Metabolomics Vol. 1, No. 1, January 2005 (© 2005)
DOI: 10.1007/s11306-005-1107-3

The origin of correlations in metabolomics data

Diogo Camacho, Alberto de la Fuente, and Pedro Mendes*

Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, MC 0477, Washington St., Blacksburg, VA, 24061, USA

Received 19 August 2004; accepted 15 September 2004

A phenomenon observed earlier in the development of metabolomics as a systems biology methodology, consists of a small but significant number of metabolites whose levels are highly correlated between biological replicates. Contrary to initial interpretations, these correlations are not necessarily only between neighboring metabolites in the metabolic network. Most metabolites that participate in common reactions are not correlated in this way, while some non-neighboring metabolites are highly correlated. Here we investigate the origin of such correlations using metabolic control analysis and computer simulation of biochemical networks. A series of cases is identified which lead to high correlation between metabolite pairs in replicate measurement. These are (1) chemical equilibrium, (2) mass conservation, (3) asymmetric control distribution, and (4) unusually high variance in the expression of a single gene. The importance of identifying metabolite correlations within a physiological state and changes of correlation between different states is discussed in the context of systems biology.

KEY WORDS: metabolomics; correlations; metabolic control analysis; systems biology.

1. Introduction

Large-scale molecular profiling methods, often referred to as "omics", are becoming predominant in molecular biology. They have facilitated the appearance of observation-driven hypothesis generation experiments, where the emphasis is predominantly in identifying new phenomena, rather than investigate a known one in detail. These technologies are also becoming important to a systems biology approach, where they are applied in the context of a solid theoretical background and through computational models (Kitano, 2002). Global transcript analysis through microarrays (Liang *et al.*, 2004) is the most commonly used technique, followed closely by large-scale protein identification and quantification (Righetti *et al.*, 2004), and analysis of protein–protein interactions (Janin and Seraphin, 2003; Cho *et al.*, 2004). There are also several approaches for metabolite identification and quantification (Raamsdonk *et al.*, 2001; Reo, 2002; Sumner *et al.*, 2003), commonly known as metabolomics or metabolite profiling, but see (Fiehn, 2001) for a clarification of terms. The techniques used in metabolomics are predominantly based on chromatography and mass spectrometry (Roessner *et al.*, 2000; Tolstikov *et al.*, 2003), although nuclear magnetic resonance (Reo, 2002), Fourier-transform infra-red spectroscopy (Oliver *et al.*, 1998; Harrigan *et al.*, 2004), and capillary electrophoresis (Baggett *et al.*, 2002; Soga *et al.*, 2002) are also commonly used. Like the other "omic" techniques, metabolomics data are usually in the form of ratios of concentrations; absolute concentrations are rarely obtained except in targeted analyses that cover only a small number of metabolites.

Metabolomics is a crucial tool in systems biology because it monitors the ultimate products of gene expression (Oliver *et al.*, 1998): organic molecules that are not directly encoded in the genome and are synthesized by a diversity of enzymes. Metabolites are produced from other metabolites resulting in a level of interdependence between their concentrations that does not exist between transcripts or proteins. These constraints result from the structure of the metabolic network (stoichiometry) and, when known, can be used to derive structural biochemical properties of those networks (e.g. Schilling *et al.*, 1999). But currently the structure of metabolic networks is limited to the primary metabolism of microbial model organisms and some mammalian tissues. Very little is known about secondary metabolism or even the primary metabolism of many organisms, resulting in the order of one hundred thousand natural products for which no synthetic pathway is known. Thus, stoichiometry-based analysis of metabolomics data is currently limited to a handful of cases. What remains to be seen is if metabolomics data can actually be used to uncover those unknown metabolic networks.

Metabolomics data can be analyzed with the same methods used in transcriptomics and proteomics, such as clustering (Roessner *et al.*, 2001), principal component analysis (Oliver *et al.*, 1998; Nicholson *et al.*, 1999), or machine learning (Kell, 2002; Ott *et al.*, 2003). Additionally, it may be possible to develop novel analyses by exploring the vast body of existing theory on

*To whom correspondence should be addressed.
E-mail: mendes@vbi.vt.edu

metabolism and its regulation (e.g. Kacser and Burns, 1973; Savageau, 1976; Atkinson, 1977; Hayashi and Sakamoto, 1986; Fell, 1996; Heinrich and Schuster, 1996). The present text describes an interpretation of certain metabolomics data structures using concepts from metabolic control analysis (Kacser and Burns, 1973; Fell, 1996; Heinrich and Schuster, 1996). Such a use of theoretical concepts is expected to result in analyses that result in a greater understanding about the underlying processes, absent from most of the methods in current use.

Even though metabolomics approaches are still sparsely reported, they are already revealing very interesting phenomena. Perhaps the most striking one was the observation in a comparison between four different *Solanum tuberosum* genotypes, that a small number of metabolite pairs displayed a remarkably high correlation among biological replicates, even though the large majority of metabolite pairs showed little or no correlation (Roessner *et al.*, 2001). Subsequent studies confirm the ubiquity of this phenomenon with different techniques (Weckwerth *et al.*, 2004) and different organisms (Fiehn, 2003; Broeckling *et al.*, 2004; Martins *et al.*, 2004). Figure 1 illustrates this phenomenon through two metabolite scatter plots, one with nearly no correlation and another with high correlation. Since the large majority of metabolite pairs do not show high correlation, and cases like the valine-methionine pair of figure 1b are rare, it becomes even more imperative to understand why such correlations exist.

A naïve interpretation of this phenomenon of metabolite correlations could be that the pairs with high correlation would be neighbors in the underlying metabolic network. If so, then observations of this phenomenon could help resolving unknown metabolic networks. Unfortunately this does not resist simple scrutiny because there are many pairs of metabolites that are neighbors in the metabolic map yet have low correlation (e.g. figure 1a), and others that are not neighbors but have high correlation (e.g. figure 1b). This has indeed been shown by theoretical and computational analyses that point to the correlations being shaped by a combination of stoichiometric and kinetic effects (Steuer *et al.*, 2003). This helped understanding that not all neighbor metabolites have high correlations (as most do not), but it did not go as far as to explain what originates the high correlations. Knowing this would allow us to infer valuable knowledge about the biochemical organization of cells. It is our aim here to further investigate the origin of these high correlations. This will be done using the established principles of metabolic control analysis and computer simulation of example biochemical networks. This analysis extends the utility of metabolite profiles to diagnose global regulatory phenomena transcending the metabolite level, which emphasizes the important role of metabolomics in systems biology approaches.

