Journal of Experimental Botany, Vol. 56, No. 417, pp. 1887–1896, July 2005
doi:10.1093/jxb/eri179 Advance Access publication 23 May, 2005
This paper is available online free of all access charges (see http://jxb.oupjournals.org/open_access.html for further details)

Journal of Experimental Botany

RESEARCH PAPER

Integrative gene-metabolite network with implemented causality deciphers informational fluxes of sulphur stress response

Victoria J. Nikiforova[1,2,*], Carsten O. Daub[1,†], Holger Hesse[1], Lothar Willmitzer[1] and Rainer Hoefgen[1]

[1] *Max Planck Institute of Molecular Plant Physiology, Department of Molecular Physiology, Am Mühlenberg 1, D-14476 Golm, Germany*

[2] *Timiryazev Institute of Plant Physiology, Russian Academy of Sciences, Botanicheskaya Str. 35, Moscow 127276, Russia*

Received 17 November 2004; Accepted 4 April 2005

Abstract

The systematic accumulation of gene expression data, although revolutionary, is insufficient in itself for an understanding of system-level physiology. In the post-genomic era, the next cognitive step is linking genes to biological processes and assembling a mosaic of data into global models of biosystem function. A dynamic network of informational flows in *Arabidopsis* plants perturbed by sulphur depletion is presented here. With the use of an original protocol, the first biosystem response network was reconstructed from a time series of transcript and metabolite profiles, which, on the one hand, integrates complex metabolic and transcript data and, on the other hand, possesses a causal relationship. Using the informational fluxes within this reconstruction, it was possible to link system perturbation to response endpoints. Robustness and stress tolerance, as consequences of scale-free network topology, and hubs, as potential controllers of homeostasis maintenance, were revealed. Communication paths of propagating system excitement directed to physiological endpoints, such as anthocyanin accumulation and enforced root formation were dissected from the network. An auxin regulatory circuit involved in the control of a hypo-sulphur stress response was uncovered.

Key words: Auxin, causality, metabolome, network, network topology, plait concept, scale-free network, sulphur metabolism, systems biology, transcriptome.

Introduction

Living organisms are complex multi-elemental, multi-functional systems existing in ever-changing environments. The viability of the system is provided via flexible and effective control circuits of multiple informational fluxes interconnecting in a dense network. This hierarchically organized network of negative feedback stimuli subordinate to superior positive feedback is so fundamental that it has been proposed as a minimal but sufficient definition of life (Korzeniewski, 2001). Unravelling such networks will allow global models of biosystem function to be characterized.

Network reconstruction and analysis are starting to be widely used to characterize and predict biosystem behaviour, giving rise to a new branch of biological knowledge, 'network genomics' (Forst, 2002). Until recently, such analyses have been limited to one level of manifestation of the genetic information, i.e. transcript networks (Thieffry *et al.*, 1998, Shen-Orr *et al.*, 2002; Rosenfeld *et al.*, 2002; Featherstone and Broadie, 2002) or metabolic networks (Schuster *et al.*, 2002; Fiehn and Weckwerth, 2003). However, changes in transcript levels are transferred to changes in metabolite levels and thereby to physiological endpoints via adaptations of physiology and homeostasis. Therefore, progressive system characterization involves integrating multiple levels of realization of the genetic information, for example, by superimposing transcript, protein, and metabolite profiles.

In a recent study, correlation analysis between the yeast transcriptome and proteome revealed that yeast sequesters sulphur during Cd^{2+} detoxification (Fauchon *et al.*, 2002). Attempts to combine transcript and metabolic data in

* To whom correspondence should be addressed in Germany. Fax: +49 331 567 8134. E-mail: nikiforova@mpimp-golm.mpg.de
† Present address: Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden.

a correlation analysis have been undertaken by Askenazi *et al.* (2003), revealing pair-wise associations between fungal genes and two metabolites important for bioengineering, and by Urbanczyk-Wochniak *et al.* (2003), revealing genes correlating reliably to nutritionally important metabolites.

