Metabolomics
DOI 10.1007/s11306-012-0399-3

ORIGINAL ARTICLE

# COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data

**Xiaoliang Sun · Wolfram Weckwerth**

Received: 5 June 2011 / Accepted: 20 January 2012


**Abstract** Metabolomics emerges as one of the cornerstones in systems biology by characterizing metabolic activities as the ultimate readout of physiological processes of biological systems thereby linking genotypes with the corresponding phenotypes. As metabolomics data are high-dimensional, statistical data analysis is complex. No single technique for statistical analysis and biological interpretation of these ultracomplex data is sufficient to reveal the full information content of the data. Therefore a combination of univariate and multivariate statistics, network topology and biochemical pathway mapping analysis is in all cases recommended. Therefore, we developed a toolbox with fully graphical user interface support in MATLAB© called *cov*ariance *in*verse (COVAIN). COVAIN provides a complete workflow including uploading data, data preprocessing, uni- and multivariate statistical analysis, Granger time-series analysis, pathway mapping, correlation network topology analysis and visualization, and finally saving results in a user-friendly way. It covers analysis of variance, principal components analysis, independent components analysis, clustering and correlation coefficient analysis and integrates new algorithms, such as Granger causality and permutation entropy analysis that are not implemented in other similar softwares. Furthermore, we provide a new algorithm to reconstruct a differential Jacobian matrix of two different metabolic conditions. The algorithm is based on the assumptions of stochastic fluctuations in the metabolic network as described by us recently. By integrating the metabolomics covariance matrix and the stoichiometric matrix $N$ of the corresponding pathways this approach allows for a systematic investigation of perturbation sites in the biochemical network based on metabolomics data. COVAIN was primarily developed for metabolomics data but can also be used for other omics data analysis. A C language programming module was integrated to handle computational intensive work for large datasets, e.g., genome-level proteomics and transcriptomics data sets which usually contain several thousand or more variables. COVAIN can perform cross analysis and integration between several datasets, which might be useful to investigate responses on different hierarchies of cellular contexts and to reveal the systems response as an integrated molecular network. The source codes can be downloaded from http://www.univie.ac.at/mosys/software.html.

**Keywords** Metabolomics · Jacobian · Inverse modelling · Genotype · Phenotype · Stoichiometric matrix · Stochastic processes · Network · Perturbation sites

## 1 Introduction

Determining metabolite levels for pathway elucidation and as a measure of metabolic fine and coarse control of pathways has a long tradition in biochemistry and physiology (Meyerhof 1927, 1947; Bassham et al. 1950; Kacser and Burns 1973; Heinrich and Rapoport 1974; Aprees 1980; Cornishbowden and Hofmeyr 1994; Giersch 1994). These measurements serve as clues to understanding pathway organization. Changes on the metabolite level are closely

**Electronic supplementary material** The online version of this article (doi:10.1007/s11306-012-0399-3) contains supplementary material, which is available to authorized users.

X. Sun · W. Weckwerth (✉)
Department of Molecular Systems Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria
e-mail: wolfram.weckwerth@univie.ac.at

X. Sun
e-mail: xiaoliang.sun@univie.ac.at

Published online: 12 February 2012

related to the microenvironment. Metabolic reaction chains are able to sense environmental stimuli within seconds and milliseconds. The results are high metabolic fluctuations. It is possible to exploit this biological variance to investigate pathway structures or the regulation of different genotypes using multivariate statistics. The application of metabolomics in systems biology gives the unique opportunity to investigate whole metabolic networks instead of single pathways in response to various environmental or developmental stimuli or for gene function analysis (Weckwerth 2003). Here, one can use metabolic markers as controls and correlate these to other processes. Most of the current metabolite profiling approaches rely on the measurement of "steady state" levels. The tests for significant changes in averages or mean levels of specific metabolites will then reveal alterations in the regulation of metabolism. This multiple univariate analysis relies on the detection of statistically significant differences between sample groups. Often we observe a high biological variation of individual compounds within a set of samples from the same genotype. This high biological variability of independent biological replicates can be exploited to go beyond the classical "state-differences-question" and can reveal systemic behaviour and biochemical regulation using multivariate statistics. We and others applied metabolite profiling to various biological systems. In all these systems we observed significant pairwise correlations between specific metabolites—termed co-regulation (Weckwerth et al. 2001, 2004a, b; Mendes et al. 2005; Morgenthal et al. 2005, 2006; Kusano et al. 2007; Wienkoop et al. 2008, 2010; Mochida et al. 2009; Fukushima et al. 2011). These correlations showed conserved or altered structures between different species (Morgenthal et al. 2006) and provided the basis for constructing connectivity networks of metabolites based on Pearson's correlation coefficient. This coefficient was then facilitated to quantify the distance of the connectivity of all the measured metabolites and enabled the construction of metabolite distance maps visualized as differential metabolite correlation networks (Weckwerth et al. 2001, 2004a, b; Weckwerth 2003). We found significant alterations of these network structures depending on the genotype and environmental perturbations (Kose et al. 2001; Weckwerth et al. 2001, 2004a, b; Steuer et al. 2003a, b; Morgenthal et al. 2005, 2006; Weckwerth and Steuer 2005). A trend in these networks is a high connectivity of only a few nodes (metabolites) whereas many nodes have only a low connectivity (Weckwerth et al. 2004a, b). Thus, the degree distribution of these networks can be investigated systematically in the context of pathway connectivity (Weckwerth and Fiehn 2002; Weckwerth 2003; Weckwerth et al. 2004a, b; Morgenthal et al. 2006; Muller-Linow et al. 2007).

Based on these empirical observations we developed a stochastic model of metabolism that can explain these phenomena and provides a reasonable framework for multivariate data mining and biological interpretation of huge metabolomic experiments (Steuer et al. 2003a, b; Weckwerth 2003). Early work of Arkin and Ross (Arkin et al. 1997, 1998; Samoilov et al. 2001; Rao et al. 2002; Vance et al. 2002) and Rascher and Lüttge (Rascher et al. 2001) demonstrated the need to introduce stochastic models for the interpretation of metabolic networks. In analogy, by introducing metabolite fluctuation using stochastic differential equations for a glycolytic pathway system the putative origin of correlations in metabolomic data was proposed in order to connect these correlations to the underlying enzymatic pathway structure (Steuer et al. 2003a, b). Using these correlation networks one is capable of revealing alterations in enzymatic activity and alterations in the differential analysis of various metabolic states (Weckwerth et al. 2004a, b; Camacho et al. 2005; Morgenthal et al. 2005). Changes in the network topology point to regulatory hubs in the biochemical network because the correlation matrix of all metabolite pairs is a fingerprint of the enzymatic and regulatory reaction network (Weckwerth 2003). It is further possible to compare the measured correlation network with the proposed underlying reaction network and the corresponding numerically resolved correlation network (Steuer et al. 2003a, b; Weckwerth 2003; Morgenthal et al. 2006). Here it becomes evident that correlations cannot be predicted only on the basis of pathway connectivity. Regulatory properties, for instance the modulation of enzyme activity serve as a source of changes in the topology of the correlation network (Weckwerth 2003; Weckwerth et al. 2004a, b; Camacho et al. 2005; Weckwerth and Steuer 2005; Morgenthal et al. 2006). In studies of other groups metabolite correlation network analyses were adapted to yeast metabolism and enzyme concentration fluctuations (Camacho et al. 2005), *Medicago truncatula* cell cultures and their response to methyljasmonate as an elicitor (Broeckling et al. 2005) or lipid metabolism in a transgenic mouse model (Clish et al. 2004). From these and our studies it became clear that the analysis of dynamic metabolic networks gives the opportunity to observe in vivo regulation of dynamic biochemical networks otherwise not accessible. Some components of the biochemical networks function also as harmonic oscillators or effectors and it will be a challenge for future applications to compare experimental fluctuations of perturbed system with computer simulations of fluctuating complex reaction pathway networks (Weckwerth 2011).

