj}])^2 S\right\}, \end{aligned} \tag{4}$$

where $S = A_{mi}(A_{mi} + A_{mj})^{-1}A_{mj}$ and $d = t_a - t_b$ is the temporal separation between two input points, (see Boyle and Frean (2004) appendix for derivation and generalization to multidimensions). Constructing

intermediate matrices C_{ij} permits the definition of a positive definite symmetric covariance matrix $\mathbf{C}$ between K variables,

$$\mathbf{C} = \begin{pmatrix} C_{11} + \sigma_1^2 \mathbf{I} & \dots & C_{1K} \\ \dots & \dots & \dots \\ C_{K1} & \dots & C_{KK} + \sigma_K^2 \mathbf{I} \end{pmatrix}_{[N \times N]}.$$

Here $N = \sum_{i=1}^{K} N_i$ is total number of observations, and N_i the number of observations of variable i. Having defined the covariance matrix, we can use the log-likelihood, which has the form (1) for the inference of the hyper-parameters $\theta = \{v_{mk}, \mu_{mk}, A_{mk}\}$. Again, following Bayesian framework, we can use the results from the GPR section to evaluate the joint predictive distribution (2) for all outputs. Alternatively, for a particular variable i, predictions can be made using the appropriate marginal distribution, which is Gaussian, with mean $\mathbf{m}_i(t')$ and variance $\mathbf{var}_i(t')$, given by

$$\begin{aligned} \mathbf{m}_i(t') &= \mathbf{k}^T \mathbf{C}^{-1} \mathbf{y}, \\ \mathbf{var}_i(t') &= \kappa - \mathbf{k}^T \mathbf{C}^{-1} \mathbf{k}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \kappa &= C_{ii}(0) + \sigma_i^2, \\ \mathbf{k}^T &= [k_1^T, \dots, k_K^T], \\ k_j^T &= [(C_{ij}(t' - t_{j,1}) \cdots C_{ij}(t' - t_{j,N_j})]. \end{aligned}$$

2.3 Derivative processes

For a GP that is derived through a linear filter, $Y(t) = h(t) \otimes X(t) + W(t)$, where $X(t)$ is a white noise GP, $h(t)$ is a kernel function and $W(t)$ is an additive noise, it is easy to formulate the expression of a derivative process. Taking a derivative of Y with respect to t, it is possible to obtain a new process U that is also a GP (Boyle, 2007),

$$U(t) \equiv \frac{d}{dt} Y(t) = \int_{-\infty}^{\infty} \left\{ \frac{d}{dt} h(t - \tau) \right\} X(\tau) d\tau = g(t) \otimes X(t),$$

Thus, it is possible to construct the derivative process by convolving a white noise GP $X(t)$ with a derivative kernel function $g(t)$. This definition enables us to consider derivative processes and the corresponding original processes as a collection of dependent GPs. This is true because the derivative processes and the original processes are derived from exactly the same input, $X(t)$.

To construct a dependent model for several related variables $\mathrm{Y} = (Y_1, \dots, Y_K)$ and their derivatives $\mathrm{U} = (U_1, \dots, U_K)$, it is necessary to define a suitable covariance structure, which in principal arises from the initial covariance function (4). For example, for a set of four dependent outputs (two original and two derivative processes), the following equations can be applied to compute the covariances (Girard, 2004; Kirk, 2011; Solak *et al.*, 2003),

- Autocovariance function of derivative process U_i

$$\mathcal{DDC}_{ii}(d) \equiv cov\left(\left. \frac{dY_i}{dt} \right|_{t=t_a}, \left. \frac{dY_i}{dt} \right|_{t=t_b} \right) = \frac{d^2}{dt_a dt_b} \mathcal{C}_{ii}(d);$$

- Cross-covariance function between two derivative processes U_i and U_j

$$\mathcal{DDC}_{ij}(d) \equiv cov\left(\left. \frac{dY_i}{dt} \right|_{t=t_a}, \left. \frac{dY_j}{dt} \right|_{t=t_b} \right) = \frac{d^2}{dt_a dt_b} \mathcal{C}_{ij}(d);$$

- Covariance between original process Y_i and corresponding derivative process U_i

$$\mathcal{DC}_{ii}(d) \equiv cov\left(Y_i, \left. \frac{dY_i}{dt} \right|_{t=t_b} \right) = \frac{d}{dt_b} \mathcal{C}_{ii}(d);$$

- Covariance between original process Y_i and derivative process U_j

$$\mathcal{DC}_{ij}(d) \equiv cov\left(Y_i, \left. \frac{dY_j}{dt} \right|_{t=t_b} \right) = \frac{d}{dt_b} \mathcal{C}_{ij}(d).$$

