Vol. 19 no. 8 2003, pages 1019–1026
DOI: 10.1093/bioinformatics/btg120

Observing and interpreting correlations in metabolomic networks

R. Steuer[1,*], J. Kurths[1], O. Fiehn[2] and W. Weckwerth[2]

[1]University Potsdam, Nonlinear Dynamics Group, Am Neuen Palais 10, 14469 Potsdam and [2]Max-Planck-Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received on September 17, 2002; revised on December 3, 2002; accepted on January 14, 2003

ABSTRACT

Motivation: Metabolite profiling aims at an unbiased identification and quantification of all the metabolites present in a biological sample. Based on their pair-wise correlations, the data obtained from metabolomic experiments are organized into metabolic correlation networks and the key challenge is to deduce unknown pathways based on the observed correlations. However, the data generated is fundamentally different from traditional biological measurements and thus the analysis is often restricted to rather pragmatic approaches, such as data mining tools, to discriminate between different metabolic phenotypes.

Methods and results: We investigate to what extent the data generated networks reflect the structure of the underlying biochemical pathways. The purpose of this work is 2-fold: Based on the theory of stochastic systems, we first introduce a framework which shows that the emergent correlations can be interpreted as a 'fingerprint' of the underlying biophysical system. This result leads to a systematic relationship between observed correlation networks and the underlying biochemical pathways. In a second step, we investigate to what extent our result is applicable to the problem of reverse engineering, i.e. to recover the underlying enzymatic reaction network from data. The implications of our findings for other bioinformatics approaches are discussed.

Contact: steuer@agnld.uni-potsdam.de

INTRODUCTION

Metabolomics has the ultimate goal of providing a comprehensive and unbiased identification and quantification of all metabolites present in a biological sample (Sauter *et al.*, 1991; Tweeddale *et al.*, 1998; Oliver *et al.*, 1998; Fiehn, 2002). Recent advances in laboratory technology allowed to automatically quantify more than 1000 distinct compounds from a single leaf extract and more than 500 compounds from potato tubers. With these capabilities at hand, the use of metabolomic methods to significantly extend and enhance the power of existing functional genomics approaches has already been demonstrated (Fiehn *et al.*, 2000). Based on the pair-wise correlation between their respective concentrations in a given sample, metabolites are integrated into metabolic correlation networks (Weckwerth and Fiehn, 2002). The resulting networks show a remarkable degree of complexity, but their relationship to biological function and biochemical pathways is as yet poorly understood (Marcotte, 2001). Consequently, the analysis is mostly restricted to rather pragmatic approaches (Roessner *et al.*, 2000; Kose *et al.*, 2001; Taylor *et al.*, 2002), which are of major biotechnological interest, but require no recourse to the actual biological or biochemical 'meaning' of the data.

In this work, we will focus on the interpretation of these data-generated networks in terms of the underlying biochemical pathways. The paper is organized as follows: After a brief description of data acquisition and pre-processing, we elucidate a possible mechanism for generating correlations within metabolic networks. As a first step, we show how unclearly related the observed pattern of correlations and the underlying metabolic pathways are. Based on the theory of stochastic systems, we propose a link and argue that the emergent pattern of correlations in a metabolic network is a direct consequence of the underlying enzymatic system. Using a linear approximation, it is possible to give this relationship explicitly in terms of the Jacobian of the system. This result provides a conceptual basis for traditional data mining tools to treat the observed pattern of correlations as a 'fingerprint' of the state of the metabolic system. In the second part of this work, we focus on smaller (sub-)systems and investigate to what extent our result enables us to reconstruct the underlying system from the measured data. In the last section, we will give the conclusions and summarize our results.

[*]To whom correspondence should be addressed.

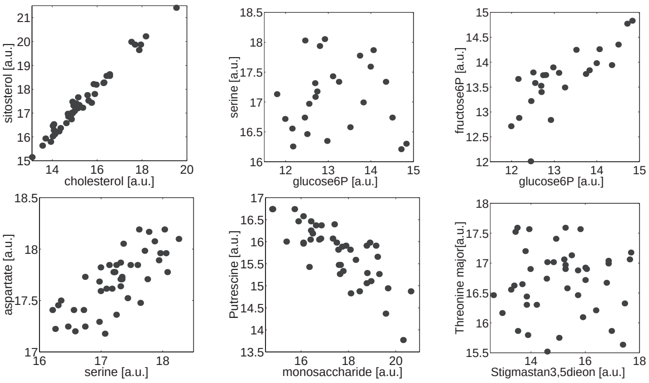

Fig. 1. Examples of metabolite:metabolite scatter-plots. Each dot corresponds to a simultaneous measurement of two metabolite concentrations (in arbitrary units) within a single sample. All samples were obtained simultaneously from tuber tissue of identical genotypes. The examples are drawn from a large data set of 657 measured metabolites with up to 43 measurements available for each metabolite. As can be easily observed, the metabolite concentrations show different degrees of correlation with each other.