2. Materials and methods

2.1. Theoretical

We make use of concepts from metabolic control analysis (Kacser and Burns, 1973; Heinrich and Rapoport, 1974; Fell, 1996; Heinrich and Schuster, 1996), in particular co-response analysis (Hofmeyr *et al.*, 1993; Hofmeyr and Cornish-Bowden, 1996).

2.2. Computational

2.2.1. Base models

All simulations were carried out with the Gepasi software (Mendes, 1993, 1997) version 3.30 on a Pentium Centrino 1.4 GHz computer (Dell Corp., Round Rock, TX) running Windows XP (Microsoft Corp., Redmond, WA). A mathematical model of yeast

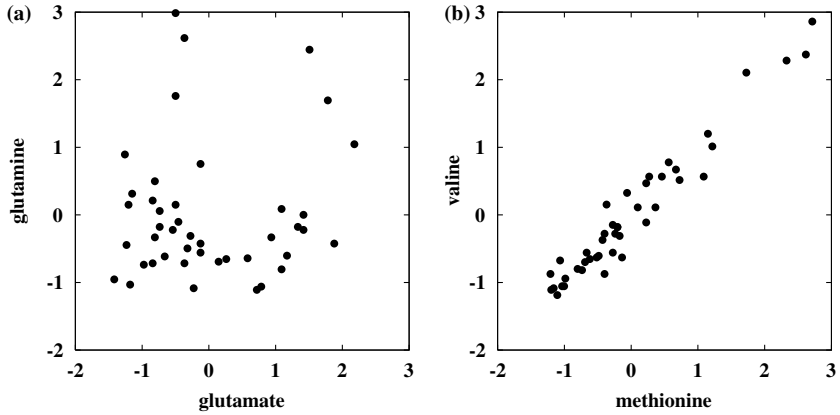

Figure 1. Metabolite pair scatter plots of replicate samples of wild type potato tubers, measured with GC-TOF-MS (data from Weckwerth *et al.*, 2004). Plots show *Z*-transformed intensities (i.e. mean centered and scaled by standard deviation). (a): glutamate-glutamine pair, which are uncorrelated ($r = 0.0243$, Spearman) despite being metabolic neighbors through glutamine synthase (EC 6.3.1.2). (b): valine-methionine, which do not participate in common reactions but have a strong correlation ($r = 0.9510$, Spearman).

glycolysis is used as an example of a metabolic network. We adopted the model of Teusink *et al.* (Teusink *et al.*, 2000), with adjustments proposed by Pritchard and Kell (2002), which provides a fairly accurate model of yeast glycolysis in the presence of 50 mM glucose in the medium. We sought a second set of parameters for this model that would correspond to a different state of the system. Unfortunately none is available at this time for yeast, so we had to proceed with creating one artificially – while this does not correspond to any real instance of yeast glycolysis, it allows us to illustrate what happens when there is a change in regulation. In order to obtain this second ''physiological'' state, we changed glucose concentration in the medium to 20 mM and the reduced enzyme limiting rates ($V_{\max}$), representing a change of enzyme concentrations. In this new state, the concentrations of the enzymes in the upper part of glycolysis were reduced to 50%, while the concentrations of the enzymes in lower glycolysis were reduced to 75% of their original values. The fixed rates of the succinate, glycogen, and trehalose branches were also reduced to 75%, while the fixed rate of the glycerol branch was reduced to 50%. This new state represents an example of changes in environment and regulation through gene expression; it does not attempt to simulate an actual physiological condition. All simulation models are available in Gepasi and SBML format (Hucka *et al.*, 2003), and can be obtained from our website (http://mendes.vbi.vt.edu/static/Metabolomics04/).

2.2.2. Biological replicates

The phenomenon discussed here is based on correlation between components of biological replicate samples. Thus it is necessary to introduce variability in the simulations but also to do it in a way that matches the differences between different organisms of the same species. In the related work of Steuer *et al.* variability was introduced by adding intrinsic noise to some metabolite concentrations, which was achieved through stochastic ordinary differential equations (Steuer *et al.*, 2003). In this way, random perturbations are constantly affecting the system, perhaps similar to what would happen in the presence of thermal noise (though the magnitude of the noise used in those simulations is arguably much larger than the results of thermal noise). A different strategy is used here to simulate variability between organisms: each sample is set to have slightly different enzyme concentrations in the initial state, but which then remain constant in time (i.e. the variability is not noise). Such different enzyme concentrations would reflect differences in expression levels and, if kept to a small relative magnitude, would represent individuals (or cultures) that are almost the same but with small differences. In the present case, each ''biological'' replicate differs from the base model by random deviations of enzyme concentrations with 90% and 110% of the base value (pseudo-random numbers drawn from a uniform distribution). In this way we are able to simulate ''replicates'' containing biological variation.

2.2.3. Co-response profiles

While biological replicates are assumed here to differ by random (but small) differences in *all* enzyme concentrations, it is important to investigate how each enzyme concentration alone affects the system. In order to do this, simulations were carried out where each enzyme concentration is varied from 90% to 110% of its base value, while keeping other enzyme concentrations fixed at their base values. Results are plotted as co-response profiles, showing how two metabolite concentrations are affected by the enzyme concentrations in log–log plots.

3. Results and discussion

3.1. Metabolic control analysis

A collection of measurements of metabolite concentrations obtained from multiple observations of replicate biological samples (figure 1) can be considered as perturbations of a mean state. The variance observed in these measurements can be considered to arise from slight differences in internal and external parameters between individuals, such as enzyme concentrations, kinetic constants, and environmental conditions. In order to express this quantitatively it is useful to introduce the concentration response coefficient, which is defined as the relative change in the steady state level of a biochemical concentration in response to a relative change in some parameter (Kacser and Burns, 1973; Heinrich and Rapoport, 1974). Formally, this is expressed by the dimensionless coefficient:

$$R_{p_k}^{X_i} = \frac{\mathrm{d}X_i}{\mathrm{d}p_k}\frac{p_k^0}{X_i^0} = \frac{\mathrm{d}\ln X_i}{\mathrm{d}\ln p_k}. \quad (1)$$

Here X_i is the concentration of interest, with X_i^0 its reference value; p_k is the parameter that changed (e.g. an enzyme concentration), with p_k^0 being the reference value.