However, the interrelation of an enzymatic reaction network and the resulting correlation matrix is still difficult to be interpreted (Steuer et al. 2003a, b; Weckwerth 2003; Camacho et al. 2005; Muller-Linow et al. 2007). Any alteration in the reaction network, inhibition of enzyme activity,

genetic suppression or enhancement of a reaction or addition of new pathways will result in different metabolite patterns and it is necessary to analyse these patterns systematically for functional interpretations (Weckwerth 2003). At the moment these patterns can not be predicted easily from modelling approaches (Weckwerth 2011). On the other hand there is a systematic relation between these patterns—represented by the data covariance matrix *C*—and the underlying biochemical network—structurally represented by the stoichiometric matrix *N*—as demonstrated in our and others recent work assuming a stochastic model of metabolism (Steuer et al. 2003a, b; Weckwerth 2003, 2011; Morgenthal et al. 2005, 2006; Muller-Linow et al. 2007). Because most of the algorithms for unsupervised or supervised data mining look for optimal variance and covariance discrimination of sample groups in data sets, this stochastic model of metabolism provides a fundamental relationship between multivariate statistics, metabolite profiling and biochemical regulation (Weckwerth and Morgenthal 2005).

However, no single statistical technique alone is sufficient to reveal the full information content of the data. For an introduction into current techniques and methods the following studies provide an excellent overview (Steuer et al. 2006; Smilde et al. 2010; Westerhuis et al. 2010; Hendrickx et al. 2011; Jansen et al. 2011). A combination of univariate and multivariate statistics, network topology, biochemical pathway mapping and inverse mapping of metabolic networks is in all cases recommended. Therefore, we developed a toolbox with fully graphical user interface (GUI) support in MATLAB© called *cova*riance *in*verse (COVAIN). As the covariance matrix is the basis for many functions of the toolbox and we tried to give an inverse functional interpretation of this covariance matrix we decided for this name. Many classical features such as principal components analysis (PCA), independent components analysis (ICA), analysis of variance (ANOVA) and correlation analysis as well as novel features such as Granger causality and permutation entropy (PE) are implemented in COVAIN allowing the user to compare different statistical methods and results for the interpretation of the original data. Moreover a new algorithm is implemented based on the assumptions of stochastic metabolic networks as discussed above. This algorithm allows for the inverse calculation of the differential biochemical Jacobian of two different states from metabolomics covariance data. Though this is mathematically a complicated inverse problem it is possible to identify putative perturbation sites in a biochemical network based on metabolomics data and stochastic modeling. Opportunities and problems of this approach are discussed in this paper. In the following sections the toolbox is described in detail.

## 2 Methods

### 2.1 GUI strategy and data structure

COVAIN synchronizes the actions of over 60 GUI components (such as pushbutton, table and list, etc.) by associating the status of each action and the resulted data uniquely with the main panel GUI handle. The GUI has two fields: (1) the status of all actions, i.e., if they have

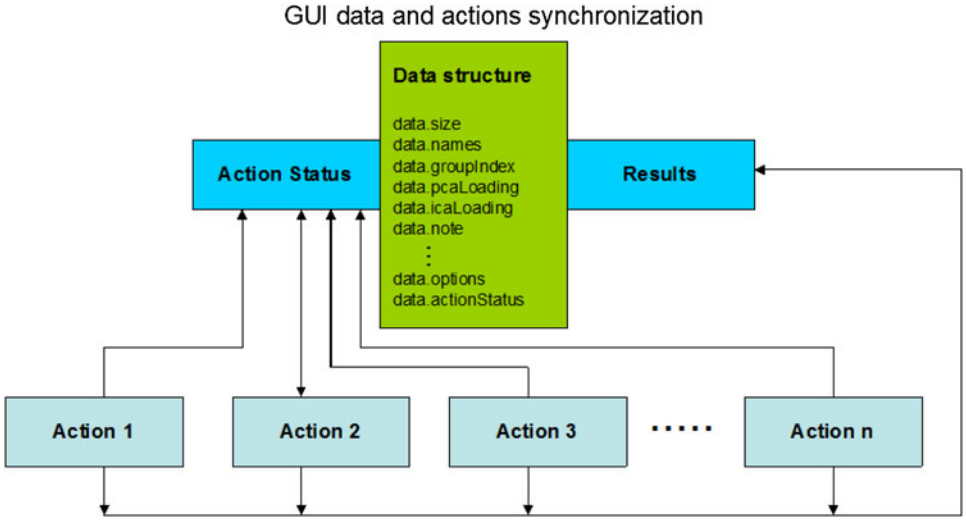



**Fig. 1** Illustration of COVAIN GUI strategy and data organization principle. Before activating a GUI action, this action needs to send a signal to the (unique) GUI root handle to check if it can be executed. For example, no analysis will be done without loading data, and no outlier adjustment will be performed without filling missing values. After executing an action, a status update signal will be sent to the GUI handle, and if this action can be executed, the result will be stored in the GUI handle associated data structure in the corresponding field

been activated or not, (2) the results storage. The original and preprocessed data, all intermediate results are saved in a structure format which consists of results of all data analysis processes. Figure 1 shows the GUI strategy and a simplified data structure.

## 2.2 Data preprocessing

### 2.2.1 Missing value imputation

The default option is using the half of the minimal value of all data to fill the missing values that are not detected by instruments. It is also optional to use prior distribution to estimate the missing values. The strategy is based on the assumption that measurements are normal-distributed.

Consider an $n$-by-$l$ data matrix $X_{nl}$, where $n$ is the number of variables in rows and $l$ is the total number of samples in columns. $p$ is the number of replicates per one condition or treatment resulting in a submatrix $X_{np}$. $x_{min}$ is the minimal value of all non-missing values in $X_{nl}$, $S_j$ is the standard deviation value of all non-missing values in the $j$-th column ($j = 1, 2,\ldots,p$). Then suppose there are $q$ missing values in the $i$-th row ($i = 1, 2,\ldots,n$) of the submatrix $X_{np}$. According to the number of non-missing values $p - q$, the missing value imputation is done as follows:

- if $p - q > 1$, the missing values are sampled from the estimated normal distribution $N(\mu_i, S_i)$, where the mean value $\mu_i$ and standard deviation value $S_i$ is estimated from the rest $p - q$ replicates.
- if $p - q \leq 1$, the missing value is replaced by half of the minimal value, $x_{min}/2$.

### 2.2.2 Outlier adjustment

The outliers are defined as measurements outside of two standard deviations of mean values for each condition of each compound. The outlier adjustment firstly proposes a prior distribution of the rest of the data and then randomly samples values from this distribution to fill outliers.

