BIOINFORMATICS

Vol. 27 ISMB 2011, pages i357–i365
doi:10.1093/bioinformatics/btr231

Detection and interpretation of metabolite–transcript coresponses using combined profiling data

Henning Redestig[1,*] and Ivan G. Costa[2,*]

[1]RIKEN Plant Science Center, Yokohama, Japan and [2]Center of Informatics, Federal University of Pernambuco, Recife, Brazil

ABSTRACT

Motivation: Studying the interplay between gene expression and metabolite levels can yield important information on the physiology of stress responses and adaptation strategies. Performing transcriptomics and metabolomics in parallel during time-series experiments represents a systematic way to gain such information. Several combined profiling datasets have been added to the public domain and they form a valuable resource for hypothesis generating studies. Unfortunately, detecting coresponses between transcript levels and metabolite abundances is non-trivial: they cannot be assumed to overlap directly with underlying biochemical pathways and they may be subject to time delays and obscured by considerable noise.

Results: Our aim was to predict pathway comemberships between metabolites and genes based on their coresponses to applied stress. We found that in the presence of strong noise and time-shifted responses, a hidden Markov model-based similarity outperforms the simpler Pearson correlation but performs comparably or worse in their absence. Therefore, we propose a supervised method that applies pathway information to summarize similarity statistics to a consensus statistic that is more informative than any of the single measures. Using four combined profiling datasets, we show that comembership between metabolites and genes can be predicted for numerous KEGG pathways; this opens opportunities for the detection of transcriptionally regulated pathways and novel metabolically related genes.

Availability: A command-line software tool is available at http://www.cin.ufpe.br/~igcf/Metabolites.

Contact: henning@psc.riken.jp; igcf@cin.ufpe.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Changes in metabolite abundances are the first molecular events that follow shifts in an organism's environment. As the steady state of metabolism is broken, gene expression changes to reestablish homeostatic control and to adapt the system to the altered living conditions. This interlocking regulation between gene expression and metabolite levels manifests in intricate patterns of coresponses where metabolite and transcript levels alternately vary consistently with the behavior of both regulators and targets (Kresnowati *et al.*, 2006).

Therefore, the elucidation of metabolite and transcript coresponses can yield important clues to the consequences of altered environmental conditions on the biochemical level as well as to the organism's coping strategies. From the perspective of biotechnology, knowing which genes affect metabolite levels is crucial for metabolic engineering since metabolism cannot, whereas gene expression can, be altered directly. Consequently, many studies using metabolomics and transcriptomics in parallel have been performed to monitor development (e.g. Carrari *et al.*, 2006; Urbanczyk-Wochniak *et al.*, 2003) and stress responses (e.g. Dutta *et al.*, 2009; Hirai *et al.*, 2005; Kaplan *et al.*, 2007) in time-series experiments. The datasets from this type of experiments are now a valuable resource with the potential to fill similar needs as the widely applied gene coexpression databases (e.g. Obayashi *et al.*, 2007) for hypothesis-generating research (Saito *et al.*, 2008).

Transcript–transcript (TT) correlations that derive from genes coding for enzymes that catalyze nearby reactions are stronger than correlations that derive from genes that catalyze metabolically distant reactions (Kharchenko *et al.*, 2005; Walther *et al.*, 2010). When the system is not at steady state, abundances of neighboring metabolites also tend to be correlated; this observation was used to approximate the glycolysis pathway from metabolite abundance measurements (Arkin *et al.*, 1997). The inverse relationship between metabolite–metabolite (MM) correlations versus pathway distance was recently shown to manifest also on the omics level during heat and cold stress responses in *Escherichia coli* (Walther *et al.*, 2010).

Although interpreting MM correlations is far from straightforward (Steuer *et al.*, 2003), metabolite–transcript (MT) correlations are arguably even more multifaceted. First, since enzymes influence the reaction speed, metabolically driven correlations are more intuitive between transcripts and metabolite fluxes than for the more easily measured abundances. Second, correlations are not expected for metabolites that are rapidly consumed by subsequent reactions. Third, other modes such as post-translational modification appear to be more common for regulating enzyme activity than transcription (Carrari *et al.*, 2006; Gibon *et al.*, 2006). However, there are many other reasons for correlations between metabolites and transcripts that do not derive from the underlying biochemical pathway. In a 2006 review, Ladurner described several mechanisms in which metabolites can tune gene expression; they include riboswitches, direct interactions with transcription factors, regulation of cofactors, chromatin remodeling, chromatin modification and hormone signaling. In plants, the last is arguably the best known since hormones play central roles in nearly all stress responses and developmental programs.

With these considerations in mind, it becomes clear that MT correlations may take many forms exhibiting various degrees of noise, time lags and conditionality on the studied response (see Fig. 1 for examples). Therefore, to detect such coresponses using

*To whom correspondence should be addressed.

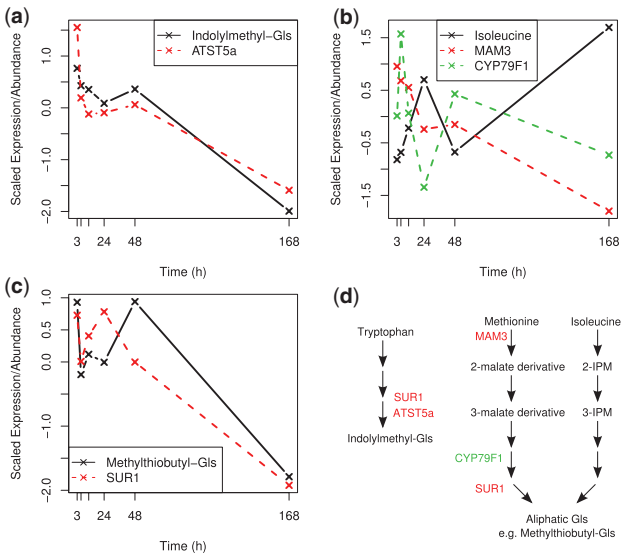

Fig. 1. Three examples of MT coresponse patterns seen in the glucosinolate (Gls) synthesis pathway during sulfur deficiency [data from Hirai *et al.* (2005)]. (**a**) ATST5a catalyzes a late step in the synthesis of Indolylmethyl-Gls and its transcript is positively correlated with the reaction product. (**b**) CYP7F1 and MAM3 catalyzes aliphatic Gls synthesis from methionine that appear anticorrelated with isoleucine, also a precursor for aliphatic Gls via a pathway that shares enzymes with the methionine → aliphatic Gls pathway. (**c**) SUR1 catalyzes a later step in aliphatic Gls synthesis and appear correlated with methylthiobutyl Gls at a positive time lag.

