Mathematical Biosciences xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Characterizability of metabolic pathway systems from time series data

Eberhard O. Voit *

The Wallace H. Coulter, Department of Biomedical Engineering at Georgia Tech. and Emory University, 313 Ferst Drive, Suite 4103, Atlanta, GA 30332-0535, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Characterizability
Dynamic flux estimation
Identifiability
Metabolic pathway
Moore–Penrose inverse
Parameter estimation

ABSTRACT

Over the past decade, the biomathematical community has devoted substantial effort to the complicated challenge of estimating parameter values for biological systems models. An even more difficult issue is the characterization of functional forms for the processes that govern these systems. Most parameter estimation approaches tacitly assume that these forms are known or can be assumed with some validity. However, this assumption is not always true. The recently proposed method of Dynamic Flux Estimation (DFE) addresses this problem in a genuinely novel fashion for metabolic pathway systems. Specifically, DFE allows the characterization of fluxes within such systems through an analysis of metabolic time series data. Its main drawback is the fact that DFE can only directly be applied if the pathway system contains as many metabolites as unknown fluxes. This situation is unfortunately rare. To overcome this roadblock, earlier work in this field had proposed strategies for augmenting the set of unknown fluxes with independent kinetic information, which however is not always available. Employing Moore–Penrose pseudo-inverse methods of linear algebra, the present article discusses an approach for characterizing fluxes from metabolic time series data that is applicable even if the pathway system is underdetermined and contains more fluxes than metabolites. Intriguingly, this approach is independent of a specific modeling framework and unaffected by noise in the experimental time series data. The results reveal whether any fluxes may be characterized and, if so, which subset is characterizable. They also help with the identification of fluxes that, if they could be determined independently, would allow the application of DFE.

1. Introduction

A central challenge of computational systems biology is the translation of biological systems into mathematical models. Addressing this challenge critically depends on two components: data of high quality and effective strategies for model design, diagnostics, and analysis. The translation process itself consists of two steps, namely the determination of suitable mathematical representations and the identification of values for the parameters in these representations. Recent years have witnessed enormous efforts in the area of parameter estimation, indicating that parameter estimation is an unavoidable and very difficult task that is not yet completely solved (*e.g.*, [1–5]). Some of its difficulties are of computational nature, while others are due to the noisiness of biological data and the fact that several computed solutions often lead to similarly good data fits [6–11].

The parameter estimation task is not only difficult; it also makes a fundamental *a priori* assumption, namely, that the mathematical structure of the model representing the given data is known. However, this assumption is seldom entirely true; in fact, one could legitimately ask whether we ever truly know the structural format of a model in biology. The choice of a particular structure for a given modeling task may be rationalized in various ways. The traditional argument has been that certain functions or models had been used frequently and successfully in a particular biological subfield and therefore had developed into default representations. A good example in ecology is the Lotka–Volterra (LV) model, which describes the time-dependent changes in population sizes by linear and binomial terms that represent interactions among the various pairs of populations [12–15]. LV models have been very successful, but no ecologist would claim that they capture the dynamics of ecosystems in their full complexity. A second example is the Michaelis–Menten function [16], which was derived from a conceptual scheme describing the enzyme catalyzed conversion of a substrate into a product under idealized conditions. Although these conditions are seldom present in real cells [17,18], this function has been used as a default in thousands of biochemical studies, and even in cases that have not much to do with enzyme catalysis, such as the uptake of nutrients through the root system of a plant [19].

The choice of an appropriate model becomes more complicated in cases where the processes to be represented are aggregates of several steps [20,21]. An example is the ubiquitous effect of an extracellular ligand or an intracellular process like gene expression. At a coarse level, this effect is direct: if the ligand is present,

* Tel.: +1 (404) 385 5057; fax: +1 (404) 894 4243.
E-mail address: eberhard.voit@bme.gatech.edu

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.01.008

a gene becomes expressed, and without a ligand, the gene is more or less silent. However, any description of this relationship in detail becomes exceedingly difficult, as it typically would have to account for physical changes in the conformation of the receptor, an entire signaling cascade, the translocation of a transcription factor, as well as the transcriptional machinery.

The alternative to an *a priori* assumption of a particular functional form is the use of a 'canonical' approximation, which is a representation that is based on theory, always leads to the same mathematical structures, and therefore permits streamlined analyses. The LV models mentioned earlier, as well as power-law models of Biochemical Systems Theory (BST; [4,18,22,23]) and linear representations fall into this category. The advantages of these representations include their guaranteed appropriateness at some operating point of choice, generality in applications, mathematical and computational tractability, and the fact that these types of models, at least initially, require very few application-specific assumptions. Nonetheless, canonical models are not necessarily the perfect answer to the question of what an appropriate representation of a process should look like, because they are not mechanistic and their parameters therefore do not have a mechanistic meaning. Also, like all other representations, they have by definition a limited range of valid approximation, and the size of this range is almost always unknown and difficult to assess.

The question thus arises whether it is possible to infer mathematical descriptions that adequately represent the true biological process without introducing too much bias. In an attempt to address this question for metabolic pathway systems, we recently proposed the method of Dynamic Flux Estimation (DFE; [24]), which is briefly reviewed in a later section. DFE uses as input the topology of a pathway system, together with time series measurements of the involved metabolites over a sufficiently wide time horizon. Of note is that DFE presupposes no knowledge or assumptions regarding the processes governing a metabolic system, but only of the topology of the network. In ideal cases, the input information is sufficient to prescribe a straightforward strategy for characterizing trends of all processes as they change over time or as they are affected by metabolites and modulators in the system. These resulting trends are not given as numerical functions, but as graphical representations. These plots, in turn, can directly be used for further analysis or allow the testing of numerically specified candidate functions. Thus, in contrast to identifiability tasks, which have the goal of determining optimal numerical settings for a model, the first phase of DFE addresses a characterizability task that precedes the identification of functional forms and parameters in the second phase.

Unfortunately, the ideal conditions allowing such an unbiased flux characterization are not often given. In particular, most metabolic pathway systems contain more fluxes than metabolite pools, and this discrepancy leads to a stoichiometric matrix of the flux system that has less than full rank (see later). Thus, unless additional information on sufficiently many fluxes is available, DFE cannot be applied. It is not even clear which fluxes would need to be identified independently to permit subsequent DFE. Discussion of this issue has led to suggestions for potentially helpful additional information, which could come from different sources. For example, in addition to the metabolic time series one might have measurements of some in- or effluxes. One might also be able to assume a flux representation from generally accepted kinetic knowledge [25]. If sufficiently many fluxes can be numerically characterized in this manner, the remaining fluxes can be computed in a point-wise fashion, as it is done in DFE with a system of full rank. If the data are rich enough, it is also sometimes possible to infer some fluxes from the data themselves [26].

This article presents an extended, general strategy for characterizing fluxes for pathway systems where the original DFE strategy is insufficient. The strategy uses a pseudo-inverse matrix method that reveals which reaction steps in a system are uniquely characterizable if time series data are available, even if the system is underdetermined. Secondly, the method permits the scanning for those reaction steps in a pathway system that, if they could be characterized independently, would be most beneficial for a subsequent DFE analysis. Intriguingly, the characterizability method proposed here is model free and uses only the topology of the pathway system, but no knowledge of regulatory features or specific time series data. The immediate result is a list of all reaction steps that could be uniquely characterized in a DFE sense if time series were available. Of course, the actual characterization of dynamic trends requires data, and a correct interpretation of these trends requires knowledge of the regulatory control patterns of the system.

2. Methods

2.1. Metabolic time series data

Modern ^{13}C- and ^{31}P-NMR methods permit the non-invasive determination of the concentrations of substrates and intracellular metabolites in living cell cultures. These measurements can be made every 30 s or even faster, thereby leading to dense metabolic time series data on the same cells and under the same conditions. In some sense, these data reflect all metabolic activities in these cultures, at least in principle. Examples of such data and their analysis can be found in [9,27–29].

Mass spectrometry (MS) has advanced to a point where very many metabolites in very small quantities can be identified simultaneously. While the method is destructive and requires the running of standards, it can be used to generate time series data as well. As an example, Kinoshita and colleagues measured metabolic time-courses of human red blood cell exposed to hypoxia, using capillary electrophoresis coupled to time-of-flight MS [30]. Other destructive methods, including liquid and gas chromatography, can similarly be used to establish metabolic profiles over relevant time horizons.

