BMC Bioinformatics

BioMed Central

Research article

Robust detection and verification of linear relationships to generate metabolic networks using estimates of technical errors

Frank Kose[1], Jan Budczies[2], Matthias Holschneider[3] and Oliver Fiehn*[4]

Address: [1]Universitaet Potsdam, D-14415 Potsdam, Germany, [2]Institute of Pathology, Charité University Hospital and provitro GmbH, D-10117 Berlin, Germany, [3]Universitaet Potsdam, Inst. f. Mathematik, D-14415 Potsdam, Germany and [4]University of California Davis, Genome Center, Davis CA 95616, USA

Email: Frank Kose - kose@likelynet.de; Jan Budczies - jb@provitro.de; Matthias Holschneider - matthias.holschneider@gmail.com; Oliver Fiehn* - ofiehn@ucdavis.edu

* Corresponding author

Published: 21 May 2007

BMC Bioinformatics 2007, **8**:162 doi:10.1186/1471-2105-8-162

Received: 25 July 2006
Accepted: 21 May 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/162

Abstract

Background: The size and magnitude of the metabolome, the ratio between individual metabolites and the response of metabolic networks is controlled by multiple cellular factors. A tight control over metabolite ratios will be reflected by a linear relationship of pairs of metabolite due to the flexibility of metabolic pathways. Hence, unbiased detection and validation of linear metabolic variance can be interpreted in terms of biological control. For robust analyses, criteria for rejecting or accepting linearities need to be developed despite technical measurement errors. The entirety of all pair wise linear metabolic relationships then yields insights into the network of cellular regulation.

Results: The Bayesian law was applied for detecting linearities that are validated by explaining the residues by the degree of technical measurement errors. Test statistics were developed and the algorithm was tested on simulated data using 3–150 samples and 0–100% technical error. Under the null hypothesis of the existence of a linear relationship, type I errors remained below 5% for data sets consisting of more than four samples, whereas the type II error rate quickly raised with increasing technical errors. Conversely, a filter was developed to balance the error rates in the opposite direction. A minimum of 20 biological replicates is recommended if technical errors remain below 20% relative standard deviation and if thresholds for false error rates are acceptable at less than 5%. The algorithm was proven to be robust against outliers, unlike Pearson's correlations.

Conclusion: The algorithm facilitates finding linear relationships in complex datasets, which is radically different from estimating linearity parameters from given linear relationships. Without filter, it provides high sensitivity and fair specificity. If the filter is activated, high specificity but only fair sensitivity is yielded. Total error rates are more favorable with deactivated filters, and hence, metabolomic networks should be generated without the filter. In addition, Bayesian likelihoods facilitate the detection of multiple linear dependencies between two variables. This property of the algorithm enables its use as a discovery tool and to generate novel hypotheses of the existence of otherwise hidden biological factors.

Background

In recent years, time course analyses of metabolic perturbations have become more important to understand metabolic networks based on experimental data [1,2]. One way to analyze metabolic networks is by systematically investigating linear relationships between all analyzed metabolites (variables) followed by constructing networks from positively identified components, and eventually comparing network topologies [3] between different physiological or genetic conditions [4,5]. Simulations of metabolic reactions have shown that even stochastic influences on metabolism may result in linear metabolic co-regulation because initial metabolic perturbations can be propagated and enhanced through the cellular biochemical network [6]. Such linear co-regulation of pairs of metabolites may point to changes in biochemical control (chemical equilibrium, mass conservation, asymmetric control distribution) as well to transcriptional regulation [7]. Variance in metabolite levels can be caused by three different factors:

(I) concentrations alter and hence increase variance due to intentionally changing the experimental conditions, for example by altering environmental parameters like external nutrients or by using different genotypes [8],

(II) metabolite data will found to vary in a stochastic manner caused by the imprecision of the analytical method [9] used for acquiring metabolite data and

(III) interestingly, even under very controlled environmental conditions, a high degree of biological variation is found for metabolite levels due to stochastic biological events that trickle through the biochemical network and thus reflect the underlying control structure at this particular biological condition [6].

Therefore, if enough biological replicates are analyzed for a given organism at a given physiological situation, the metabolic phenotype can be investigated not only by its corresponding average metabolic values, but also by a snapshot of its corresponding metabolic network. However, biologists often do not know the inherent biological variability in advance and hence tend to use just a few independent biological replicates based on preliminary power analysis. Resulting data may be sufficient to estimate arithmetic means of metabolic levels but do not enable analyzing the linear control structure between different biological conditions. One of the challenges for calculating linearity networks is to compute the likelihood or significance of the presence of a truly linear relationship, with the aim of excluding both false negative and false positive detections of linearities.

Estimating optimal linearity parameters has been solved decades ago for cases, for which linear dependence of variables could be reasoned based on background knowledge. However, in metabolic data sets, the control structure of metabolites is unknown *a priori*. Therefore, two fundamental questions need to be answered:

(a) For which pairs of variables can a linear relationship be hypothesized?

(b) Are there sub sets of data that reflect differences in linear behavior of variables? For example, linearity may be given for only a group of data but absent in another group, or the linearity parameters between these groups may be different.

An unbiased analysis of linear relationships between pairs of variables needs to test whether there is one or more valid linear hypotheses that could explain data in complex data sets. This procedure defines a novel approach for testing biological data: instead testing pre-defined hypotheses [10], the likelihood of hypotheses is calculated that may be used to explain complex process. This hypothesis-building is fundamentally different from estimating the best parameters of an assumed linear relationship by regression equations [11-13] for which various software packages exist, and solutions for estimating parameters for multiple linear relationships[14]. However, regressions do not test the probability of the presence of linear relationships, especially in high-dimensional data sets. Instead, regressions are founded on the presence of linearity that is justified by background knowledge. In biochemistry, the existence of linear relationships cannot generally be assumed trivial but must receive thorough statistical evaluations. In addition, regressions usually do not account for technical errors [15] that are critical in practice. All measurements comprise technical errors which are due to inadequacy of the total chemical-analytical method, specifically the extraction, sample preparation and the instrumental data acquisition. Hence, the degree of technical errors will vary between the chemical nature of the metabolites, their absolute concentrations and influences of different sample matrices. Furthermore, outliers and missing data further obscure detection of linear hypotheses. For regressions, on the other hand, the impact of outliers has been studied extensively, and multiple measures to assess and weigh the influence of outliers have been developed. Assumptions on the degree of technical errors may further refine weighting factors in regression analysis, and such factors can be optimized for example using the EM-algorithm. Nevertheless, regression is not an explorative tool for data analysis. Additionally, metabolomic data do not distinguish between dependent and independent variables. All variables are subject to varying degree of noise (analytical-chemical measurement

errors, i.e. technical errors). No controlled observations of supposedly independent variables can be acquired [16,17]. Consequently, the control structure of metabolic linearity networks can only be assessed with a tool that solves the following tasks:

(1) Linear relationships must be detected in an unbiased and observer-independent manner.