Let $\mathbf{R}$ denote a block matrix,

$$\mathbf{R} = \begin{pmatrix} C_{11} & C_{12} & DC_{11} & DC_{12} \\ C_{21} & C_{22} & DC_{21} & DC_{22} \end{pmatrix}, \quad \mathbf{L} = \mathbf{R}^T,$$

which describes the correlations between observations $\mathbf{Y} = (Y_1, Y_2)$ and their 'function' values $\mathbf{Z} = (Z_1, Z_2)$, and corresponding derivative variables $\mathbf{U} = (U_1, U_2)$ evaluated at any finite number of test points $t_1, \dots, t_r$. In a similar manner, let $\mathbf{H}$ denote

$$\mathbf{H} = \begin{pmatrix} \tilde{C}_{11} & \tilde{C}_{12} & \widetilde{DC}_{11} & \widetilde{DC}_{12} \\ \tilde{C}_{21} & \tilde{C}_{22} & \widetilde{DC}_{21} & \widetilde{DC}_{22} \\ \widetilde{DC}_{11} & \widetilde{DC}_{12} & DDC_{11} & DDC_{12} \\ \widetilde{DC}_{21} & \widetilde{DC}_{22} & DDC_{12} & DDC_{21} \end{pmatrix},$$

where the $\tilde{C}_{ij}$ matrices contain the correlations between functions Z_1 and Z_2 evaluated at a finite set of test points $t_1, \dots, t_r$; $\widetilde{DC}_{ij}$ the correlations between functions $\mathbf{Z} = (Z_1, Z_2)$ and derivative variables $\mathbf{U} = (U_1, U_2)$ evaluated at the same test points; and finally, DDC_{ij} consists of auto/cross-correlations between derivative variables U_1 and U_2. The matrices $\mathbf{R}$, $\mathbf{L}$ and $\mathbf{H}$ are building components of the overall covariance matrix $\mathbf{K}$, which is symmetric and positive definite,

$$\mathbf{K} = \begin{pmatrix} \mathbf{C} + \sigma^2 \mathbf{I} & \mathbf{R} \\ \mathbf{L} & \mathbf{H} \end{pmatrix}.$$

At a finite number of input points $t_1, \dots, t_r$, the matrix $\mathbf{K}$ allows us to place a joint prior over observations $\mathbf{Y}$, functions $\mathbf{Z}$ and derivatives $\mathbf{U}$,

$$[\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}_1, \mathbf{U}_2] \sim \mathcal{N}(\mathbf{0}, \mathbf{K}).$$

Evaluating a GP posterior

$$[\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{U}_1, \mathbf{U}_2] | [\mathbf{Y}_1, \mathbf{Y}_2] \sim \mathcal{N}(\mathbf{m}_{post}, \mathbf{K}_{post}), \tag{6}$$

where

$$\mathbf{m}_{post} = \mathbf{L}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{R} \quad \text{and} \quad \mathbf{K}_{post} = \mathbf{H} - \mathbf{L}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{Y},$$

enables us to make joint predictions for the original and derivative processes simultaneously. Alternatively, if there is no need to sample from the posterior process, we can use marginal Gaussian distributions to make predictions for individual output. The marginal distributions for output i and its derivative process at any input point $t^\star$,

$$\begin{aligned} \mathbf{m}_{Y_i}(t^\star) &= \mathbf{k}_{Y_i}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{Y}, \\ \mathbf{m}_{U_i}(t^\star) &= \mathbf{k}_{Z_i}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{Y}, \\ \mathbf{var}_{Y_i}(t^\star) &= \kappa - \mathbf{k}_{Y_i}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{k}_{Y_i}^T, \\ \mathbf{var}_{U_i}(t^\star) &= \eta - \mathbf{k}_{Z_i}\left[C + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{k}_{Z_i}^T, \end{aligned} \tag{7}$$

where $\mathbf{m}_{Y_i}$ is the mean of the original process, $\mathbf{m}_{U_i}$ the mean of the derivative process, $\mathbf{var}_{Y_i}$ the variance of the original process and $\mathbf{var}_{U_i}$ the variance of the derivative process, and furthermore

$$\kappa = \mathcal{C}_{ii}(0) + \sigma_i^2, \qquad \eta = \mathcal{DDC}_{ii}(0)$$

$$\mathbf{k}_{Y_i} = \begin{pmatrix} \mathcal{C}_{i1}(t^\star - t_{1,1}) \\ \dots \\ \mathcal{C}_{i1}(t^\star - t_{1,N_1}) \\ \mathcal{C}_{i2}(t^\star - t_{2,1}) \\ \dots \\ \mathcal{C}_{i2}(t^\star - t_{2,N_2}) \end{pmatrix}, \quad \mathbf{k}_{U_i} = \begin{pmatrix} \mathcal{DC}_{i1}(t^\star - t_{1,1}) \\ \dots \\ \mathcal{DC}_{i1}(t^\star - t_{1,N_1}) \\ \mathcal{DC}_{i2}(t^\star - t_{2,1}) \\ \dots \\ \mathcal{DC}_{i2}(t^\star - t_{2,N_2}) \end{pmatrix}$$

Equations (6) and (7) can easily be extended to make predictions about any number of variables.