2.2. A brief review of dynamic flux estimation (DFE)

2.2.1. Rationale

The generic format of ODE models for metabolic systems is

$$\frac{d\mathbf{X}}{dt} = \dot{\mathbf{X}} = \mathbf{N} \cdot \mathbf{R}. \tag{1}$$

In this generic formulation, **X** is the vector of metabolite concentrations, **N** is the stoichiometric matrix, and **R** is a vector containing the specific reactions in the pathways. The stoichiometric matrix describes which variables are involved in which reaction [31–33]. An example is the branched pathway system in Fig. 1, which consists of one independent variable (X_0), three dependent variables (X_1,X_2,X_3) and five reactions (v_1,...,v_5), and has the stoichiometric matrix

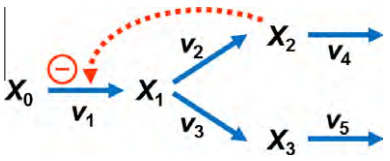

Fig. 1. Branched pathway with one feedback signal.

Please cite this article in press as: E.O. Voit, Characterizability of metabolic pathway systems from time series data, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.01.008

$$N = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix} \quad (2)$$

The positive and negative entries in **N** correspond to influxes and effluxes associated with each metabolite pool, respectively. Thus, the second row indicates that X_2 receives input through reaction v_2 (element $N_{2,2}$) and loses material through v_4 (element $N_{2,4}$). Independent variables and regulatory signals are not explicit in **N**.

Of particular interest is that the formulation in Eq. (1) separates the topology of the pathway, which is represented by **N**, from all numerical information, including regulatory effects; this information is exclusively taken into account by the vector of reactions, **R**. In particular, an explicit formulation of **R** exhibits the mathematical format of the representations of reactions, while **N** does not. For instance, each v_i in **R** could be a Michaelis–Menten function, a power-law function, or have some other mathematical structure.

DFE capitalizes on the linearity of the system description in Eq. (1) with respect to the reactions. Specifically, at any given point in time, Eq. (1) is a system of linear equations, where the derivative on the left-hand side equals a sum or difference of flux values at the same time point. The derivative can be estimated from the time series data, because it is the slope of the corresponding metabolite at the same time point [34–36]. As a result, the slope of a variable at a given time point is equal to a linear function of flux values.

2.2.2. Illustration example

For a simple illustration of DFE, consider the linear pathway with feedback shown in Fig. 2. Suppose that, in a laboratory experiment, the substrate X_0 had been supplied externally and that the uptake v_1 had been measured at several time points. Suppose further that time series data had been obtained for X_1 and X_2 and that slopes of the time courses were estimated with some accuracy. Idealized, noise-free data of this type are shown in Fig. A.1 and Table A.1 of the Appendix.

For this demonstration, artificial data were generated with the following system

$$\begin{aligned} \dot{X}_0 &= -0.2X_0 \\ \dot{X}_1 &= 0.2X_0 - 0.6X_1^{0.5}X_2^{-0.2} \\ \dot{X}_2 &= 0.6X_1^{0.5}X_2^{-0.2} - 0.75X_2^{0.8} \end{aligned} \quad (3)$$

and the initial values $X_0 = 10$, $X_1 = 2$, $X_2 = 1$. In reality, just the data in Fig. A.1 would be available, whereas the format and parameterization of the equations in (3) would be unknown.

The time series data are first used to estimate the slopes S_1 and S_2, of X_1 and X_2 respectively, at a series of time points t_i, $i = 1,\ldots,k$ (*cf.* [34–36]). With these estimates (Table 1), the system can be reformulated symbolically as k pairs of equations of the form

$$\begin{aligned} S_1(t_i) &= v_1(t_i) - v_2(t_i) \\ S_2(t_i) &= v_2(t_i) - v_3(t_i) \end{aligned} \quad (4)$$

for $i = 1,\ldots,k$. For this simple case, it is easy to express the fluxes as functions of measured slopes; in a realistic application, this step would require a matrix inversion. For the illustration example, one obtains

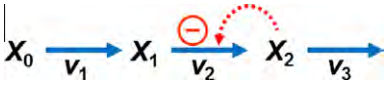

Fig. 2. Linear pathway with an external substrate input, two internal metabolites, and one feedback inhibition signal.

$$\begin{aligned} v_2(t_i) &= v_1(t_i) - S_1(t_i) \\ v_3(t_i) &= v_1(t_i) - S_1(t_i) - S_2(t_i) \end{aligned} \quad (5)$$

so that the fluxes are expressed in terms of the slopes and the measured values of v_1. The direct results of DFE are thus point plots of v_2 and v_3 against time or against X_1 and X_2. As an example, Fig. 3 shows connected plots of v_2 against X_1 and X_2. The structural and numerical identification of functions that best fit these plots is still to be done, and general guidelines do not exist for this task. Nonetheless, the task is much simpler now, because the identification of explicit functions can be done from plots that graph these functions, rather than from combined processes that govern the dynamic time courses of the ODE system. By characterizing the fluxes individually, many problems associated with error compensation and extrapolability are mitigated [24].

2.2.3. Shortcomings of DFE

The Achilles heel of DFE is the requirement that the system contain as many reactions as metabolites, which is usually not the case. If equality is not given, a direct application of DFE stalls. To ameliorate the situation, it is sometimes possible to characterize a sufficient number of fluxes outside DFE. For instance, the kinetic characteristics of a reaction may be known, and since metabolite time series data are available, one may convert this information into a functional representation of the reaction [25]. By thus augmenting the system, the system matrix might eventually have full rank, so that DFE can be applied. While this augmentation fills the rank, it introduces a certain degree of bias, due to the assumption of a functional format and appropriate kinetic parameters.

If the data are non-monotonic or if sufficiently many datasets are available, it is sometimes possible to augment DFE purely based on the time series data. The method is somewhat cumbersome, but leads to success if the data allow [26].

2.3. Application of the Moore–Penrose pseudo-inverse to DFE

This article does not address an augmentation of DFE but the question of whether at least some reactions can be characterized even if the matrix is non-square or does not have full rank, and if so, which reactions are characterizable and which are not. This information is very useful for selecting possible augmentation schemes. The method is directly based on pseudo-inverse matrices.

Elementary linear algebra tells us that we can invert a matrix if it is square and has full rank. The premier application of such an inversion is the task of solving linear algebraic equations. However, if the system matrix is "tall," *i.e.*, if there are more equations than unknowns, the system tends to be over-determined, and a true solution cannot be found. The best outcome is a solution that does not satisfy all equations but is optimal in the sense of linear regression; that is, it minimizes the overall squared-residual error. If the matrix is "wide," because the system contains fewer equations than unknowns, the solution tends to be under-determined, and in most cases infinitely many solutions can be found that satisfy all equations. As discussed before, this situation is typical for DFE and of most interest here.

A method for dealing with the inversion of systems that do or do not permit a unique solution uses the Moore–Penrose pseudo-inverse, which is a generalization of the matrix inverse of elementary linear algebra. The literature is quite rich (*e.g.*, [37–41]), and even Wikipedia presents a good introduction. The pseudo-inverse of matrix **N** is commonly called $\mathbf{N}^+$. Two features associated with the pseudo-inverse are of pertinence here. First, for an underdetermined system $\mathbf{Nv} = \mathbf{b}$, the solution $\mathbf{z} = \mathbf{N}^+\mathbf{b}$ is the solution of the system that satisfies the Euclidean 2-norm; that is, it corresponds to the smallest distance from the origin among

Table 1
Metabolite concentrations and slopes associated with the simple pathway system in Eq. (3).