### 2.2.3 Data transformation

COVAIN provides two data transformations tailored for subsequent data analysis, log-transformation ($\log_2$ and $\log_{10}$) and $z$ score transformation. Log-transformation improves normal distribution of the data. $z$ score transformation is one of the most important options to allow for intra-comparison of values basically transforming covariance analysis into correlation analysis. However, the user has to be aware that these transformations will affect the outcome of the statistical analysis dramatically. $z$ score transformation will allow the comparison of compounds which have completely different scales or concentration ranges, i.e. the influence of a very low concentrated on a very high concentrated compound and vice versa.

## 2.3 Multivariate statistics

### 2.3.1 Correlation analysis

The correlation analysis provides two methods: Pearson's moment correlation (the default) and Spearman's rank correlation. Mean values for each condition are used for calculation. For small datasets (the number of variables is less than one hundred), the correlation coefficients across all conditions are shown in a heat map and saved; for medium datasets (the number of variables is hundred to several hundred), no graphs are shown but the correlation coefficients data are saved; for large datasets, the correlation coefficients are calculated by an external C program via a MEX file to achieve efficient calculation time and stored in a separate txt file to prevent a too large size of the GUI (which will affect the software efficiency).

The correlation analysis also provides results for network analysis (for further details of metabolite correlation analysis see Weckwerth and Fiehn 2002; Steuer et al. 2003a, b; Weckwerth 2003; Weckwerth et al. 2004a, b; Morgenthal et al. 2005, 2006). Two compounds are considered connected if their pair-wise absolute correlation coefficient is larger than a threshold value. COVAIN predefines three threshold values as 0.5, 0.8 and 0.95, $p$ value 0.01. It collects the number of connections of a compound to other compounds and saves this information.

### 2.3.2 PCA and ICA

The PCA function uses the MATLAB© Statistics Toolbox function *princomp* while ICA uses the method published in Scholz et al. (2004). The user needs to define how many principal components (PCs) and independent components (ICs) to be shown in the scatter plots of pair-wise PCs and ICs. The PCA generates four figures: the first figure shows the variance occupancy (in percentage) of each PC, the second shows the loadings plot of pair-wise PCs, the third and the fourth show the scores plot of all samples in a 2-D (PC1 and PC2 as $x$- and $y$-axis) and 3-D (PC1, PC2 and PC3 as $x$-, $y$- and $z$-axis) graph, respectively. The ICA also provides four figures: the first figure shows kurtosis of every IC, the second to the fourth figures show scores and loadings. COVAIN uses MATLAB© Statistics Toolbox *gname* function to label the names of compounds and conditions. COVAIN provides two separate buttons for labeling PCA and ICA, respectively.

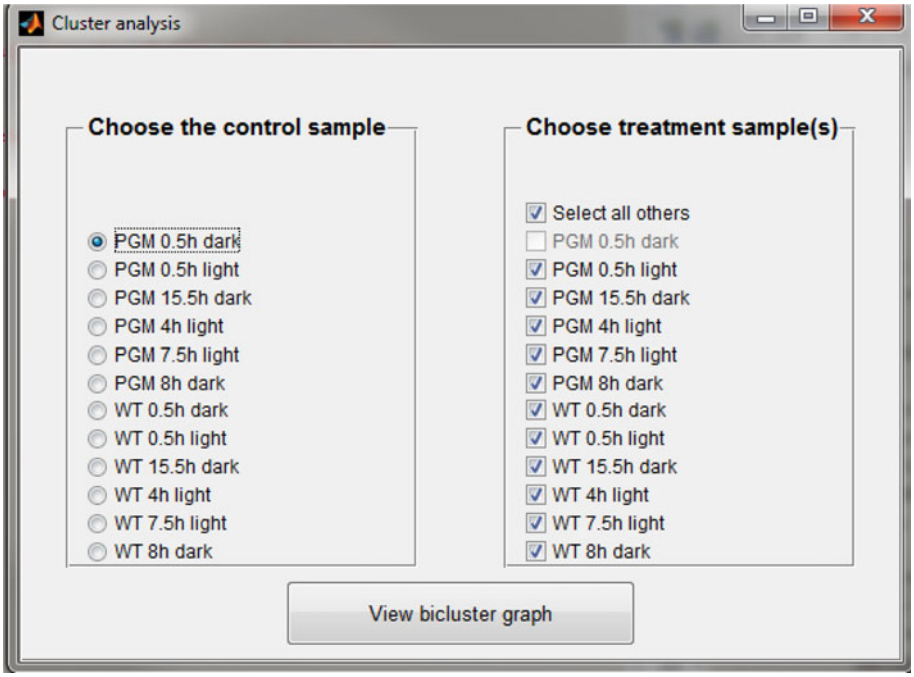



**Fig. 2** An example of the case–control clustering analysis. In the pop-up window, users need to choose one control condition in the *left side* of panel, and the default cases are all other conditions. It is optional to choose several other case conditions

#### 2.3.3 Clustering

The clustering analysis is doing bi-hierarchical clustering for both compounds and conditions. The case–control analysis pops up a window to let the user define which condition is the control and cluster the difference upon the control condition, while the "all conditions" analysis biclusters all conditions on their measurements. Figure 2 shows an example of the case–control analysis.

Basic parameters and descriptions of the procedures are found in the options and help functions.

### 2.4 Time-series analysis

#### 2.4.1 Data selection

Time-series analysis requires the user to define the indices of the conditions in a specific format. This is important, because user's data may contain time points whose names are differentially labeled, for instance, "20 min", "0.5 h", "1 day". The COVAIN will read the sequence of time points by sorting their names as "0.5 h", "1 day" and "20 min", which is incorrect. Therefore the user has to define the time-series data in a logical way, i.e. in minutes: 20, 30 and 1,440 min.

#### 2.4.2 PE analysis

PE is originated from dynamic systems theory (Bandt and Pompe 2002). It calculates the entropy of ordered patterns of time-series, shown in Eq. 1.

$$PE_m = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

where $p_i$ is the probability of occurrence of order pattern $i$ in three subsequent time points measured for metabolite $m$ under one condition and computed as the relative frequency of pattern $i$ in the total of $n - 2$ consecutive three-point order patterns.

High entropy indicates many response actions in the time-series, thereby being more complex or more unstable. Recently, we applied PE to gene expression data from AtGenExpress (Kilian et al. 2007) and found house-keeping genes have lower PE and abiotic response genes have higher PE (Sun et al. 2010). Therefore, PE can be used to predict the functions of unknown genes. Here, the PE can be applied for the first time in metabolomics data analysis.

#### 2.4.3 Granger causation analysis

Consider the time-series of variable $X$ and $Y$. They can be formulized as Eq. 2

$$\begin{aligned} X(t) &= \sum_{i=1}^{d} C_{X,i}\, X(t-i) + \sum_{i=1}^{d} C_{XY,i}\, Y(t-i) + R_X(t) \\ Y(t) &= \sum_{i=1}^{d} C_{YX,i}\, X(t-i) + \sum_{i=1}^{d} C_{Y,i}\, Y(t-i) + R_Y(t) \end{aligned} \tag{2}$$

with $C_{x,i}$ regression coefficient between $x(t)$ with $x(t-i)$ and $C_{xy,i}$ regression coefficient between $x(t)$ with $y(t-i)$. $X(t)$ and $Y(t)$ are the conditions at time point $t$, $R$ is the

residual error and $d$ is the maximal time lag. In COVAIN, an association between $X$ and $Y$ is considered existing if the $p$ value of the $F$ test on the cross-coefficients is less than 0.01. For short time-series of metabolomics data, $d$ is set to 1. However, this feature can be adjusted according to the number of time points measured.