3 APPLICATIONS AND RESULTS

To demonstrate the performance of derivative processes, we consider two simulation examples—a system of two oscillating signals and a simple model of linear metabolic pathway—before turning to a more complicated metabolic process and, finally, some real metabolic network data. The derivative processes can be used to address the flux estimation problem from time course data. Here GPs describe the dynamics of metabolites, and the corresponding derivative processes capture the functional forms of the associated fluxes. Below, all examples were implemented using the free statistical computing platform *R* www.r-project.org.

3.1 Oscillating signals

A simple oscillating signal can be expressed as $z(t) = A\sin(\omega t + \phi)$, where A is the amplitude, $\omega = 2\pi f$ the angular frequency and ϕ the phase angle. This is a particularly useful example because it is easy to evaluate the performance of derivative processes, as the derivative signals have a known analytic form. We consider a simple system that consists of two oscillating signals, $z_1(t)$ and $z_2(t)$,

$$z_1(t) = \sin(2t), \quad \Rightarrow \quad \dot{z}_1(t) = 2\cos(2t),$$
$$z_2(t) = \sin\left(2t + \frac{\pi}{4}\right), \Rightarrow \quad \dot{z}_2(t), = 2\cos\left(2t + \frac{\pi}{4}\right),$$

with $t \in [0, 4\pi]$. To model real experimental measurements, we add random noise to the simulated trajectories, $Y_1(t) = z_1(t) + \epsilon_1$, $Y_1(t) = z_2(t) + \epsilon_2$, where $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$; we have observations of both signals at regular time intervals, $D_1 = \{t_{1,i}, Y_{1,i}\}_{i=1}^{N_1=10}$ and $D_2 = \{t_{2,j}, Y_{2,j}\}_{j=1}^{N_2} = 10$. To build a single model that captures the relationship between the two signals, we apply the dependent GP framework (3) (K = 2) on a combined dataset $\mathbf{D} = \{D_1, D_2\}$; each signal can be expressed as a superposition of three GPs—two of which are constructed via convolution between a noise source and a Gaussian kernel, and the third one is an additive noise. We set parameters A_i of each Gaussian kernel to be $\exp(f_i)$ and noise levels to $\sigma_1 = \exp(\eta_1)$, $\sigma_2 = \exp(\eta_2)$, leading to a set of hyper-parameters $\theta = (v_i, f_i, \mu_1, \mu_2, \eta_1, \eta_2)$, $i = 1, \ldots, 4$. To build the model the following priors are chosen: $v_i, f_i \sim (1, 2^2)$, $\eta_j \sim \mathcal{N}(-2, 2^2)$ and $\mu_j \sim \mathcal{N}(0.5, 1^2)$, $j = 1, 2$; the maximum a posteriori (MAP) estimate $\hat{\theta}$ is determined using a multistarting Nelder–Mead optimization algorithm (Nelder and Mead, 1965). Dependent GP posteriors (6) allow us to make joint predictions about both signals and their derivative processes at any finite number of input points, and the resulting posterior processes are summarized in Figure 1. From these posterior processes, it can be seen that the mean behaviour of our model agrees with trajectories of underlying noiseless signals, and to make predictions about derivative processes, it is enough to consider only samples from the original sinusoidal trajectories.