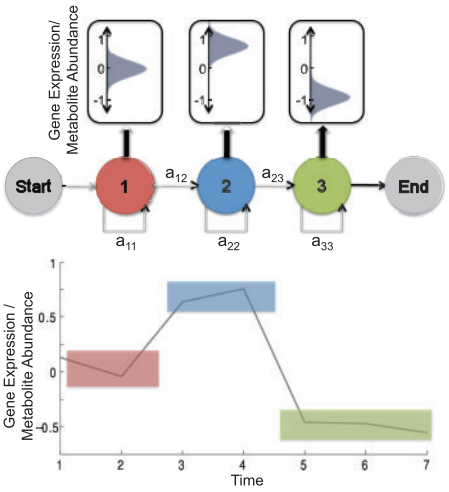

Fig. 2. Examples of three-state HMM time courses with an up and down gene expression/metabolite abundance pattern (top). The state mean and the transitions determine the expected expression/abundance value for a particular time interval. For example, we depict in the bottom a time series with high similarity to the example HMM. The colored boxes correspond to the most probable expression/abundance intervals defined for each of the three states.

combined profiling data, methods that can handle these issues and deliver interpretable results are needed.

1.1 Previous approaches

The most commonly applied method has been to rank all transcripts according to their Pearson correlation with different metabolites (Urbanczyk-Wochniak *et al.*, 2003). Hirai *et al.* (2005) coclustered transcripts and metabolites and identified relevant patterns by manually inspecting the resulting clusters. Bradley *et al.* (2009) used a Bayes net to calculate an association statistic by conditioning the Pearson correlation on four broad classes of metabolites and the studied stress response. Although it has proven highly useful, the Pearson correlation has the considerable disadvantage that it is sensitive to noise and that it cannot detect time-lagged responses.

To cope with delayed responses, lagged covariances/Pearson correlations have been applied to study the heat and cold stress responses in *Saccharomyces cerevisiae* (Walther *et al.*, 2010) and *E.coli* (Takahashi *et al.*, 2011) and for predicting transcription factor targets in *Arabidopsis thaliana* (Redestig *et al.*, 2007). However, the introduction of the lags truncates the already typically short time-series and may exacerbate the problem with noise sensitivity.

Another serious problem with the direct study of pairwise MT correlations (regardless of how they are calculated) is that since data are scarce, the large number of false-positives renders interpretation very difficult. Even with a very good measure, finding relevant patterns among millions of correlations is a staggering task, especially when done in a completely unsupervised fashion. One approach that addresses this issue is to use multivariate regression via for example O2PLS (Bylesjö *et al.*, 2007) or PLS (Pir *et al.*, 2006). These methods relate the entire transcriptomics and metabolomics data blocks with each other in a single model; this is useful for interpreting large-scale trends but cannot easily be used for identifying specific MT coresponses.

1.2 Our approach

In this study, we aimed to design a general method for detecting MT coresponses that addresses these issues. Specifically, we demanded that the method should (i) be able to work with time-series data with few time-points, that it be robust to (ii) noise and (iii) time lags and (iv) that it deliver directly interpretable results to facilitate hypothesis generation. To this end, we introduced two novel methodological developments: (i) a hidden Markov model-based similarity measure and (ii) an approach that summarizes several similarity statistics on the pathway level.

1.2.1 Hidden Markov Model based similarity We propose here a novel Hidden Markov Model (HMM)-based similarity for accessing the temporal coresponse of MT pairs. We used a particular linear HMM topology defined by Schliep *et al.* (2003, 2005). Such models have been successfully applied in distinct gene expression time course applications such as querying (Schliep *et al.*, 2003), model-based clustering (Costa *et al.*, 2005; Schliep *et al.*, 2005) and classifying treatment responses (Costa *et al.*, 2009; Hafemeister *et al.*, 2011). The linear HMM can be interpreted as a segmentation method, where each state defines an expression range that a time-series follows, e.g. low or high expression, during a particular time interval. For example, the model in Figure 2 defines a prototypical up and down expression behavior. Given its stochastic nature, the linear HMMs have been shown to be robust to modeling noisy and lagged time courses (Costa *et al.*, 2009; Schliep *et al.*, 2005).

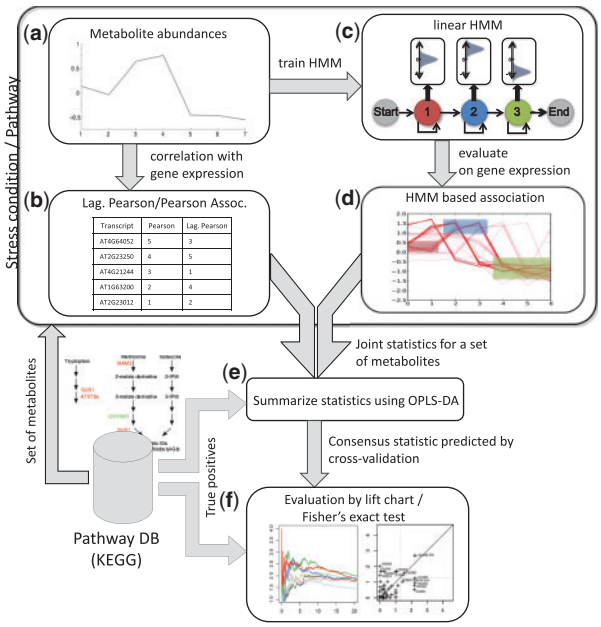

Fig. 3. Schematic of the proposed methodology. For the metabolites in a given pathway (Step a), we calculate their (direct and lagged) Pearson correlation with the expression of all genes (Step b). Additionally, we train a linear HMM for each metabolite (Step c) and evaluate gene expression in the obtained models (Step d). Steps (a–d) are repeated for all metabolites in a given pathway for a particular dataset. To combine multiple coresponse statistics, we perform OPLS-DA using a set of true positives from a pathway database (Step e). Finally, we evaluate the results using an enrichment test and lift-charts on consensus statistics predicted during cross-validation (Step f).

1.2.2 Pathway-based coresponse summarization Considering the complex interdependency between metabolites and transcripts, we reasoned that regulatory patterns may be easier to detect and interpret on the level of isolated pathways such as glycolysis or the tricarboxylic acid cycle than when considering individual MT pairs. Therefore, we propose the use of orthogonal projections to latent structures discriminant analysis (OPLS-DA) (Bylesjö *et al.*, 2006) to combine the similarity statistics for metabolites in a particular pathway. Moreover, this approach can be used to combine different types of similarity statistics. The results can be used to rank the transcripts after determining the strength of their association with a given pathway based on their coresponses with metabolites in that pathway.

We depict in Figure 3 the method schematic. For a metabolite's abundance profile at a particular stress condition (a) we apply traditional approaches by measuring the Pearson and lagged Pearson correlation with all transcripts (b). Additionally, we train one HMM to capture the metabolite's prototypical temporal abundance (c). Next we evaluate the likelihood of all gene expression time-courses with each HMM (d). A high likelihood indicates a high similarity/coresponse between the metabolite and the transcript. Note that the HMM also assigns high scores to time courses with similar but lagged expression patterns. Steps (a)–(d) are repeated for all metabolites measured in a particular pathway/stress condition. Next, we use OPLS-DA to summarize the statistics for all metabolites in the corresponding pathway using the annotated transcripts as true positives (Step e). To evaluate the found coresponses, we use both Fisher's exact test on pathway enrichments and lift charts on consensus statistics calculated in cross-validation fashion.