The Granger causality analysis (1969) aims to analyze if the time-series of one variable is "controlled" by time-lagged values of other variables. The theory was recently applied in a metabolism study (Walther et al. 2010). The lag of time point is set to 1 for time-series with only a few data points (<10), however, can be chosen flexible. COVAIN identifies "causality" and a correlation between two compounds if the $p$ value of the $F$ test on the regression coefficients is smaller than 0.01. The Granger causality analysis results are saved.

Basic parameters and descriptions of the procedures are found in the options and help functions.

### 2.5 Inverse calculation of the Jacobian

For a dynamical system, the Jacobian matrix characterizes the local dynamics around the steady state. The dynamic representation of a metabolic pathway consisting of $n$ metabolites can be a set of differential equations (Eq. 3), where $(f_1, f_2,\ldots,f_n)$ are the functions of metabolite concentrations $(S_1, S_2,\ldots,S_n)$ over time. The corresponding Jacobian is the matrix of all first-order partial derivatives of all functions $f_i$ on all metabolites $S_j$, shown in Eq. 4. In this way, the Jacobian describes the influence on the change of each metabolite upon the changes of other metabolites.

$$\begin{cases} \frac{dS_1}{dt} = f_1(S_1, S_2, \ldots, S_n) \\ \frac{dS_2}{dt} = f_2(S_1, S_2, \ldots, S_n) \\ \qquad\vdots \\ \frac{dS_n}{dt} = f_n(S_1, S_2, \ldots, S_n) \end{cases} \tag{3}$$

$$Jacobian = \begin{pmatrix} \frac{\partial f_1}{\partial S_1} & \frac{\partial f_1}{\partial S_2} & \cdots & \frac{\partial f_1}{\partial S_n} \\ \frac{\partial f_2}{\partial S_1} & \frac{\partial f_2}{\partial S_2} & \cdots & \frac{\partial f_2}{\partial S_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial S_1} & \frac{\partial f_n}{\partial S_2} & \cdots & \frac{\partial f_n}{\partial S_n} \end{pmatrix}_{n\times n} \tag{4}$$

Recently, we developed an approach that linked the Jacobian of a metabolic system with the covariance of the involved metabolite concentration data represented by Eq. 5 (Steuer et al. 2003a, b; Weckwerth 2003, 2011):

$$CJ^T + JC = -2D. \tag{5}$$

Here, $J$ is the Jacobian matrix. $D$ is the fluctuation matrix in which the diagonal entries characterize the fluctuation magnitude of each metabolite. $C$ is the covariance matrix of metabolites.

By connecting the covariance matrix and the Jacobian matrix with the fluctuation matrix, Eq. 5 links the statistical features of the data with dynamical properties of the system, while taking into account the noise in the data captured by the fluctuation matrix. The generic type of Eq. 5 is widely used in control systems and is known as the "Lyapunov Equation" (Paulsson 2005). Later, van Kampen (1992) expanded it for general stochastic systems.

Despite this theoretical basis, the solution of the Jacobian cannot be obtained directly, because $J$ has more independent variables than the symmetric covariance matrix, or in other words, the equations are underdetermined. For example, an $n$-by-$n$ covariance matrix has $n * (n + 1)/2$ independent values due to its symmetry, while there are $n^2$ values in the Jacobian that need to be determined. The authors (Steuer et al. 2003a, b) suggested using parameterized solutions to eliminate such under-determination. However, as the parameter space for uncertain parameters are large, the actual Jacobian may not be easily obtained by such parameterization.

The elements of the Jacobian matrix, the square matrix with elements pertaining to the partial derivative of the rate of change of every metabolite with regard to all metabolites in the system (Eq. 4), can be expanded into separate terms according to reversible and irreversible reactions (Eq. 6),

$$J = N_d \frac{\partial f}{\partial S} \tag{6}$$

where $N_d$ is the directed stoichiometric matrix containing reversible and irreversible enzymatic reactions near the steady state. The information for the reversible and irreversible reactions between $S_i$ and $S_j$ can be obtained by genome-scale network reconstruction based on publicly accessible database such as KEGG (Kanehisa et al. 2008) and BioCyc (Karp et al. 2005).

As introduced above, the covariance matrix is symmetric and the Eq. 5 is therefore under-determined. We then can circumvent this problem by introducing the stoichiometric matrix ($N$) of a metabolic network, which is typically very sparse (Weckwerth 2011). If we integrate the reversibility and irreversibility of the reactions with $N$, we have a "directed stoichiometric matrix". We therefore label these directional information in the traditional $N$ and name it as $N_d$ (Eq. 6) and then determine non-zero entries in the Jacobian ($J$). Sometimes, there exists regulation between metabolites without substances consumption, which is reflected in $J$ but not in the $N$. For such cases, we need additional knowledge from literature and databases to assign these non-zero entries in $J$.

$N$ is very sparse. A typical metabolic network reconstructed recently for *Arabidopsis thaliana* has 1,567 reactions among 1,748 metabolites (Dal'Molin et al. 2010).

Furthermore, we searched metabolic network models in the BioModels database (Le Novere et al. 2006) and found no exceptions to the rule that the number of non-zero entries in $N$ exceeds the independent entries in $C$.

Another case is if the number of zero entries in $J$ exceeds the $n * (n + 1)/2$ independent entries in $C$. It is the most common case since metabolic networks are sparse. In this situation, Eq. 5 is over-determined. Then, an approximate solution ($J_{reverse}$) can be obtained by using regularization methods (Engl et al. 1996, 2009). To solve the inverse Jacobian from the covariance matrix we apply Eq. 5 and a total least square algorithm (Markovsky and Van Huffel 2007). The diagonal entries of the fluctuation matrix $D$ are randomly sampled, the amplitude can be chosen in the options.

## 3 Results and discussion

### 3.1 General features of COVAIN

COVAIN starts with uploading a dataset in Microsoft Excel or tab-separated text-like formats. In the data preprocessing step it recognizes the missing values and outliers in the dataset. It is optional to fill missing values or adjust outliers (see below). Data can be transformed in log or $z$ score scaling. The preprocessed data can be visualized by error bar plot and ANOVA.

COVAIN consists of four modules. (1) Multivariate statistics, including correlation analysis, PCA, ICA and clustering analysis, (2) time-series analysis, including correlation analysis, PCA, clustering, PE and Granger causation analysis, (3) network analysis, including topological analysis, KEGG pathway mapping and network visualization, (4) inverse Jacobian calculation, that compute the Jacobian matrix and plot the differential Jacobian corresponding to two different metabolic conditions associated with the data. The network analysis and inverse Jacobian analysis use the results from multivariate statistics and time-series analysis as input.

Finally, COVAIN collects all the results, saves them to a MATLAB worksapce (.mat format) and exports the preprocessed data, the correlation coefficients and the loadings of PCA and/or ICA in Excel or tab-separated text format. Figure 3 shows the software structure and strategy.