3.2 Linear pathway

Next we consider a linear metabolic pathway with two regulatory signals (see Goel *et al.* (2008) Supplementary Material for details), which is summarized in Figure 3a. Here the flow from x_1 to x_2 is negatively regulated by metabolite x_3, and x_3 increases

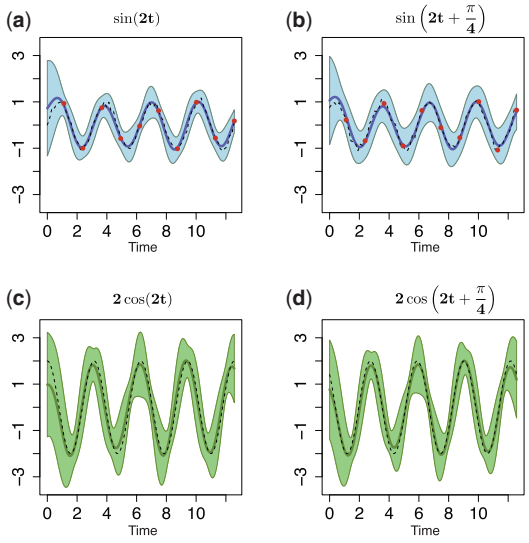

Fig. 1. Predictions with MGPs model for two oscillating signals. (**a** and **b**) Dashed lines represent true behaviour of noiseless *sin*(·) trajectories; dots correspond to the noisy observations for both signals (data); solid lines are the mean behaviour of the MGPs model (predictions with original GPs); light areas correspond to two standard deviations at each prediction point. (**c** and **d**) Dashed lines represent true behaviour of noiseless *cos*(·) trajectories; solid lines show the mean behaviour of the MGPs model (predictions with derivative processes); light areas correspond to two standard deviations at each prediction point

the transformation of x_2 into x_3. A set of ordinary differential equations (ODEs) can be used to describe the dynamics of these two metabolites, x_2 and x_3 (x_1 is the constant external input),

$$\dot{x}_2 = \frac{x_1 V_{max}}{K_m(1 + \frac{x_3}{K_i}) + x_1} - x_2^{0.5} x_3,$$
$$\dot{x}_3 = x_2^{0.5} x_3 - x_3^{0.5}. \tag{8}$$

To apply the derivative process approach, we simulate the ODE model with the following parameter values $(V_{max}, K_m, K_i) = (18.6819, 9.7821, 0.5992)$ and initial conditions $x_2(0) = 1$, $x_3(0) = 1$. In this model, the concentration of x_1 is assumed to be constant and equal to 2. The dataset consists of selected points from simulated trajectories with added Gaussian noise $\mathcal{N}(0, 0.05^2)$. Again we combine the 'noisy' measurements, and fit the dependent GP model to make predictions about the original trajectories and their derivatives. To obtain a functional expressions for fluxes v_1 and v_2 we need to estimate a dynamical variations of metabolic, x_2, x_3, derivatives. The derivative processes provide the predictions for the left side of Equation (8) at any finite number of time points, whereas the original GPs describe the solution on the same ODE (8). This enables us to link the metabolite measurements to metabolic fluxes. Figure 2 illustrates the predictions with posterior processes, where solid blue lines correspond to the mean behaviour of the model, dashed lines to the original x_2 and x_3 trajectories and solid green lines to their derivatives. In addition, if we assume that we are able to measure flux $v_3 = x_3^{0.5}$, we can obtain the functional expressions for fluxes v_1 and v_2 that are summarized in Figure 2c and d. The dark pink lines illustrate predicted fluxes from noisy metabolite

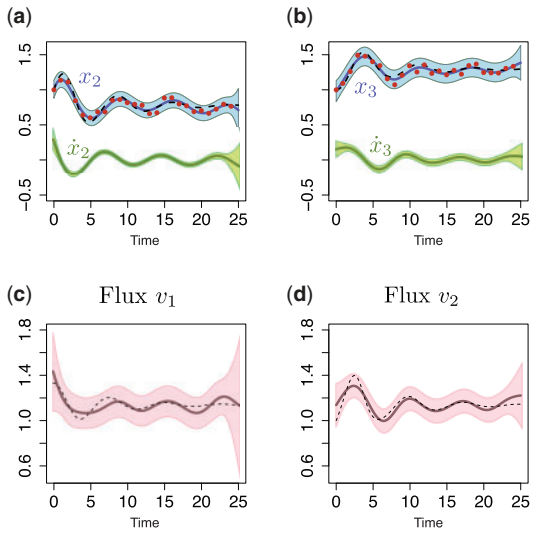

Fig. 2. Predictions with MGPs model for linear metabolic pathway. (**a** and **b**) Dashed lines represent a simulated x_2 and x_3 trajectories from ODE model; dots correspond to the sparse noisy observations for x_2 and x_3 (data); solid blue/green lines are the mean behaviour of the MGPs model (blue, predictions with original GPs; green, predictions with derivative process); light areas correspond to two standard deviations at each prediction point. (**c** and **d**) Dark lines are predicted fluxes, light areas correspond to the confidence region, and dashed lines represent true behaviour of noise-free fluxes v_1 and v_2 (calculated from ODE system)