DATA ACQUISITION AND ANALYSIS

Numerous techniques exist for metabolite detection and feasibility studies for applying profiling techniques to plant metabolism have been reported in the literature (Fiehn *et al.*, 2000). Since these provide the starting point of our considerations, we will give a brief outline of metabolomic data acquisition and the current status of data analysis in the following. The experimental setup to follow the relative amounts of hundreds of compounds in potato plant extracts (*Solanum tuberosum*) has already been presented elsewhere and we encourage the reader to consult the experimental literature for details (Roessner *et al.*, 2000; Fiehn *et al.*, 2000). In the present study, we premise that our semi-automated data acquisition approach allows an accurate identification and quantitation of plant metabolites. This implies that the intensity of any metabolite can be directly compared to the intensity of this metabolite in another sample.

The starting point of a metabolomic analysis is thus a $M \times N$ matrix of metabolite concentrations where M denotes the number of metabolites and N the number of samples. Figure 1 shows a visualization of the acquired data in metabolite:metabolite scatter-plots. As already visible, the concentration of a given metabolite does not vary independently, but may correlate with other metabolites. The observed correlations between two metabolite concentrations can be quantified by the correlation coefficient C_{ij} (sometimes referred to as Pearson Correlation)

$$C_{ij} = \frac{\Gamma_{ij}}{\sqrt{\Gamma_{ii}\,\Gamma_{jj}}} \tag{1}$$

where Γ_{ij} denotes the co-variance of two metabolite concentrations S_i and S_j.

$$\Gamma_{ij} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle \qquad i, j = 1, \ldots, M \tag{2}$$

The correlation coefficient implicitly defines a ‘distance’ between metabolites and usually serves as a starting point for many data mining tools, such as various types of clustering algorithms (D’haeseleer *et al.*, 2000).

A different approach to visualize the correlation matrix is to organize the metabolites into metabolic correlation networks (Weckwerth *et al.*, 2001; Kose *et al.*, 2001). Depending on, whether their correlation $\left|C_{ij}\right|$ exceeds a given threshold C^{t}, two metabolites S_i and S_j are connected with a ‘link’. The resulting network may then be interpreted as a graph whose vertices are given by the metabolites and whose edges depend on whether two metabolites are correlated or not. The only free parameter in this procedure is the threshold C^{t}, which is usually chosen in such a way that the correlations with $C_{ij} > C^{\mathrm{t}}$ may be treated as significant with respect to a certain probability. A graphical representation of the network is obtained by assigning each metabolite to coordinates in a two-dimensional plane, such that the pairwise distances approximately reflect the ‘similarity’ given by the correlation matrix (Weckwerth *et al.*, 2001; Arkin and Ross, 1995). Note that in the following the essential input of our method is always the observed co-variance matrix Equation (2), from which both, the correlation matrix and the correlation network, may be deduced.

In this work, we focus on the interpretation of these data-generated networks in terms of the biochemical pathways and will study the following questions: Is there a straightforward connection between the underlying enzymatic system and the observed correlations? Can we deduce novel pathways, based on the observed co-variance matrix?

THE INTERPRETATION OF CORRELATIONS

To answer the above described questions even partially, we start by pointing out some facts that are crucial for our further analysis: All samples were obtained simultaneously from an ensemble of identical genotypes. Neither of the experiments included the application of particular stress factors (such as water deficiency). Still, the metabolite concentrations varied considerably. We argue that this variability must have biological causes, reflecting the (intrinsic) flexibility of metabolic networks in the studied populations (Weckwerth and Fiehn, 2002). This view is supported by the observation that metabolite concentrations do not vary independently, but show strong correlations with the concentrations of other metabolites. Thus, our starting assumptions are: (i) the observation of correlations implies biological variability. If all metabolites remain at their steady state level and the variability is just given by measurement errors[†], no correlations could be observed. (ii) the observation of correlations shows that the metabolite concentrations are dependent on each other, this dependence must be strongly connected to the underlying biophysical system.

In the subsequent sections, we will adopt the following view as a working hypothesis: Cell metabolism constitutes a complex dynamical system, which is continuously subject to fluctuations. These fluctuations arise from a continuously changing environment, as well as from complex patterns of regulation, generated by the network itself. These fluctuations induce variability in certain metabolites, propagate through the network and generate an emergent pattern of correlations.

An example

To illustrate these ideas more clearly, we will now consider a numerical example. We assume that a reaction mechanism relies on the supply of certain metabolites. Further, we assume that this supply is not constant but a fluctuating quantity. These variations will affect the concentration of other metabolites and finally create an observable pattern of correlations.

To simulate this numerically, we use a rather detailed model of glycolysis, as described by Hynne *et al.* (2001). This example will also serve to demonstrate that there is no straightforward connection between observed correlations and the underlying reaction network. A schematic view of the model is given in Figure 2, for further details, in particular the rate mechanisms and constants, we refer the reader to the original publication (Hynne *et al.*, 2001).