### 3.2 Network topological analysis, KEGG mapping and visualization

The network topology is inferred from correlation analysis or by Granger causality analysis as described in the Sect. 2. For pathway mapping and corresponding network analysis

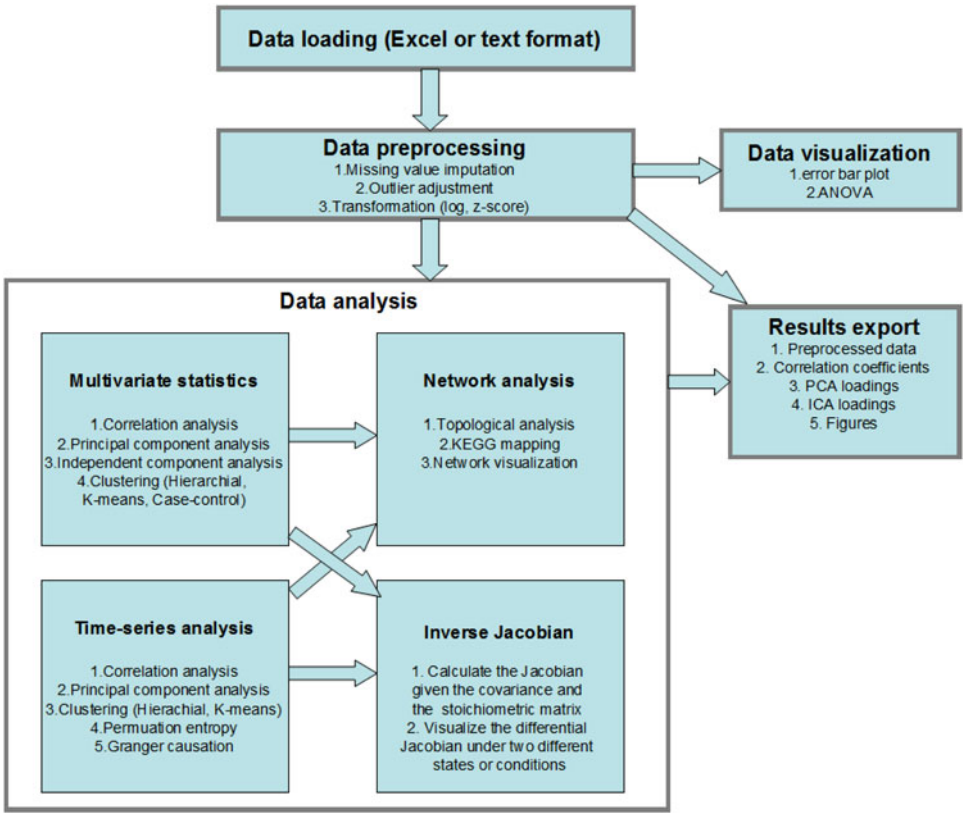



**Fig. 3** Illustration of COVAIN workflow: data loading, preprocessing to analysis, visualization and results exporting

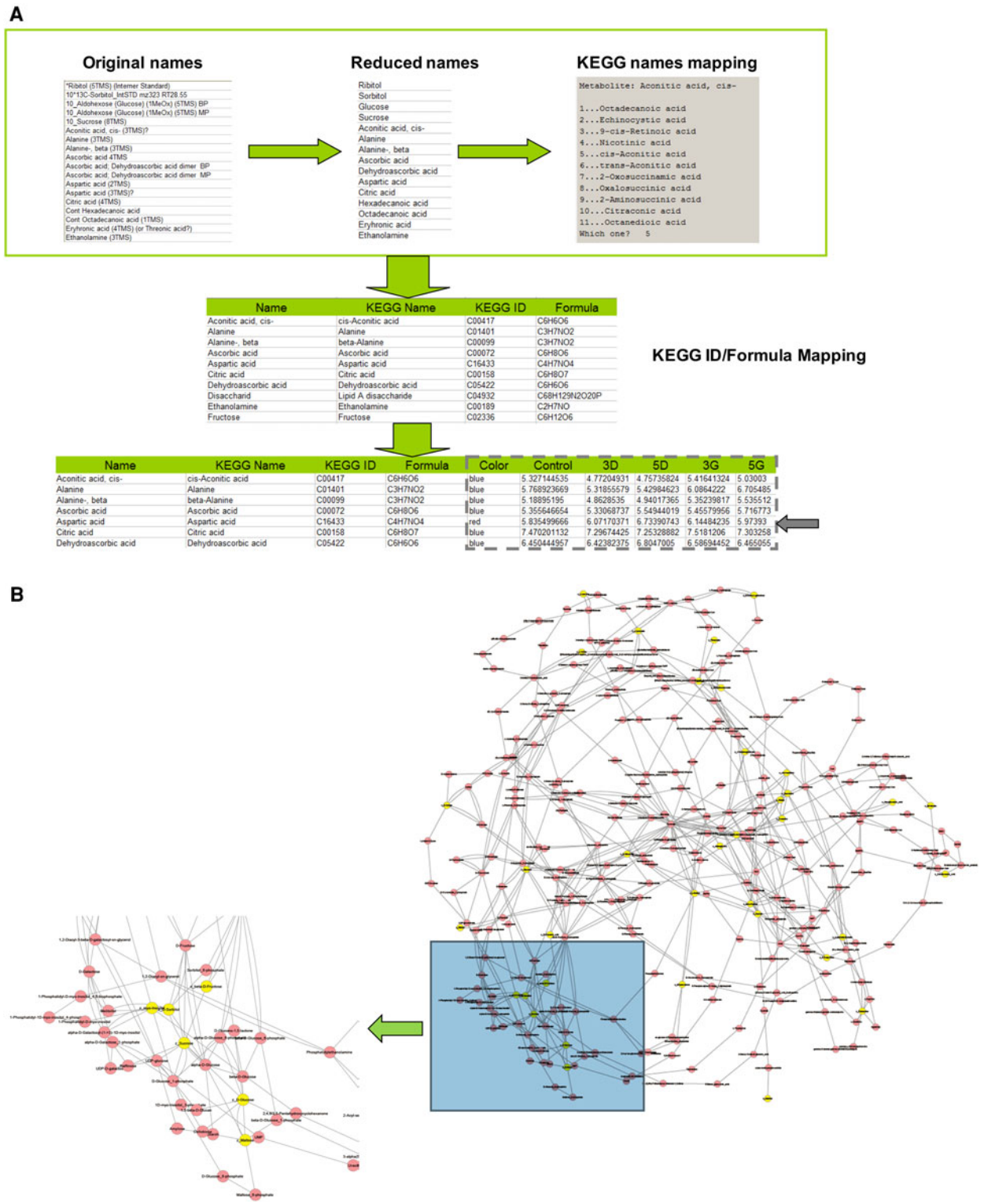
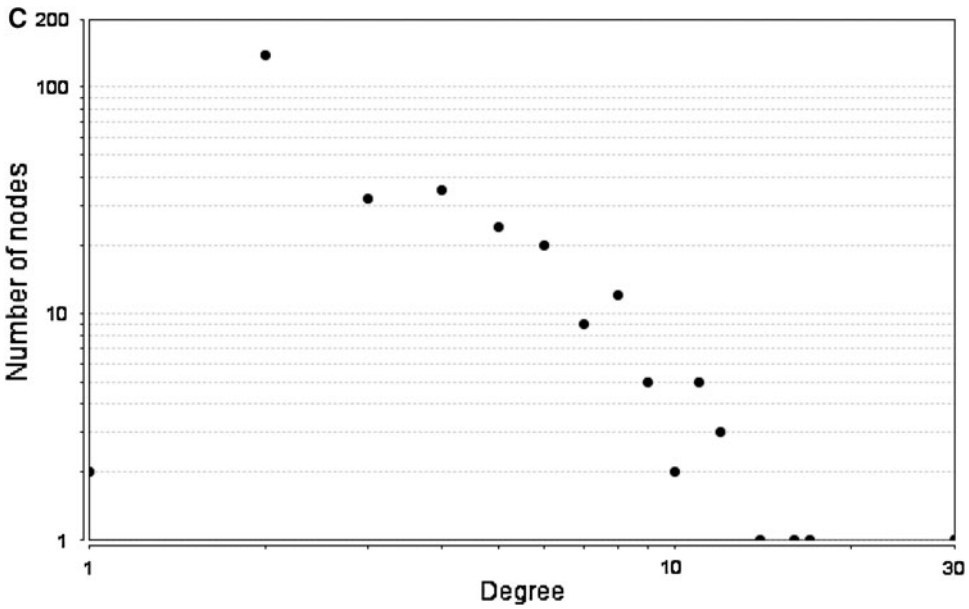