We applied the proposed methodology to study responses to sulfur deficiency (root and leaf tissues) (Hirai *et al.*, 2003), cold stress (Kaplan *et al.*, 2004) and elevated CO_2 stress (Dutta *et al.*, 2009). We also made use of simulated data to evaluate the individual similarity statistics under controlled conditions.

2 METHODS

2.1 Notation

- $g \in \{1,...,N\}$: be a running index for the genes,
- $m \in \{1,...,M\}$: be a running index for the metabolites,
- $t \in \{1,...,T\}$: denote the time points,
- $X \in \mathbb{R}^{N,T}$: expression matrix of all genes and time-points in a particular treatment/stress,
- $Y \in \mathbb{R}^{M,T}$: abundance matrix of all metabolites and time-points in a particular treatment/stress,
- $x_g \in \mathbb{R}^T$: expression time course of gene g,
- $y_m \in \mathbb{R}^T$: abundance time course of metabolite m,
- $x_{g,t} \in \mathbb{R}$: expression of gene g at time t,
- $y_{m,t} \in \mathbb{R}$: abundance of metabolite m at time t.

2.2 Detection of metabolite–transcript coresponses

The application problem concerns scoring the coresponse between a pair of transcripts and metabolites, (x_g, y_m) in a given stress response. For this, we use a series of statistics that measure the similarity between temporal expression/abundances. Moreover, as distinct similarity statistics measure distinct similarity properties of a pair of time courses, we investigate the use of either individual statistics or their combination, as described below.

2.2.1 Pearson correlation For a pair of gene expression and metabolite abundance time courses (x_g, y_m), the Pearson correlation can be estimated as

$$\mathrm{PC}(x_g, y_m) = \frac{\sum_{t=1}^{T}(x_{g,t}-\mu_{x_g})(y_{m,t}-\mu_{y_m})}{(n-1)S_{x_g}S_{y_m}} \quad (1)$$

where $\mu_{x_g} = \sum_{t=1}^{T} x_{g,t}/T$, $\mu_{y_m} = \sum_{t=1}^{T} y_{m,t}/T$, $S_{x_g} = \sum_{t=1}^{T}(x_{g,t}-\mu_{x_g})^2/T$ and $S_{y_m} = \sum_{t=1}^{T}(y_{m,t}-\mu_{y_m})^2/T$. $\mathrm{PC}(x_g, y_m)$ will have values in the range $[-1, 1]$, where 0 indicates no correlation, 1 a perfect positive correlation and -1 a perfect negative correlation.

2.2.2 Lagged Pearson correlation The lagged Pearson correlation can be estimated by introducing shifts in the original time courses and estimating the Pearson correlation with the above-defined formula. More formally, the shifted pair $(x_g^{(l)}, y_m^{(l)})$ with a lag l is obtained for positive lags as $y_m^{(l)} = \{y_{m,l+1},...,y_{m,T}\}$ and $x_g^{(l)} = \{x_{m,1},...,x_{m,T-l}\}$ and for negative lags as $y_g^{(l)} = \{y_{m,1},...,y_{m,T-l}\}$ and $x_m^{(l)} = \{x_{m,l+1},...,x_{m,T}\}$. The lag Pearson correlation (LPC) is estimated as

$$\mathrm{LPC}(x_g, y_m) = \max_{l=\{\lfloor -t/2 \rfloor,...,\lfloor t/2 \rfloor\}} \mathrm{PC}(x_g^{(l)}, y_m^{(l)}). \quad (2)$$

Correspondingly, one can estimate the optimal time lag as

$$\mathrm{Lag\ LPC}(x_g, y_m) = \operatorname*{arg\,max}_{l=\{\lfloor -t/2 \rfloor,...,\lfloor t/2 \rfloor\}} \mathrm{PC}(x_g^{(l)}, y_m^{(l)}). \quad (3)$$

2.2.3 Linear HMMs An HMM is a probabilistic function of a Markov Chain. It is defined by a set of states S, the transition probabilities a_{ij} for moving from state i to j and the emission densities f_i attached to the i-th state (see Rabiner, 1989, for details). We use HMM with a linear topology, i.e. each

state has a self-transition and a transition to the next state (Fig. 2). The model also has a start and end state to force the time-courses to visit all states at least once. The emission densities are based on univariate Gaussians with free parameters μ_i, σ_i. To support missing values and noise, we use a mixture model as the emission function (see Costa *et al.*, 2009; Schliep *et al.*, 2005, for details). A HMM is parametrized as $\lambda = (A, (\mu_1 \ldots, \mu_S), (\sigma_1, \ldots, \sigma_S))$, where A is a $S \times S$ matrix with entry a_{ij} defining the transition from state i to state j. The linear HMM can be interpreted as a segmentation method, where each state defines an expression range that a time-series follow, e.g. low or high expression, during a particular time interval. Here, the intensity of the expression is parametrized by the mean value of the emission density μ_s and the length of the interval is parametrized by the transition probability s_{ii}. For example, the model in Figure 2 defines a prototypical up and down expression behavior.

2.2.4 HMM-based similarity To obtain an HMM-based similarity for a pair (y_m, x_g), we perform the following procedure. For a given metabolite m, we estimate an HMM λ_m with the Baum–Welch algorithm using y_m as the training data. Next, we use the forward algorithm to obtain the likelihood over observation x_g under the model λ_m (Rabiner, 1989). The likelihood reflects the similarity between the HMM trained with y_m and that evaluated with x_g, that is

$$\mathrm{HMM}(x_g, y_m) = \mathbb{P}(x_g | \lambda_m). \quad (4)$$

This statistic will have high values for time-series with high similarities and low values for time courses with no similarity. Note that the Pearson correlation also captures negative correlation, i.e. time courses that have a similar but inverted pattern of expression. Given that the time-series are normalized, i.e. mean zero and SD 1, an HMM-based inverted similarity can be obtained by inverting the signal of the gene time course before evaluating Equation (4). We will call the normal and inverted similarities HMM^+ and HMM^-, respectively.

Overall, the only parameter to be defined by the user is the number of states. For our previous applications, we chose the number of states to be at maximum 40% of the number of time points. Given the short duration of the time courses analyzed in this work, we only consider HMMs with two or three states, called HMM2 and HMM3, respectively. We also performed analysis for four states (Supplementary Figs. S1–S3 and S8); however, the results were either worse or similar to HMM3.