One can estimate the displacement of the concentration of interest from its reference steady state, when caused by a known parameter change:

$$\Delta\ln X_i \approx R_{p_k}^{X_i}\Delta\ln p_k. \quad (2)$$

The displacement from the reference state is proportional to the parameter change, with the proportionality constant being the response coefficient. In most cases of interest this is, in fact, an approximation because the responses are non-linear. Expanding from the previous situation, where a single parameter was changed, to the general case when n parameters are changed, the displacement in concentration can be written as a sum of n terms, each corresponding to the effect of a single parameter:

$$\Delta \ln X_i \approx \sum_k^n R_{p_k}^{X_i} \Delta \ln p_k. \tag{3}$$

When the concentrations of two metabolites, X_i and X_j, that suffered multiple perturbations are plotted against each other, the coordinates in the logarithmic plane of each observation relative to the reference are determined by

$$\frac{\Delta \ln X_i}{\Delta \ln X_j} \approx \frac{\sum_k^n R_{p_k}^{X_i} \Delta \ln p_k}{\sum_k^n R_{p_k}^{X_j} \Delta \ln p_k} \tag{4}$$

In the special case where a single parameter has been changed equation (4) reduces to

$$\frac{\Delta \ln X_i}{\Delta \ln X_j} \approx \frac{R_{p_1}^{X_i} \Delta \ln p_k}{R_{p_1}^{X_j} \Delta \ln p_k} = \frac{R_{p_k}^{X_i}}{R_{p_k}^{X_j}} \equiv {}^{p_k} O_{X_j}^{X_i}, \tag{5}$$

where ${}^{p_k} O_{X_j}^{X_i}$ is the ratio of two response coefficients, and is known as a co-response coefficient (Hofmeyr and Cornish-Bowden, 1996; Hofmeyr *et al.*, 1993). In this special case of a single parameter change, the metabolite scatter plots discussed here are the same as co-response profiles. All observations lie approximately on a straight line, with slope ${}^{p_k} O_{X_j}^{X_i}$ and the correlation coefficient would approximate $+1$ or -1, depending on the slope being positive or negative, respectively. Technical variance, reflected as error in the measurements, would reduce these values, though this is expected to be a small effect. The length of the line defined by the observations depends on the size of the parameter perturbations and on the sensitivity of the metabolites towards that parameter. That length can be calculated using the Pythagorean theorem:

$$d = \left[\left(R_{p_k}^{X_i} \Delta \ln p_k^{\max} - R_{p_k}^{X_i} \Delta \ln p_k^{\min} \right)^2 + \left(R_{p_k}^{X_j} \Delta \ln p_k^{\max} - R_{p_k}^{X_j} \Delta \ln p_k^{\min} \right)^2 \right]^{\frac{1}{2}}, \tag{6}$$

where $\Delta \ln p_k^{\max}$ and $\Delta \ln p_k^{\min}$ correspond to the two extreme perturbations of the parameter.

In general, the underlying differences between replicate samples arise from multiple parameters, and thus the metabolite scatter plots obtained from them are *not* the same as co-response profiles, unlike what had been suggested earlier (Weckwerth and Fiehn, 2002). Because the difference in parameter values between each pair of replicates is expected to be random, the slope defined by the concentrations of two metabolites in each pair of samples is expected to be different, generating a cloud of points. The shape of that cloud of points is determined by axes whose slope corresponds to the individual co-response profiles for each of the parameters, and whose length depends on the size of each parameter fluctuations and the sensitivity of the metabolites towards them (equation 6); this is demonstrated below with simulations (figures 3 and 4). When there are many parameters varying there will be many such axes, potentially with widely different slopes, and in general the correlation is not expected to be high. However, there are several special cases in which many parameters can vary and still yield high correlation between two variables. A few of these cases are examined here.

3.1.1. *One parameter dominates*

One special case occurs when only one term in equation (4) dominates. This occurs when $\left| R_{p_k}^{X_i} \Delta \ln p_k \right| >> \left| R_{p_l}^{X_i} \Delta \ln p_l \right|$ and $\left| R_{p_k}^{X_j} \Delta \ln p_k \right| >> \left| R_{p_l}^{X_j} \Delta \ln p_l \right|$ for all $l \neq k$. This special case can occur when (a) the variability of one parameter is much higher than any of the others parameters, or (b) the responses of the two concentrations towards one parameter are much higher than towards any other (a third case combining these two for the same parameter would result in an even greater effect, except if they happen for two different parameters, when their effects would cancel). In these circumstances the contribution of one parameter on the concentration of the two metabolites dominates and the other parameters become negligible. One of the axes is much longer than the others, dominating the shape of the scatter plot and most points will be aligned with this axis yielding a high correlation. Effectively only one parameter varies, reducing equation (4) to (5).

An interesting case may arise here: if two concentrations, A and B, fall in this situation (high correlation due to dominance of a single parameter), and one of them also with a third concentration, say A and C, then by necessity, B and C will also have high correlation. This is because if a single parameter dominates the correlation of A with B, then that same parameter must be the same that correlates A with C, and therefore also B with C. Such a phenomenon may lead to groups of metabolites whose concentrations are tightly correlated, forming a clique. An analysis of cliques of correlated metabolites (Kose *et al.*, 2001) is likely to uncover such cases.