As another approach to integrate transcriptome and metabolome, data are mapped on known metabolic pathways (Grosu *et al.*, 2002; Hirai *et al.*, 2004). Among software tools suitable for such visualizations, AraCyc (http://www.arabidopsis.org/tools/aracyc/, Mueller *et al.*, 2003) and MapMan (http://gabi.rzpd.de/projects/MapMan/, Thimm *et al.*, 2004) are designed especially for *Arabidopsis*. The latter approach, however, is based on legacy and static metabolic data, so it generally does not allow previously unknown biochemical or regulatory pathways to be identified. The elucidation of informational fluxes linking input signals with response endpoints must be possible from a network which integrates transcriptional and metabolic changes in an unbiased way. In the course of the study presented here, an unbiased gene-metabolite network of correlations was reconstructed from transcript and metabolite profiles and the resulting informational fluxes controlling the systems response to sulphur deprivation in *Arabidopsis* were analysed.

Materials and methods

Physiological experimenting: transcript and metabolite profiling

For sulphur-starvation experiments, *Arabidopsis thaliana* genotype *Col-0* plants were grown on a solidified agarose medium in sterile Petri dishes (half-normal Murashige-Skoog salts for control sulphur-sufficient medium, 89% less sulphur for sulphur-depleted medium). Sulphur depletion was applied as constitutive stress (germination and growth on sulphur-deficient medium for 10 d or 13 d) and induced stress (germination on normal medium for 8 d, then the transfer of seedlings to a sulphur-deficient medium for 6 d or 10 d); seedlings grown on sulphur-sufficient medium were used as the four corresponding controls. Material for each of eight experimental points was collected as whole seedlings, in five repetitive pools, containing 500–600 plants in each pool. The experiment is described in detail previously (Nikiforova *et al.*, 2003). The same plant material was used for RNA isolation followed by the transcript profiling technique, as well as for metabolite profiling, elemental sulphur, thiol, and anthocyanin measurements (using GC-MS, HPLS, ICP-AES, and spectrophotometry). As a result of these studies, a sulphur-deficient transcriptome (Nikiforova *et al.*, 2003) and metabolome (Nikiforova *et al.*, 2005) was described, providing a basis for the reconstruction of an integrated gene-metabolite network.

Network reconstruction protocol

To reconstruct a causal gene-metabolite network of statistically significant correlations, the original algorithm was elaborated (described and discussed in detail below). A condensed overview on the applied techniques is provided here. First, the transcript and metabolite profiles were combined in dataset 1 which consisted of 6454 non-redundant genes and 81 non-redundant chemical compounds, or 'metabolites', all containing relative concentration levels at eight experimental points. Then, the following step-by-step protocol was used to reconstruct the response gene-metabolite network from the transcript and metabolite profiles.

(i) From dataset 1 those genes were selected which correlated to sulphur and sulphur-responding metabolites (Fig. 1A). For this: (a) a Pearson correlation coefficient r between each of the sulphur-responding metabolites and the whole set of genes from dataset 1 was calculated. (b) The gene expression part of dataset 1 was shuffled 1000 times. (c) With the produced 1000 shuffled datasets step (i, a) was repeated 1000 times for each of the sulphur-responding metabolites. Distributions of r for the original and shuffled datasets were analysed. (d) Genes, correlating reliably to sulphur-responding metabolites, were determined from the comparisons of r distributions in the original and shuffled datasets, if they appeared in a histogram bin, that contains at least two times more genes from the original dataset than from shuffled datasets (Fig. 1A, lower graph).
(ii) Transcript levels for a set of genes, purified from noise in step (i), were combined with a dataset of 81 metabolites into dataset 2.
(iii) The Pearson correlation coefficient (r) distance matrix and the mutual information (MI) distance matrix of dataset 2 were calculated, transformed for compatibility (tr and tMI), and plotted (Fig. 1B, blue dots).
(iv) Dataset 2 was shuffled.
(v) r distance matrix and MI distance matrix of shuffled dataset 2 were calculated, transformed for compatibility (tr and tMI), and plotted (Fig. 1B, purple dots).
(vi) tr and tMI distance matrices of the original dataset 2 were overlaid with tr and tMI distance matrices of the shuffled dataset 2 (Fig. 1B).
(vii) Noise correlations were cut off by setting graphically a threshold for overlaid matrices (Fig. 1B, dotted lines) in a border of the area strongly covered by associations from the shuffled dataset; as a result significant pair associations, gene/gene, gene/metabolite, or metabolite/metabolite, that were acceptable for network reconstruction were determined (Fig. 1B, shadowed area).
(viii) A response gene-metabolite network was formed by pair-wise connection of genes or/and metabolites from all significant associations.