We shall emphasize that the choice of this particular model is not a quest for realism, but is owed to the fact that it contains non-linear rate laws based on the biochemical literature, and considers explicitly the production and consumption of co-factors, such as ATP. The parameters of the model were optimized to describe the behaviour near the Hopf-bifurcation. In our simulations, however, the external glucose concentration was shifted below the bifurcation point, to ensure the existence of a single steady state. Also, it should be emphasized that this model does not necessarily give a realistic representation of glycolysis in potato tubers, but merely stands for an arbitrary, but reasonably complex, reaction mechanism.

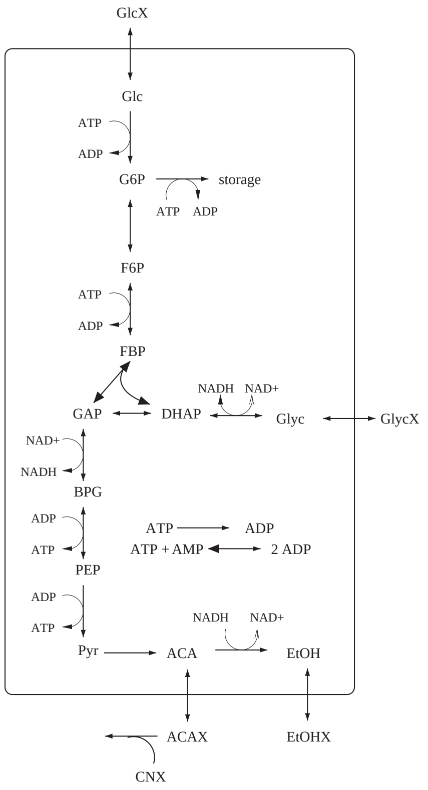

Fig. 2. A scheme of the glycolysis pathway corresponding to the model of Hynne *et al.* (2001). The model consists of 24 reactions, whose parameters have been fitted to experimental data. The model was optimized to describe glycolyic oscillations in yeast. In our simulation, we shift the parameters to $[GlcX_0] = 30$ a.u. and $[CNX_0] = 25$ a.u. (arbitrary units) to ensure the existence of a non-oscillating steady state.

In order to introduce variability in the network, the external glucose is considered to be a time-dependent (stochastic) variable $GlcX(t)$. The time evolution of $GlcX(t)$ can thus be modeled by a stochastic differential equation,

$$\frac{d[\mathrm{GlcX}]}{\mathrm{d}t} = f[\ldots] + \sqrt{2D}\xi(t) \tag{3}$$

where $f[\ldots]$ denotes the original deterministic equation and $\xi(t)$ Gaussian white[‡] noise with zero mean and unit variance.

$$\langle \xi(t) \rangle = 0 \qquad \langle \xi(t)\xi(t') \rangle = \delta(t - t') \tag{4}$$

At this point, we have to clarify our use of the term ‘fluctuation’. Recently, there has been much interest in fluctuations and stochastic mechanisms within molecular networks (Rao *et al.*, 2002; Thattai and van Oudenaarden, 2001; Morton-Firth and Bray, 1998; McAdams and Arkin, 1997). Therein, the term fluctuation refers to the *internal* or *intrinsic* noise caused by the fact that the system consists of (a low number of) discrete particles. These

[†] For a discussion of the typical precision of the measurement tools and other experimental aspects, see (Roessner *et al.*, 2000; Fiehn *et al.*, 2000).

[‡] With respect to relevant timescales.

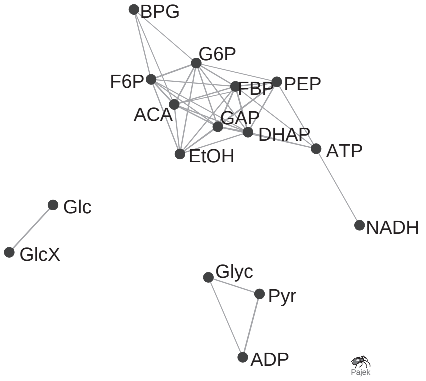

Fig. 3. A metabolic correlation network based on the simulated output of the glycolysis model. The position of the vertices were determined with the Kamada–Kawai algorithm, as implemented in the software package *Pajek* (Batagelj and Mrvar, 1998), using the full correlation matrix. Edges indicate a correlation coefficient $\|C_{ij}\| > 0.5$. Can we relate this network to the underlying pathways, depicted in Figure 2?

molecular fluctuations are inherent in the mechanism by which the system evolves (van Kampen, 1992). In contrast to that, our measurements consist of an average over tens of thousands of cells. Fluctuations on the molecular level are therefore averaged out.

In this work, the term ‘fluctuations’ thus refers to macroscopic fluctuations, affecting a large number of cells simultaneously. These may be generated by a (continuously) changing environment, as well as by complex regulation in (other parts of) the metabolic network. It should be noted that the formulation of irregular functions of time as stochastic processes is widely utilized in physics, also when the fluctuations are of unspecified or unknown provenance (van Kampen, 1992).