2.3 Combining similarity statistics

We summarize the coresponse statistics for all metabolites in a given pathway, p, to obtain a single statistic, b, that measures the coresponse between each transcript and that pathway in general. For this purpose, we chose OPLS-DA as it sets no constraints regarding colinearity or the distribution of the input data (Bylesjö *et al.*, 2006). Furthermore, a wide range of diagnostic tools and plots exist for this type of models and this facilitates interpretation. Let Z be the scaled and centered $g \times m \in p$ matrix with the similarity statistics for the metabolites in p. Further, let $\vec{v}$ be a column vector with ones for the transcripts that are annotated to p and zeroes for those that are not (as defined by a pathway database such as KEGG). Then we use OPLS-DA to decompose Z into a $\vec{v}$-correlated part and a $\vec{v}$-uncorrelated part by:

$$Z = \vec{b}\vec{w}^T + B_O P_O + E, \quad (5)$$

where $B_O P_O$ approximates the $\vec{v}$-uncorrelated variance in Z and E is the residual. Our consensus statistic is given by the $\vec{v}$-correlated score vector $\vec{b}$. The model is predictive and $\vec{b}$ can be estimated for a new Z as well. We estimate the complexity of the model (number columns in P_O) by 10-fold class-balanced cross-validation and use separate round of cross-validation to obtain independent observations of $\vec{b}$ to use for evaluation purposes.

The same approach can be used to summarize any number of statistics by simply augmenting Z with other coresponse statistics. Thereby, we provide a general way to combine complementary methods for scoring MT coresponses. In particular, we use this approach for combining the HMM^+ statistics with the inverted similarities from HMM^- or for combining Pearson with HMM.

2.4 Datasets

2.4.1 Arabidopsis stress response data Four public combined profiling datasets were used to evaluate our approach. The 'sulfur root' and 'sulfur leaf' were introduced by Hirai *et al.* (2004, 2005) for studying the transcriptome and metabolome in leaves and roots during sulfur deficiency stress. The transcriptomics data were obtained from custom Agilent arrays and the metabolomics data were obtained by untargeted FTMS and targeted high performance liquid chromatography (HPLC) and capillary electrophoresis (CE) analysis [42 (root) and 28 (leaf) annotated metabolites and 7342 transcript probes after filtering, 6 time-points in both datasets]. The third dataset is from a study of cold stress response reported by Kaplan *et al.* (2007) who used Affymetrix ATH1 array and GC-TOF/MS analysis, respectively (87 annotated metabolites, 6680 filtered transcript probesets, 7 time-points). The fourth dataset comes from a study of the response to elevated CO_2 concentration in leaves by Dutta *et al.* (2009); it was performed using TIGR arrays and GC-TOF/MS (76 annotated metabolites and 7138 filtered transcript probesets, 9 time-points). All metabolites and transcripts were standardized to zero-mean and unit-variance. Metabolite identifiers were unified using the MetMask tool (Redestig *et al.*, 2010).

2.4.2 Simulated data We resort to simulated data to investigate the performance of the similarity statistics under controlled conditions. We use the impulse model (Chechik and Koller, 2009) as a smooth function of the gene expression/metabolite abundance response to stress/treatment conditions. For a particular simulated stress/treatment condition, we define an impulse model for each metabolite. We sample a time course from this function at regular intervals to represent metabolite abundance. Gene expression, associated with this metabolite, was sampled from the same function (50 samples). In this case, we also allowed for time lags, that is, observations can be drawn at previous or later time points than the original metabolite time course. Noise sampled from a normal distribution with mean zero and σ_{noise} is added independently to each time point from all time courses. We repeat the same procedure for five distinct impulse models representing five distinct metabolites. We allow time lags from the range $[-2, -1, 0, 1, 2]$; they are chosen with probability $[0.1, 0.2, 0.4, 0.2, 0, 1]$. At the end, we obtain for each simulated datum the temporal abundance of 5 metabolites and the expression of 250 genes. As the gold standard for evaluation, gene–metabolite pairs are associated if they were sampled from the same impulse model. To investigate how the distinct method works in distinct conditions, we changed the random noise parameter σ_{noise} from 0.1 to 0.5 and the number of time points from 7 to 14. For each parameter selection, we generate 10 datasets. All software and processed datasets can be found at http://www.cin.ufpe.br/~igcf/Metabolites.

3 RESULTS

3.1 Evaluation of individual statistics on simulated data

To evaluate the individual similarity statistics under controlled conditions, we applied HMM, Pearson and lagged Pearson approaches on simulated data. We are particularly interested in evaluating the methods under distinct amounts of injected noise, numbers of time points and the presence of time lags. If we consider only time courses without lag, the Pearson correlation is overall the best method (Supplementary Fig. S1). Nevertheless, for the most degenerate scenario—high noise and few time points—the distinct HMM topologies display similar performance as the Pearson correlation. In the presence of time courses with time lag (Fig. 4), lagged Pearson performs best in the presence of few time points and low noise, but rather poorly under other conditions. The

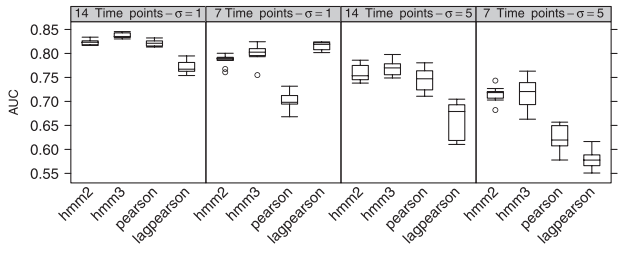

Fig. 4. Boxplots with area under the curve (AUC) estimates for the prediction of true MT coresponses on simulated data using HMMs, Pearson and lagged Pearson. Experiments were based on time courses with 7 or 14 time points and low- ($\sigma_{\text{noise}} = 0.1$) or high-injected noise ($\sigma_{\text{noise}} = 0.5$).

Pearson correlation performs best when the number of time points is equal to 14, or in the lower values of gene calls (Supplementary Figs. S1–S3). This is mostly because of the deficiency of Pearson in the detection of the coresponse of pairs that contains lags in the time courses (see Supplementary Fig. S2 for evaluation of only lagged time courses). The two HMM topologies present consistent results over all scenarios. In particular, they outperform Pearson and lagged Pearson correlations in the scenario of high injected noise and few time points (Fig. 4, right-most panel). Note that few time points and the presence of high amounts of noise are characteristic of the majority of actual experimental data. These results reinforce the importance of the HMM-based distance for detecting coresponses in noisy, undersampled time courses with time-lagged MT coresponses.

3.2 Unsupervised prediction of specific metabolite–transcript coresponses

A good measure for coresponses should give higher values for truly associated than for unrelated MT pairs. A direct way to determine whether a coresponse measure lives up to this expectation is to compare the values for close neighbors in the underlying metabolic pathways with values for distant pairs. To this end, we generated an undirected aggregated bipartite graph for all *A.thaliana*-specific KEGG pathways by connecting metabolites to enzymes that catalyze their formation or breakdown. Similar to the methodology of Walther *et al.* (2010), we used this graph to classify TT, MM and MT pairs at a maximum distance of three steps as cognate pairs and metabolites at a minimum distance of eight steps as non-cognates. We then evaluate the performance of the different coresponse measures for discriminating cognate from non-cognate pairs by calculating the area under the ROC curve (AUC) (Fig. 5). Notably, with all AUC values close to 0.5 as expected from a random classifier for all MT pairs, none appeared informative for this classification, indicating that expression/abundance profiles contain little information about immediate associations in the metabolic reaction graph.