t	X_0	v_1	X_1	X_2	S_1	S_2
0	10	2	2	1	1.151472	0.09852814
1	8.187308	1.637462	2.880906	1.142217	0.6457941	0.1574867
2	6.703201	1.34064	3.342895	1.286703	0.2975619	0.1254981
3	5.488117	1.097623	3.507693	1.388293	0.04525962	0.07727053
4	4.49329	0.898658	3.456303	1.442177	−0.1380429	0.03144606
5	3.678795	0.7357589	3.24938	1.453391	−0.2678741	−0.007870263
6	3.011942	0.6023885	2.934854	1.428643	−0.3547169	−0.04059536
7	2.46597	0.493194	2.551702	1.374082	−0.4062285	−0.06767775
8	2.018965	0.4037931	2.132225	1.294808	−0.4282133	−0.09019441
9	1.652989	0.3305978	1.703618	1.1949	−0.4251398	−0.1090841
10	1.353352	0.2706705	1.289138	1.077603	−0.4004636	−0.1250771
11	1.108031	0.2216061	0.9089883	0.945547	−0.3568811	−0.1386591
12	0.9071783	0.1814357	0.5809344	0.8010357	−0.2966265	−0.1499712
13	0.7427346	0.1485469	0.3205111	0.6465308	−0.2220953	−0.1584515
14	0.6080996	0.1216199	0.1400759	0.4858637	−0.137816	−0.1615551
15	0.4978698	0.09957397	0.04385477	0.3280041	−0.05745645	−0.1504126
16	0.4076214	0.08152427	0.01288757	0.1956741	−0.01286664	−0.1089796
17	0.3337321	0.06674643	0.005798956	0.1125757	−0.003972747	−0.05996322
18	0.2732368	0.05464735	0.003036819	0.06852463	−0.001871244	−0.03133076
19	0.2237073	0.04474147	0.001679863	0.04511861	−0.000958626	−0.01718472
20	0.1831561	0.03663122	0.0009648202	0.03176125	−0.00052187	−0.01033442

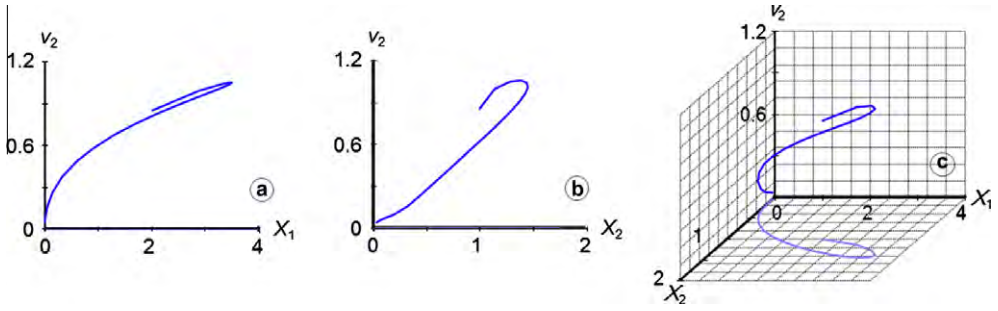

Fig. 3. Plots, derived from DFE, of v_2 against X_1, X_2, and both.

all possible solutions[1]. Second, to find the entire solution of an underdetermined system **Nv = b**, one computes a particular solution, which is given as $\mathbf{N}^+\mathbf{b}$, and computes

$$\mathbf{v} = \mathbf{N}^+\mathbf{b} + [\mathbf{I} - \mathbf{N}^+\mathbf{N}]\mathbf{w} \tag{6}$$

where **I** is the identity matrix and **w** is an arbitrary vector, both of the appropriate dimensions. For simplicity of the following discussion, we denote with **D** the matrix of the differences between the identity and the product $\mathbf{N}^+\mathbf{N}$:

$$\mathbf{D} = \mathbf{I} - \mathbf{N}^+\mathbf{N} \tag{7}$$

D spans the same space as the kernel of **N**. $\mathbf{N}^+$ and **D** are easily computed in Matlab with the pseudo-inverse function pinv(**N**). Instead of computing **D**, one may also use the Matlab function null(**N**), which computes an orthonormal basis of the null space of **N**.

One should note that the Moore–Penrose pseudo-inverse is not the only possible mechanism for identifying a generalized matrix inverse [42]. However, it satisfies our needs here and is therefore considered sufficient.

3. Results

3.1. General insights

Both $\mathbf{N}^+$ and **D** can be applied directly to DFE, with **D** actually being more informative than $\mathbf{N}^+$. The minimum-norm solution, given by $\mathbf{N}^+$, does provide a solution, but unless the system happens to have a unique solution, some fluxes in $\mathbf{v} = \mathbf{N}^+\mathbf{b}$ are bound to be negative, and therefore biologically meaningless. Thus, $\mathbf{N}^+$ contains pertinent information, but this information is more easily interpreted in the matrix **D**, which is derived from $\mathbf{N}^+$, or from the kernel null(**N**), whose computation leads to the same conclusions, as we will see in the examples.

In the special case where **D** consists entirely of 0's, $\mathbf{N}^+$ actually is identical to the regular inverse and the solution is uniquely given by the minimum-norm solution, which in this case is biologically meaningful. This result is not surprising. More interesting, if **D** contains a row of 0's, the corresponding flux in the general solution is unaffected by the arbitrary vector **w**. Thus, given time series data, this flux can be characterized uniquely, in a DFE sense where it can be plotted against time or against its contributing metabolites. If several rows exclusively contain 0's, all corresponding fluxes are uniquely characterizable.

If two rows of **D** are identical, but not 0, the corresponding fluxes are affected by the arbitrary vector **w** in the same manner. These fluxes tend to correlate with each other and are often found to be associated with the same variable or with a chain of reactions within the pathway. It seems difficult to extract more pertinent information from the identity of non-zero rows (see numerical example at the end of the Results section).

Intriguingly, the characterizability analysis does not require knowledge of the vector **b** in Eq. (6). In DFE, this vector contains the left-hand sides of the differential equations, which consist of slope estimates. In addition, the vector may contain known flux values within the system, which as numerical values can be moved from the right to the left side of Eq. (6). Thus, the above statements

[1] Traditionally, linear algebra uses the notation **A x = b** for these types of analysis. Here, all considerations refer to stoichiometric systems, which are more intuitively represented as **N v**, and using the notation **A x** seems more confusing than helpful.

regarding **D** are independent of actual time series data, so that issues of noise are immaterial at this point. Expressed differently, the analysis of flux characterizability can be performed without actual data, because it is exclusively based on the topology of the pathway.

As a consequence, the statements above hold even if the fluxes are regulated by metabolites in the system, for instance, through feedback inhibition. This result is due to the fact that all information regarding the metabolic regulation of the pathway exclusively affects **b**, but not **N**, $\mathbf{N}^+$, or **D**, so that regulatory features are not relevant for this characterizability analysis.

Finally, if **D** does not have desirable features, its computational analysis may help identify additional experiments that, for instance, would render rows of **D** to become zero. Specifically, one may scan the system matrix **N** for non-zero elements, set these to zero, one at a time or two at a time, and study whether these changes would result in rows of zeros in **D**. The interpretation of this strategy is the following: setting a non-zero entry of **N** to zero means characterizing the corresponding reaction step over the time points of the experiment with independent means. Because the measured flux values can be merged with the slopes in the corresponding equation(s), they move from **N**, $\mathbf{N}^+$, and **D** to **b**. An example is provided in the section *Larger Systems*. Other examples are discussed in [25].

3.2. Illustration examples

The power of the analysis of **D** becomes evident most easily in small examples. They consist of pathways with a few variables and fluxes. It is immaterial in these examples whether the fluxes are regulated by any of the metabolites, as was discussed before. This regulation is of course very important, but it exclusively enters the numerical values in **b** and does not affect **N**, $\mathbf{N}^+$, or **D**.

3.2.1. Branched pathway with accumulation

Consider the pathway in Fig. 4a which consists of three variables and four reactions. The corresponding flux matrix is

$$\mathbf{N} = \begin{pmatrix} 1 & -1 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{8}$$

$\mathbf{N}^+$ is computed in Matlab as pinv(**N**), and the result is

$$\mathbf{N}^+ = \begin{pmatrix} 0.5 & 0.5 & 0.5 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ -0.5 & -0.5 & -0.5 \end{pmatrix} \tag{9}$$

Furthermore we obtain **D** directly as $\mathbf{D} = \mathbf{I} - \mathbf{N}^+\mathbf{N}$:

$$\mathbf{D} = \begin{pmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix} \tag{10}$$

The second and third rows consist of zeros, indicating that time series data would allow the unique characterization of v_2 and v_3. Furthermore, the first and fourth rows are identical, which implies that v_1 and v_4 are affected in the same fashion by an arbitrary vector **w**. These findings make intuitive sense: the flux v_3 is characterizable from time measurements on X_3, and v_2 is uniquely determined by v_3 and the measured levels of X_2. By contrast, any increase or decrease in v_1 can be compensated by a corresponding change in v_4, without changing X_1, and the two are therefore coupled. This coupling is generally complicated and nonlinear, and details depend on the numerical features of the pathway and its regulation (see numerical example at the end of the Results section).