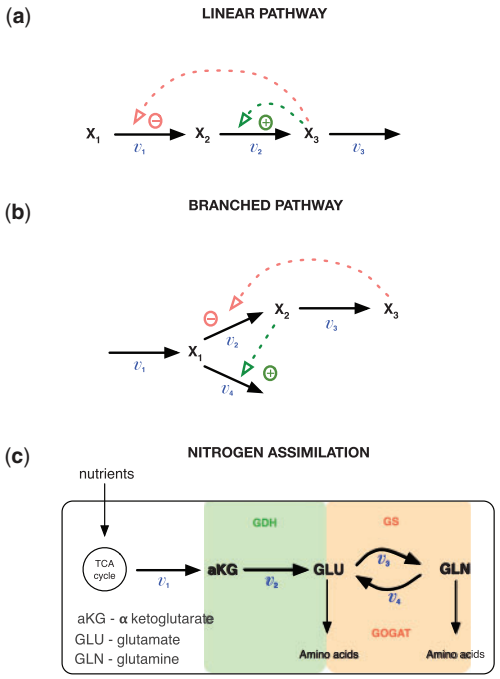

Fig. 3. Pathway information. (**a**) A simple linear metabolic pathway; red and green dashed lines correspond to the inhibition and activation signals. (**b**) Illustrates a branched pathway with positive (green) and negative (red) regulatory signals. (**c**) Illustrates a metabolic pathway in *E.coli*, here v_i, $i = 1...4$ denote the fluxes; αKG, GLU and GLN correspond to the metabolites; TCA is a short notation for the citrate cycle in *E.coli*

measurements, dashed lines are real fluxes (calculated from ODEs (8)) and light pink area corresponds to the confidence region.

3.3 Branched pathway

We now turn to an example of metabolic pathway that was originally proposed by Voit (2013) (see *Example of actual characterization*); Figure 3b illustrates a schematic representation of a branched pathway with two regulatory responses, where x_3 inhibits the conversions of x_1 into x_2, and x_2 positively regulates reaction v_4. The following ODE model describes the dynamics of the metabolites that are involved in this pathway,

$$\begin{aligned}\dot{x}_1 &= 0.05 - 1.1x_1^{0.5}x_3^{-0.75} - 2.8x_1^{0.8}x_2^{0.4},\\ \dot{x}_2 &= 1.1x_1^{0.5}x_3^{-0.75} - 1.1x_2^{0.6},\\ \dot{x}_3 &= 1.1x_2^{0.6},\end{aligned} \tag{9}$$

where x_1, x_2, x_3 denote the metabolites. For a given pathway (Fig. 3b), the change in metabolite concentration can be described by the differences between incoming and outgoing fluxes. For this reason, we are able to obtain the following expressions for fluxes v_1, v_2, v_3 and v_4,

$$\begin{aligned}\dot{x}_1 &= v_1 - v_2 - v_4, \qquad & v_1 - v_4 &= \dot{x}_1 + v_2,\\ \dot{x}_2 &= v_2 - v_3, \qquad \Rightarrow & v_2 &= \dot{x}_2 + \dot{x}_3,\\ \dot{x}_3 &= v_3, & v_3 &= \dot{x}_3.\end{aligned} \tag{10}$$

These expressions define a system of linear equations that is underdetermined, as we have more fluxes to estimate than available equations, and it cannot be solved using standard Gaussian elimination techniques. For this reason, additional information is required to uniquely determine fluxes v_1 and v_4. In this example, we will focus only on estimation of fluxes v_2 and v_3 from available data rather than try to address a uniqueness problem of v_1 and v_4.

The above ODE model enables us to generate simulated time course data using the initial conditions $x_1(0) = 4$, $x_2(0) = 1$ and $x_3(0) = 2$. Next, we apply the dependent GP framework (3) ($K = 2$) on the combined dataset $\mathbf{D} = \{D_1, D_2\}$, where $D_1 = \{t_{2,i}, x_{2,i}\}_{i=1}^{N_1=20}$ and $D_2 = \{t_{3,i}, x_{3,i}\}_{i=1}^{N_1=20}$ contains the measurements of metabolites x_2 and x_3 with added random Gaussian noise $\mathcal{N}(0, 0.01^2)$ (we chose a low noise level so that predictions with derivative processes could be easily compared with the original fluxes in the example in Voit (2013). For a set of model hyper-parameters $\theta = (v_i, f_i, \eta_1, \eta_2, \mu)$, $i = 1, \ldots, 4$ we use the following priors, $v_i \sim (2, 2^2)$, $f_i \sim (-3, 2^2)$, $\eta_j \sim \mathcal{N}(-2, 2^2)$, $j = 1, 2$ and $\mu \sim \mathcal{N}(0.5, 1^2)$, and calculate the MAP estimate $\hat{\theta}$ as before. Figure 4 illustrates the predictions with posterior processes using Equation (7); (a and b) graphs summarize metabolite data. The dark blue lines correspond to the mean behaviour of the original GPs and agree well with simulated x_2 and x_3 dynamics; the green lines describe the derivatives of the same metabolites and can be understood as a slope estimates. In Figure 4c and d, dark pink lines illustrate the predicted metabolic fluxes v_2 and v_3 under consideration of pathway Figure 3b. From ODE model (9), we can calculate original fluxes over the time (in real situations this