3.3 Supervised prediction of pathway level metabolite–transcript coresponses

The above results confirm the complex interdependency and difficulty in detecting MT coresponses. We investigated whether regulatory patterns may be easier to detect and interpret on the level of whole pathways than when considering individual MT pairs. Using the summarization approach introduced in Section 2.3, we therefore calculate consensus statistics for each transcript and KEGG

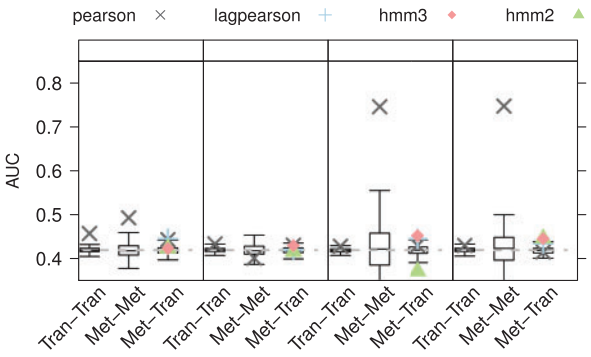

Fig. 5. AUC for the classification of cognate pairs of TT, MM and MT (maximum three steps in the metabolic pathway) from non-cognate pairs (minimum of eight steps). The box shows the expected distribution of the AUC under H_0 as estimated by randomization and the dot indicates the observed AUC. While Pearson correlation shows higher coresponses between MM pairs in CO_2 and sulfur; and with TT pairs in CO_2, none of the considered that MT coresponse measures are informative for performing this classification.

pathway using each of the considered coresponse measures or their combination. As a result, we obtain ranks of the transcripts according to how strongly associated they are with the different pathways given their metabolite coresponse patterns. It should be noted that summarized statistics can only be calculated for pathway-level coresponses since the computation requires a set of true positives. All statistics were calculated in cross-validation fashion to minimize the generalization error.

To evaluate these rankings, we first used the Fisher's exact test to obtain a list of the pathways for which a significant enrichment of truly associated transcripts was seen among the top 1-, 3-, 5- and 10% transcripts for each method and dataset. Of the 119 pathways (and pathway maps), 50 were significantly enriched for at least one method and dataset at the 10% cut-off level.

To visualize the enrichment at all possible cutoffs for these pathways, we use lift charts. The lift value is defined as the observed ratio of true positives divided by the total ratio of true positives (expected invariant over the cut-off rate for a random classifier). We use lift values since they emphasize performance in lower positive call rates; this is important in this application. Furthermore, because they indicate the relative benefit, they can be compared across different pathways even though the pathways contain different numbers of genes and we therefore summarize them by averaging (Fig. 6). In the sulfur root dataset, HMM2 and HMM3 performed slightly better than Pearson which in turn outperformed lagged Pearson. All combined approaches performed better than their individual counterparts and the best results were obtained with HMM3+Pearson. For the sulfur leaf dataset, the Pearson correlation gives the worst- and HMM2+Pearson the best result. On the cold stress dataset, the Pearson correlation performed comparable to HMM2+Pearson and better than the other HMM-based measures. On the CO_2 dataset, all the combined methods performed comparably with a slight preference for HMM2+Pearson and HMM3+Pearson. Of the individual methods, HMM2 was overall best but it notably underperformed on the cold stress dataset. In general, the Pearson correlation and HMM-based methods appeared complementary and provided the overall best performance when used together.

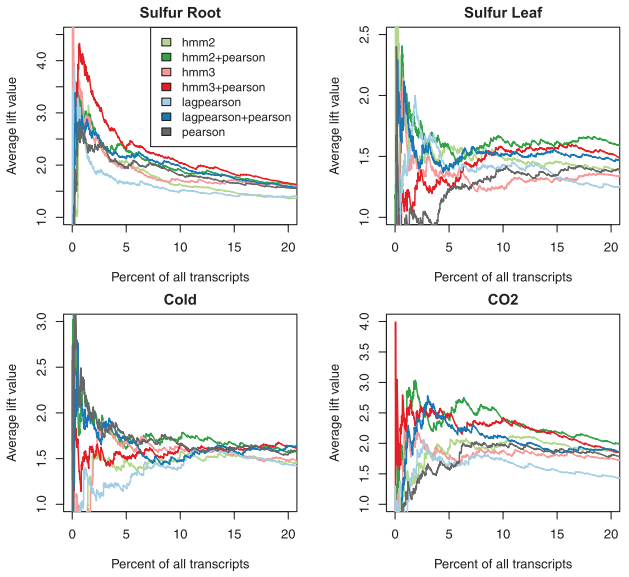

Fig. 6. Average lift charts for classifying transcripts to 50 different KEGG pathways. The lift value indicates how many times better the actual coresponse statistic is compared to using a random classifier at the corresponding cut-off level.

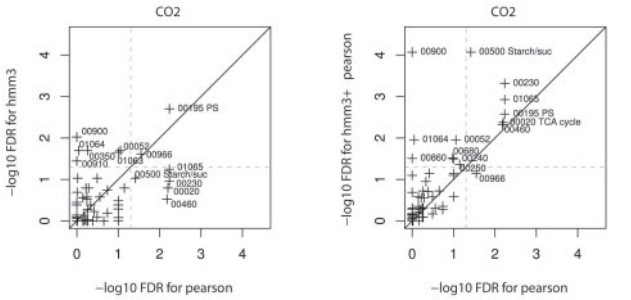

Fig. 7. *P*-value scatter plots comparing pathway enrichments at the 10% cut-off level for the CO_2 dataset for (**a**) HMM3 and Pearson, (**b**) combined HMM3+Pearson and Pearson. The gray line corresponds to FDR < 0.05. Points in the upper-left part of the graph indicate a higher enrichment for the method on the *y*-axis; points in the bottom-right corner indicate a higher enrichment for the method on the *x*-axis. The farther the points are from the diagonal, the higher is the enrichment gain of the best method. In the left plot, we observe the presence of pathways with higher enrichment for either HMM or Pearson. This indicates that each method recovers coresponses on distinct pathways. In the right plots, we see that the combination of Pearson and HMM has an overall higher enrichment of pathways than Pearson alone. PS: photosynthesis; Starch/suc: starch and sucrose synthesis pathway.