Using the Matlab function null(**N**), instead of pinv(**N**) and **D**, reveals that the basis of the kernel consists of the vector $(0.7071, 0, 0, 0.7071)^{tr}$, thus indicating again that v_2 and v_3 are characterizable from time series measurements.

Variation 1. Assume that v_4 does not exist or is obtainable numerically with other means, so that it can be merged with the estimated slopes $S_1(t_i)$ (Fig. 4b). The only difference to the prototype above is that matrix element $N_{1,4}$ equals 0. Except for possible slight numerical inaccuracies, it is immaterial whether the earlier 4×3 matrix is used and $N_{1,4}$ is changed from -1 to 0 or if one defines a 3×3 matrix, which consists of the first three columns of the earlier matrix. In the first case, $\mathbf{N}^+$ has four rows, with the fourth row consisting of 0's. In the second case, $\mathbf{N}^+$ only contains the first three rows:

$$\mathbf{N}^+ = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{11}$$

D consists entirely of zero's, within machine precision. Thus, $\mathbf{N}^+$ is identical to the regular inverse, and the system is fully identifiable. This again makes intuitive sense, as the fluxes may be identified in backwards order, starting with v_3. Computing null(**N**) leads to the same conclusion.

Variation 2. Based on Variation 1, assume that there is an efflux from X_3 (Fig. 4c). Now the stoichiometric matrix is

$$\mathbf{N} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \tag{12}$$

and we obtain

$$\mathbf{N}^+ = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ -0.25 & 0.5 & 0.25 \\ -0.25 & -0.5 & 0.25 \\ -0.25 & -0.5 & -0.75 \end{pmatrix} \tag{13}$$

and

$$\mathbf{D} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \tag{14}$$

No row of **D** is zero, and therefore no flux can be characterized uniquely. At the same time, all rows are identical, and the same transformation by **w** is to be applied to all reactions steps. Given the strict linear nature of the pathway, the result says that all fluxes have to increase or decrease in lockstep.

Variation 3. Let's return to the original pathway with two branches at X_1, but assume that there is no influx to X_1 (Fig. 4d) or that it can be measured independently. The stoichiometric matrix and its pseudo-inverse are

$$\mathbf{N} = \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 1 \end{pmatrix} \tag{15}$$

and

$$\mathbf{N}^+ = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & -1 \\ 0 & 0 & 1 \end{pmatrix} \tag{16}$$

Please cite this article in press as: E.O. Voit, Characterizability of metabolic pathway systems from time series data, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.01.008

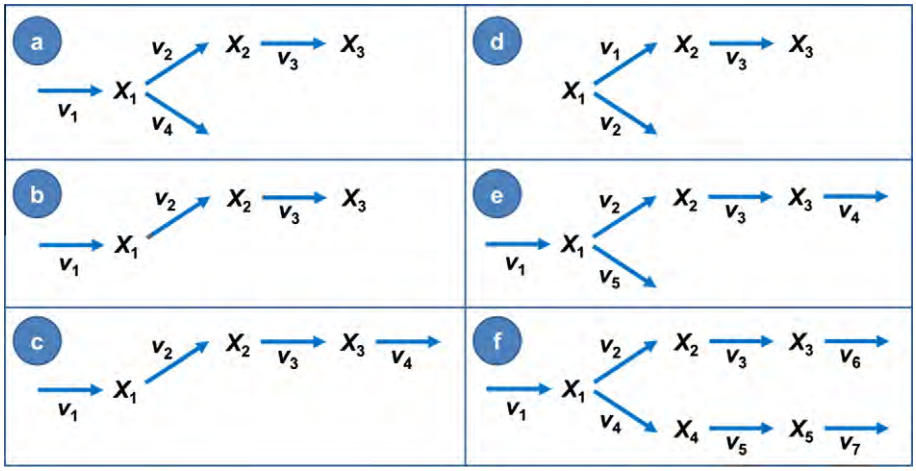

Fig. 4. Different variants of a branched pathway. For the characterizability analysis proposed here, it is immaterial whether any of the reaction steps are regulated by metabolites of the system.

D is zero. The interpretation is that, since there is no influx to the system, the change in X_1, which gave the system one degree of freedom before, is now governed by v_1 and v_3, and v_2 receives all material not flowing into the direction of v_1.

Variation 4. We start again with the original pathway but account for an efflux from X_3 (Fig. 4e). Now the system has three variables and five fluxes. Its stoichiometric matrix is

$$\mathbf{N} = \begin{pmatrix} 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}. \tag{17}$$

The pseudo-inverse is

$$\mathbf{N}^+ = \begin{pmatrix} 0.4286 & 0.2857 & 0.1429 \\ -0.1429 & 0.5714 & 0.2857 \\ -0.1429 & 0.4286 & 0.2857 \\ -0.1429 & 0.4286 & -0.7143 \\ -0.4286 & -0.2857 & -0.1429 \end{pmatrix} \tag{18}$$

and we obtain

$$\mathbf{D} = \begin{pmatrix} 0.5714 & 0.1429 & 0.1429 & 0.1429 & 0.4286 \\ 0.1429 & 0.2857 & 0.2857 & 0.2857 & -0.1429 \\ 0.1429 & 0.2857 & 0.2857 & 0.2857 & -0.1429 \\ 0.1429 & 0.2857 & 0.2857 & 0.2857 & -0.1429 \\ 0.4286 & -0.1429 & -0.1429 & -0.1429 & 0.5714 \end{pmatrix} \tag{19}$$

Rows 2, 3, and 4 of **D** are identical, which implies parallel shifts in rates v_2, v_3, and v_4. Rows 1 and 5 are not identical, but row 5 is equal to the difference between row 1 and row 2 (or 3 or 4), thus pointing to a rank of 2 and thus to two degrees of freedom, namely in the amount of influx, and the subsequent split between v_2 and v_5.

3.2.2. Pathway with reversible reaction

The pathway in Fig. 5a is linear but contains one reversible reaction. If we can justify the replacement of the forward and reverse steps with a single net reaction step, the system reduces to a fully determined example, as it was discussed before. However, if this substitution is not *a priori* justified, the stoichiometric matrix must account for three metabolites and four reactions. It has the form

$$\mathbf{N} = \begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{20}$$

$\mathbf{N}^+$ and **D** are

$$\mathbf{N}^+ = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & 1 \end{pmatrix} \tag{21}$$

and

$$\mathbf{D} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{22}$$

The interpretation of **D** is straightforward: reactions v_1 and v_4 are characterizable from time series data, while v_2 and v_3 are not fixed but have to be changed at the same rate. Using null(**N**) rather than computing **D**, we obtain the basis column vector $[0, 0.7071, 0.7071, 0]^{tr}$, which leads to the same conclusion.

Variation 1. Suppose the same system does not have an influx (Fig. 5b) or that this influx is measurable with other means, so that it can be merged with slope estimates. Now the system has three variables and three reactions, and one may be led to assume a unique solution. The stoichiometric matrix is

$$\mathbf{N} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{23}$$

$\mathbf{N}^+$ and **D** are

$$\mathbf{N}^+ = \begin{pmatrix} -0.3333 & 0.1667 & 0.1667 \\ 0.3333 & -0.1667 & -0.1667 \\ -0.3333 & -0.3333 & 0.6667 \end{pmatrix} \tag{24}$$

and

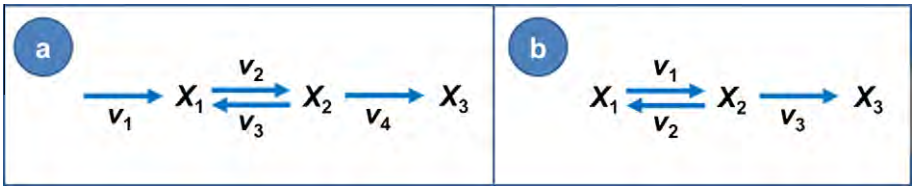

Fig. 5. Linear pathways with one reversible reaction.

$$\mathbf{D} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{25}$$

Although there are three metabolites and three reactions, only one reaction is characterizable, whereas the other two are not: they permit equal scaling. Even if v_3 could be independently measured, it would not ameliorate the situation. By contrast, if v_1, v_2 or their ratio could be measured, DFE could be applied.