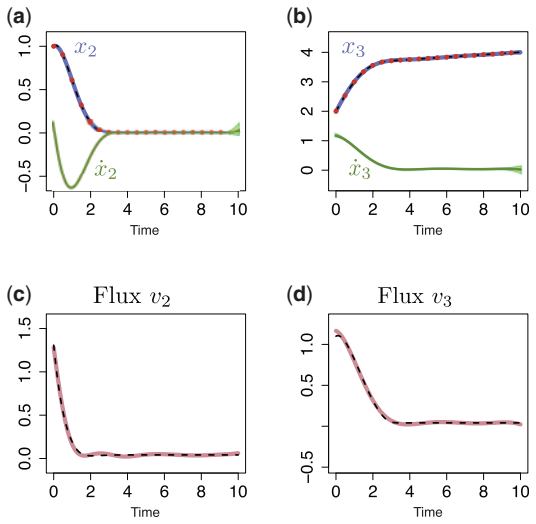

Fig. 4. Predictions with MGP model for a branched metabolic pathway. (**a** and **b**) Dashed lines represent simulated x_2 and x_3 trajectories from the ODE model; red dots correspond to the sparse observations for x_2 and x_3 (data); solid lines are the mean behaviour of the MGPs model (blue, predictions with original GPs; green, predictions with derivative process); light areas correspond to two standard deviations at each prediction point. (**c** and **d**) Dark lines are predicted fluxes; dashed lines represent true behaviour of fluxes v_2 and v_3 (calculated from the ODE system)

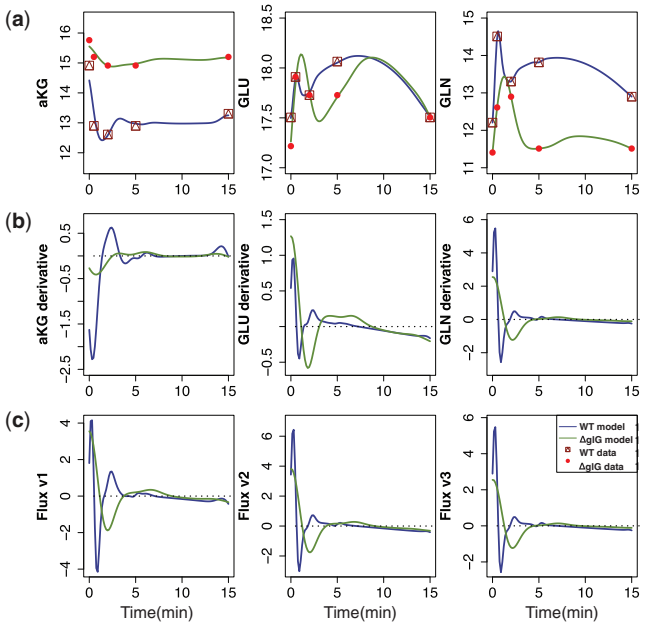

Fig. 5. Predictions with MGPs model for *E. coli* (WT and ΔglnG). (**a**) The symbols indicate experimentally measured concentrations of *αKG*, *GLU* and *GLN* metabolites (dots for WT, squares for ΔglnG). Solid lines correspond to the mean behaviour of dependent GPs model. (**b**) Predicted derivative behaviour for *αKG*, *GLU* and *GLN* metabolites, where solid lines correspond to the mean behaviour of dependent derivative processes. (**c**) Predicted fluxes v_1, v_2 and v_3 for convenience, dotted line illustrates horizontal 0-axis

would not be possible). Figure 4c and d shows a good agreement between predicted and original fluxes.