**Fig. 4** Pathway mapping and network topology analysis. **a** Metabolomics data from *Arabidopsis thaliana* cell cultures were mapped to KEGG-pathways (Lena Fragner, Yanmei Chen, Xiaoliang Sun and Wolfram Weckwerth, unpublished data). **b** In a second step a minimal graph network was defined from the pathways. This network was visualized with Cytoscape (Shannon et al. 2003). The *yellow nodes* represent the experimentally detected metabolites. The *red nodes* represent a minimal set of metabolites complementing the set of experimentally detected metabolites. **c** The connectivity structure can be analysed by Cytoscape and reveals behavior similar to a power-law (Color figure online)

**Fig. 4** continued




we have implemented several algorithms (see Fig. 4a). First, the metabolite names stemming from a typical metabolomics analysis are mapped against KEGG names and formulas to identify all involved pathways. In a second step a minimal interconnected graph model, i.e., each node can be reached from every other node, is defined. The resulting pathway network corresponding to the experimentally detected metabolites and a minimal set of interconnecting metabolites is exported and visualized with Cytoscape (Fig. 4b).

COVAIN uses MATLAB© Bioinformatics Toolbox function *biograph* to visualize the network with the "equilibrium" layout. At the same time, COVAIN produces one.net format file and one.sif format file to allow visualization of the network using Pajek (Batagelj and Mrvar 2004) or Cytoscape, respectively (Shannon et al. 2003). In particular, Cytoscape version 2.8 has integrated a "NetworkAnalyzer" plugin that can analyze network properties, such as degree distribution, betweenness and shortest path distribution and other features (Smoot et al. 2011).

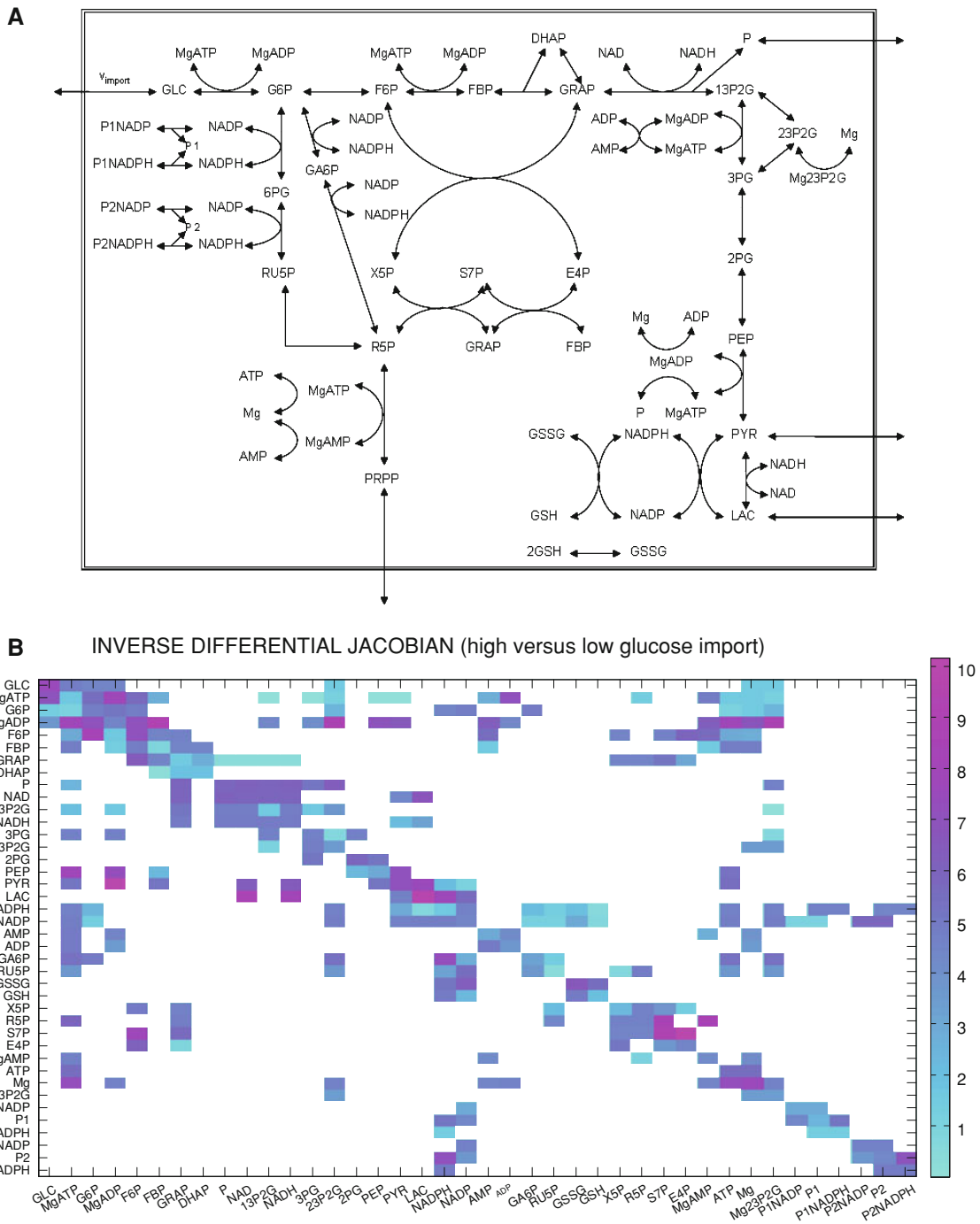
### 3.3 Inverse calculation of a differential Jacobian

Typically, a metabolomics experiment produces a large data matrix with many samples (conditions) and many variables (metabolites). A complete workflow is described elsewhere (Weckwerth and Morgenthal 2005; Weckwerth 2011). Recently, we have derived an equation which connects the covariance matrix $C$ of such an experiment with the Jacobian $J$ of the underlying network of biochemical reactions and regulations (Steuer et al. 2003a, b). The equation can be solved if the stoichiometric matrix $N$ of a metabolic network is exploited (Weckwerth 2011). This equation can be used for inverse calculation of the Jacobian from metabolomics covariance data. The method is described in detail in the Sect. 2. To reveal perturbation sites between two different metabolic states we introduce now the differential Jacobian. Suppose multiple repeat measurements under two different treatments $a$ and $b$ are available. For both treatment conditions, the associated Jacobian $J_a$ and $J_b$ can be calculated separately. The differential Jacobian matrix, $\mathrm{d}J_{ij}$, is defined by the relative change between the Jacobian $A$ and $B$ for every element $i, j$, as defined by Eq. 7. Note that we use the $\log_2$ ratio to center on zero and render the ratios symmetric around zero.

$$\mathrm{d}J_{ij} = \log_2\left(abs\left(\frac{J_{a,ij}}{J_{b,ij}}\right)\right) \tag{7}$$

The matrix of all elements $\mathrm{d}J_{ij}$ is a square matrix as are the original Jacobian matrices for conditions $a$ and $b$.