(2) Sub sets of data need to be grouped according to presence of (multiple) linear relationships.

(3) Criteria have to be applied that verify linear hypotheses based on test statistics.

(4) Technical errors: varying degree of analytical-chemical measurement errors and missing data have to be accounted for.

Especially, the potential presence of multiple linear relationships and independence of both variables poses problems for simple regression analyses. As a substitute for regression, the degree of correlation has been used for detecting linear relationships despite the fact that correlation only relates the covariance to the total variance, but does not verify genuine linearities. Moreover, Pearsons' correlation coefficients lack robustness against outliers, especially for multivariate datasets, and a number of different approaches have been suggested to link estimates to better test statistics [18]. In practice, however, empirical or heuristic thresholds are taken to distinguish strong or weak correlations, but no mathematical basis exists on which such thresholds can safely be founded. In some cases, Student's statistics *p*-values have been taken in an effort to validate Pearson's correlations [19]. Unfortunately, such *p*-values only describe the significance of the non-randomness of data pairs but do not test hypotheses if data pairs can be described by a (single or multiple) linear functions. Consequently, correlation networks based on Pearsons correlations may be strongly distorted [20].

A further approach has been taken using partial correlations that deconvolute contributions by additional parameters in order to reduce the list of correlations to basic dependencies [21] which may present a link from correlation to causality [22,23]. This method is valuable to investigate the control structure within a given correlation network but it does not remove the principle robustness problem of correlation estimates. Simple correlations coefficients always decline with increasing variance that is introduced by method errors during data acquisition. In contrary, partial correlation coefficients may be increasing, decreasing or even change the algebraic signs with increasing method errors [20]. In order to remedy this situation, scientists tend to select high Pearson's correlation thresholds [24] which imply that the variance caused by method errors is small in relation to the biological variance. The latter assumption is often true when comparing widely different metabolic phenotypes such as certain mutant genotypes, or severe stress conditions such as acute (metabolic) diseases in comparison to healthy states. However, metabolic theory predicts that even incremental changes in enzymatic properties can have large effects on metabolic control, especially when multiple enzymes are affected [25]. Such changes might be too subtle to cause large differences in average concentrations but would still effect the pathway control structure and hence, linearities in pairs of metabolite data. Consequently, the metabolic control structure can only be assessed with a robust tool for linearity detection.

We here present a different approach. Using the Bayesian law [26], a likelihood formula is derived that is based on information about the measurement error using a specific technical method. This formula is then transformed in way that allows searching for local maxima of linear parameters within the total hypothesis space. Such likelihood maxima are subsequently assessed for residuals of the corresponding linearity parameters using simulated test statistics. We demonstrate the power of this approach using a synthetic data set with a given set of true linear relationships which are subsequently subjected to both increasing technical errors and increasing number of samples.

Results and Discussion

(1) A model for the technical error in metabolomic data

Let $\{x_{ij}\}$ denote the entirety of n metabolite measurements in a collection of m samples. The measurements can be arranged in a matrix x_{ij} with rows i = 1, ..., n that refer to the metabolites and columns j = 1, ..., m that refer to the samples. Each measurement results from the sum of the true metabolite content x'_{ij} and a technical error,

$$x_{ij} = x'_{ij} + e_{ij}. \quad (1)$$

The technical errors e_{ij} include the chemical-analytical error, but can also include a contribution from different storage manners or times of the biomaterial after its extraction. The technical variance of the *j*th measurement can be derived by a probability density function ρ_j that reflects knowledge about sample storage and data acquisition. For missing data, it is only known that these can be expected in a defined range but with uniform probability distribution. For non-missing data, the technical error is modeled by a multivariate normal distribution that is centered around zero. More precisely, the probability density for the technical error $e_j = (e_{1\,j}, ..., e_{nj})^t$ of the *j*th metabolic profile is given by

$$\rho_j(e_j) = N \exp\left(-\frac{1}{2} e_j{}^t \sum_j{}^{-1} e_j \right). \quad (2)$$

In principle, the variance matrices Σ_j can be estimated from the covariance matrix of replicated measurements of the same biomaterial. In practice, correlations between the technical errors of different metabolites are often disregarded, leading to a model with diagonal variance matrices $\Sigma_j = \mathrm{diag}\left(\sigma_{1j}^2, ..., \sigma_{nj}^2\right)$. Further one often works with fixed absolute or fixed relative errors,

$$\sigma_{ij}^{(\mathrm{abs.})} = a_i \text{ or } \sigma_{ij}^{(\mathrm{rel.})} = r_i x_{ij}, \quad (3)$$

respectively.

(2) Maximum Likelihood (ML) function for a general linear problem

Using the Bayesian law, the likelihood for parameters of a given linear hypothesis can be calculated. The Bayesian law allows to interconvert the conditional probabilities of cause (linear relationship) and effect (measured data value) [27,28],

$$p(A \mid B) = \frac{p(B \mid A) p(A)}{p(B)} \quad (4)$$

The general form of a linear relationship in the metabolomics data is

$$\alpha_1 x'_{ij} + ... + \alpha_n x'_{nj} = \beta \quad \text{for all samples} \quad j = 1, ..., m. \quad (5)$$