3.2.3. Larger systems

The proposed extensions of DFE via pseudo-inverses are in principle independent of scale. However, one must expect that larger systems will often have higher degrees of freedom and therefore elude characterization. Nonetheless, the analysis may point to key reaction steps which, if they could be determined independently, would lead to a higher degree of characterizability. As an example, consider the branched pathway with five metabolites and seven reaction steps, which is an extension of the first set of examples (Fig. 4f). The stoichiometric matrix is thus 7×5 (not shown) and **D** is 7×7:

$$\mathbf{D} = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2667 & 0.2667 & -0.0667 & -0.0667 & 0.2667 & -0.0667 \\ 0.2 & 0.2667 & 0.2667 & -0.0667 & -0.0667 & 0.2667 & -0.0667 \\ 0.2 & -0.0667 & -0.0667 & 0.2667 & 0.2667 & -0.0667 & 0.2667 \\ 0.2 & -0.0667 & -0.0667 & 0.2667 & 0.2667 & -0.0667 & 0.2667 \\ 0.2 & 0.2667 & 0.2667 & -0.0667 & -0.0667 & 0.2667 & -0.0667 \\ 0.2 & -0.0667 & -0.0667 & 0.2667 & 0.2667 & -0.0667 & 0.2667 \end{pmatrix} \tag{26}$$

Inspection of **D** indicates that rows 2, 3, and 6 are identical and that 4, 5, and 7 are identical. Thus, the corresponding reactions increase or decrease in parallel. Furthermore, there is a linear dependence, for instance, in the form of Row1 = Row2 + Row4. If the effluxes v_6 and v_7 could be measured, columns 6 and 7 of the stoichiometric matrix would consist only of zeros, and so would the first five rows of **D**. Thus, the entire system would be characterizable.

As an alternative to **D**, we may compute the null space of **N** with the Matlab function null(**N**), which directly yields a result with the same information content, but which is less redundant, namely

$$null(N) = \begin{bmatrix} 0.3313 & 0.5387 \\ 0.5134 & 0.0555 \\ 0.5134 & 0.0555 \\ -0.1821 & 0.4832 \\ -0.1821 & 0.4832 \\ 0.5134 & 0.0555 \\ -0.1821 & 0.4832 \end{bmatrix} \tag{27}$$

As in **D**, rows 2, 3, and 6 are identical, and rows 4, 5, and 7 are identical. Furthermore, row1 = row2 + row4. Using the same arguments as before, we can conclude that **v** is characterizable.

In more complicated cases, the null space may be subjected to additional analyses. For instance, it might be useful to perform the so-called varimax rotation [43], which is commonly used in principal component analysis [44], where it assists in the identification of a new coordinate system in which a few variables are dominant and all others are close to zero. This change of coordinates and identification of near-zero variables could lead to additional insights into the structure of the null space and thus the metabolic system.

As a different scenario, suppose v_2 could be measured independently. The characterization would in effect split the pathway into two smaller pathways, and it is clear that v_3 and v_6 could directly be characterized. Furthermore, the remaining fluxes ($v_1 - v_2$), v_4, v_5, and v_7 would be similarly scaled by any vector **w**, and additional independent determination of v_1 would make the entire system characterizable from time series data.

As a second example of affecting the stoichiometric matrix, consider a pathway with an internal loop, as shown in Fig. 6. The stoichiometric matrix is

$$\mathbf{N} = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix} \tag{28}$$

and **D** is

$$\mathbf{D} = \begin{pmatrix} 0.375 & 0.25 & -0.125 & -0.125 & 0.375 \\ 0.25 & 0.5 & 0.25 & 0.25 & 0.25 \\ -0.125 & 0.25 & 0.375 & 0.375 & -0.125 \\ -0.125 & 0.25 & 0.375 & 0.375 & -0.125 \\ 0.375 & 0.25 & -0.125 & -0.125 & 0.375 \end{pmatrix} \tag{29}$$

Rows 1 and 5, and 3 and 4, are identical and imply parallel scaling, which is consistent with intuition. Furthermore, Row2 = -Row1 + Row3. If the details of the internal loop are not of particular interest, the metabolite pools of X_2, X_3, and X_4 could be merged, and the overall influx and efflux would be v_1 and v_5, which would still allow one remaining degree of parallel scaling.

3.2.4. Example of actual characterization

Suppose the pathway under study is similar to the prototypes in Fig. 4, but contains two regulatory signals (Fig. 7). For illustration purposes we assume that the system can be modeled with a Generalized Mass Action within Biochemical Systems Theory [23], but except for simulating time series data, we pretend not to know the governing equations. As a numerical illustration, and without any particular reason for the choice of parameter values, the describing model is shown in Eq. (30)

$$\begin{aligned} \dot{X}_1 &= 0.05 - 1.1X_1^{0.5}X_3^{-0.75} - 2.8X_1^{0.8}X_2^{0.4} \\ \dot{X}_2 &= 1.1X_1^{0.5}X_3^{-0.75} - 1.1X_2^{0.6} \\ \dot{X}_3 &= 1.1X_2^{0.6} \end{aligned} \tag{30}$$

The pathway is easily simulated once initial values are chosen; for instance, $(X_1(0), X_2(0), X_3(0)) = (4,1,2)$. The resulting time courses are shown in Fig. 8. In reality, this would be the only input

information, outside the pathway topology. Also in reality, one would now estimate slopes for all variables at sufficiently many time points. Here we can simply compute these slopes.

The stoichiometric matrix, the pseudo-inverse, and **D** are identical to those in the original branched pathway example (Eqs. (8)–(10)). The earlier analysis showed that v_2 and v_4 are characterizable, while v_1 and v_4 are coupled, probably in a nonlinear fashion. Suppose we had estimated slopes from data following the trends in Fig. 8. At $t = 0$, the vector of slopes for X_1, X_2, and X_3, which in this illustration was taken from the simulation, is $\mathbf{S} = (-9.746, 0.2081, 1.1)$; earlier, this vector was generically called **b**. In this example, **b** exists exclusively of slopes, while it could, in other cases, also contain numerically known fluxes. $\mathbf{N}^+$ and **S** allow us to compute the minimum-norm solution, which is

$$\mathbf{v}(0) = \mathbf{N}^+\mathbf{S}(0) = \begin{pmatrix} -4.219 \\ 1.308 \\ 1.100 \\ 4.219 \end{pmatrix}. \tag{31}$$

The complete solution at time 0 is

$$\mathbf{v}(0) = \mathbf{N}^+\mathbf{S}(0) + \mathbf{D}\cdot\mathbf{w}(0) = \begin{pmatrix} 0.5 & 0.5 & 0.5 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ -.5 & -.5 & -.5 \end{pmatrix}\cdot\mathbf{S}(0) + \begin{pmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix}\cdot\mathbf{w}(0) \tag{32}$$

D indicates that v_2 and v_3 are characterizable: no vector **w** can change the values of v_2 and v_3 in **v**(0). Indeed, the values in Eq. (31) are exactly the same as the fluxes values of $1.1X_1^{0.5}X_3^{-0.75}$ and $1.1X_2^{0.6}$ in the dynamical system (Eq. (30)), when the metabolite concentrations at $t = 0$ are entered. Namely, substituting the initial values, one obtains $1.1 \cdot 4^{0.5} \cdot 2^{-0.75} = 1.308$ and $1.1 \cdot 1^{0.6} = 1.1$.