Equations and algorithms

To make calculations of Student's t-test and the Pearson correlation coefficient r, an algorithm incorporated into the Microsoft Excel 2000 software program was used.

To calculate mutual information (MI), the algorithm described by Steuer *et al.* (2002) was used, incorporated into the MetaGeneAlyse web tool (Daub *et al.*, 2003), and improved with B-spline functions (Daub *et al.*, 2004).

To threshold noise correlations, original datasets were shuffled with an over-sampling rate of 1000 with the use of an original algorithm, incorporated into the MetaGeneAlyse web tool (Daub *et al.*, 2003).

To produce compatible correlation matrices, equations for r and MI calculations were transformed in a way to have the strongest correlations approaching 0 (tMI for transformed MI and tr for transformed r), according to the following equations, incorporated into the MetaGeneAlyse web tool was used:

$$tr = \sqrt{2(1 - |r|)} \quad (1)$$

$$tMI = 1 - MI/MI_{\max} \quad (2)$$

To visualize the reconstructed network, the Pajek computer program for large network analysis (http://vlado.fmf.uni-lj.si/pub/networks/pajek/; Batagelj and Mrvar, 1998) was used.

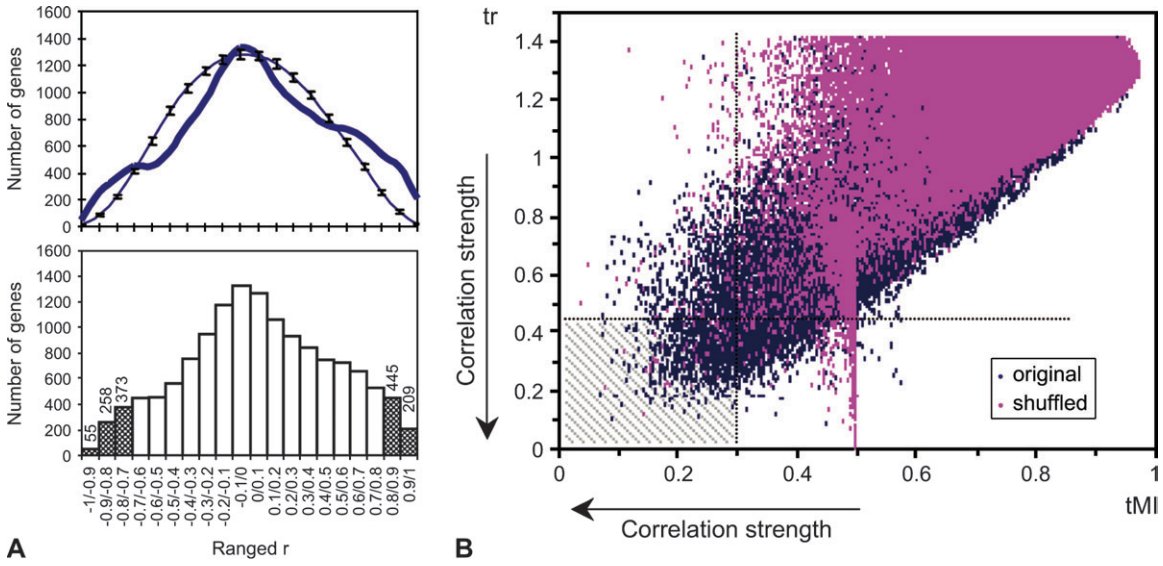

Fig. 1. Response network reconstruction. (A) Determining genes correlating reliably to serine as an example of sulphur-responding metabolites. The upper graph: distribution of Pearson correlation coefficient *r* for genes from original (thick dark-blue curves) and shuffled (thin curves) datasets; error bars are standard deviations calculated from 1000 shufflings. The lower graph: histograms showing the distribution of Pearson correlation coefficient *r*. Those bins containing at least two times more genes from the original dataset than from shuffled datasets are shadowed; numbers of genes in these bins are depicted. Genes appeared in these bins were combined in a new dataset 2. (B) Scatter plot of associations, gene/gene, gene/metabolite, or metabolite/metabolite, plotted as dots with corresponding values for transformed Pearson correlation coefficient (*tr*) and transformed mutual information (*tMI*), for the original (blue) and shuffled (red) datasets. Arrows indicate the increasing correlation strength, with the strongest correlations approaching 0 (due to transformation). Threshold limits are depicted with dotted lines. The shaded area contains associations selected for network reconstruction.

Results and discussion

Integrating transcript and metabolite profiles into one dataset

For an integrated analysis, total RNA and hydrophilic metabolites were isolated from the same samples (five repetitions each; each repetition a pool of 500–600 seedlings) and used to obtain transcript profiles by array hybridization (Nikiforova *et al.*, 2003) and metabolite profiles by gas chromatography-mass spectrometry (GC-MS) analysis (Nikiforova *et al.*, 2005). Relative transcript amounts for 6454 non-redundant genes and relative concentrations of 81 non-redundant chemical compounds (78 detected by GC-MS plus sulphur, glutathione, and anthocyanins measured using other methods) were determined at eight experimental points under conditions of constitutive and induced sulphur starvation, which lasted for 6–13 d. The entire compilation of these experimental data points is subsequently called dataset 1.