In what follows we collect the coefficients of the above equation in a vector $\alpha = (\alpha_1, ..., \alpha_n)^t$ and express the linear relationship as $\alpha^t x_{\bullet j} = \beta$. By $x_{\bullet j}$ we denote the metabolic profile of the jth sample. The entirety of the parameters α and β results in A. The probability $p(A)$ is the *a priori* probability of the parameters α and β before the measurement has been performed. Because no preference can be given, the *a priori* probability is constant for all A. The same is true for $p(B)$ with B representing the measured metabolite concentrations. Therefore, $p(A \mid B)$ is the likelihood for parameter A if the pair of variables B is given. We have used an unbiased approach here assuming random and unrelated technical errors, and we cannot know beforehand if a certain metabolite will be detectable or not, and how large the concentration of such metabolite could be. These assumptions result in a constant probability for p(A) and p(B), because otherwise certain values for A and B would be more likely than others. From the Bayesian law it can be concluded

$$p(A \mid B) = c \cdot p(B \mid A). \quad (6)$$

The constant value c can be neglected because the objective is to compare different linear hypotheses. Consequently, a hypothetical metabolic profile has the same probability at a given linear hypothesis as a hypothetical linear hypothesis at a given metabolic profile. The expression $p(B|A)$ is therefore the probability that the measured metabolic profile B is determined at a given set of parameters A, which can be calculated using the function which describes the probability distribution of the technical error. We are now in position to state the following general theorem:

Let $x = (x_1, ..., x_n)^t$ include the measurements of n metabolites in a biological sample with technical errors that follow a Gaussian distribution with covariance matrix Σ. Let a (n-N)-dimensional surface in the n-dimensional metabolite space be defined by the equations

$$\alpha_{k1} x_1 + ... + \alpha_{kn} x_n = \beta_k \quad \text{for } k = 1, ..., N. \quad (7)$$

The coefficients of these equations comprise a matrix $\alpha = (\alpha_{ki})$ and a vector $\beta = (\beta_1, ..., \beta_N)^t$. The matrix elements can also be arranged in vectors $\alpha_k : (\alpha_{k1}, ..., \alpha_{kn})^t$. It is assumed the hyperplanes defined by (8) are orthogonal in pairs with respect to the covariance matrix, i.e.

$$\alpha_k{}^t \Sigma\, \alpha_l = 0 \quad \text{for } k, l = 1, ..., N \text{ and } k \neq l. \quad (8)$$

Then, the likelihood for the metabolite concentrations to lie on the (n-k)-dimensional surface is given by

$$p(x \mid \alpha, \beta) = \exp\left(-\frac{1}{2} \sum_{k=1}^{N} \frac{\left(\alpha_k{}^t x - \beta_k \right)^2}{\alpha_k{}^t \Sigma \alpha_k} \right). \quad (9)$$

The theorem is proven in Additional File 1. However, the result for the likelihood has a simple interpretation: it is proportional to the density of the normal distribution taken at the distance of the measurement from the surface. This distance has to be calculated by taking the covariance matrix of the technical errors as metric of the metabolite space. Next, let us illustrate the theorem by a special case that is especially interesting for applications: Consider two metabolite concentrations that are measured with technical standard deviations σ_1, σ_2 and technical covariance σ_{12}. Then, the likelihood for the metabolite concentration to lie on the straight line $\alpha_1 x_1 + \alpha_2 x_2 = \beta$ is given by

$$p(x_1, x_2 \mid \alpha_1, \alpha_2, \beta) = \exp\left(-\frac{1}{2} \frac{(\alpha_1 x_1 + \alpha_2 x_2 - \beta)^2}{\sigma_1^2 \alpha_1^2 + 2\sigma_{12}\alpha_1\alpha_2 + \sigma_2^2 \alpha_2^2} \right). \quad (10)$$

Returning to the general line of the text and the Bayesian reasoning we obtain for the likelihood for a linear relationship described by $A = \{\alpha_{ki}, \beta_k\}$ after measurement of the metabolomic data $B = \{x_{ij}\}$ the result

$$p(A \mid B) = \prod_{j=1}^{m} l_j(\alpha, \beta \mid x_{\bullet j}) \quad (13)$$

with

$$l_j(\alpha, \beta \mid x_{\bullet j}) = p(x_{\bullet j} \mid \alpha, \beta) = \exp\left(-\frac{1}{2} \sum_{k=1}^{N} \frac{\left(\alpha_k{}^t x_{\bullet j} - \beta_k\right)^2}{\alpha_k{}^t \Sigma_j \alpha_k} \right) \quad (14)$$

The corresponding likelihood function is a sum of contributions from each of the biological samples,

$$L(A \mid B) = \ln p(A \mid B) = \sum_{j=1}^{m} \ln l_j(\alpha, \beta \mid x_{\bullet j}) \quad (15)$$

Maximizing $p\ (A|B)$ or $L\ (A|B)$ gives the maximum likelihood. The resulting estimator for the considered linear relationship is called the simple ML-estimator. The likelihood takes values between 0 and 1. Likelihoods are different to probabilities [29] with respect to $p\ (B|A)$ which is maximized in order to find the most likely parameters for a given hypothesis, here: a linear hypothesis.

(3) An adapted Maximum Likelihood estimator for robust verification of linear hypotheses

The product $p\ (B|A)$ can never become larger than one of its factors, and it comprises exactly one global maximum. Consequently, just a single outlier may decrease $p\ (B|A)$ significantly. Unfortunately, outlier data are frequently found in biological data sets due to both the multitude of factors in biological cells and the complexity of data acquisition methods that may result in false positive data points. Furthermore, it is still unclear, how many linear relationships exist for a given pair of variables. Both questions are reflected by introducing a decoupling term, the constant c, to the likelihood function,

$$L_c(A \mid B) = \ln p_c(A \mid B) = \sum_{j=1}^{m} \ln(l_j(\alpha, \beta \mid x_{\bullet j}) + c) \quad (16)$$

Additional File 2 gives the impact of the magnitude of the constant c on the total likelihood. It is demonstrated in an empirical way that the total likelihood is not decreased by any $c \geq 1$. In what follows we fix the constant at the value $c = 1$ und consider an adapted ML-estimator that is constructed by maximization of $L_1(A \mid B)$.