We have used a published model of the red blood cell from BioModels database (Le Novere et al. 2006) to illustrate the concept of the inverse differential Jacobian (Fig. 5a). Based on this model metabolite data are simulated which are comparable with a typical experimental data set. Two different states (high and low import of glucose, reaction $v_{import}$) and the corresponding covariance matrices are calculated. For the inverse approach we have used the simulated covariance data for each state (high and low glucose import) and Eq. 5 (see Method description 2) to calculate the Jacobian from the data. This corresponds to




the typical situation of the experimentalist: in the experimental design for the investigation of a biological system—different genotypes/phenotypes, growth conditions or treatments—a preliminary knowledge or estimate of the structure of the underlying biochemical network is assumed. This preknowledge can be translated into a stoichiometric matrix $N$ (for further details see Weckwerth 2011). What is missing is the correct localization of one or more perturbation sites in the biochemical network. Using this approach it is possible to systematically search for differential regulations of several reactions as the response to changed metabolite concentrations (Weckwerth 2011). This is exemplified for the perturbed red blood cell model in Fig. 5b. The differential Jacobian shows—as expected—strong perturbation sites at the first reactions where imported glucose is participating (upper left corner of the differential Jacobian in Fig. 5b). A drawback is that many other reactions seem to show responses as well, and the

◀ **Fig. 5** Inverse calculation of a differential Jacobian. **a** To illustrate the approach a metabolic model network was used corresponding to the metabolism of the red blood cell from the BioModels database (http://www.ebi.ac.uk/biomodels-main/BIOMD0000000070). From the model metabolite data were simulated with high and low glucose import. **b** The simulated metabolite data of the red blood cell model are used to calculate the inverse differential Jacobian for two different states, high and low glucose import, respectively. Equation 5 was used to calculate the corresponding Jacobian_High and Jacobian_Low. The differential Jacobian of both conditions shows a high perturbation in the corresponding glucose import reaction $v_0$ (*left upper corner*), however, also perturbations in other reactions. The reproducibility of the calculation is low due to a stiff problem (Jia et al. 2011). The approach needs further refinement in the future, however, demonstrates the principal feasibility to calculate differential Jacobians from experimental data to systematically investigate genome-scale metabolic networks—defined by the stoichiometric matrix *N* and putative perturbation sites in the corresponding biochemical regulation network (for further details see Weckwerth 2011 and a step-by-step tutorial in supplementary materials). *GLC* glucose in, *MgATP* MgATP, *G6P* glucose 6-phosphate, *MgADP* MgADP, *F6P* fructose 6-phosphate, *FBP* fructose 1,6-phosphate, *GRAP* glyceraldehyde 3-phosphate, *DHAP* dihydroxyacetone phosphate, *P* phosphate, *NAD* NAD, *13P2G* 1,3-bisphospho-D-glycerate, *NADH* NADH, *3PG* 3-phospho-D-glycerate, *23P2G* 2,3-bisphospho-D-glycerate, *2PG* 2-phospho-D-glycerate, *PEP* phosphoenolpyruvate, *PYR* pyruvate, *LAC* lactate, *NADPH* NADPH, *NADP* NADP, *AMP* AMP, *ADP* ADP, *GA6P* phospho-D-glucono-1,5-lactone, *RU5P* ribulose 5-phosphate, *GSSG* oxidized glutathione, *GSH* reduced glutathione, *X5P* xylulose 5-phosphate, *R5P* ribose 5-phosphate, *S7P* sedoheptulose 7-phosphate, *E4P* erythrose 4-phosphate, *MgAMP* MgAMP, *ATP* ATP, *Mg* Mg, *Mg23P2G* Mg 2,3-bisphospho-D-glycerate, *P1NADP* protein1 bound NADP, *P1* protein1, *P1NADPH* protein1 bound NADPH, *P2NADP* protein2 bound NADP, *P2* protein2, *P2NADPH* protein2 bound NADPH

reproducibility of the calculation needs to be tested. The inverse calculation is highly sensitive to noisy data and will result in low reproducibility of the Jacobian calculations. This is a typical inverse problem (Engl et al. 1996, 2009) and will be addressed by us further for the optimization of the inverse calculation of the Jacobian from noisy and incomplete data.

## 4 Conclusion

Here, a Metabolomics toolbox is presented which combines classical features and novel tools for data mining in metabolomics data sets. A focus is covariance and related correlation network analysis, as well as projection of data into pathways and the relation of metabolite dynamics in form of the covariance matrix *C* to the corresponding pathways in form of the stoichiometric matrix $N_d$ and the Jacobian *J*. Initial algorithms provide the convenient and direct linkage of metabolite data and the underlying biochemical network. This linkage enables the investigation of regulatory and biochemical perturbation sites in a complex metabolic network. Future work will extend these methods to genome-scale applications. Drawbacks of the methods are the incompleteness of Metabolomics data sets. Major efforts in the future will include the improvement of the metabolomics techniques to cover more metabolites and the adjustment of measured metabolites with the entries of the stoichiometric matrix (Weckwerth 2011). Generic properties of the covariance–Jacobian relation will be investigated in the future. Furthermore, the influence of stochastic fluctuation on regulatory properties within metabolic networks can be investigated systematically using the metabolomics toolbox COVAIN.

**Acknowledgments** The authors thank especially Dirk Walther, Lena Fragner and Stefanie Wienkoop for their helpful suggestions.

## References

Aprees, T. (1980). Integration of pathways of synthesis and degradation of hexose phosphates. In J. Preiss (Ed.), *The biochemistry of plants* (Vol. 3, pp. 1–29). New York: Academic Press.

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

Arkin, A., Ross, J., & McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics, 149*, 1633–1648.

Bandt, C., & Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Physical Review Letters, 88*, 174102.

Bassham, J. A., Benson, A. A., & Calvin, M. (1950). The path of carbon in photosynthesis. *Journal of Biological Chemistry, 185*, 781–787.

Batagelj, V., & Mrvar, A. (2004). Pajek—analysis and visualization of large networks. *Graph Drawing Software, 378*, 77–103.

Broeckling, C. D., Huhman, D. V., Farag, M. A., et al. (2005). Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *Journal of Experimental Botany, 56*, 323–336.

Camacho, D., Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics, 1*, 53–63.

Clish, C. B., Davidov, E., Oresic, M., et al. (2004). Integrative biological analysis of the APOE*3-Leiden transgenic mouse. *Omics: A Journal of Integrative Biology, 8*, 3–13.

Cornishbowden, A., & Hofmeyr, J. H. S. (1994). Determination of control coefficients in intact metabolic systems. *Biochemical Journal, 298*, 367–375.