Pathway level analysis provides a direct way to search for interpretable coresponse patterns between metabolites and transcripts. To further visualize these results, we performed side-by-side comparison of the Fisher exact test *P*-values at the 10% cut-off level using Pearson and the HMM method of choice for each pathway (Fig. 7 and Supplementary Figs. S4–S7). Similar to the results seen for the lift charts, it is clear that HMM and Pearson are complementary, i.e. they show enrichment for distinct pathways. Moreover, this complementarity is successfully summarized using our OPLS-DA-based approach, which shows an overall higher enrichment than Pearson alone.

3.4 Examples of metabolite–transcript coregulation

Although phosphate was the only detected metabolite member of the metabolic map for photosynthesis (map 00195), the corresponding transcripts were already highly enriched in the top 3% of the gene lists from HMM3 with the cold, CO_2 and sulfur root datasets (phosphate was not present in the sulfur leaf dataset and plants were grown on transparent media; this may explain the expression of photosynthesis genes in the root). At the same cutoff, the Pearson correlation-based gene lists showed high enrichments of photosynthesis genes for the cold stress dataset but markedly lower enrichment than HMM3 in the CO_2 and sulfur root datasets (Fig. 7 and Supplementary Figs. S4–S7). Considering the abundance/expression profiles for phosphate and the photosynthesis-related genes, it becomes clear how HMM3 and Pearson are distinguished in their priorization of different patterns (Fig. 8a, b and c). Expectedly, genes that are ranked highly by the Pearson correlation have patterns nearly identical (or inversely identical) to phosphate itself. This leads to a high enrichment of photosynthetic genes under cold stress but much less enrichment under CO_2 and sulfur stress conditions where the response of the photosynthesis genes exhibits a more complex dependency on the phosphate levels. HMM3 on the other hand is mostly concerned with the overall shape of the expression patterns and therefore ranks the photosynthesis genes highly in all three cases despite their inverted pattern and the presence of noise. Moreover, delay lag is noticeable in the gene expression response in the sulfur root and CO_2 datasets.

The only detected metabolite in the zeatin synthesis pathway (Fig. 8d) is *O*-acetyl-l-serine (OAS). A significant proportion of the zeatin synthesis-related genes are enriched at the very top of HMM3-based gene list but not on Pearson. In particular, the expression profile of *ATCKX5* and *DOGT1*, which were not detected by Pearson, precede the fluctuations in OAS abundance.

Another informative example is the starch and sucrose synthesis pathway (pathway 00500) under elevated CO_2 conditions. Six different metabolites in this pathway were detected and our summarization approach integrates their coresponse patterns with the transcripts to form a single consensus statistic. This statistic significantly enriches the relevant transcripts at the top of the list when using either Pearson or HMM3 (Fig. 7). However, HMM3 and Pearson reveal different coresponse patterns; using them together results in a greatly improved recovery of the transcripts annotated to this pathway (Fig. 9). Since the summarization is performed with OPLS-DA, the importance of different metabolites can be investigated by looking at the model's loadings [$\vec{w}$ in Equation (5)]. In this example, HMM3 mainly detects coresponse with maltose and glucose-6-phosphate, whereas the Pearson correlation is higher with trehalose and sucrose. When using both HMM3 and Pearson together, these signals become integrated resulting in a high recovery of a large proportion of the starch and sucrose synthesis-related transcripts.

3.5 *De novo* detection of metabolite–transcript coregulation

The summarization step requires a set of true positives for weighting the different statistics according to how well they capture

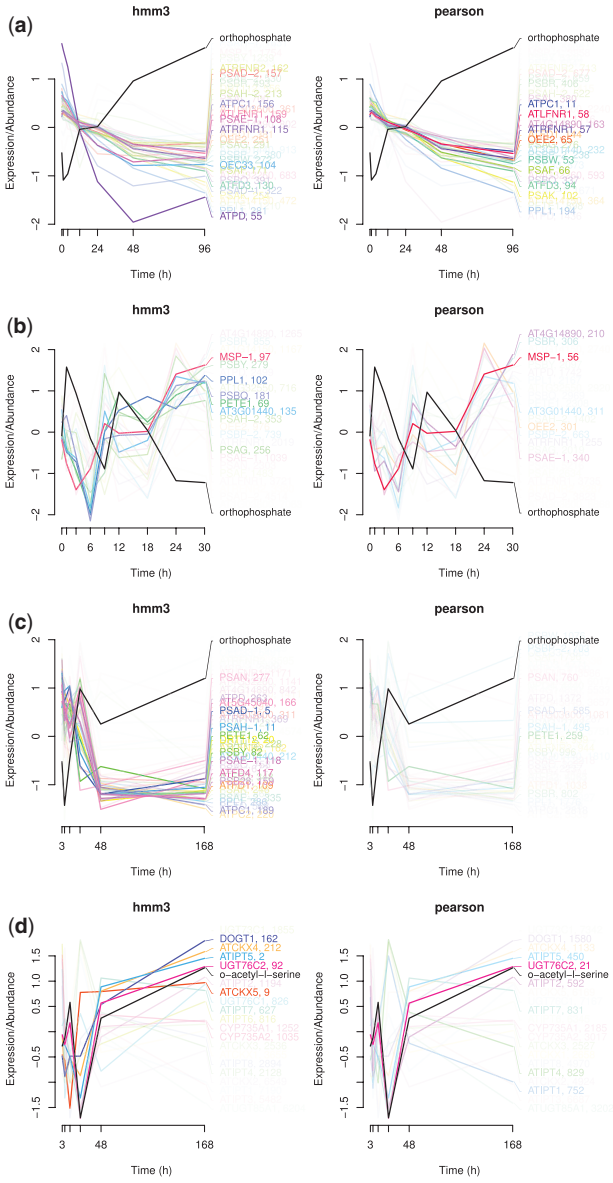

Fig. 8. Coresponse patterns between metabolites and transcripts as detected by HMM3 and the Pearson correlation. (**a–c**) Coresponse between phosphate and transcripts in the photosynthesis pathway map (00195) under (**a**) cold stress, (**b**) elevated CO_2 and (**c**) sulfur deficiency (root). (**d**) Coresponse between OAS and transcripts in the zeatin synthesis pathway (00908) under sulfur deficiency (root). The color intensity is proportional to the rank. Numbers to the right of the gene names correspond to the ranking of the gene in the entire dataset (max rank 7342 for sulfur-, 6680 for the cold stress- and 7138 for the CO_2 dataset).