The adapted likelihood function is a sum, to which every data point adds a contribution between zero and a maximum value of ln 2 if it coincides with the linear hypothesis that is under investigation. This step alters the impact of the Bayesian law. It results in assessing each individual variable pair by a likelihood of contribution to a (linear) hypothesis, and not by assessing the entirety of all variable pairs. Consequently, the contribution of outliers is evanescent as demonstrated in figure 1 and is limited to a reduction of L of ln 2 in the worst case. Figure 1 shows two plots, each representing 30 data pairs. In the upper panels, the 30 data pairs shall follow a hypothetical linear function with an additional modeled analytical measurement error. The lower panels represent a data set in which 15 of the 30 data points are transposed by a constant value, so that two linear functionalities exist. For each of the two examples, likelihood distributions are given across a part of the hypothesis space for the simple ML estimator (mid panels B) and the adapted ML estimator (right hand panels C). For the case of a dataset that comprised two likely linear functions, the simple ML estimator only recognizes a shift in the local maximum but fails to detect two local maxima according to the two linearities. In contrary, the adapted ML estimator correctly identifies both local maxima and is thus able to detect the most likely parameters for both linear functions. In fact, the local likelihood maximum of the original single linear function does not shift for the adapted ML estimator when 15 data points are shifted by a constant factor but it just leads to a decrease of the maximal possible likelihood. If all measured data are assigned to different linear hypotheses according to their corresponding likelihoods, the criterion (2) as given above in the section 'background' is fulfilled: Sub sets of data are now grouped according to presence of (multiple) linear relationships.

Additional considerations are outlined for the case of missing data (NANs, not-a-number) which are often found in metabolomic data sets. In such cases, the probability function $p\ (B|A)$ needs to be adapted. This function then represents the information about the missing value: for example, data could get lost due to measurement instrument malfunctions or the variable (i.e. a metabolite level) might be below the limit of detection in a given biological situation. In both cases, the probability density function is uniform, i.e. the probability is constant in a certain range. For the case of 'below detection limit', the probability density is limited, for the case of 'instrument

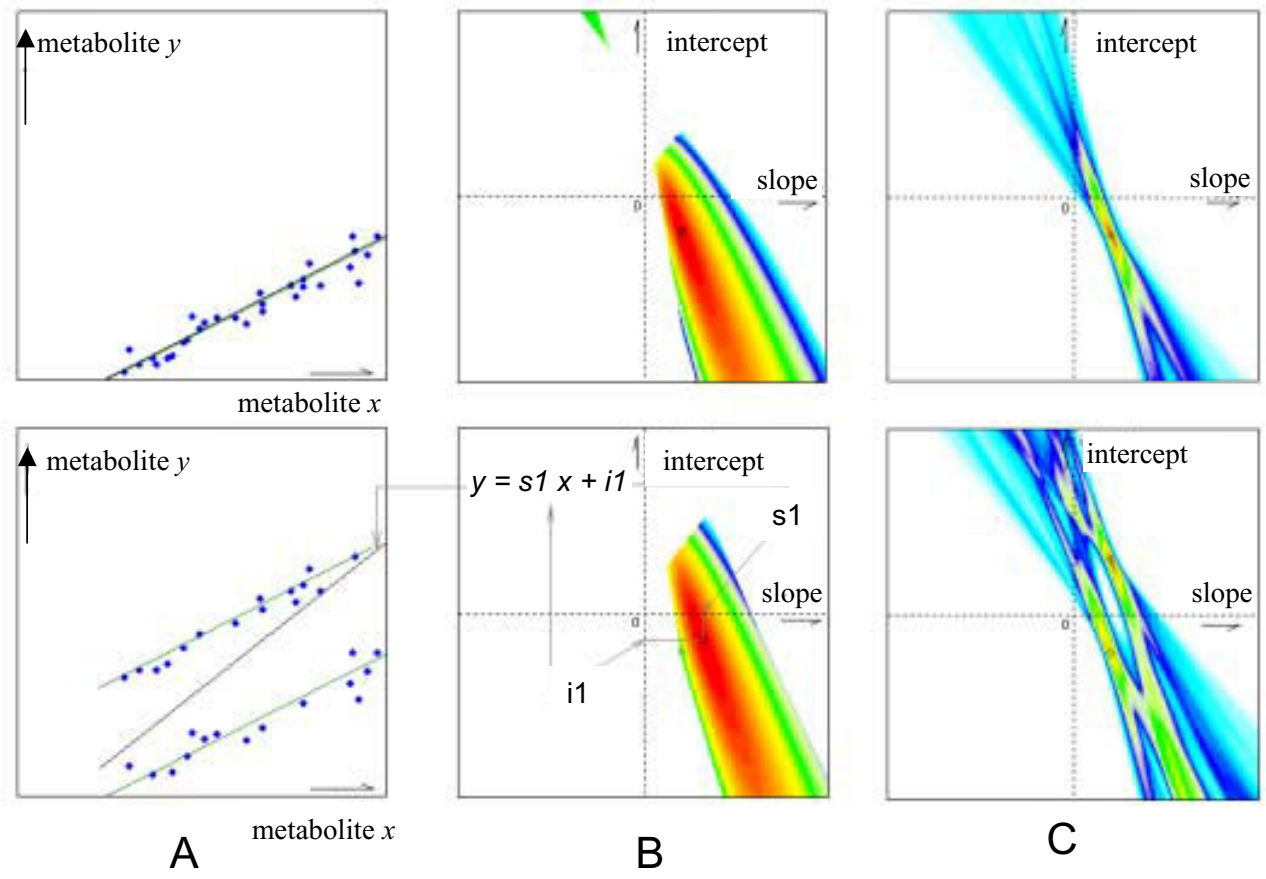

Figure 1
Comparison between simple and adapted maximum likelihood estimation. *Graph A*. The upper panel represents a set of 30 covariate pairs (‚samples') which can be described by a linear function. Deviation from this function is due to a simulated technical error. The lower panel comprises 30 samples for which half of the data were shifted for a constant value. *Graph B* Likelihood distribution for the hypotheses space using the simple maximum likelihood estimator using data from upper and lower panel from graph A. For any given linearity parameter (slope and intercept), the estimated likelihood is increasing from white to cyan, blue, gree, yellow, orange and red. Upper panel: For a single linearity, the global maximum (black circle) matches with the linearity parameters of the simulated function (green circle). Lower panel: The simple maximum likelihood estimator fails to detect and represent the presence of two linear functions. The global maximum is calculated for a single linearity which is depicted in graph A, lower panel. *Graph C*: Likelihood distribution for the hypotheses space using the simple maximum likelihood estimator using the same data set as in graphs A and B. A single linearity is correctly identified (upper panel). Importantly, data sets comprising more than one linear function are also correctly matched reporting both slope and intercept parameters.

malfunction' the probability density is zero at all levels. However, for both cases, the undefined integral is one (we must assume a false negative metabolite detection). If we knew the true cause of the missing value (i.e. either false negative or true negative), the correct probability density function could be modeled. For now, however, we need to assume an infinite technical error which demands to add the maximal likelihood of ln2 to these data points. Accordingly, missing data do not have a diminishing impact on *L*. An extreme case of data set that exclusively comprises missing data would result in a maximal likelihood for arbitrary linear hypotheses. This interpretation is correct because all hypotheses would be equally probable and could not be denied, which, however, results to an interpretive power of zero. Consequently, for real cases a maximal number of missing values needs to be defined in order to deny any linear hypothesis that might be due to missing explanatory power. The upper limit of the number of such missing data has to be set by the user who may call in further biological or analytical background information for individual metabolite pairs.