3.4 *Escherichia coli* nitrogen assimilation

Finally, we apply our technique to the experimental data from *E. coli*, where we have measurements of the abundance of several key metabolites involved in nitrogen assimilation. Nitrogen is one of the key chemical elements that acts as a nutrient for the cells; ammonium is a preferred source of nitrogen for *E. coli* growth (Schumacher *et al.*, 2013; van Heeswijk *et al.*, 2013). In *E. coli*, ammonium can be absorbed via two pathways: glutamate dehydrogenase (GDH) that operates during cell growth in ammonium-rich environments, and glutamine synthetase-glutamate synthase (GS-GOGAT) that operates during cell growth in low-ammonium conditions (van Heeswijk *et al.*, 2013). Here, we are focussing on experimental conditions, where after a period of nitrogen starvation, the bacterial cultures are spiked with ammonium (Schumacher *et al.*, 2013); Figure 5a shows experimentally obtained measurements for *α-ketoglutarade* (*αKG*), *glutamate* (*GLU*) and *glutamine* (GLN) metabolites over the time after ammonium spike; red dots correspond to a wild-type (WT) *E. coli* metabolic measurements, and in squares—isogenic glnG deletion (ΔglnG) measurements. Below we focus on the pathway summarized in Figure 3c, which includes both GDH and GS-GOGAT. For modelling purposes, we assume that fluxes v_3 and v_4 can be summarized by the overall flux v_3 that describes the flow from GLU to GLN, as there is not enough information to discriminate between them. From the pathway, we can construct a system of linear equations that describe the dependencies between fluxes and metabolites,

$$\begin{aligned} \alpha\dot{K}G &= v_1 - v_2, & v_1 &= \alpha\dot{K}G + G\dot{L}U + G\dot{L}N, \\ G\dot{L}U &= v_2 - v_3, \quad \Rightarrow & v_2 &= G\dot{L}U + G\dot{L}N, \\ G\dot{L}N &= v_3, & v_3 &= G\dot{L}N. \end{aligned} \tag{11}$$

We fit a dependent GP model (3) (K = 3) to the WT data and then to ΔglnG data (collected from a strain where glnG is absent). In the model, *αKG* is expressed as a sum of three GPs: the first GP describes *αKG*, the second expresses the relationship between *αKG* and *GLU* and the third one describes additive noise; *GLN* is modelled similarly. However, *GLU* is modelled as the sum of four GPs, where the first three describe *GLU*; the dependence between *GLU* and *αKG*; the dependence between *GLU* and *GLN*; and the fourth is an additive noise. Choosing kernel functions to be Gaussian $h_k(t) = v_k \exp\{-\frac{1}{2}t^2 A_k\}$, we obtain the MAP estimate for all hyper-parameters (17 in total). The predictions with posterior process (7) are summarized in Figure 5, where solid blue lines describe predictions with dependent GP models for WT *E. coli*, and green lines for ΔglnG. Using the relationship (11), we can estimate fluxes v_1, v_2 and v_3 (Fig. 5c).

To evaluate our predictions, we can compare flux v_3 and GS protein levels in WT and ΔglnG *E. coli* (see Supplementary Fig. S1). In *E. coli*, glnG encodes the transcription factor, NtrC (nitrogen regulator) that controls GS expression levels, and in its active form, GS catalyses glutamine synthesis (van Heeswijk

et al., 2013). Experimentally, it was observed that in ΔglnG case protein, GS levels were significantly lower compared with the GS levels in WT *E. coli* (see Supplementary Fig. S1C and D). Because there is less enzyme available to catalyse the reaction in ΔglnG, the flux v_3 in the mutant will be noticeably reduced compared with the WT flux v_3 (see Supplementary Fig. S1A and B).

4 DISCUSSION AND CONCLUSIONS

Flux estimation has become central to many analyses into the metabolic processes and mechanisms. Typically, the estimates for a set of fluxes are obtained in a point-wise manner at discrete time points. It is clear that this fails to capture the temporal behaviour of the fluxes and additional consideration of parametric models is compulsory to fully explain the fluxes; further, this approach is susceptible to noise that is present in experimentally measured metabolite data.

Here we have addressed these problems and proposed a novel non-parametric Bayesian approach to modelling metabolic fluxes. This is based on MGPs that enable the construction of derivative processes. Because the derivative processes and original processes share the same input source, we can complement the dependent GP model and make joint predictions about original and derivative processes at any finite number of input points. Such derivative processes can be applied to characterize the temporal behaviour of metabolic fluxes from time course data—without having to make reference, e.g. transcriptomic data, to explain temporal variation—and here we have demonstrated the applicability on simple models and a real-world example.

GPs, including our approach, propagate uncertainty in line with the assumed covariance structures. This can lead to large confidence intervals, especially if the dependencies among different observations are not considered explicitly. With increasing number of metabolic species within the pathway, the derivative process approach might become computationally costly due to the inference of a large number of hyper-parameters and a matrix inversion step; however, this limitation potentially might be addressed by considering a sparse approximation for the full covariance matrix of all metabolic species (Alvarez and Lawrence, 2009). These can in principle deal with genome-level data.