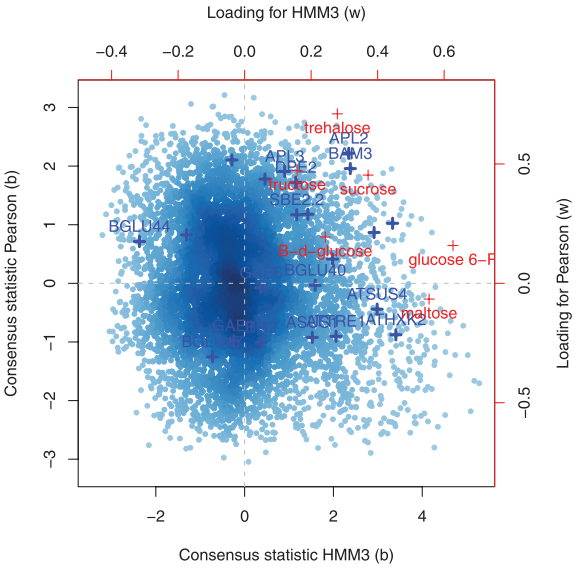

Fig. 9. HMM3 and Pearson consensus statistics for scoring coresponses patterns between genes and metabolites in the starch and sucrose synthesis pathway under elevated CO_2 conditions. Blue crosses indicate genes that are pathway members and blue circles other genes. Red crosses correspond to the metabolites and their corresponding loadings (importance for separating pathway members from other genes) are given on top and on right axes.

Table 1. Example of pathways, where interacting proteins as described by AtPIN, also were enriched among the top 10% in the ranked gene lists

Dataset/method	Pathway	Description	KEGG	AtPIN
Cold	00020	TCA cycle	0.001	0.025
hmm2 + Pearson	00195	Photosynthesis	2 (−22)	0.005
	00230	Purine metabolism	6 (−04)	0.0013
	00240	Pyrimidine metabolism	0.13	0.00086
	01061	BS of phenylprop.	0.07	0.012
	01064	Alkaloids BS (ornit, lys, nic)	0.23	0.039
	01065	Alkaloids BS (his., purine)	0.0018	0.005
	00290	Val., Leu. and Ile BS	0.77	0.0068
	00640	Propanoate metabolism	0.3	0.031
CO_2	00230	Purine metabolism	2.3 (−05)	0.033
hmm3 + Pearson	00240	Pyrimidine metabolism	0.0057	0.0083

Numbers correspond to the FDR values from Fisher's exact test.

the observed coresponse patterns in a given pathway. Although this requirement precludes complete *de novo* discovery of MT coresponses on the pathway level, unknown genes can still be scored when a set of guiding examples is available. To examine whether relevant genes outside the KEGG pathways also receive high rankings in our examples, we extracted genes that code for proteins that are predicted by AtPIN (Brandão *et al.*, 2009) to interact with the enzymes in each pathway. Using the Fisher's exact test, we then checked whether the transcripts for the interacting proteins were overrepresented among the top 10% in the ranked gene lists. For the cold and CO_2 datasets, several such scenarios could be seen (Table 1). As a notable example, AT2G46820 (P subunit of photosystem I), AT4G21280 (a locus coding for *PsbQ*), and AT4G39710 (photosystem I interacting NADH dehydrogenase, Peng *et al.*, 2009) are predicted to interact with the proteins in the photosynthesis map and are among the top 2% of the gene list for the cold stress datasets. These genes were not among the KEGG annotations originally used (the org.At.Tair.db package from Bioconductor), indicating that our approach can be used to mine for

novel metabolic properties and also for genes outside the set used for calculating the consensus statistic.

4 DISCUSSION

We presented an approach for detecting coresponses between metabolites and transcripts using combined profiling time-series data. In previous studies, the direct or lagged Pearson correlation was used for identifying such patterns but as demonstrated by our simulations, this approach does not work well for short time-series with strong noise and time-lagged responses (Fig. 4). Our HMM-based similarity measure performed well also in this most degenerate scenario but, as expected, not better than the Pearson correlation in the more favorable scenarios.

In general, both MM and TT correlations are expected to decrease with increasing distance in the underlying metabolic pathways. By looking at differences in the correlations between cognate and non-cognate transcript and metabolite pairs, we could reproduce earlier findings (Kharchenko *et al.*, 2005; Walther *et al.*, 2010) with the CO_2 dataset for TT and MM correlations, and for MM correlations with the sulfur datasets (Fig. 5). However, different from Walther *et al.* (2010), we could not see a dependency between metabolic distance and MT correlations for any of the studied datasets regardless of the used measure. Circumstances such as differences in experiment designs and analytical platforms may be causative of this discrepancy. Possibly, since *A.thaliana* is a much more complex organism than yeast, it could also show less obvious dependency on the pathways that have been deciphered primarily in single-cell organisms. Nonetheless, with the inevitably high number of false positives among pairwise correlations, we contend that coresponses on the level of whole pathways are both more likely to be detectable and easier to interpret.

In actual stress response data, we saw similar tendencies as in our simulation; Pearson was better at detecting direct, clear responses (Fig. 8a) and HMM was better at more complex patterns (Fig. 8b–d). In combined profiling data, we both expected and saw many types of coresponses between metabolites and transcripts. However, the nature of these responses—lagged or direct, noisy or clear—is not enough for judging their relevance because they have many possible origins (Ladurner, 2006): each with its own expected pattern. Therefore, we propose to combine different types of measures and then to summarize them by looking at a set of true associations as guiding examples. This approach worked well; the combined HMM+Pearson methods were best for predicting pathway comemberships between the metabolites and transcripts on all four datasets (Fig. 6).

The richness of the interplay between metabolism and gene expression requires attentiveness when interpreting observed patterns. While direct coregulation is conceptually possible, pleiotropic and indirect causes are presumably more likely. However, observed coresponses may still be informative for understanding the perception of stress and the events that follow on a molecular level. We presented three examples of this.

(i) In plant stress responses, photosynthesis is strongly downregulated to avoid the production of detrimental reactive oxygen species (Mittler, 2002) while metabolism undergoes large-scale changes. With decreasing energy production, the ATP/ADP ratio drops and free phosphate levels increase. These are the same patterns we observe under cold stress and sulfur deficiency (Fig. 8a and c). However, under conditions of elevated CO_2, which may initially be beneficial to the plant (Kanani *et al.*, 2010), the photosynthesis rate increases and consequently, phosphate levels start to decrease (Fig. 8b). This pattern was also detected with our approach although it is overloaded by circadian rhythmic expression (Harmer *et al.*, 2000).

(ii) OAS is a key metabolite in sulfur assimilation and also a member of the cytokinin (zeatin) synthesis pathway (Hirai *et al.*, 2005). Cytokinins themselves are well-known repressors of sulfur uptake (Ohkama *et al.*, 2002). The observation of a strong increase in *ATCKX5* expression before the increase in the OAS-levels may therefore be a manifestation of upregulated sulfur transport and of changes that follow sulfur assimilation. The coresponses in the glucosinolate synthesis pathway discussed by Hirai *et al.* (2005) were also clearly detected in both root and shoot datasets (Supplementary Figs. S6 and S7; pathway 00966).

(iii) During elevated CO_2 conditions, as the expression of photosynthesis genes increases, carbon assimilation also changes. As it is a central metabolism pathway, metabolites involved in starch and sucrose synthesis show complex behavior but our method still highlighted coresponses between them and related transcripts. Interestingly, both the Pearson correlation and HMM3 consensus statistics were informative here but completely orthogonal (Fig. 9). Furthermore, the OPLS-DA model prioritizes different metabolites, highlighting that different types of coresponse patterns are present and that Pearson and HMM complement each other.