Concluding, the following properties are observed for the adapted ML-estimator:

(i) The adapted ML-estimator considers technical errors.

(ii) The adapted ML-estimator detects linear patterns and groups sub sets of data accordingly.

(iii) The adapted ML-estimator is robust against outliers.

(iv) The adapted ML-estimator relies on background information on missing values and therefore does not distort interpretations.

Therefore, the adapted ML-estimator realizes a solution to several of the challenges of unbiased and robust detection of multiple linear hypotheses in complex data sets.

(4) Algorithm for the detection of linear relationships

In order to assign measured data to a hypothetical linear relationship without contradictions, corresponding residues have to be analyzed. One condition is that these residues are randomly distributed; otherwise, additional systematic errors would have to be assumed. Secondly, the residues have to be explainable by the technical errors in a statistical manner. The adapted ML-estimator already realizes a measure for agreement between (linear) model and data under consideration of the corresponding technical errors. Thresholds can now be determined for rejecting specific linear hypotheses using the distribution of L, resulting in false discovery rates for which limits can be set. The core of the algorithm determines the local maximum of a likelihood distribution which is subsequently compared to limits of a test statistics. It can be assumed that this maximum will be the global maximum since all measured data will be explained by the tested linear relationship. However, outliers will reduce the likelihood drastically. Consequently, data are only considered if residues are small to the hypothetical linearity. The 2 σ interval was chosen to exclude outliers at 95% confidence. The data inside the 2 σ confidence interval is denoted by B_{return}. The likelihood is subsequently normalized to the number m_{return} of this data. The parameter m_{return} comprises the number of samples that were returned to belong to a linear function despite deviation that is due to the contribution of unrelated variance. Each data point contributes a value of ln 2 to the likelihood function, resulting in the normalized likelihood

$$L_{\text{return}} = \frac{L_1\left(A_{\max} \mid B_{\text{return}}\right)}{m_{\text{return}} \times \ln 2} \tag{17}$$

We now have two parameters, m_{return} and L_{return}, for which test statistics can be determined based on randomly selected true linear relationships. $A_{\max}$ denotes the parameters for which the maximum likelihood is assumed. The distributions of m_{return} and L_{return} were assessed by Monte Carlo simulations: For each sample size ranging from 3 to 150 data points we have generated 25,000 random data sets, and test statistics were derived for each sample size m. The data set were generated by selecting an arbitrary linear function and a random selection of data points corresponding to this linear function. Technical errors were sampled from Gaussian distributions and added to the data points. After localizing the maximum value of L_1 one determines all samples which belong to the corresponding linear function. Based on L_1 and m_{return}, the value of L_{return} is determined as given above. The frequency distributions of L_{return} for different values of m_{return} are shown for the example of m = 20 samples (figure 2). Figure 2 demonstrates that the L_{return} distributions varied for different m_{return} values, and consequently, corresponding test statistics were established that set the limits for rejecting the null-hypotheses at a false negative error rate of ≤ 5% for each of the m_{return} values.

(5) Determination of false positive and false negative error rates

The degree of noise can be described in terms of the reliability that is defined as ratio of biological variance and total variance. The later is just the sum of biological and technical variance if both variances are not correlated. In that case the reliability of the measurement of metabolite x is given as

$$\rho_x = \frac{\sigma^2_{biol}(x)}{\sigma^2_{biol}(x) + \sigma^2_{tech}(x)} \tag{18}$$

The average reliability can easily be obtained from the simulations, and hence, the degree of noise can well be described as

$$R = 1 - \bar{\rho}_x. \tag{19}$$

We here assume that linear relationships between two metabolites are only confused by technical errors, but not by other biological factors, so the degree of noise here is only induced by technical errors. In order to test the algorithm described above, a data set was simulated that closely describes the problem. This model data set comprised 200 variables which were grouped into 20 clusters of equal size. All variables within a cluster were described by a linear relationship $y = ax + b$, but between clusters, no linearity was modeled apart from random relationships. For each test, a different number of samples was taken to assume experimental data from metabolomic snapshots, with further and various levels of technical errors that were added to the modeled measurements. Technical errors were assumed to follow a Gaussian distribution. In total, the total data set was investigated for 19,900 pairwise relationships of which 900 were modeled to be described by linear relationships in order to assess false positive and false negative error rates. The parameters

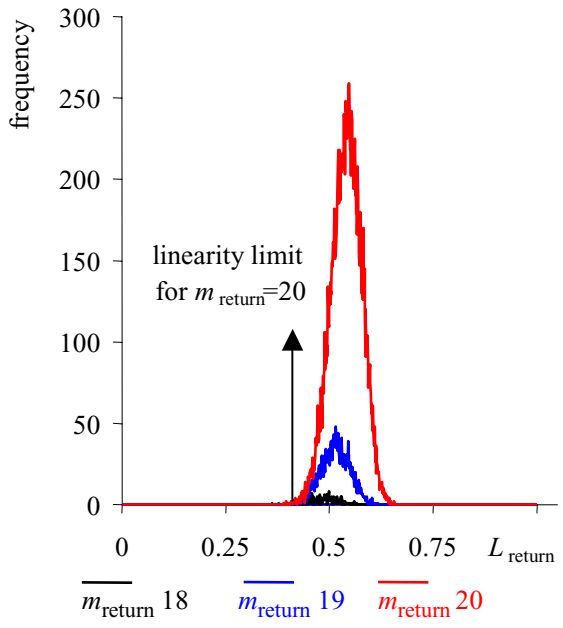

Figure 2
Determination of the linearity rejection region by Monte Carlo simulations. 3–150 samples were used from linear functions which were imposed by additional Gaussian noise. The example for m = 20 is shown, for which in some cases, fewer than 20 samples were returned due to outliers that were caused by the imposed technical error. For each of these m_{return} values, adapted maximum likelihood limits were determined for which the null hypothesis, the existence of a linearity, would need to be rejected.