In our example, the fluctuations in the external glucose concentration propagate down the pathway. To obtain a hypothetical ‘measurement’, the system is integrated numerically[§] and (after excluding transients) the concentrations of all metabolites are recorded simultaneously at a given point in time. Further ‘measurements’ are obtained by repeating this procedure with different realizations of the fluctuations[¶]. Figure 3 shows a visualization of the resulting correlation matrix as a metabolic correlation network. Examples of metabolite:metabolite scatterplots are shown in Figure 4.

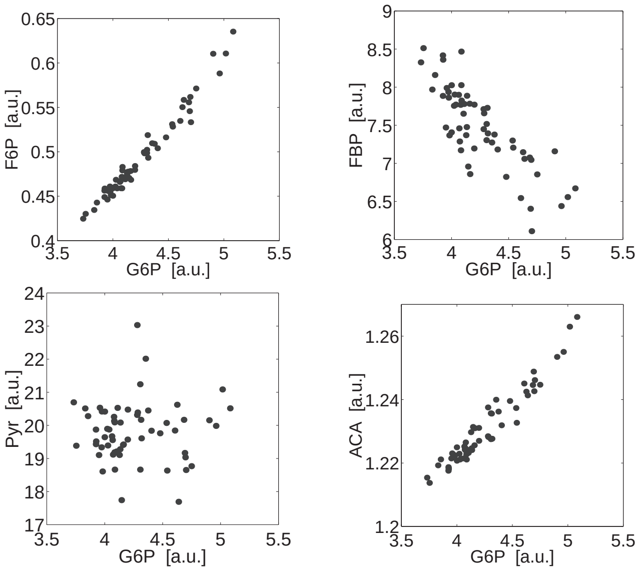

Fig. 4. Examples of simulated metabolite concentrations, obtained numerically from the model depicted in Figure 2. Shown are metabolite:metabolite scatterplots of $G6P$ with other metabolites further down the glycolytic chain. In the simulation, the amplitude of the fluctuations was set to $D = 0.5$.

We observe a strong correlation between glucose-6P (G6P) and fructose-6P (F6P). Between G6P and FBP, a negative correlation is observed. Next, there seems to be almost no correlation between G6P and pyruvate (Pyr), but, surprisingly, again a very strong correlation between G6P and acetaldehyde (ACA). The numerical values are given in Table 1. Obviously, at a first glance, the observed pattern of correlations has no clear connection to the underlying pathway and an intuitive interpretation of the observed correlations must almost unavoidably fail. Moreover, even with detailed knowledge about the reaction mechanism, the prediction of expected correlations does not seem a trivial task. In the next section, we will therefore develop a systematic approach, which allows to deduce the correlation matrix from a given reaction scheme.

A SYSTEMATIC APPROACH

Not only metabolomic network analysis, but also many other bioinformatics algorithms, rely on the interpretation of observed correlations. Thus a prediction of the expected correlation matrix, given a certain reaction system, is potentially of importance for further improvements of various algorithms. Here, we consider an arbitrary metabolic reaction network, in which certain metabolites are subject to fluctuations. The system is given by a set of non-linear

[§] We used a modified Runge–Kutta 2nd order algorithm. Note that the numerical simulation of stochastic differential equations involves some additional difficulties, compared to the deterministic case. See e.g. (Mannella, 2000) for more details and a short overview over various algorithms.

[¶] This is equivalent to simultaneously recording the concentrations of all metabolites at successive timepoints, provided that the time interval between two ‘measurements’ is sufficiently long.

Table 1. The values of the correlation coefficient C_{XY} obtained numerically from the model. The correlation coefficient was estimated using 100 datapoints and averaged over 50 realizations with σ_C denoting the standard deviation

Metabolite X	Metabolite Y	Average C_{XY}	Standard deviation σ_C
G6P	F6P	0.982	0.003
G6P	FBP	−0.86	0.03
G6P	Pyr	−0.02	0.08
G6P	ACA	0.97	0.03

differential equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{S} = \mathbf{N}\nu(\mathbf{S}) = \mathbf{f}(\mathbf{S}) \tag{5}$$

where $\mathbf{S}$ denotes the vector of metabolite concentrations, $\mathbf{N}$ the stoichiometric matrix and $\nu(\mathbf{S})$ the (non-linear) vector of fluxes (Heinrich and Schuster, 1996). For simplicity, we focus on fluctuations near the (stable) steady state $\mathbf{S^0}$

$$\mathbf{X}(t) = \mathbf{S}(t) - \mathbf{S^0} \tag{6}$$

and use a local linear approximation of Equation (5).