D reveals that v_1 and v_4 are not characterizable. In line with this insight, the first entry of **v**(0) is negative (see Eq. (31)), which is a reflection of the fact that **v**(0) satisfies the minimum norm of the system. Biologically, the solution is not meaningful and requires (unknown) scaling with some vector **w**. Specifically, we obtain the equations

$$\begin{aligned} v_1(0) &= -4.219 + 0.5(w_1(0) + w_4(0)) \\ v_4(0) &= +4.219 + 0.5(w_1(0) + w_4(0)) \end{aligned} \tag{33}$$

which show that these two rates are affected in the same manner by any changes in **w**(0). If $(w_1(0) + w_4(0))$ happens to be 8.538, which we can only know from the details of Eq. (30), the solution of the dynamic model is retrieved ($v_1(0) = 0.05$; $v_4(0) = 8.488$). However, this value is not identifiable without knowledge of the system. Furthermore, the entries of **w** are generally different for every time point. The only justified conclusion is that v_1 and v_4 depend on each other with a positive, but presumably nonlinear relationship. This relationship depends on the values of the metabolites affecting v_1 and v_4, and also on the numerical features of the functions representing the fluxes. In this particular case, v_1 is constant, but v_4 ($=2.8X_1^{0.8}X_2^{0.4}$) depends on X_1 and X_2 in a highly nonlinear fashion (see Eq. (30)). Eq. (33) also reflects the fact that, if we knew v_1 numerically for all time points t_i, we could compute $w_1(t_i) + w_4(t_i)$ and infer v_4, or *vice versa*.

In the same fashion as for $t = 0$, other flux values can be computed by evaluating $\mathbf{v}(t_i) = \mathbf{N}^+\mathbf{S}(t_i)$. In all cases, v_2 and v_3 directly return the correct values. For example, if $t = 1$, we obtain $v_2 = 0.1912$ and $v_3 = 0.8129$. Once v_2 and v_3 have been determined for sufficiently many time points, they may be plotted against time or against the contributing metabolites. Figs. 9 and 10 show these types of plots. While the time plots are somewhat interesting, the plots against metabolites are probably more informative as they provide clues regarding the dependence of a reaction step on its substrate(s) and modulator(s). In particular, the positive and negative kinetic orders of X_1 and X_3 in reaction v_2 are reflected with corresponding trends in Fig. 10. While these graphical relationships between a reaction and time, or between a reaction and its contributing metabolites, result directly from the DFE analysis, actual parameterized representations are not revealed, thus leaving the task of fitting these plots. This step is not trivial. However, it is much easier than any attempt to identify functional forms directly from the complete set of time series data.

It is even possible that a wrong functional format fits the plots relatively well, especially if noise in the time series data is considered. For example, suppose one assumed that v_2 had the format

$$v_2 = \frac{V_{max}X_1}{K_M + X_1}\cdot\exp(-p\cdot X_3). \tag{34}$$

It is not difficult to parameterize this function against the graphical plots in Fig. 10 against. With $V_{max} = 8$, $K_M = 0.3$, and $p = 0.9$ one obtains the fits in Fig. 11, and substituting this representation of v_2 into Eq. (30) yields time courses that are essentially indistinguishable from those in Fig. 8 (results not shown). The probability of incorrect formats decreases with the availability of additional datasets that shed light on different ranges of the dependent variables.

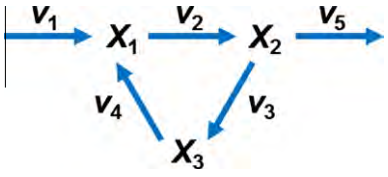

Fig. 6. Pathway with a loop.

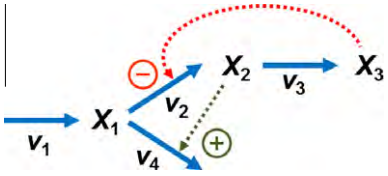

Fig. 7. Branched pathway with two regulatory signals.

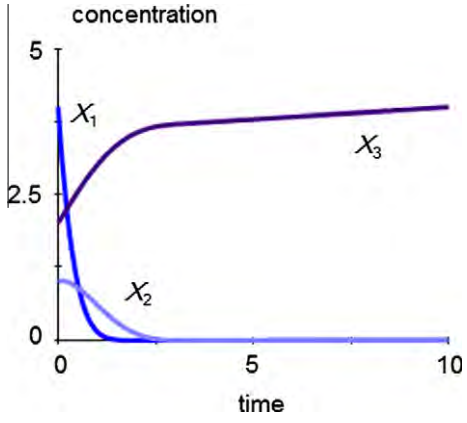

Fig. 8. Numerical simulation of the pathway in Fig. 7 and Eq. (30).

The reactions v_1 and v_4 are not characterizable from $\mathbf{N}^+$ and $\mathbf{D}$. Both are always shifted by $0.5\cdot(w_1 + w_4)$, but this expression generally has a different value for every t_i. While it is theoretically possible to compute vectors $\mathbf{w}$ that make the flux vector completely non-negative for all time points, this computation is not always easy in practical situations when the system is large.

4. Discussion

Arguably the most crucial task of computational systems biology is the translation of a biological phenomenon into a computational model that can be analyzed in lieu of its counterpart in the real world. This translation requires (1) the setting up of equations and (2) the fitting and validation of the parameters in these equations so that simulations with the model match experimental results. An interesting—yet often untested—challenge is the appropriate choice of the mathematical format of the equations, which only rarely can be truly justified, because the mechanisms to be described by the equations are often complicated and seldom known in sufficient detail. Canonical approximations may be used instead, but they have their own limitations.

Nature has not provided us with guidelines steering us toward the correct functional forms. Furthermore, the total repertoire of mathematical functions is infinite, and even if we can determine a function that models a certain dataset perfectly, there is no proof that the format of the function is correct. Good examples are high-order polynomials that can be constructed to match every data point precisely but can utterly fail in extrapolations to slightly changed conditions.

In the context of metabolic pathway systems, Dynamic Flux Estimation (DFE) addresses this conundrum in a two-phase strategy [24]. In the first phase, one attempts to establish the trend of each flux as a graphical plot against time or against its contributing metabolites. The result is significant, because it shows what each flux looks like, while essentially eliminating redundancies between fluxes affecting the same or different metabolites. The second phase attempts to convert the flux plots into functional forms. As in other identification tasks, these are entirely unknown, but the problem is vastly simplified because each flux can be analyzed individually and because this identification requires only explicit functions, rather than systems of differential equations. Nevertheless, the identification in the second phase of DFE is not guaranteed to be correct, because the selected function could be under- or over-parameterized and conceivably only apply to the specific ranges of values for which data were measured. Indeed, the validity of the results of the second phase of DFE can only be tested with additional data.

While DFE allows a significant leap forward in the identification of metabolic models, it has the unfortunate drawback that it requires a matrix inversion that is only directly possible if the pathway system contains as many fluxes as metabolites. Earlier extensions of DFE called for additional kinetic information on some fluxes or depended on rich datasets with quite particular features [25,26]. Here, we use pseudo-inverses of matrices to solve as much

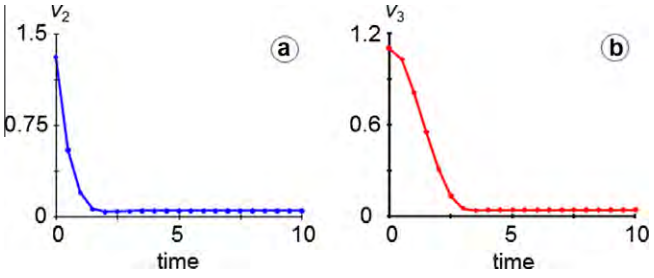

Fig. 9. Time plots v_2 and v_3 against time, respectively, reconstructed from $\mathbf{N}^+$ in cases where D has rows of 0's.