ACKNOWLEDGEMENT

The authors thank Jake Bundy and Volker Behrends for the *E. coli* metabolite data.

Funding: Leverhulme Trust (to J.Ž. and M.P.H.S.), the Royal Society (to J.P. and M.P.H.S.), HFSP (to P.K. and M.P.H.S.) and BBSRC (to T.T. and M.P.H.S.).

Conflict of Interest: none declared.

REFERENCES

Alvarez,A.M. and Lawrence,D.N. (2009) Sparse convolved Gaussian processes for multi-output regression. *Adv. Neural Inf. Process. Syst*., **21**, 57–64.

Blank,L.M. and Ebert,B.E. (2012) From measurement to implementation of metabolic fluxes. *Curr. Opin. Biotechnol*., **24**, 13–21.

Boyle,P. (2007) Gaussian processes for regression and optimisation. Doctoral dissertation. Victoria University of Wellington.

Boyle,P. and Frean,M. (2004) Multiple-output Gaussian process regression. In: *Technical report*. Victoria University of Wellington.

Boyle,P. and Frean,M. (2005) Dependent Gaussian processes. *Adv. Neural Inf. Process. Syst*., **17**, 217–224.

Bryant,W.A. *et al*. (2013) Analysis of metabolic evolution in bacteria using whole-genome metabolic models. *J. Comp. Biol*., **20**, 755–764.

Chou,I-C. and Voit,E.O. (2012) Estimation of dynamic flux profiles from metabolic time series data. *BMC Syst. Biol*., **6**, 84.

Colijn,C. *et al*. (2009) Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol*., **5**, 1–14.

Goel,G. *et al*. (2008) System estimation from metabolic time-series data. *Bioinfromatics*, **24**, 2505–2511.

Girard,A. (2004) Approximate methods for propagation of uncertainty with Gaussian process models. Doctoral dissertation. University of Glasgow.

Haykin,S. and Moher,M. (2010) *Communication Systems*. 5th edn. Wiley, Asia.

van Heeswijk,W.C. *et al*. (2013) Nitrogen assimilation in *Escherichia coli*: puttin molecular data into a systems perspective. *Microbiol. Mol. Rev*., **77**, 628–695.

Honkela,A. *et al*. (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.

Jia,G. *et al*. (2011) Parameter estimation of kinetic models from metabolic profiles: two-phase dynamic decoupling method. *Bioinformatics*, **27**, 1964–1970.

Kirk,P. (2011) Inferential stability in systems biology. Doctoral dissertation. Imperial College London.

Kirk,P. and Stumpf,M.P.H. (2009) Gaussian process regression bootstrapping. *Bioinformatics*, **25**, 1300–1306.

Klamt,S. and Stelling,J. (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol*., **21**, 64–69.

McKay,D.J.C. (1998) Introduction to Gassign processes. In: Bishop,C.M. (ed.) *Neural Networks and Machine Learning. NATO ASI Series*. Springer-Verlag, pp. 133–165.

Neal,R.M. (1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Arxiv Preprint Physics/9701026*, Technical report 9702, Department of Statistics, University of Toronto.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Comp. J*., **199**, 133–154.

Orth,J.D. *et al*. (2010) What is flux balance analysis? *Nat. Biotechnol*., **28**, 245–248.

Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning*. 1st edn. The MIT Press, Cambridge.

Roberts,M.J. (2008) *Fundamentals of Signals and Systems*. 1st edn. Mc Graw Hill, New York.

Rossell,S. *et al*. (2013) Inferring metabolic state in uncharacterized environments using gene-expression measurements. *PLoS Comput. Biol*., **9**, 1–11.

Schwartz,J-M. and Kanehisa,M. (2006) Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*, **7**, 186.

Solak,E. *et al*. (2003) Derivative observations in Gaussian process models of dynamic systems. *Adv. Neural Inf. Process. Syst*., **15**, 1033–1040.

Schumacher,J. *et al*. (2013) Nitrogen and carbon status are integrated at the transcriptional level by the nitrogen regulator NtrC *in vivo*. *MBio*, **4**, 1–9.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*., **17**, 53–60.

Voit,E.O. and Almeida,J. (2004) Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, **20**, 1670–1681.

Voit,E.O. (2013) Characterizability of metabolic pathway systems from time series data. *Math. Biosci*., **5**, 1–11.

Zamboni,N. (2011) ^{13}C metabolic flux analysis in complex systems. *Curr. Opin. Biotechnol*., **22**, 103–108.