Reconstructing a statistically significant response network

During the developing hypo-sulphur stress, concentration levels of sulphur-responding genes and metabolites changed differentially in a time-dependent manner, forming distinct alteration patterns. Assuming that the more similar the pattern, the shorter the distance between genes and metabolites in the communication network, existing in a plant as a biosystem, the combined gene/metabolite distance matrix was calculated for dataset 1 using Pearson correlation coefficient *r*.

For the number of transcripts (6454) and metabolites (81) determined, the number of all possible pair-wise correlations, constituting a correlation matrix, is very large ($((6454+81)^2/2$ equals about 21.4 million). Inherently this contains a correspondingly large noise component. To decrease dataset 1 by its putative noise component, only those genes which correlate reliably to sulphur and sulphur-related metabolites glutathione, anthocyanins, allantoin, *O*-acetyl-serine, putrescine, raffinose, serine, tryptophan, and uric acid (assigned as being significantly altered by sulphur depletion) were kept for further analysis. To detect these genes, the original dataset 1 was first shuffled (or randomized). From the shuffling of the dataset and recalculation of a distance matrix based on the shuffled dataset it is possible to estimate a significance threshold. The smallest distance calculated from the shuffled dataset can be regarded as the upper significance threshold for the unshuffled case. The applied shuffling algorithm of the MetaGeneAlyse tool (Daub *et al.*, 2003) works on a dataset row by row. Within each row two values are randomly chosen and exchanged. For n values in a row an over-sampling rate of o exchanges $n \times o$ pairs of values. For dataset 1, where each row represented one gene or metabolite with corresponding eight values, an over-sampling rate of 1000 was used. The distribution of values within a row was conserved. By this procedure, 1000 shuffled datasets were generated and

assumed to contain no sense information except noise. Then, correlation matrices were calculated from these 1000 shuffled datasets and compared with the correlation matrix of the original dataset (step 1 of the 'Network reconstruction protocol' above). A case example is shown in Fig. 1A. There, the correlation of serine with respect to all genes is shown as a function of the number of genes displaying a certain correlation coefficient. As expected, the vast majority of genes displays low *r*-values between −0.8 and 0.8. Distance matrices of all shuffled datasets, although not identical, produced highly similar distributions of *r*, which, however, were clearly different from those produced by original distance matrices, derived from the experimental data (Fig. 1A, upper graph). This observation was taken as an important argument for the robustness of the method with respect to noise.