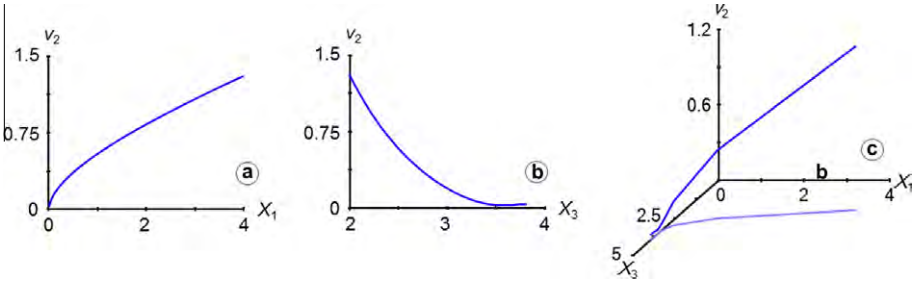

Fig. 10. Plots of reaction v_2 against its substrate (X_1), its inhibitor (X_3), and both. The light blue line in panel c shows the projection in the (X_1–X_3) plane. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

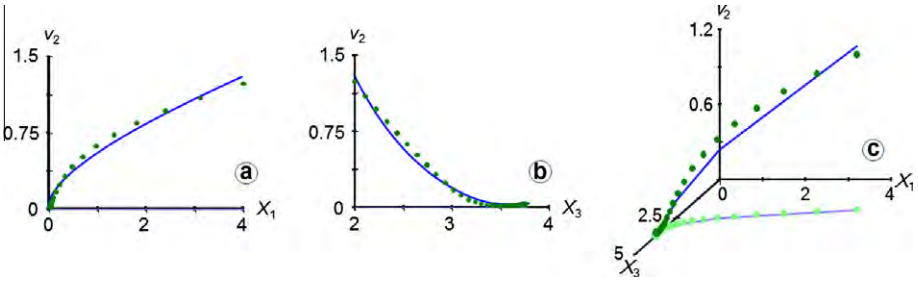

Fig. 11. No guidance can be given for selecting a parameterized representation of fluxes, and it is possible that a wrong representation nevertheless fits the flux profiles relatively well, especially if the input data are noisy. Shown here is a representation of v_2, where the dependence on X_1 is formulated as a Michaelis–Menten function and the dependence on X_3 as a negative exponential (green dots). The blue lines are copied from Fig. 10. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: E.O. Voit, Characterizability of metabolic pathway systems from time series data, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.01.008

of the DFE approach as is possible for a given pathway topology. Specifically, the proposed analysis reveals which fluxes in a pathway system could be characterized, if metabolic time series data were available. The analysis does not identify the numerical features of any of these fluxes and is therefore called 'characterizability analysis' rather than 'identifiability analysis'. This distinction is important and quite interesting. The characterizability analysis is much more limited than a full flux identification procedure, because its result is something like an existence proof, rather than a prescription for the mathematical format of the fluxes in the system. Ideally, characterizability leads to graphical representations of all fluxes and suggests functional formats that are subsequently to be parameterized. Therefore, characterizability may be considered a precursor to identifiability. However, it is also possible that a pathway model is not characterizable, but that the assumption of a specific flux format permits unique parameter identification. Thus, characterizability and identifiability are distinct aspects of the general task of formulating parametric representations of biological processes.

Two of its intriguing features of the proposed method are the following. First, the analysis is entirely model free. No choice of functions or a particular modeling framework is required, and there is consequently no issue with limited validity or accuracy of some candidate model. Second, the analysis is not affected by the quality of experimental data, missing data, noise, or the existence of known or unknown regulatory signals. The characterization is based purely on the topology of flux distribution within the pathway system, that is, on the stoichiometric matrix. A corollary of these insights is that additional data of the same or a similar type do not make a DFE analysis possible if it was not possible before. For instance, if new data with slightly altered inputs were generated for an underdetermined system, the collective data would still be underdetermined. The data could possibly improve the estimation of slopes, fill gaps in the data, or smooth out plots of fluxes, but they would not make DFE executable without additional information.

Thus, the first step of an underdetermined DFE task might be to compute from the stoichiometric matrix of a system which fluxes are directly characterizable. If all fluxes are characterizable, one could proceed and measure metabolic time series data, and if these could be obtained with sufficient accuracy, the trends in fluxes could be computed as plots against time or against the metabolites contributing to the dynamics of the flux. If some fluxes are not characterizable, it is clear that a parametric representation of the reactions in the system is not possible without the infusion of addition information. In this case, one could explore what information would be needed to augment the rank of the system sufficiently to execute DFE. This exploration could consist of screening the system in the following fashion. Assume that unknown flux UF1 can be identified outside DFE. If so, it can be merged with the vector **b**, which contains the slopes of metabolites at different time points. In effect, this merger reduces the number of non-zero entries in the stoichiometric matrix **N** (for example, compare Eqs. (20) and (23)). Compute **D** or null(**N**) to test whether the system is now characterizable. If so, independent information on UF1 would suffice to solve the problem. If not, study the next unknown fluxes UF2, UF3, etc. If none of these analyses succeeds, study combinations of unknown fluxes. Because each step requires only the computation of **D** or null(**N**), which is very fast, it should be possible to devise an automated screening algorithm, which would result in sets of fluxes to be determined to make DFE feasible. Ultimately the screening process will terminate, although not necessarily with the only possible solution, because if almost all fluxes were known, the remaining few fluxes could certainly be characterized.

In a similar vein, **N**, **D** and null(**N**) may be explored to gauge whether the pooling of metabolites or the merging of fluxes would make a pathway system characterizable. As typical examples shown here, the two directions of a reversible reaction were merged into a net flux, and a small cyclic module was condensed into a single pool, thereby reducing the discrepancy between the numbers of fluxes and metabolites. Similarly, other subsystems could be pooled, which would lead to a loss of detail within the subsystems, but could make the overall system characterizable.

Once the system is characterizable, metabolic time series data need to be obtained and all fluxes are graphically identified. Subsequently, the graphical trends are to be converted into mathematical representations. These steps are highly dependent on the repertoire of candidate functions and the quality of the metabolic times series data. Related issues of experimental noise and uncertainty, missing datasets, and mass leaks from the system were discussed elsewhere [25], but remain to be issues worth pursuing.

Overall, the procedure may appear to be complicated. However, one has to consider what is being achieved if DFE can be applied. Namely, it is possible to "see" from the graphical representations what the individual fluxes governing the system look like in dependence on time or on the metabolites that affect them. It appears that this "insight" is not achievable with other existing methods.

Acknowledgements

I would like to thank Luis L. Fonseca, Po-Wei Chen, two anonymous reviewers and an editor for valuable, constructive comments. This work was funded in part by the National Science Foundation (Projects MCB-0946595 (PI: EOV) and ARRA-0928196 (PI: E. Mann)), the National Institutes of Health (Projects NIH-GM063265 (PI: Y.A. Hannun) and P01-ES016731 (PI: G. W. Miller)), the BioEnergy Science Center (BESC; PI: Paul Gilna), a US. Department of Energy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science, and the Georgia Research Alliance. The funding agencies are not responsible for the content of this article.

Appendix A

Numerical. characteristics of artificial data associated with the example in Eq. (3)

The illustration of DFE with Fig. 2 and Eq. (3) requires metabolic time series data, which are presented in Table A.1. In a real example, these data would have been extracted from experimental time courses as they are shown in the idealized graphs of Fig. A.1.

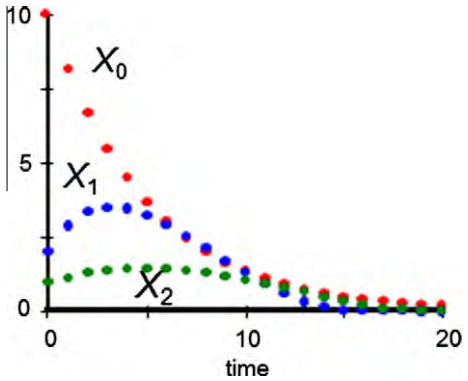

Fig. A.1. Artificial, noise-free metabolic time series data, computed here from Eq. (3), but in a real application corresponding to an experiment with the pathway in Fig. 2, in which a bolus of X_0 is supplied externally and used up within about 20 time units.