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{X} \approx \mathbf{J}\,\mathbf{X} \qquad \text{with} \quad \mathbf{J} := \mathbf{N}\left.\frac{\partial \nu}{\partial \mathbf{S}}\right|_{\mathbf{S^0}} \tag{7}$$

Note that the entries in the Jacobian $\mathbf{J}$ are related to the elasticity coefficients, as used within Metabolic Control Analysis (MCA; Heinrich and Schuster, 1996). Some of the metabolites are subject to external fluctuations, modeled by a Langevin-type equation

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \sum_j J_{ij}\,X_j + \sqrt{2\,D_i}\,\xi_i(t) \tag{8}$$

with ξ_i being Gaussian white noise, with zero mean and unit variance. The (stationary) solution $P(\mathbf{x})$ of Equation (8) is known to be a multivariate Gaussian distribution (van Kampen, 1992), thus fully characterized by its first and second moments. To obtain the covariance matrix Γ, we use the corresponding (stationary) Fokker–Planck equation for $P(\mathbf{x})$.

$$0 = -\sum_{ij} J_{ij}\frac{\partial}{\partial x_i}x_j P + \sum_{ij} D_{ij}\frac{\partial^2 P}{\partial x_i \partial x_j} \tag{9}$$

where $D_{ij} = \delta_{ij} D_i$ is diagonal. We are interested in the quantities

$$\langle X_k X_l\rangle = \int x_k x_l\; P(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{10}$$

Multiplying Equation (9) with $x_k x_l$ and integrating yields (van Kampen, 1992; Honerkamp, 1990)

$$0 = \sum_j J_{kj}\langle X_l X_j\rangle + \sum_j J_{lj}\langle X_k X_j\rangle + 2D_{kl} \tag{11}$$

Equation (11) may be rewritten in terms of the co-variance matrix Γ, as defined in Equation (2) (van Kampen, 1992)

$$\mathbf{J}\,\Gamma + \Gamma\mathbf{J}^{\mathrm{T}} = -2\mathbf{D} \tag{12}$$

where $\mathbf{J}^{\mathrm{T}}$ denotes the transpose of $\mathbf{J}$. Equation (12) establishes a fundamental relationship between the observed co-variance (and hence the correlations) and the underlying reaction network. Given an arbitrary Jacobian $\mathbf{J}$ and the fluctuation matrix $\mathbf{D}$, the elements of Γ are given as the solution of a linear equation.

We may now apply Equation (12) to our earlier described example of glycolysis. The construction of the Jacobian from the rate laws is straightforward[‖]. By inserting it into Equation (12), we obtain an expression for the co-variance matrix Γ, which then results in the correlation matrix $\mathbf{C}$ (Equation 1)[**]. Between G6P and F6P, we obtain $C_{\mathrm{G6P,F6P}} \approx 0.98$, between G6P and FBP the correlation is $C_{\mathrm{G6P,FBP}} \approx -0.88$, further $C_{\mathrm{G6P,Pyr}} \approx -0.08$ and $C_{\mathrm{G6P,ACA}} \approx 0.99$. This is in good correspondence with the pattern of correlations given in Table 1. Note that the values in Table 1 were found numerically using the full non-linear model. Small deviations must therefore be attributed to the linear approximation.

In a situation where the linear approximation does not apply (e.g. oscillations), it is sometimes possible to solve Equation (10) using a numerical solution of the (non-linear) Fokker–Planck equation. While this is computationally demanding, it emphasizes that even in a more general setting, the observed correlations can be interpreted as a direct consequence of the underlying system. However, we expect the linear approximation to hold in many cases. This is supported by the fact that MCA, which is a genuinely linear theory, was found to yield reasonable results for many metabolic systems (Heinrich and Schuster, 1996). We note that in our interpretation, even within the linear approximation, the observed correlation networks must not necessarily be static, but can also describe dynamic entities. This can be included by considering a time dependent Jacobian $\mathbf{J}(t)$[††].

REVERSE ENGINEERING

Having established a relationship between the observed co-variance and the underlying dynamical system, the

[‖] For the model, depicted in Figure 2, the Jacobian has 82 nonzero entries. We estimate it numerically, using the MATLAB routine `numjac`.

[**] To solve Eq (12) for an (moderately large) Jacobian, we use the software package MATHEMATICA.

[††] In this case, one obtains a differential equation for the co-variance matrix: $d\Gamma/dt = \mathbf{J}(t)\,\Gamma + \Gamma\mathbf{J}^{\mathrm{T}}(t) + 2\mathbf{D}$.

crucial question is: Are we able to deduce properties of the underlying system from the observed co-variance matrix? Is it possible to reconstruct the Jacobian given a measured co-variance matrix Γ?

In order to address the above-mentioned questions, we shift the focus towards smaller (sub-)system. In particular, we refer to the work of Arkin *et al.* (1997). Therein the response of a reaction pathway to random changes (fluctuations) in the concentration of a subset of 'input' species was recorded experimentally. The resulting correlations were shown to yield an informative non-causal diagram representing the regulatory structure of the system. Within this setup, the fluctuations are now introduced deliberately. Thus, in the following, we assume the fluctuation matrix **D** to be known.