were varied in an exhaustive permutation of the sample size from 3–150 samples and relative technical errors from 0.1–100%. Error rates were determined four times for each combination of ‚number of samples' and ‚technical errors', and the average of these four determinations was taken. Technical errors may be divided into the absolute error and a relative error. The absolute error, for example, is constituted by the resolution of an analytical instrument or constant background through chemical impurities of reagents and solvents. Such errors can be carefully controlled in validated chemical procedures and are usually less important than relative errors. In most cases in metabolomics, the total technical error is dominated by relative errors that relate to the true value, originating for example by sample storage, extraction and sample preparation procedures, and by cross-contamination and carry over between samples. Technical errors can be estimated by reproducing all sample preparation steps multiple times from small aliquots of a larger homogenized pool, and subsequent data acquisition. The magnitude of relative errors varies by the vulnerability of the compounds to be altered during the sample preparation and measurement process. However, for the sake of clarity, identical technical errors were used in the simulations for each pair of metabolites.

The error rate of the algorithm is exemplified for selected sample numbers in figure 3 (upper panel), using the algorithm described so far. The null hypothesis used here was assuming the existence of a linear relationship between any pair of metabolites. This is in opposite to classical use of null-hypotheses, reversing the meaning of false positives and false negatives in our work. Therefore, in our case rejecting the null hypothesis when it is actually true means rejecting true linearities or generating false negative errors. It is important to note that the count for false negative detections (type I errors) stays below 5% except for sample numbers smaller than five. The minimal error rate corresponds to the limits that resulted from constructing the test statistics. For the false positive error rates (type II errors), a different trend is observed. Except for very low technical errors or large numbers of samples, the type II errors quickly exceed the 5% error thresholds. Generally, the sample size required to cope with higher technical errors rapidly increases for maintaining acceptable false positive rates. For high technical errors, the pattern between any two variables resembles a scatter around a constant value. In such cases, the number of false positive linearity detections increases because any constant value can be explained by a discretionary linear function. If higher limits were used for the parameters L_{return} and m_{return}, the 5% threshold for the false positive rates would be reached at higher technical error rates. However, simultaneously, the minimal error rate for of false negatives would increase. Consequently, type I and type II error rates could in principle be balanced by adapting the thresholds for L_{return} and m_{return} in a qualitative manner. Nevertheless, the total error rate can only be influenced by decreasing the technical error or increasing the number of samples taken into account.

As outlined above, increasing levels of technical errors cause higher false positive error rates of detections of linearities. However, the number of false positive detections can be shifted towards false negative error rates, if desired for a specific biological study. Therefore, a filter has been developed that filters out all potential false positives (Additional File 4). The filter has been tested on the same simulated data as the algorithm before. Results are shown in figure 3 (lower panel) for false negative and false positive linearity detections. Compared to figure 3 (upper panel), a reverse order for false linearity discoveries was observed. Technical errors in metabolomics are usually in the range around 20–25%, although for certain compound classes, these may be lower. For a given biological situation, experimental biologists rarely use more than 10