3.1.2. *Equilibrium and mass conservation*

Another special case occurs when all co-response profiles lie on top of each other, i.e. both variables respond in the same direction towards all parameters. In this case all co-response coefficients are equal

$$\frac{R_{p_k}^{X_i}}{R_{p_k}^{X_j}} \equiv {}^{p_k} O_{X_j}^{X_i} = \alpha, \quad \text{for all } k. \tag{7}$$

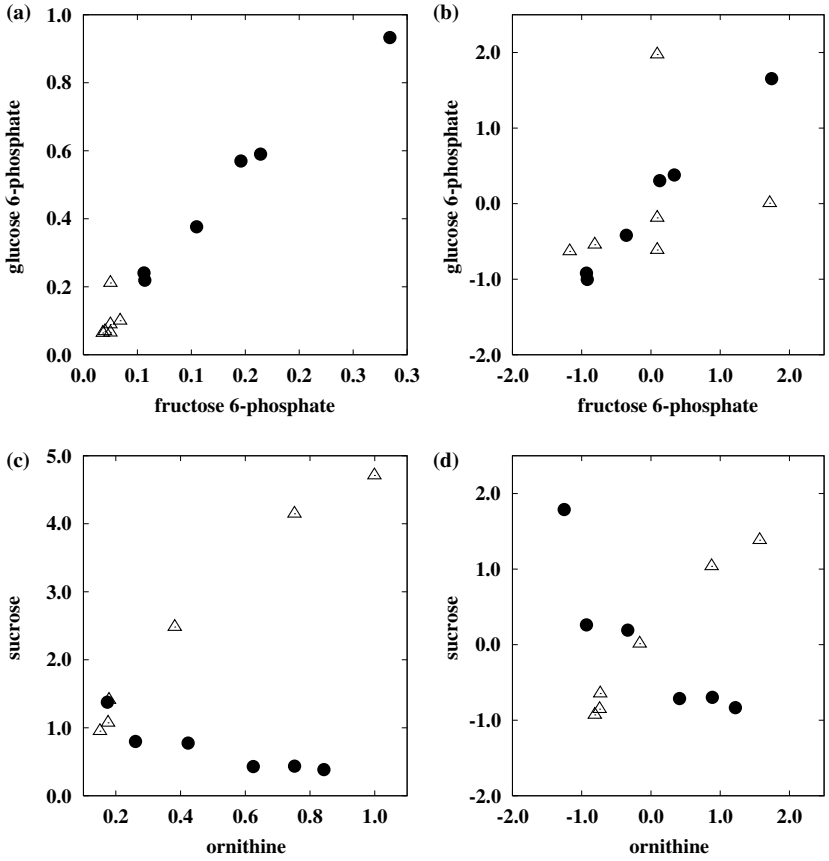

Figure 2. Comparison of correlations between two metabolite pairs in a wild type (wt) and transgenic potato tubers, measured with GC-MS (data from Roessner *et al*., 2001). Pearson correlations are calculated for each set of replicate samples from the same genotype (rank correlation estimates would be very inaccurate for this small sample size). (a) and (b): Fructose 6-phosphate and glucose 6-phosphate in wt (empty triangles) and INV-42 (filled circles); the two metabolites have low correlation in the wt ($r = 0.2597$) but high in INV-42 ($r = 0.9947$), with $p < 9.3 \times 10^{-4}$ for the comparison. (c) and (d): Ornithine and sucrose in wt (empty triangles) and INV-33 (filled circles); the correlation changes from $r = 0.9875$ in the wt to $r = -0.8965$ in the transgenic ($p < 1.05 \times 10^{-6}$). Note the change in sign in the correlation, indicating a considerably different regulation mode in the transgenic. Plots (a) and (c) represent metabolite levels as determined from GC-MS (corrected for sample size and internal standard). Plots (b) and (d) represent the same data after subtracting the mean and dividing by the standard deviation. In (a) one can easily recognize that the metabolite levels in INV-42 are larger than in the wt, but (b) allows for a better assessment of the low level of correlation in the wt. Both transgenic lines express a yeast invertase, further details in the original publication (data from Roessner *et al*., 2001).

We can write $R_{p_k}^{X_i} = \alpha R_{p_k}^{X_j}$ for all k, and express equation (4) solely in terms of responses of variable j:

$$\frac{\Delta \ln X_i}{\Delta \ln X_j} \approx \frac{\sum_k^n \alpha R_{p_k}^{X_j} \Delta \ln p_k}{\sum_k^n R_{p_k}^{X_j} \Delta \ln p_k} = \alpha. \qquad (8)$$

All observations will be approximately aligned on a line with slope α, equal to all the co-response coefficients, and consequently the correlation will be high. This situation can happen in two different cases: (a) when two metabolites are in (or close to) chemical equilibrium, or (b) when they share a conserved moiety (Hofmeyr *et al*., 1986). In the case of equilibrium α is a positive constant, and in the case of moiety conservation it is negative, with the correlations being close to $+1$ and -1, respectively.

3.1.3. *Different physiological states*

It is important to note that this interpretation of correlation in metabolite scatter plots based on response coefficients is only valid for replicates of a single physiological state. This is due to the fact that response coefficients are linear approximations around a reference state, but biochemical systems are governed by non-linear interactions. If samples from two different physiological states are used, then they will likely not align, resulting in a low correlation, even if in each of the states alone there was high correlation. On the other hand, if the mean value of the concentrations of two states differ by a large amount, then the points in the scatter plot will exist in two clusters wide apart and this may result in a spurious high correlation. This will happen even if there was no correlation in each of the states. Thus, one should be careful to only combine data from a single physiological state.