To make the calculation more transparent, we consider a very simple example, as shown in Figure 5. All reactions are modeled as irreversible first-order mass-action. We start with constructing the correlation matrix. The Jacobian is given as

$$\mathbf{J} = \begin{pmatrix} -(k_{12}+k_{13}) & 0 & 0 \\ +k_{12} & -k_{23} & 0 \\ +k_{13} & +k_{23} & -k_{\text{out}} \end{pmatrix} \tag{13}$$

The fluctuations shall affect only the metabolite S_1, thus $D_{11} = D$, while all other entries in **D** are zero. The matrix Γ is symmetric.

$$\mathbf{D} = \begin{pmatrix} D & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} \\ \Gamma_{12} & \Gamma_{22} & \Gamma_{23} \\ \Gamma_{13} & \Gamma_{23} & \Gamma_{33} \end{pmatrix}$$

Inserting this into Equation (12) yields a linear system of equations, specifying the $M(M+1)/2$ independent entries of Γ. We choose $k_{12} = k_{13} = k_{\text{out}} = 1$, $k_{23} = 2$ and, after a simple calculation, obtain

$$\Gamma = \begin{pmatrix} \frac{D}{2} & \frac{D}{8} & \frac{D}{4} \\ & \frac{D}{16} & \frac{D}{6} \\ & & \frac{7D}{12} \end{pmatrix} \tag{14}$$

The resulting correlation coefficients are $C_{12} = \sqrt{1/2} \approx 0.71$, $C_{13} = \sqrt{3/14} \approx 0.46$ and $C_{23} = \sqrt{16/21} \approx 0.87$. Figure 6 shows a comparison with a numerical estimation, which is in good agreement with the theoretical prediction. Note that within a linear approximation, the correlation coefficients do not depend on the amplitude D of the fluctuations.

We may now address, whether or not we are able to reconstruct the Jacobian from the observed covariance matrix. In general, this is, of course, not the case. The matrix Γ is symmetric, for a system of M metabolites it contains only $M(M+1)/2$ independent entries. In contrast to that the Jacobian **J** has M^2 entries, leaving $M(M-1)/2$ entries unspecified. Still, as we shall see below, an observed co-variance matrix does hold *some* information about the underlying system.

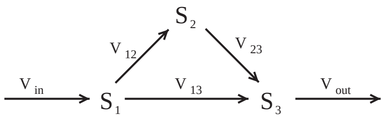

Fig. 5. A simple reaction mechanism. All reactions are modeled as irreversible first-order mass-action kinetics: $\nu_{ij} = k_{ij}\,S_i$, $\nu_{\text{out}} = k_{\text{out}}\,S_3$ and $\nu_{\text{in}} = const$. The steady state is $S_1^0 = \nu_{\text{in}}/(k_{12}+k_{13})$, $S_2^0 = k_{12}S_1^0/k_{23}$ and $S_3^0 = \nu_{\text{in}}/k_{\text{out}}$. For the simulation the parameters were $\nu_{\text{in}} = 5$, $k_{12} = k_{13} = k_{\text{out}} = 1$ and $k_{23} = 2$.

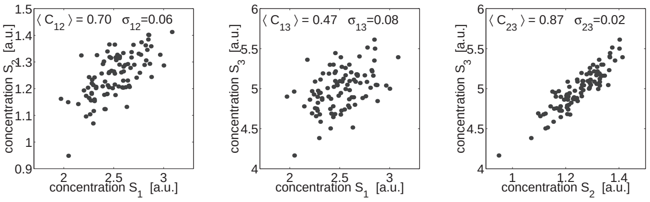

Fig. 6. A numerical analysis of the reaction scheme shown in Figure 5. The metabolite S_1 is subject to fluctuations, which then propagate through the network. The concentrations fluctuate around the steady-state $S_1^0 = 2.5$, $S_2^0 = 1.25$ and $S_3^0 = 5.0$. Also shown is the observed correlation coefficient C_{ij}, averaged over 100 realizations of $N = 100$ data points with σ denoting the standard deviation.

For simplicity, we assume that the covariance matrix Γ is known exactly. If we insert the expression for Γ (Equation (14)) into Equation (12) and take the fluctuation matrix **D** as known, we get a linear system of equations for the entries of the Jacobian **J**. However, there are only six (since $M = 3$) independent equations for the nine unknowns J_{ij}. Consequently, it is only possible to give a parametric solution $\hat{\mathbf{J}} = \hat{\mathbf{J}}[\lambda_1, \lambda_2, \lambda_3]$.