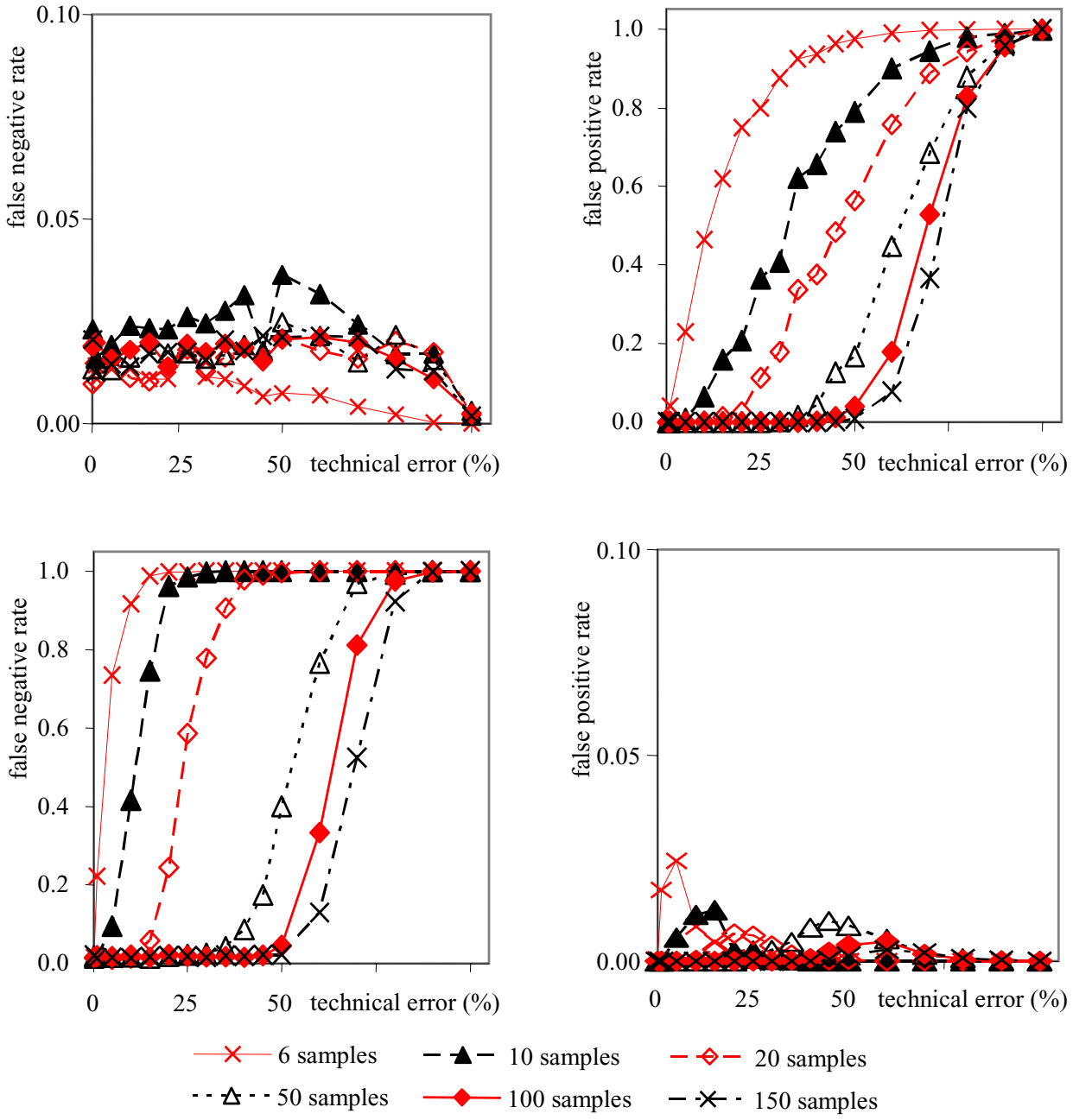

Figure 3
False negative and false positive error rates of the algorithm tested on simulated data in relation to the number of samples and the assumed technical errors, in % of the total variance. 900 pair-wise linear relationships between 200 metabolites were defined that were tested against the total of 19,900 potential linearities. Upper panel: Error rates without filter. Lower panel: error rates with filter.

biological replicates ('samples'), often even less. With other words, for such a combination of technical errors and low number of replicates, the algorithm yields either the accurate identification of all true linearities (without filter), but for the price of a high number of false positives, or the algorithm results in the full deletion of false posi-

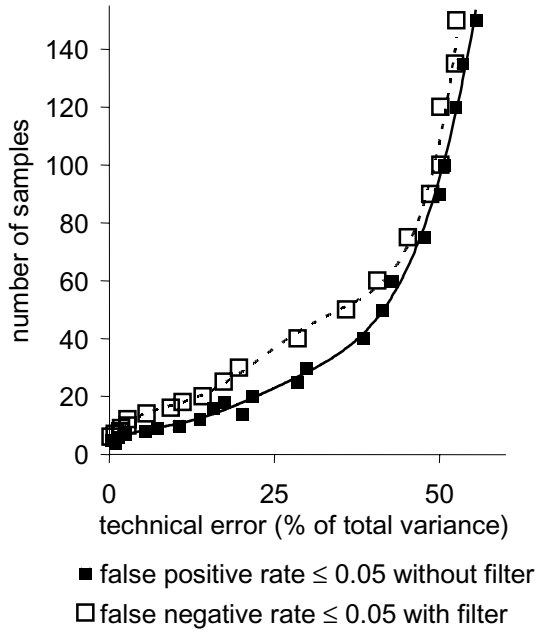

Figure 4
Number of samples required in relation to the assumed relative technical errors, if both false positive and false negative error rates are to remain below 0.05 (i.e. 5%). 900 pair-wise linear relationships between 200 metabolites were defined that were tested against the total of 19,900 potential linearities.

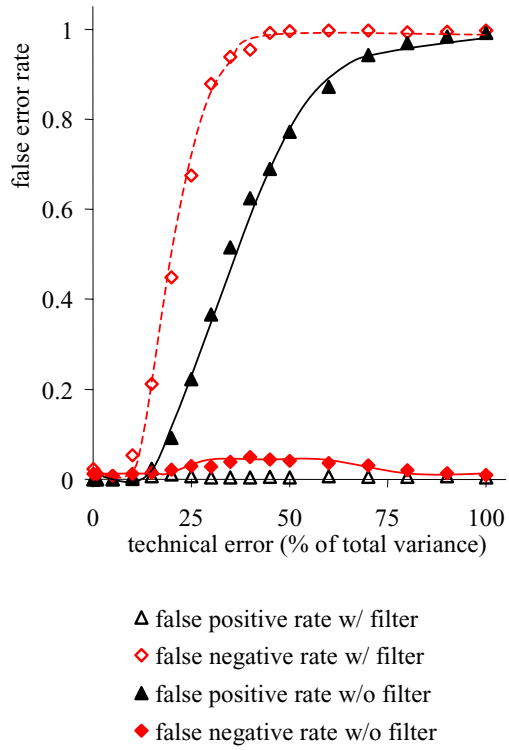

Figure 5
Influence of outliers on false negative and false positive error rates on a sample size of m = 20. In a similar manner to figure (3), simulated data were imposed by technical errors, but in addition, by the presence of outliers that were located in a 2–1000 σ distance from the linear function. The algorithm proved to be robust against such outliers.

tives (with filter), but for the price of not detecting a high number of the true linearities. We therefore have compared the results with and without filter in order to determine, how many samples would be needed to stay reliably at a false error rate ≤ 0.05 (i.e. 5%) for both false negatives and false positives. Figure 4 demonstrates that false positive detections without activating the filter obey a more favourable response to the combination of sample size and technical errors than the false negative error rates with active filter. Consequently, at technical errors of 20%, a minimum of about 20 samples is needed to stay reliably below 5% error rates for both false positive and false negative detections of linear metabolic relationships. This is an important result for practical use of this algorithm for robust generation of metabolomic networks. As default, the filter is not needed to be applied unless researches want to be very strict on the false positive error rates. The simulation presented here demonstrates that it is possible to remain at a total error rate of less than 5% if more than 20 samples are analyzed at the 20% technical error rate. The entirety of linear relationships of metabolites may subsequently be visualized as network graphs. Such graphs can be compared between different physiological or genetic conditions, in order to generate novel functional hypothesis on regulation of metabolic networks in a robust manner.

The robustness of the algorithm was tested on a model dataset with 20 samples (figure 5). The thresholds for L_{return} and m_{return} were adjusted to tolerate an outlier rate of 5% (one of 20 samples). Outliers were modelled with a distance from 2 σ to 1000 σ away from the true linearity. It was found that outliers were generally easier recognized when these were very distant from the linearity. Despite the additional outliers, false positive and false negative error rates were found to almost identical as in figure 3. The number of false negative linearity detections remained below 5% without filter in all cases, and conversely, the false positive rate remained unchanged with activated filter. Hence, the use of Bayesian likelihood estimations enables robust detection and verification of linear relationships in an unbiased way and in complex

datasets. If more outliers are present in a data set, these may actually constitute several local likelihood maxima as shown in figure 1C (lower panel) or in the figures in Additional File 2. Such additional linear relationships may be revealed if several outliers follow a different linear function and hence yield unexpected hypotheses of cellular regulation. Finding and validating such multiple linearities has so far been hard to accomplish with classical tools but is now amenable with the algorithm presented here. The algorithm has been implemented in a stand alone software solution. For data sets of a size of 200 variables × 150 samples, robust linearity networks are generated in around 10 min computing time using a 512 MB RAM and 3.5 GHz personal computer. The actual computing time will vary from 3.5–15 min, depending on the actual linearity structure of the data set. Improved implementations of the algorithm, specifically for the search of global likelihood maxima, may certainly be worked out more effectively with respect to computational run time. However, acquiring metabolomic data of the size of 150 samples (from growth of biological organisms, harvesting, sample processing, data acquisition to data processing) will take time on the order of weeks which surely justifies computational efforts on standard personal computers on the order of minutes.