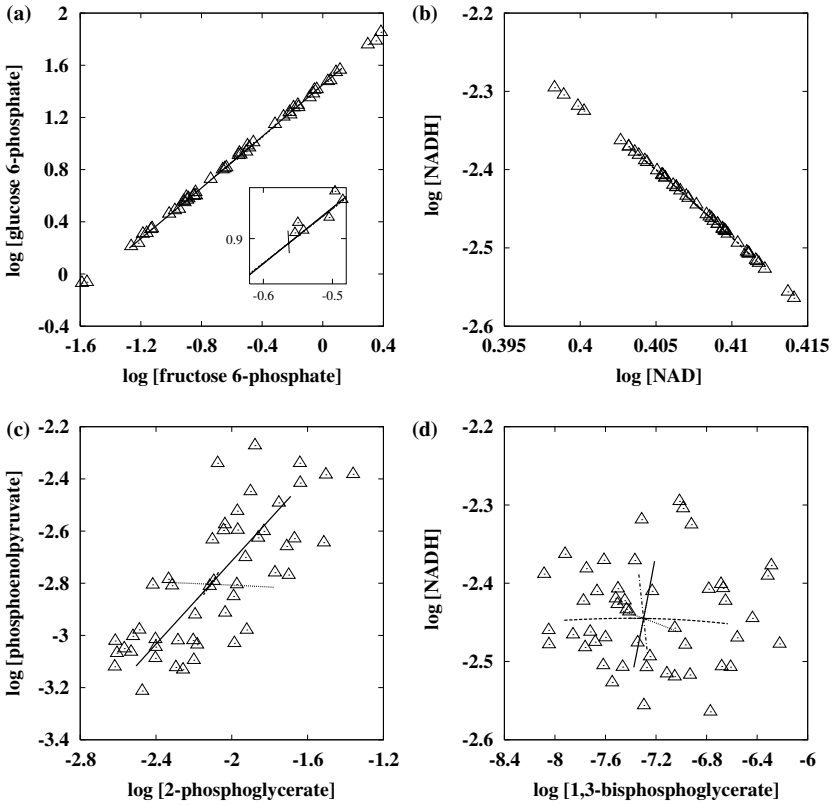

Figure 3. Relationship between metabolite scatter plots and co-response profiles. Triangles correspond to "biological" replicates obtained by simulation of a yeast glycolysis model where each replicate differs from each other by a small random change in all enzyme concentrations (±10%, see Methods). Lines correspond to co-responses towards single enzymes, and were determined by simulation, changing the concentration of each enzyme (±10%) while keeping all others constant. Values plotted are the logarithm (base 10) of the concentrations. Solid lines: hexose transporter (HXT); dashed lines: glycerol 3-phosphate dehydrogenase (G3PDH); dotted lines: enolase (ENO); dashed-dotted lines: alcohol dehydrogenase (ADH). (a): glucose 6-phosphate and fructose 6-phosphate ($r = 0.9988$), which are near equilibrium. Note that the co-response for phosphoglucoisomerase (PGI), their isomerase, is very small but perpendicular to HXT (see inset), all others are superimposed with HXT. (b): NAD^+ and NADH ($r = -0.9997$), which form a moiety conservation cycle. Co-response to all enzymes is overlapping and in the same direction as the scatter. (c): 2-phosphoglycerate and phospho*enol*pyruvate ($r = 0.7741$), that are linked by a single enzyme (ENO). (d): 1,3-bisphosphoglycerate and NADH ($r = -0.0398$), both products of the same enzyme (glyceraldehyde 3-phosphate dehydrogenase, GAPDH).

More relevant than combining samples of different states in a single scatter plot, is to compare scatter plots obtained with each single state. This will be particularly important when in one physiological state there is high correlation, but not in the other, or when the correlations are both high but of opposite sign. Examples of this state-dependent correlation are shown below with simulations. Although it is not usually possible to decompose a metabolite scatter plot into the independent contributions of the parameters (co-response profiles), their interpretation in terms of response and co-response coefficients yields insight into the regulation of the system, since these coefficients are indeed global measures of regulation.

3.2. Scatter plots and correlation

Scatter plots have been traditionally displayed using metabolite levels obtained directly from profiling experiments (Roessner *et al.*, 2001; Fiehn, 2003; Weckwerth *et al.*, 2004). These levels are peak areas corrected by the sample mass and by the area of an internal standard (Roessner *et al.*, 2000). Figure 2a and c provide a demonstration of plots using metabolite levels directly; it should become obvious that these plots are best used to compare the magnitudes of the levels between states and between the two metabolites. These plots can, however, be hard to interpret in terms of correlation, due to the inherent problem of scale. In figure 2a the collection of points in the two states lie along the same regression line, and one could be tempted to think that the two metabolites are correlated in all of the genotypes. As it turns out, in the wild type the correlation among replicates is low (0.26), and it only appears to be in line with the transgenic data because the average level of the two metabolites lies in the same line delineated by the transgenic samples. The reason why this plot is inappropriate to judge correlation is because it is a measure that is dependent of scale. Figure 2b

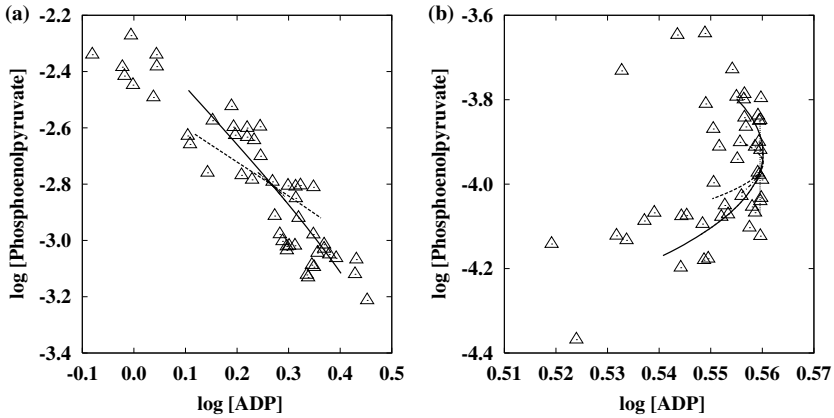

Figure 4. Changes in correlation reflect changes in the co-response profile of the pair ADP and phospho*enol*pyruvate (both are substrates of pyruvate kinase, PYK). Triangles correspond to "biological" replicates, simulated as in figure 3 (see also Methods). Lines correspond to changes in each enzyme, while keeping all others constant. Solid lines: hexose transporter (HXT); dashed lines: ATPase; dotted lines: PYK; dashed-dotted lines: glycogen branch. The plot is of the logarithm (base 10) of the concentrations. A: yeast model with high glucose concentration in the medium (model of Teusink *et al.*, 2000), $r = 0.9079$. B: same yeast model glucose concentration and changed basal concentration of enzymes (see Methods for further details). Rank correlation is now $r = 0.4113$, but note the non-linear nature of the co-response profile. Interestingly PYK, the enzyme that has the two metabolites as substrates, only affects the concentration of phospho*enol*pyruvate.

presents the same data after applying the transformation of equation (9):

$$X_i' = \frac{X_i - \bar{X}}{\sigma_X}, \tag{9}$$

where X_i' are now the values plotted, $\bar{X}$ is the mean value, and σ_X is the standard deviation of the concentration values of metabolite X. After this transformation, all linear relationships have a slope of $+1$ or -1. These plots are adimensional, but their units can be interpreted in terms of standard deviations. This transformation is similar to the process of autoscaling, which is used for normalizing spectra and giving peaks equal "importance". Figures 1, 2b and d demonstrate their use. Note that from figure 2b it is much more obvious that the wild type has lower correlation. Additionally, the plot also makes obvious that the low correlation value is mostly dependent on a single replicate that has a high value of fructose 6-phosphate (2 standard deviations away from the mean). An important point to make, then, is that more data would be required to confirm that these two metabolites have low correlation in the wild type potato tuber. Nevertheless the probability that the true correlation between the two samples of figure 2B is the same is around 10^{-4} (i.e. about 1 in 10,000 such correlation comparisons would fail to identify two equal correlations, and in this data set there are only 4900 comparisons). The idea that these two sugar phosphates are more correlated in the genotype overexpressing an invertase gene than in the wild type does seem to make sense, giving some support to this observation.