Now, in order to filter the genes correlating reliably to sulphur and sulphur-related metabolites, the following discrimination criterion was introduced: the correlations were assumed to be reliable if the number of correlating genes in an analytical bin in a histogram of the distribution of analytical and virtual distances (Fig. 1A, lower graph) is at least 2 times higher than in a virtual bin of the shuffled dataset. Based on this assumption, those bins which were shaded in Fig. 1A were considered to contain reliably correlating items. Genes fulfilling these selection criteria were combined within a new noise-free dataset (dataset 2), which was used for response network reconstruction.

However, uneven distribution of the measured values of sulphur and some metabolites resulted in a weak relevance based on the Pearson correlation coefficient. Therefore, in addition, new distance matrices were calculated for the same dataset 2 (original and shuffled), using mutual information (*MI*) as another correlation measure, independent of the uneven value distribution (Butte and Kohane, 2000; Steuer *et al.*, 2002; Daub *et al.*, 2004). Finally, to identify significantly correlating associations acceptable for network reconstruction, it was necessary to estimate the threshold for correlation reliability by superimposing the distance matrices obtained for both approaches (i.e. Pearson correlation *r* and *MI*, for original and shuffled datasets) with the idea to consider only those associations for which both methods pointed to a reliable correlation. However, different maxima for the strongest correlations calculated with the use of *r* and *MI* prevented the direct comparison of both matrices. Therefore, to produce compatible matrices, the equations for *r* and *MI* calculations were transformed (see 'Equations and algorithms' above). After transformation, high correlations resulted in small values for both matrices, with the strongest correlations approaching 0 (*tr* for transformed *r*, equation 1, and *tMI* for transformed *MI*, equation 2). Now, by overlaying the calculated analytical *tMI* and *tr* matrices (Fig. 1B, blue dots) and virtual *tMI* and *tr* matrices of the shuffled dataset (Fig. 1B, purple dots), the threshold for correlation reliability was estimated graphically. The peak with *tMI* around 0.5, which is present in both original and shuffled datasets, contains associations with outlying values, which would be detected falsely as positive using the Pearson correlation coefficient alone. From the overlaying matrices, the area strongly covered by associations from the shuffled dataset was cut off. As a result, acceptable values for correlations to be considered in the network reconstruction were finally found as follows: $tr<0.45$ ($|r| \geqslant 0.9$) and $tMI<0.3$ (Fig. 1B, shaded area).

Now, based on the central assumption on item co-behaviour as introduced above (i.e. the more similar the pattern, the shorter the distance between genes and metabolites in the communication network), a response communication network was reconstructed by establishing a connecting link between paired elements, gene/gene, gene/metabolite, or metabolite/metabolite, if an association of these two elements appears among those defined as significantly correlated (shaded area in Fig. 1B). Under the chosen connectivity thresholds, the reconstructed network contained 541 elements (vertices) and 5212 associations (edges).