Conclusion

Use of the technical error concomitant with a maximum likelihood assessment of linearity parameters and verification by simulated test statistics enables a robust detection and verification of liner relationships in complex data sets. An implementation of this algorithm will enable biologists to calculate and compare linearity networks in metabolomic or other multivariate data sets, from which biological hypotheses may be derived. The algorithm can be modified with respect to the ratio of type I and type II errors depending on the biological focus of a study. It is highly advised to use more than 20 biological replicates for each condition that is to be tested in a biological experimental design of *genotypes x environments* (*G x E*), unless advances in analytical chemistry and instrumentation decrease the overall technical error to very low levels, i.e. below 5%. Even the existence of more than one linear relationship per pair of variables can be detected using the maximum likelihood algorithm, which has so far been hard to compute with classical approaches.

Authors' contributions

FK has worked out, tested and implemented the algorithm. MH had initially advised on the mathematics of likelihood estimations. JB eventually revised and improved the mathematical description of the algorithm and contributed to writing the paper. OF conceived the study, participated in developing and testing the algorithm and drafted and wrote the manuscript.

Additional material

Additional File 1

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-162-S1.pdf]

Additional File 2

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-162-S2.pdf]

Additional File 3

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-162-S3.pdf]

Additional file 4

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-162-S4.pdf]

Acknowledgements

The work was funded by the NIEHS through the R01 project ES13932 granted to OF and by a fellowship granted to FK by the Max-Planck Society, Germany. Helpful comments by Joachim Selbig are appreciated.

References

1. Morgenthal K, Weckwerth W, Steuer R: **Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation.** *Biosystems* 2006, **83(2–3):**108-117.
2. Ratcliffe RG, Shachar-Hill Y: **Measuring multiple fluxes through plant metabolic networks.** *Plant Journal* 2006, **45(4):**490-511.
3. Kose F, Weckwerth W, Linke T, Fiehn O: **Visualizing plant metabolomic correlation networks using clique-metabolite matrices.** *Bioinformatics* 2001, **17(12):**1198-1208.
4. Fiehn O: **Metabolic networks of Cucurbita maxima phloem.** *Phytochemistry* 2003, **62(6):**875-886.
5. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O: **Differential metabolic networks unravel the effects of silent plant phenotypes.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101(20):**7809-7814.
6. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks.** *Bioinformatics* 2003, **19(8):**1019-1026.
7. Camacho D, de la Fuente A, Mendes P: **The origin of correlations in metabolomics data.** *Metabolomics* 2005, **1:**53-63.
8. Lin H, Bennett GN, San KY: **Chemostat culture characterization of Escherichia coli mutant strains metabolically engineered for aerobic succinate production: A study of the modified metabolic network based on metabolite profile, enzyme activity, and gene expression profile.** *Metabolic Engineering* 2005, **7(5–6):**337-352.
9. Grubbs FE: **Errors of Measurement, Precision, Accuracy and Statistical Comparison of Measuring-Instruments.** *Technometrics* 1973, **15(1):**53-66.
10. Tocher JF: **Pigmentation survey of school children in Scotland.** *Biometrika* 1908, **6:**A1-A67.
11. Horton NJ, Laird NM: **Maximum likelihood analysis of generalized linear models with missing covariates.** *Statistical Methods in Medical Research* 1999, **8(1):**37-50.
12. Lindsey KJ: **Applying generalized linear models.** 1st edition. New York: Springer; 1997.
13. Davis PL: **Aspects of robust linear regression.** *Annals of Statistics* 1993, **21(4):**1843-1899.

14. Andrews DF: **Robust method for multiple linear-regression.** *Technometrics* 1974, **16(4):**523-531.
15. Wald A: **The fitting of straight lines if both variables are subject to error.** *Annals of Mathematical Statistics* 1940, **11:**284-300.
16. Berkson J: **Are There 2 Regressions.** *Journal of the American Statistical Association* 1950, **45(250):**164-180.
17. Scheffe H: **Fitting Straight-Lines When One Variable Is Controlled.** *Journal of the American Statistical Association* 1958, **53(281):**106-117.
18. Cressie N, Read TRC: **Pearsons-X2 and the Loglikelihood Ratio Statistic-G2 – a Comparative Review.** *International Statistical Review* 1989, **57(1):**19-43.
19. Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR: **Parallel analysis of transcript and metabolic profiles: a new approach in systems biology.** *Embo Reports* 2003, **4(10):**989-993.
20. Chen YP, Popovich PM: **Correlation: Parametric and nonparametric measures.** 1st edition. Sage Publications; 2002.
21. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients.** *Bioinformatics* 2004, **20(18):**3565-3574.
22. Perl J: **Causality: Models, reasoning and inference.** 1st edition. New York: Cambridge University Press; 2000.
23. Wright S: **Correlation and causation Part I. Method of path coefficients.** *Journal of Agricultural Research* 1920, **20:**0557-0585.
24. Morgenthal K, Wienkoop S, Scholz M, Selbig J, Weckwerth W: **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* 2005, **1:**109-121.
25. Thomas S, Fell DA: **The role of multiple enzyme activation in metabolic flux control.** *Advances in Enzyme Regulation* 1998, **38:**65-85.
26. Bayes T: **An Essay Towards Solving a Problem in the Doctrine of Chances.** *Biometrika* 1958, **45(3-4):**296-315.
27. Lee PM: **Bayesian statistics: An introduction.** New York: Oxford University Press; 1989.
28. Box GEP, Tiao GC: **Bayesian inference in statistical analysis.** Reading, MA: Addison-Wesley Publishing Company; 1973.
29. Fisher RA: **On the 'probable error' of a coefficient of correlation deduced from a small sample.** *Metron* 1921, **1:**1-32.