In the case when a high correlation originates from the dominance of one parameter, then it would be interesting to construct the plot in log–log space, as this would match the way in which co-response profiles are usually plotted. Below we present simulation results using such log-transformed plots (figures 3 and 4).

A word of caution is warranted for the estimation of correlations from small sample sizes, since correlation is very dependent on sample size (small samples providing bad estimates). The contrast of figure 1a ($n = 43$) and the wild type data in figure 2b ($n = 6$) should make this obvious. A rule of thumb would be that correlation should not be calculated with $n < 10$, though only a few studies (Fiehn, 2003; Weckwerth *et al.*, 2004) present sample sizes larger than this. Another consideration concerns the use of Pearson correlation, as this assumes linearity and will provide bad estimates if the relationship between variables is curved, as is common in biochemical data. Spearman (rank) correlation provides better estimates, as it does not depend on linearity, only on monotonicity, and is more robust towards outliers. Unfortunately Spearman correlation seems even more dependent on large sample sizes than Pearson correlation. The problem of small sample size can be partially overcome if one adopts a tight significance level to comparing correlations. But beware that with sample sizes as small as those of figure 2b, even correlations as large as $r = 0.95$ are most likely not significant. An alternative way to measure "association" could be effected using mutual information, however this is even more sensitive to sample size than rank correlation, and suffers from the problem that data needs to be discretized.

3.3. Simulations

The use of computer simulations complements the analytical results presented above, and enables investigating lines of thought for which it would be hard

to carry out experiments. We use a well-described model of the yeast glycolytic pathway (Teusink *et al.*, 2000) to investigate the relation between co-response and correlation in biological replicates. Figure 3 depicts simulation results corresponding to the four special cases discussed above. Figure 4 depicts an illustration of how correlation between biological replicates reflects changes in overall system regulation.

3.3.1. Equilibrium

Figure 3a corresponds to the case in which two metabolites are near chemical equilibrium (mass-action ratio of 0.20 for an equilibrium constant of 0.29). The majority of the co-response lines caused by each of the enzymes are aligned on top of each other and the "biological" replicates do align over the same line. The protein towards which the response of the concentration of glucose 6-phosphate and fructose 6-phosphate is higher is the hexose transporter (HXT), and is reflected in the plot as the longest of the segments. Interestingly, the co-response of the metabolites towards the enzyme that inter-converts them, phosphoglucoisomerase (PGI, EC 5.3.1.9), is in a direction almost perpendicular to the direction of the HXT co-response and the data points, despite being a weak co-response (the line segment is very short, see zoomed inset of figure 3a). This is in contrast to the explanation provided for the correlation of the same metabolite pair in *Cucurbita maxima* (Fiehn, 2003). Another pair of metabolites that is in equilibrium consists of glyceraldehyde 3-phosphate and glycerone phosphate and produces with even higher correlation than the one in figure 3a (data available on supplementary information). The fructose 6-phosphate:glucose 6-phosphate pair has been observed with very high correlation in a wide variety of metabolomics studies (Roessner *et al.*, 2001; Fiehn, 2003; Weckwerth *et al.*, 2004; Martins *et al.*, 2004; Broeckling *et al.*, 2004). We believe that this correlation happens when there is a high glycolytic flux and the reaction is near equilibrium. For lower glycolytic flux, this correlation seems to be lost, at least in wild type potato tubers (Roessner *et al.*, 2001) and yeast cultures in post-diauxic phase (Martins *et al.*, 2004).

3.3.2. Mass conservation

Mass conservation relations are ubiquitous in metabolism, the most common being represented by moiety-conserved cycles. These cycles are formed by a small number of molecules that carry a common moiety whose degradation or synthesis is much slower than the reactions of the cycle. Examples of such cycles are those formed by the NAD moiety to carry reductive equivalents and by the adenosine moiety to carry free energy. As discussed above, members of these moiety-conserved cycles are expected to be highly correlated, and at least one of them should have negative correlation with the others (due to the mass conservation constraint). In the yeast glycolysis model there are two such cycles: NAD-NADH and AMP-ADP-ATP. Figure 3b displays the scatter plot of the NAD-NADH pair, with a rank correlation of −0.9997. All of the co-response lines are superimposed on the same direction as the replicates, but unlike the previous case, this pair is not in equilibrium (so it is distinct from the previous case). The co-response profile is dominated by the steps that interconvert the NAD moiety between oxidized and reduced forms, namely glycerol 3-phosphate dehydrogenase (G3PDH, EC 1.1.99.5), alcohol dehydrogenase (ADH, EC 1.1.1.1), glyceraldehyde 3-phosphate dehydrogenase (GAPDH, EC 1.2.1.12), and the succinate branch.

It is important to not confuse the profiles of equilibrium and mass conservation: in the former, the correlation is always positive, while in the latter there has to be a negative correlation. And while there may be other reasons for negative correlations other than mass conservations, their presence in a metabolite profile should raise questions if those metabolites are not possibly in a mass conservation relation.