Global properties of the reconstructed response network suggest robustness and stress tolerance of the underlying biological system

The systems response develops by propagating through the network of informational flows reconstructed in the course of the present study. The key aspects of network functionality can be predicted directly from its structure (Albert *et al.*, 2000; Jeong *et al.*, 2000; Stelling *et al.*, 2002). The network connectivities k (6.6 in average) were distributed extremely non-homogeneously. The distribution probability $N(k)$ had a linear tail for large k, following a power law $N(k) \sim k^{-\gamma}$, with the exponent γ=2.27 (Fig. 2A). This value lies within the γ range found for large networks of different nature, that is, between 2.1 and 4 (Barabasi and Albert, 1999). Such non-homogeneously-wired networks, called 'scale-free networks' (Barabasi and Albert, 1999), possess a number of universal characteristics. One of them is high tolerance to errors due to high robustness (Albert *et al.*, 2000). Local errors/changes rarely lead to the loss of the global information-carrying ability of the network, as was demonstrated for both *in silico* and *in vivo* mutagenesis studies for the *E. coli* metabolic network (Edwards and Palsson, 2000). However, highly connected nodes (hubs) are the sites of system vulnerability in scale-free networks (McCabe, 2002; Clipsham *et al.*, 2002). Critically important for network stability, hubs can be considered as putative controllers of homeostasis maintenance. In Table 1, vertices with the highest number of connectors (20–32) are listed. When analysing which of the different functional categories of genes shows the highest number of connectors (Table 2), it becomes clear that nucleotide metabolism, protein destination, and intracellular transport are the three

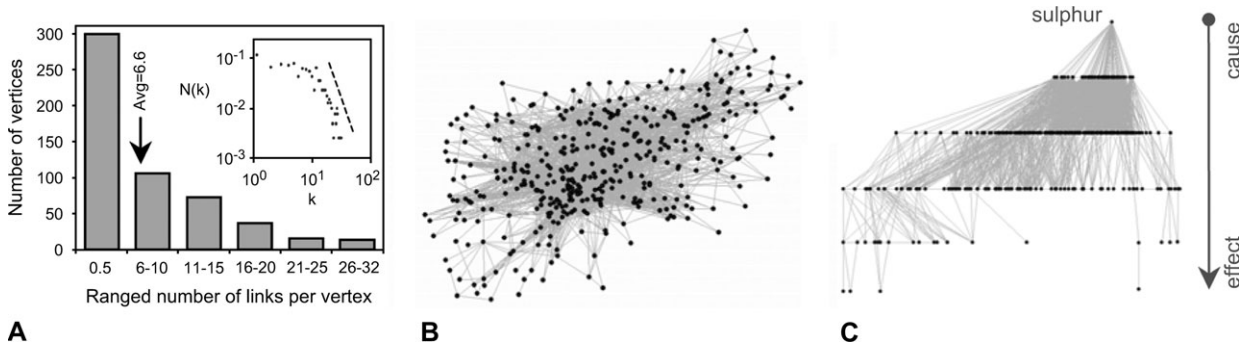

Fig. 2. Global properties of the reconstructed network. (A) Inhomogeneous distribution of vertex connectivities; the average number of links per vertex is depicted with an arrow. On the embedded graph the probability distribution function $N(k)$ of k connectivities (log-log scale, base 10) has a power-law tail with an exponent γ=2.27 (the slope of the dashed line). (B) Original response network profile; (C) network profile with centralized S provides 'cause-to-effect' relationship.

Table 1. *List of network elements (vertices) with the highest number of connectors*