Table 2 shows the reconstructed Jacobian as a function of the parameters λ_i. By inserting the true values $\lambda_1 = -2$, $\lambda_2 = 0$ and $\lambda_3 = -2$, we might easily verify, that Table 2 is indeed a correct parameterization of the true Jacobian, as given in Equation (13).

$$\hat{\mathbf{J}} = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix} = \begin{pmatrix} -2 & 0 & 0 \\ 1 & -2 & 0 \\ 1 & 2 & -1 \end{pmatrix}$$

As shown, knowledge about the co-variance matrix puts $M(M+1)/2$ linear constraints on the M^2 elements of the Jacobian. Thus, our method does not lead to a

Table 2. The reconstructed Jacobian $\{J_{ij}\}$ parameterized by the unknown values λ_1, λ_2 and λ_3. Note that if the amplitude D of the fluctuations in Equation (14) is not known, this would give an additional free parameter

$J_{11} = \lambda_1$	$J_{13} = -\frac{1}{2}\,(8 + 4\lambda_1 + \lambda_2)$
$J_{12} = \lambda_2$	$J_{21} = \frac{1}{30}\,(64 + 20\lambda_1 + 2\lambda_2 - 3\lambda_3)$
$J_{22} = \lambda_3$	$J_{23} = -\frac{1}{20}\,(32 + 10\lambda_1 - \lambda_2 - 6\lambda_3)$
	$J_{31} = \frac{1}{15}\,(67 + 35\lambda_1 + 6\lambda_2 - 9\lambda_3)$
	$J_{32} = \frac{1}{5}\,(54 - 3\lambda_2 + 22\lambda_3)$
	$J_{33} = -5 - \lambda_1 - \lambda_3$

complete reconstruction of the system, even when the co-variance matrix is known exactly. However, there are some improvements possible, which shall be discussed shortly: The number of independent entries in **J** is usually considerably smaller than M^2: For methods that exploit redundancy in **J** see (Díaz-Sierra *et al.*, 1999; Klamt *et al.*, 2002). Also, in a realistic setting, most of the entries might already be known, the task being thus to specify the remaining entries. Even, if we have no particular information about the underlying system, we know that metabolic networks are, in general, sparse (Jeong *et al.*, 2000). Thus, similar to the approach developed by Yeung *et al.* (2002), one may optimize the solution to provide a Jacobian with a maximal number of 'zero' elements.

Still, the strongest limitation in a realistic example would be that the estimation of the co-variance matrix is usually affected by considerable errors. As could be observed in Figure 6, this holds in particular for the case of weak correlations. Here again, some improvements are possible: If we follow the work of Arkin *et al.* (1997), the (now deliberately) introduced fluctuations can be very slow, corresponding to a (slowly varying) constant input. In this case, the system is always at a steady state and the metabolite:metabolite scatterplots reduce to a straight line (linear relationship). Within this setting, our method resembles a co-response analysis, as introduced in Cornish-Bowden and Hofmeyr (1994) and Giersch (1995) and a more detailed analysis of this approach is left for future work.

DISCUSSION AND CONCLUSIONS

In this work, we have discussed the analysis and interpretation of metabolomic data sets acquired by high-throughput measurements. All samples were taken from plants with identical genotypes and grown under uniform conditions. Still, the concentrations of metabolites showed a remarkable degree of variability. More important, metabolite concentrations do not vary independently, but are highly interconnected via metabolic correlation networks. We argued that the observed correlations between metabolite concentrations are a result of the underlying enzymatic reaction network. According to our hypothesis, plant metabolism is a highly dynamical process which is, even under stationary and uniform conditions, continuously changing under the influence of fluctuations. These fluctuations propagate through the network and induce a specific pattern of correlations, which is then observed experimentally.

To investigate this, we have presented a systematic approach, which connects the underlying dynamical system to the observed correlation matrix for arbitrary external fluctuations. Using a linear approximation it is possible to give this relationship explicitly in terms of the Jacobian of the system. Based on this view, the pair-wise correlation network represents a snapshot of the physiological state of the plant at a given point in time.

This provides a fundamental conceptual basis for previous and future analysis of metabolomic data sets. Even if it is not possible to specify the precise origin of the fluctuations in large-scale experiments, our analysis enables to treat the observed correlations as a 'fingerprint' of the underlying biophysical system. In this way, the organization of metabolites in complex correlation networks exploits the intrinsic flexibility of metabolism to gain additional information about the state of a molecular system. Also, we can draw two important consequences from our analysis: (i) The relationship between the observed pattern of correlations and the underlying pathways is complex and cannot be given in terms of simple heuristic principles; and (ii) more importantly, a mutant plant would manifest itself by having a (slightly) altered enzymatic reaction network, thus a slightly different Jacobian. Consequently, we must expect mutants not only to show different steady state concentrations, but also a different co-variance, hence correlation matrix. And this is indeed what is observed experimentally (Weckwerth *et al.*, 2001).

As a second step, we have discussed the applicability of our results to the problem of reverse engineering. It was shown that the covariance matrix does not hold enough information to fully specify a system. However, this may be remedied if additional knowledge about the system is available. Since our present experimental setup does not allow the recording of time-series, we have restricted the analysis to instantaneous correlations. However, as a further step, one may also obtain an expression for the full time-lagged co-variance matrix $\Gamma(\tau)$. This would provide further information about the dynamical system and would also supplement the method introduced by Arkin *et al.* (1997) with a sound theoretical foundation.