3.3.3. Moderate correlation

We now analyze the case of moderate correlations, i.e. $0.6 < |r| < 0.8$. In our simulation there are many cases of moderate correlation; figure 3c displays the case of 2-phosphoglycerate and phospho*enol*pyruvate, which happen to be linked by the enzyme enolase (ENO, EC 4.2.1.11), but which have only a moderate correlation ($r = 0.7741$, rank). In this case the co-response profile has three enzymes with decreasing strength and different directions: HXT has the strongest effect, followed by G3PDH, and then ENO (the enzyme that links them). The angle formed by HXT and G3PDH is marked and defines the scatter in the "biological" replicates, with the largest scatter along the direction of HXT, and less scatter in the direction of G3PDH. Indeed, the scatter is generated by variance in each of the enzymes, with the direction set by the co-response of the metabolites towards the enzyme, and the amount of scatter proportional to the response of each of them to that enzyme. If a pair of metabolites is controlled essentially by a single enzyme, then their replicates will be correlated no matter how close they maybe in the metabolic map; this correlation would originate from an amplification of the variance for that enzyme. Due to the summation theorem for concentration control, which states that the sum of all concentration control coefficients towards the same metabolite is zero (Heinrich and Rapoport, 1974), for one enzyme to dominate (e.g. 10-fold larger than the others), the others must be small and have opposite sign. We predict that this situation is not common, though possible. Another possibility for the high correlation is for the concentration of one of the enzymes to vary widely between replicates, and this would only happen if

there was no tight control of its expression. In this case, the variance is still due to that enzyme, but not by amplification. A summary of the origin of metabolite correlation among replicates is that it is due to large variation of a single enzyme, or through differential amplification of the variance of a single enzyme. In any case the observed correlations are properties of the whole system, not of any particular metabolite, enzyme, or reaction!

3.3.4. Low correlation

The final special case is the most common: when two metabolites are poorly correlated. Figure 3d displays the pair 1,3-bisphosphoglycerate and NADH, both products of the reaction catalyzed by GAPDH. Despite this metabolic constraint, they display a very low correlation in this model ($r = -0.0398$, rank). The co-response profile has four enzymes with relatively similar strength and with different angles. Accordingly to our hypothesis, the scatter of "biological" replicates is anisotropic resulting in the low correlation. Even though the co-response is equally dominated by four enzymes (HXT, ADH, G3PDH, and ENO), we believe that two of them alone would have been able to generate low correlation, as long as they formed an angle close to 90°, as is the case with the pairs HXT:ADH or G3PDH:ENO. This example strengthens the notion that neighboring metabolites may have little or no correlation – not because they are not related, but because the variance in the enzymes that control them affects them in equal amounts and different directions. Overall, this is what happens to the majority of metabolite pairs, and is a consequence of the systemic nature of metabolic control (Kacser and Burns, 1973). Indeed it is another manifestation of the same principle that results in dominant mutations being rare (Kacser and Burns, 1981).

3.3.5. Comparing correlations

One of the important aspects that can be pursued based on the thesis that correlation of replicates and associated scatter plots reveal aspects of regulation, concerns comparing correlations between two distinct physiological states or genotypes. Early experimental data from potato tubers (Roessner *et al.*, 2001) already displays this phenomenon as is depicted in figure 2, even though not highlighted in the original publication. To demonstrate how these changes relate to the co-response profiles, we now analyze an example from our simulations. Figure 4 depicts the relationships between ADP and phospho*enol*pyruvate in the yeast glycolysis model in two distinct states (see Methods for details). In terms of correlation alone, the pair is moderately correlated in the high glucose state ($r = -0.9079$) but in the low glucose state the correlation decreases considerably ($r = 0.4113$). Inspection of the scatter plots (figure 4) reveals that the co-response profile is dominated by HXT in the high glucose state; in the low glucose state, even though HXT still has the largest magnitude of control, several other enzymes now also have comparable control levels on the two metabolites. The correlation becomes low in the low glucose state due to some spread of the replicates but also to the non-linearity of the relationships (here even rank correlation fails to identify the relation because there is a change in derivative in the curve). The change in regulation between the high and low glucose states (see Methods section) is well reflected in both the scatter plots as well as on the rank correlation. Note that the two correlations are very significantly different ($p < 3 \times 10^{-21}$).

When this principle is applied to different genotypes, it becomes similar to the method that has became known as FANCY (Teusink *et al.*, 1998; Raamsdonk *et al.*, 2001), which relies on similarities in co-response to identify mutants that act on similar parts of metabolism. However, in FANCY one is interested in the actual values of co-response, which would be determined from the mutant phenotypes, while here we are restricted to observing correlations of replicate samples that do originate from biological variation filtered by the complete co-response profile. In this case there is no attempt at determining the co-response profile, which is arguably hard to do experimentally and it is why we have resorted to simulation here. The two approaches are related and could be complementary.

3.6. General guidelines

The mathematical and computational analyses presented above, paired with examples from previously published data, lead us to formulate a few general guidelines towards a better interpretation of metabolite correlations in replicate profiles:

1. The scatter of replicates is defined by the individual co-response profiles of the two metabolites by all of the enzymes in the system. Thus the correlation of metabolites in replicate samples is a systemic property.
2. Correlation should be calculated for replicates of the same state, as they are expected to change with changes in regulation. In order to combine different states for the same correlation calculation one needs to show that they are equivalent (e.g. by ANOVA or *t*-tests).
3. If a metabolite pair has at least two enzymes whose co-responses form an angle close to 90°, and response coefficients of similar magnitude, then they will have low correlation. This happens even if the metabolites are directly interacting in the metabolic network.
4. Metabolites in chemical equilibrium will have nearly perfect positive correlation. As a consequence, metabolites with negative correlation are *not* in equilibrium. In this case, the correlation does not

originate from the enzyme that catalyzes the equilibrium reaction, as the metabolites have very small response towards it.
5. At least two metabolites that belong to a moiety-conserved cycle, or other mass conservation relations, will have a negative correlation.
6. When two metabolites are moderately correlated (or more), it may be due to large concentration response coefficient towards a common enzyme, or an enzyme that has carries unusually high variance. But to identify this enzyme requires further data.
7. Changes in correlation between sets of replicates in different physiological states indicate changes in the regulation of the metabolites in question.
8. Metabolite correlations should be determined with large samples (e.g. larger than $n=10$) and using Spearman rank correlation. In case there are fewer samples, then Pearson correlation should be used but only extremely strong correlations should be trusted. The significance level (p value) of the correlation estimates should be explicitly calculated and used as a guide to eliminate unreliable estimates (we suggest $p < 10^{-4}$ or lower).
9. Metabolite scatter plots should be examined both in terms of direct concentration values (or its log–log variant) *and* normalized (i.e. mean subtracted and divided by standard deviation). Correlation should only be judged visually through normalized plots, while linear or log–log plots are more useful to judge relative variation of each metabolite.