Table A.1
Metabolite concentrations and slopes associated with the simple pathway system in Eq. (3).

t	X_0	v_1	X_1	X_2	S_1	S_2
0	10	2	2	1	1.151472	0.09852814
1	8.187308	1.637462	2.880906	1.142217	0.6457941	0.1574867
2	6.703201	1.34064	3.342895	1.286703	0.2975619	0.1254981
3	5.488117	1.097623	3.507693	1.388293	0.04525962	0.07727053
4	4.49329	0.898658	3.456303	1.442177	−0.1380429	0.03144606
5	3.678795	0.7357589	3.24938	1.453391	−0.2678741	−0.007870263
6	3.011942	0.6023885	2.934854	1.428643	−0.3547169	−0.04059536
7	2.46597	0.493194	2.551702	1.374082	−0.4062285	−0.06767775
8	2.018965	0.4037931	2.132225	1.294808	−0.4282133	−0.09019441
9	1.652989	0.3305978	1.703618	1.1949	−0.4251398	−0.1090841
10	1.353352	0.2706705	1.289138	1.077603	−0.4004636	−0.1250771
11	1.108031	0.2216061	0.9089883	0.945547	−0.3568811	−0.1386591
12	0.9071783	0.1814357	0.5809344	0.8010357	−0.2966265	−0.1499712
13	0.7427346	0.1485469	0.3205111	0.6465308	−0.2220953	−0.1584515
14	0.6080996	0.1216199	0.1400759	0.4858637	−0.137816	−0.1615551
15	0.4978698	0.09957397	0.04385477	0.3280041	−0.05745645	−0.1504126
16	0.4076214	0.08152427	0.01288757	0.1956741	−0.01286664	−0.1089796
17	0.3337321	0.06674643	0.005798956	0.1125757	−0.003972747	−0.05996322
18	0.2732368	0.05464735	0.003036819	0.06852463	−0.001871244	−0.03133076
19	0.2237073	0.04474147	0.001679863	0.04511861	−0.000958626	−0.01718472
20	0.1831561	0.03663122	0.0009648202	0.03176125	−0.00052187	−0.01033442

References

[1] I.-C. Chou, E.O. Voit, Recent developments in parameter estimation and structure identification of biochemical and genomic systems, Math. Biosci. 219 (2009) 57.
[2] E.J. Crampin, S. Schnell, P.E. McSharry, Mathematical and computational techniques to deduce complex biochemical reaction mechanisms, Prog. Biophys. Mol. Biol. 86 (2004) 77.
[3] P. Gennemark, D. Wedelin, Efficient algorithms for ordinary differential equation model identification of biological systems, IET Syst. Biol. 1 (2007) 120.
[4] E.O. Voit, Biochemical Systems Theory (BST): A review, International Scholarly Research Network (ISRN) – Biomathematics (2013) 1–53, Article 897658.
[5] S. Gugushvili, C.A.J. Klaassen, $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing, Bernoulli 18 (2012) 1061.
[6] R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, J.P. Sethna, Universally sloppy parameter sensitivities in systems biology models, PLoS Comput. Biol. 3 (2007) 1871.
[7] G. Jia, G.N. Stephanopoulos, R. Gunawan, Parameter estimation of kinetic models from metabolic profiles: two-phase dynamic decoupling method, Bioinformatics 27 (2011) 1964.
[8] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics 25 (2009) 1923.
[9] S. Srinath, R. Gunawan, Parameter identifiability of power-law biochemical system models, J. Biotechnol. 149 (2010) 132.
[10] M. Vilela, S. Vinga, M.A. Maia, E.O. Voit, J.S. Almeida, Identification of neutral sets of biochemical network models from time series data, BMC Syst. Biol. 3 (2009) 47.
[11] E.O. Voit, What if the fit is unfit? Criteria for biological systems estimation beyond residual errors, in: M. Dehmer, F. Emmert-Streib, A. Salvador (Eds.), Applied Statistics for Biological Networks, J. Wiley and Sons, New York, 2011, pp. 183–200.
[12] A. Lotka, Elements of Physical Biology, Williams and Wilkins, Baltimore, 1924. Reprinted as 'Elements of Mathematical Biology', Dover, New York, 1956.
[13] V. Volterra, Variazioni e fluttuazioni del numero d'individui in specie animali conviventi, Mem. R. Accad. dei Lincei 2 (1926) 31.
[14] R.M. May, Stability and Complexity in Model Ecosystems, Princeton University Press, 1973.
[15] M. Peschel, W. Mende, The Predator-Prey Model: Do we Live in a Volterra World?, Akademie-Verlag, Berlin, 1986
[16] L. Michaelis, M.L. Menten, Die Kinetik der Invertinwirkung, Biochemische Zeitschrift 49 (1913) 333.
[17] M.A. Savageau, Michaelis-Menten mechanism reconsidered: implications of fractal kinetics, J. Theor. Biol. 176 (1995) 115.
[18] N.V. Torres, E.O. Voit, Pathway Analysis and Optimization in Metabolic Engineering, Cambridge University Press, Cambridge, UK, 2002.
[19] E.O. Voit, P.J. Sands, Modeling forest growth I. Canonical approach, Ecol. Model. 86 (1996) 51.
[20] E.O. Voit, Modelling metabolic networks using power-laws and S-systems, Essays Biochem. 45 (2008) 29.
[21] E.O. Voit, A First Course in Systems Biology, Garland Science, New York, NY, 2012.
[22] M.A. Savageau, Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology, Addison-Wesley Pub. Co., 1976. Advanced Book Program (reprinted 2009), Reading, Mass.
[23] E.O. Voit, Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists, Cambridge University Press, Cambridge, UK, 2000.
[24] G. Goel, I.C. Chou, E.O. Voit, System estimation from metabolic time-series data, Bioinformatics 24 (2008) 2505.
[25] E.O. Voit, G. Goel, I.-C. Chou, L. da Fonseca, Estimation of metabolic pathway systems from different data sources, IET Syst. Biol. 3 (2009) 513.
[26] I.-C. Chou, E.O. Voit, Estimation of dynamic flux profiles from metabolic time series data, BMC Syst. Biol. 6 (2012).
[27] L.L. Fonseca, C. Sánchez, H. Santos, E.O. Voit, Complex coordination of multi-scale cellular responses to environmental stress, Mol. BioSyst. 7 (2011) 731.
[28] E.O. Voit, J.S. Almeida, S. Marino, R. Lall, G. Goel, A.R. Neves, H. Santos, Regulation of glycolysis in *Lactococcus lactis*: an unfinished systems biological case study, IEE Proc. Syst. Biol. 153 (2006) 286.
[29] J. Srividhya, E.J. Crampin, P.E. McSharry, S. Schnell, Reconstructing biochemical pathways from time course data, Proteomics 7 (2007) 828.
[30] A. Kinoshita, K. Tsukada, T. Soga, T. Hishiki, Y. Ueno, Y. Nakayama, M. Tomita, M. Suematsu, Roles of hemoglobin allostery in hypoxia-induced metabolic alterations in erythrocytes: simulation and its verification by metabolome analysis, J. Biol. Chem. 282 (2007) 10731.
[31] T. Heinrich, S. Schuster, The Regulation of Cellular Systems, Chapman and Hall, New York, 1996.
[32] G.R. Gavalas, Nonlinear Differential Equations of Chemically Reacting Systems, Springer-Verlag, Berlin, 1968.
[33] B.Ø. Palsson, Systems Biology: Properties of Reconstructed Networks, Cambridge University Press, New York, 2006.
[34] E.O. Voit, M.A. Savageau, Power-law approach to modeling biological systems. III Methods of analysis, J. Ferment. Technol. 60 (1982) 223.
[35] E.O. Voit, J. Almeida, Decoupling dynamical systems for pathway identification from metabolic profiles, Bioinformatics 20 (2004) 1670.
[36] J.M. Varah, A spline least squares method for numerical parameter estimation in differential equations, SIAM J. Sci. Stat. Comput. 3 (1982) 28.
[37] A. Bjerhammar, Application of calculus of matrices to method of least squares; with special references to geodetic calculations, Trans. Roy. Inst. Tech. Stockholm 49 (1951) 1.
[38] E.H. Moore, On the reciprocal of the general algebraic matrix, Bull. Am. Math. Soc. 26 (1920) 394.
[39] R. Penrose, A generalized inverse for matrices, Proc. Cambridge Philos. Soc. 51 (1955) 406.
[40] A. Albert, Regression and the Moore–Penrose Pseudoinverse, Academic Press, New York, London, 1972.
[41] R.C. Aster, B. Borchers, C.H. Thurber, Parameter Estimation and Inverse Problems, Academic Press, Amsterdam, 2013.
[42] A. Ben-Israel, T.N.E. Greville, Generalized Inverses: Theory and Applications, Springer, New York, 2003.
[43] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187.
[44] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.

Please cite this article in press as: E.O. Voit, Characterizability of metabolic pathway systems from time series data, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.01.008