Dal'Molin, C. G. D., Quek, L. E., Palfreyman, R. W., et al. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiology, 152*, 579–589.

Engl, H. W., Hanke, M., & Neubauer, A. (Eds.). (1996). *Regularization of inverse problems* (Vol. 375). Dordrecht: Kluwer.

Engl, H. W., Flamm C., Kugler P., et al. (2009). Inverse problems in systems biology. *Inverse Problems, 25*. doi:10.1088/0266-5611/1025/1012/123014.

Fukushima, A., Kusano, M., Redestig, H., et al. (2011). Metabolomic correlation-network modules in *Arabidopsis* based on a graph-clustering approach. *BMC Systems Biology, 5*, 1.

Giersch, C. (1994). Determining elasticities from multiple measurements of steady-state flux rates and metabolite concentrations—theory. *Journal of Theoretical Biology, 169*, 89–99.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 414–426.

Heinrich, R., & Rapoport, T. A. (1974). Linear steady-state treatment of enzymatic chains—general properties, control and effector strength. *European Journal of Biochemistry, 42*, 89–95.

Hendrickx, D. M., Hendriks, M., Eilers, P. H. C., et al. (2011). Reverse engineering of metabolic networks, a critical assessment. *Molecular Biosystems, 7*, 511–520.

Jansen, J. J., Szymanska, E., Hoefsloot, H. C. J., et al. (2011). Between metabolite relationships: An essential aspect of metabolic change. *Metabolomics*. doi:10.1007/s11306-011-0316-1.

Jia, G., Stephanopoulos, G. N., & Gunawan, R. (2011). Parameter estimation of kinetic models from metabolic profiles: Two-phase dynamic decoupling method. *Bioinformatics, 27*, 1964–1970.

Kacser, H., & Burns, J. A. (1973). The control of flux. *Symposia of the Society for Experimental Biology, 27*, 65–104.

Kanehisa, M., Araki, M., Goto, S., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research, 36*, D480–D484.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., et al. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research, 33*, 6083–6089.

Kilian, J., Whitehead, D., Horak, J., et al. (2007). The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant Journal, 50*, 347–363.

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

Kusano, M., Fukushima, A., Arita, M., et al. (2007). Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in *Arabidopsis thaliana*. *BMC Systems Biology, 1*, 17.

Le Novere, N., Bornstein, B., Broicher, A., et al. (2006). BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research, 34*, D689–D691.

Markovsky, I., & Van Huffel, S. (2007). Overview of total least squares methods. *Signal Processing, 87*, 2283–2302.

Mendes, P., Camacho, D., & de la Fuente, A. (2005). Modelling and simulation for metabolomics data analysis. *Biochemical Society Transactions, 33*, 1427–1429.

Meyerhof, O. (1927). Recent investigations on the aerobic and anaerobic metabolism of carbohydrates. *Journal of General Physiology, 8*, 531–542.

Meyerhof, O. (1947). The rates of glycolysis of glucose and fructose in extracts of brain. *Archives of Biochemistry, 13*, 485–487.

Mochida, K., Furuta, T., Ebana, K., et al. (2009). Correlation exploration of metabolic and genomic diversity in rice. *BMC Genomics, 10*, 568.

Morgenthal, K., Wienkoop, S., Scholz, M., et al. (2005). Correlative GC-TOF-MS based metabolite profiling and LC-MS based protein profiling reveal time-related systemic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics, 1*, 109–121.

Morgenthal, K., Weckwerth, W., & Steuer, R. (2006). Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems, 83*, 108–117.

Muller-Linow, M., Weckwerth, W., & Hutt, M. T. (2007). Consistency analysis of metabolic correlation networks. *BMC Systems Biology, 1*, 44–56.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews, 2*, 157–175.

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

Rascher, U., Hutt, M. T., Siebke, K., et al. (2001). Spatiotemporal variation of metabolism in a plant circadian rhythm: The biological clock as an assembly of coupled individual oscillators. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 11801–11805.

Samoilov, M., Arkin, A., & Ross, J. (2001). On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos, 11*, 108–114.

Scholz, M., Gatzek, S., Sterling, A., et al. (2004). Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics, 20*, 2447–2454.

Shannon, P., Markiel, A., Ozier, O., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*, 2498–2504.

Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., et al. (2010). Dynamic metabolomic data analysis: A tutorial review. *Metabolomics, 6*, 3–17.

Smoot, M. E., Ono, K., Ruscheinski, J., et al. (2011). Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics, 27*, 431–432.

Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003a). Interpreting correlations in metabolomic networks. *Biochemical Society Transactions, 31*, 1476–1478.

Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003b). Observing and interpreting correlations in metabolomic networks. *Bioinformatics, 19*, 1019–1026.

Steuer, R., Morgenthal, K., Weckwerth, W., & Selbig, J. (2006). A gentle guide to the analysis of metabolomic data. *Methods in Molecular Biology, 358*, 105–126.

Sun, X., Zou, Y., Nikiforova, V., et al. (2010). The complexity of gene expression dynamics revealed by permutation entropy. *BMC Bioinformatics, 11*, 607.

van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*. Amsterdam: Elsevier.

Vance, W., Arkin, A., & Ross, J. (2002). Determination of causal connectivities of species in reaction networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 5816–5821.

Walther, D., Strassburg, K., Durek, P., & Kopka, J. (2010). Metabolic pathway relationships revealed by an integrative analysis of the transcriptional and metabolic temperature stress–response dynamics in yeast. *Omics: A Journal of Integrative Biology, 14*, 261–274.

Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology, 54*, 669–689.

Weckwerth, W. (2011). Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Analytical and Bioanalytical Chemistry, 400*, 1967–1978.

Weckwerth, W., & Fiehn, O. (2002). Can we discover novel pathways using metabolomic analysis? *Current Opinion in Biotechnology, 13*, 156–160.

Weckwerth, W., & Morgenthal, K. (2005). Metabolomics: From pattern recognition to biological interpretation. *Drug Discovery Today, 10*, 1551–1558.

Weckwerth, W., & Steuer, R. (2005). Metabolomic networks: From experiment to biological interpretation. In S. Vaidyanathan, G. G. Harrigan, & R. Goodacre (Eds.), *Metabolomics*. New York: Springer.

Weckwerth, W., Tolstikov V., & Fiehn O. (2001). Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. *Proceedings of the 49th ASMS conference on mass spectrometry and allied topics* (pp. 1–2).

Weckwerth, W., Loureiro, M. E., Wenzel, K., & Fiehn, O. (2004a). Differential metabolic networks unravel the effects of silent

plant phenotypes. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7809–7814.

Weckwerth, W., Wenzel, K., & Fiehn, O. (2004b). Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics, 4*, 78–83.

Westerhuis, J. A., van Velzen, E. J., Hoefsloot, H. C., & Smilde, A. K. (2010). Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. *Metabolomics, 6*, 119–128.

Wienkoop, S., Morgenthal, K., Wolschin, F., et al. (2008). Integration of metabolomic and proteomic phenotypes: Analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana*. *Molecular and Cellular Proteomics, 7*, 1725–1736.

Wienkoop, S., Weiss, J., May, P., et al. (2010). Targeted proteomics for *Chlamydomonas reinhardtii* combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses. *Molecular Biosystems, 6*, 1018–1031.