4. Concluding remarks

Measures of correlation between metabolites in replicate profiles can be very informative about the underlying biological system. An earlier study established that metabolite correlations do not necessarily correspond to proximity in the biochemical network (Steuer *et al.*, 2003). This analysis described four different regulatory configurations that are expected to be the origin of metabolite correlation. Simulations suggest that when the correlations are very strong, they are likely due to chemical equilibrium. An interesting prediction, still to be confirmed, is that metabolites sharing conserved moieties should have high correlations, and at least one of them being negatively correlated with the others. But most high correlations may be due to either (1) stronger mutual control by a single enzyme, or (2) variation of a single enzyme level much above others. In both cases it is impossible to identify the responsible enzyme from these data alone, though hints can be obtained from the set of metabolites forming correlation cliques (Kose *et al.*, 2001). Ultimately, further data is required for resolve the responsible regulatory events with protein profiles being the most promising for this effect. It is hoped that the concepts introduced here will enable better analysis of metabolomics data in the context of systems biology.

Acknowledgments

We are grateful to Bharat Mehrotra, Lloyd Sumner, Corey Brockling, and Ralf Steuer for helpful discussions. We thank Ute Roessner-Tunali and Oliver Fiehn for providing previously published data in suitable formats. This work was financially supported by grants DBI-109732 from the National Science Foundation, and R01 GM068947 from the National Institute for General Medical Sciences.

References

Atkinson, D.E. (1977). *Cellular Energy Metabolism and its Regulation* Academic Press, New York.

Baggett, B.R., Cooper, J.D., Hogan, E.T., Carper, J., Paiva, N.L. and Smith, J.T. (2002). Profiling isoflavonoids found in legume root extracts using capillary electrophoresis. *Electrophoresis* **23**, 1642–1651.

Broeckling, C.D., Huhman, D.V., Farag, M., *et al.* (2004). Metabolic profiling of Medicago truncatula cell cultures reveals effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* in press..

Cho, S., Park, S.G., Lee do, H. and Park, B.C. (2004). Protein-protein interaction networks: from interactions to networks. *J. Biochem. Mol. Biol.* **37**, 45–52.

Fell, D.A.(1996). *Understanding the Control of Metabolism*. Portland Press, London.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genom.* **2**, 155–168.

Fiehn, O. (2003). Metabolic networks of Cucurbita maxima phloem. *Phytochemistry* **62**, 875–886.

Harrigan, G.G., LaPlante, R.H., Cosma, G.N., *et al.* (2004). Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicol. Lett.* **146**, 197–205.

Hayashi, K. and Sakamoto, N. (1986). *Dynamic Analysis of Enzyme Systems. An Introduction*. Springer-Verlag, Berlin.

Heinrich, R. and Rapoport, T.A. (1974). A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* **42**, 89–95.

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

Hofmeyr, J.H. and Cornish-Bowden, A. (1996). Co-response analysis: a new experimental strategy for metabolic control analysis. *J. Theor. Biol.* **182**, 371–380.

Hofmeyr, J.H.S., Cornish-Bowden, A. and Rohwer, J.M. (1993). Taking enzyme kinetics out of control – putting control into regulation. *Eur. J. Biochem.* **212**, 833–837.

Hofmeyr, J.H., Kacser, H. and van der Merwe, K.J. (1986). Metabolic control analysis of moiety-conserved cycles. *Eur. J. Biochem.* **155**, 631–641.

Hucka, M., Finney, A., Sauro, H.M., *et al.* (2003). The systems biology markup language (SBML) a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.

Janin, J. and Seraphin, B. (2003). Genome-wide studies of protein-protein interaction. *Curr. Opin. Struct. Biol.* **13**, 383–388.

Kacser, H. and Burns, J.A. (1973). The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104.

Kacser, H. and Burns, J.A. (1981). The molecular basis of dominance. *Genetics* **97**, 639–666.

Kell, D.B. (2002). Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* **29**, 237–241.

Kitano, H. (2002). Computational systems biology. *Nature* **420**, 206–210.

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

Liang, M., Cowley, A.W. and Greene, A.S. (2004). High throughput gene expression profiling: a molecular approach to integrative physiology. *J. Physiol.* **554**, 22–30.

Martins, A.M., Camacho, D., Shuman, J., Sha, W., Mendes, P. and Shulaev, V. (2004) A systems biology study of two distinct growth phases of *Saccharomyces cerevisiae* cultures. *Curr. Genomics*. in press.

Mendes, P. (1993). GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.* **9**, 563–571.

Mendes, P. (1997). Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**, 361–363.

Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181–1189.

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

Ott, K.H., Aranibar, N., Singh, B. and Stockton, G.W. (2003). Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **62**, 971–985.

Pritchard, L. and Kell, D.B. (2002). Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. *Eur. J. Biochem.* **269**, 3894–3904.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., *et al.* (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnol.* **19**, 45–50.

Reo, N.V. (2002). NMR-based metabolomics. *Drugs Chem. Toxicol.* **25**, 375–382.

Righetti, P.G., Campostrini, N., Pascali, J., Hamdan, M. and Astner, H. (2004). Quantitative proteomics: a review of different methodologies. *Eur. J. Mass Spectrom.* **10**, 335–348.

Roessner, U., Luedemann, A., Brust, D., *et al.* (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29.

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

Savageau, M.A. (1976). *Biochemical Systems Analysis*. Addison-Wesley, Reading, MA.

Schilling, C.H., Schuster, S., Palsson, B.O. and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* **15**, 296–303.

Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M. and Nishioka, T. (2002). Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.* **74**, 2233–2239.

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

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochem.* **62**, 817–836.

Teusink, B., Baganz, F., Westerhoff, H.V. and Oliver, S.G. (1998). Metabolic control analysis as a tool in the elucidation of the function of novel genes. *Meth. Microbiol.* **26**, 297–336.

Teusink, B., Passarge, J., Reijenga, C.A., *et al.* (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* **267**, 5313–5329.

Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N. and Fiehn, O. (2003). Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* **75**, 6737–6740.

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

Weckwerth, W., Loureiro, M.E., Wenzel, K. and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA* **101**, 7809–7814.

Westerhoff, H.V. and Chen, Y.-D. (1984). How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. *Eur. J. Biochem.* **142**, 425–430.