Bradley *et al.* (2009) predicted MT associations by combining several stress response datasets using a Bayesian approach. As that methodology was based on the use of Pearson correlations and does not handle time lags, it would profit from the incorporation of the coresponse statistics proposed here. The presented photosynthesis example indicates that combining datasets may be useful also in our case. However, this appeared to be more of an exception than a rule since the significant pathways overall differed strongly between different datasets (Supplementary Figs. S4–S7). Furthermore, as a difficulty in combining datasets from completely different experiments, there typically is a very small overlap of the detected metabolites. For example, only seven metabolites were measured in all the datasets we studied. Moreover, combining data will only capture MT coresponses in pathways affected in the majority of the studied stress conditions. This may worsen predictions for stress-specific metabolite–transcript associations.

Extrapolation of the predictions with additional sources of biological data, as shown with AtPIN, indicates the power of the methodology for finding *de novo* associations. One interesting future aspect is the inclusion of further interaction data for validation purposes and integrative analysis. Moreover, we are currently developing a web tool over an expanded collection of datasets that includes genome-wide predictions of MT associations.

ACKNOWLEDGEMENTS

H.R. thanks Kazuki Saito, Masami Y. Hirai and Masanori Arita for discussions and helpful comments and I.G.C. thanks Christoph Hafemeister for providing scripts for generating simulated data.

Funding: Fundacão de Amparo a Pesquisa de Pernambuco (to I.G.C. in part); Conselho de Desenvolvimento Científico e Tecnológico

(Brazil) (to I.G.C. in part); travel grant from the Japan Society for the Promotion of Science (JSPS) (to H.R. in part).

Conflict of Interest: none declared.

REFERENCES

Arkin,A. *et al*. (1997) A test case of correlation metric construction of a reaction Pathway from Measurements. *Science*, **277**, 1275–1279.

Bradley,P.H. *et al*. (2009) Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS Comput. Biol.*, **5**, e1000270.

Brandão,M.M. *et al*. (2009) AtPIN: Arabidopsis thaliana protein interaction network. *BMC Bioinformatics*, **10**, 454.

Bylesjö,M. *et al*. (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemomet.*, **20**, 341–351.

Bylesjö,M. *et al*. (2007) Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J.*, **52**, 1181–1191.

Carrari,F. *et al*. (2006) Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiol.*, **142**, 1380–1396.

Chechik,G. and Koller,D. (2009) Timing of gene expression responses to environmental changes. *J. Comput. Biol.*, **16**, 279–290.

Costa,I.G. *et al*. (2005) The graphical query language: a tool for analysis of gene expression time-courses. *Bioinformatics*, **21**, 2544–2545.

Costa,I.G. *et al*. (2009) Constrained mixture estimation for analysis and robust classification of clinical time series. *Bioinformatics*, **25**, i6–i14.

Dutta,B. *et al*. (2009) Time-series integrated 'omic' analyses to elucidate short-term stress-induced responses in plant liquid cultures. *Biotechnol. Bioeng.*, **102**, 264–279.

Gibon,Y. *et al*. (2006) Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Genome Biol.*, **7**, R76.

Hafemeister,C. *et al*. (2011) Classifying short gene expression time-courses with bayesian estimation of piecewise constant functions. *Bioinformatics*, **27**, 946–952.

Harmer,S.L. *et al*. (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science*, **290**, 2110–2113.

Hirai,M.Y. *et al*. (2003) Global expression profiling of sulfur-starved Arabidopsis by DNA macroarray reveals the role of O-acetyl-l-serine as a general regulator of gene expression in response to sulfur nutrition. *Plant J.*, **33**, 651–663.

Hirai,M.Y. *et al*. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **101**, 10205–10210.

Hirai,M.Y. *et al*. (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J. Biol. Chem.*, **280**, 25590–25595.

Kanani,H. *et al*. (2010) Individual vs. combinatorial effect of elevated CO2 conditions and salinity stress on Arabidopsis thaliana liquid cultures: Comparing the early molecular response using time-series transcriptomic and metabolomic analyses. *BMC Syst. Biol.*, **4**, 177.

Kaplan,F. *et al*. (2004) Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiol.*, **136**, 4159–4168.

Kaplan,F. *et al*. (2007) Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant J.*, **50**, 967–981.

Kharchenko,P. *et al*. (2005) Expression dynamics of a cellular metabolic network. *Mol. Syst. Biol.*, **1**, 2005.0016.

Kresnowati,M.T.A.P. *et al*. (2006) When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Mol. Syst. Biol.*, **2**, 49.

Ladurner,A.G. (2006) Rheostat control of gene expression by metabolites. *Mol. Cell*, **24**, 1–11.

Mittler,R. (2002) Oxidative stress, antioxidants and stress tolerance. *Trends Plant Sci.*, **7**, 405–410.

Obayashi,T. *et al*. (2007) ATTED-II: a database of coexpressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.

Ohkama,N. *et al*. (2002) Regulation of sulfur-responsive gene expression by exogenously applied cytokinins in Arabidopsis thaliana. *Plant Cell Physiol.*, **43**, 1493–1501.

Peng,L. *et al*. (2009) Efficient operation of NAD(P)H dehydrogenase requires supercomplex formation with photosystem I via minor LHCI in Arabidopsis. *Plant Cell*, **21**, 3623–3640.

Pir,P. *et al*. (2006) Integrative investigation of metabolic and transcriptomic data. *BMC Bioinformatics*, **7**, 203.

Rabiner,L.R. (1989) A tutorial on Hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

Redestig,H. *et al*. (2007) Transcription factor target prediction using multiple short expression time series from Arabidopsis thaliana. *BMC Bioinformatics*, **8**, 454.

Redestig,H. *et al*. (2010) Consolidating metabolite identifiers to enable contextual and multi-platform metabolomics. *BMC Bioinformatics*, **11**, 214.

Saito,K. *et al*. (2008) Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. *Trends Plant Sci.*, **13**, 36–43.

Schliep,A. *et al*. (2003) Using Hidden Markov Models to analyze gene expression time course data. *Bioinformatics*, **19** (Suppl. 1), i255–i263.

Schliep,A. *et al*. (2005) Analyzing gene expression time-courses. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 179–193.

Steuer,R. *et al*. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019–1026.

Takahashi,H. *et al*. (2011) Dynamics of time-lagged gene-to-metabolite networks of Escherichia coli elucidated by integrative omics approach. *OMICS*, **15**, 15–23.

Urbanczyk-Wochniak,E. *et al*. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.*, **4**, 989–993.

Walther,D. *et al*. (2010) Metabolic pathway relationships revealed by an integrative analysis of the transcriptional and metabolic temperature stress-response dynamics in yeast. *OMICS*, **14**, 261–274.