Finally, we like to note that there are a vast number of possible further developments and applications of our approach. First of all, there is an increasing interest in the large-scale organization of metabolic networks, which are believed to have some, evolutionary favorable, characteristic properties (Jeong *et al.*, 2000). The large-scale metabolomic measurements might thus enable us to study these characteristics experimentally, by connecting the network structure with the structure of the observed correlation networks. Further, the traditional analysis of metabolism is almost exclusively restricted to deterministic rate functions. The response of metabolic networks to fluctuating inputs might thus reveal new, previously unknown, schemes of regulation. In addition, it should be emphasized that the study of fluctuations is not restricted to metabolic networks. Many bioinformatics algorithms rely on the correlations matrix as their essential input. In particular, the clustering of gene-expression implicitly assumes that 'co-regulated' (correlated) genes have something in common in their regulator mechanism. Our approach might facilitate further understanding of this assumption and put such principles on firmer ground.

ACKNOWLEDGEMENTS

The authors would like to thank S. Clodong, A. Sitz, B. Blasius, U. Schwarz, Cs. Zhou, A. Flöter (all *Univ. Potsdam*) and J. Selbig (*MPIMP*) for stimulating discussion and M. McKenzie for proofreading the manuscript. R. S. is supported by the HSP-N grant of the the state of Brandenburg.

REFERENCES

Arkin,A. and Ross,J. (1995) Statistical construction of chemical reaction mechanisms from measured time-series. *J. Phys. Chem.*, **99**, 970–979.

Arkin,A., Shen,P. and Ross,J. (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**, 1275–1279.

Batagelj,V. and Mrvar,A. (1998) Pajek—program for large network analysis. *Connections*, **21**, 47–57.

Cornish-Bowden,A. and Hofmeyr,J.-H.S. (1994) Determination of control coefficients in intact metabolic systems. *Biochem. J.*, **298**, 367–375.

D'haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

D́iaz-Sierra,R., Lozano,J.B. and Fairén,V. (1999) Deduction of chemical mechanisms from the linear response around steady state. *J. Phys. Chem. A*, **103**, 337–343.

Fiehn,O. (2002) Metabolomics—the link between genotype and phenotype. *Plant Mol. Biol.*, **48**, 155–171.

Fiehn,O., Kopka,J., Dörmann,P., Altmann,T., Trethewey,R.T. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Giersch,C. (1995) Determing elasticities from multiple measurements of flux rates and metabolite concentrations. *Eur. J. Biochem.*, **227**, 194–201.

Heinrich,R. and Schuster,S. (1996) *The Regulation of Cellular Systems*. Chapman & Hall, New York.

Honerkamp,J. (1990) *Stochastische Dynamische Systeme*. VCH, Weinheim, (in German).

Hynne,F., Danø,S. and Sørensen,P.G. (2001) Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biochem. J.*, **94**, 121–163.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic network. *Nature*, **407**, 651–654.

Klamt,S., Schuster,S. and Gilles,E.D. (2002) Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnol. Bioeng.*, **77**, 734–751.

Kose,F., Weckwerth,W., Linke,T. and Fiehn,O. (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, **17**, 1198–1208.

Mannella,R. (2000) A gentle introduction to the integration of stochastic differential equations. In Freund,J.A. and Pöschel,T. (eds), *Stochastic Processes in Physics, Chemistry and Biology*. Springer, Berlin, pp. 353ff.

Marcotte,E.M. (2001) The path not taken. *Nat. Biotechnol.*, **19**, 626–627.

McAdams,H. and Arkin,A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA*, **94**, 814–819.

Morton-Firth,C.J. and Bray,D. (1998) Predicting temporal fluctuations in a intracellular signalling pathway. *J. Theor. Biol.*, **192**, 117–128.

Oliver,S.G., Winson,M.K., Kell,D.B. and Baganz,F. (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, **16**, 373–378.

Rao,C.V., Wolf,D.M. and Arkin,A.P. (2002) Control, exploitation and tolerance of intracellular noise. *Nature*, **420**, 231–237.

Roessner,U., Wagner,C., Kopka,J., Trethewey,R.N. and Willmitzer,L. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.*, **23**, 131–142.

Sauter,H., Lauer,M. and Fritsch,H. (1991) Metabolic profiling of plants—a new diagnostic-technique. *Acs. Symposium Series*, **443**, 288–299.

Taylor,J., King,R.D., Altmann,T. and Fiehn,O. (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, **18**, S241–S248.

Thattai,M. and van Oudenaarden,A. (2001) Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA*, **98**, 8614–8619.

Tweeddale,H., Notley-McRobb,L. and Ferenci,T. (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ('metabolome') analysis. *J. Bacteriol.*, **180**, 5109–5116.

van Kampen,N.G. (1992) *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam.

Weckwerth,W. and Fiehn,O. (2002) Can we discover novel pathways using metabolomic analysis. *Curr. Opin. Biotechnol.*, **13**, 156–160.

Weckwerth,W., Tolstikov,V. and Fiehn,O. (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS. In *Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics*. pp. 1–2.

Yeung,M.K.S., Tegnér,J. and Collins,J.J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl Acad. Sci. USA*, **99**, 6163–6168.