Network element	Number of connections	Functional annotation (for genes)[a]
At4g02080	32	SAR1/GTP-binding secretory factor
Sulphur	31	
At3g49580	30	Putative protein
At4g36760	29	Aminopeptidase-like protein
At5g63600	28	1-Aminocyclopropane-1-carboxylic acid oxidase-like
At5g48000	28	Cytochrome P450-like protein
At5g25890	28	IAA28
At5g13800	27	Hydrolase, alpha/beta fold family (http://www.tigr.org)
Serine	27	
At3g01420	27	Feebly-like protein
At3g14210	27	Myrosinase-associated protein, putative
At1g36370	26	Putative hydroxymethyltransferase
At2g17190	26	Putative ubiquitin-like protein
At3g05160	25	Putative sugar transporter
At2g27530	25	60S ribosomal protein L10A
At5g64350	25	Immunophilin
At2g05840	24	20S proteasome subunit (PAA2)
At2g47650	23	Putative nucleotide-sugar dehydratase
At4g11320	23	Drought-inducible cysteine proteinase RD21A precursor
At5g11670	22	NADP dependent malic enzyme-like protein
At4g29040	22	26S proteasome subunit 4-like protein
At2g27710	22	60S acidic ribosomal protein P2
At5g52240	22	Progesterone-binding protein-like
At5g37600	21	Glutamate–ammonia ligase
At1g02000	21	Nucleotide sugar epimerase, putative
At3g27830	21	50S ribosomal protein L12-A
At5g67560	21	ADP-ribosylation factor-like protein
At5g55190	21	Small Ras-like GTP-binding protein
At4g13930	20	Hydroxymethyltransferase
At5g19440	20	Cinnamyl-alcohol dehydrogenase-like protein
At5g53350	20	ATP-dependent Clp protease regulatory subunit CLPX
At2g21870	20	Putative ATP synthase
At1g75270	20	GSH-dependent dehydroascorbate reductase 1, putative

[a] Gene annotations were derived from the *Arabidopsis thaliana* database (MATDB) at the server of Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de/proj/thal/db).

most hub-rich functional categories. This finding resembles gene expression networks in *E. coli* (Thieffry *et al.*, 1998), where the largest group of hubs is represented by protein synthesis and destination genes. By contrast, genes involved in protein synthesis and subclass translation displayed the lowest overall connectivity (2.5), which may result from the precise adjustment of the enzymes involved in the regulation of translation, whereas ribosomal proteins, which are able to associate with a variety of proteins, provide higher than average connectivity in the network (8.1) (Table 2).

Yeast's general ability to compensate for individual mutations has been proposed to be largely a result of the scale-free properties of its gene expression network (Featherstone and Broadie, 2002). This conclusion has been extended here to a gene-metabolite communication network of a multicellular organism maintaining homeostasis during nutritional stress. The scale-free network topology suggests that the stability of the newly formed homeostasis under sulphur-limiting conditions is due to a high redundancy of gene/metabolite communication paths, triggered by critical changes in non-redundant excitement-accumulating hubs.

Implementing a causal relationship to the reconstructed response network

The characteristic feature for both applied methods of calculating correlation distance matrices (the Pearson correlation coefficient and mutual information) is non-causality: nothing in the definition of correlation implies that the relation between two variables is one of cause and effect. For the network reconstructed from these matrices, this implies that the direction of informational flows between elements is not defined (Fig. 2B). However, it is assumed here from *a priori* knowledge of the primary cause of system excitement, i.e. depletion of sulphate from the medium, that the changed sulphur level is the starting point, or excitation, for the response development. In keeping

Table 2. *Average numbers of connections for genes from the reconstructed network, belonging to different functional categories*

Functional category class/subclass[a]	Average number of connections
Metabolism (01)	**6.9**
Amino-acid metabolism (01.01)	6.2
Nitrogen and sulphur metabolism (01.02)	4.8
Nucleotide metabolism (01.03)	17.4
Phosphate metabolism (01.04)	6.4
C-compound and carbohydrate metabolism (01.05)	6.5
Lipid, fatty-acid and isoprenoid metabolism (01.06)	5.7
Secondary metabolism (01.20)	8.1
Energy (02)	**8.3**
Electron transport and energy conservation (02.11)	8.9
Respiration (02.13)	8.0
Cell growth, cell division and dna synthesis (03)	**6.2**
DNA synthesis and replication (03.16)	4.0
Cell cycle control and mitosis (03.22)	6.5
Transcription (04)	**5.9**
rRNA processing (04.01.04)	4.0
mRNA transcription (04.05)	5.7
Protein synthesis (05)	**6.9**
Ribosomal proteins (05.01)	8.1
Translation (initiation, elongation and termination) (05.04)	2.5
Protein destination (06)	**11.0**
Protein folding and stabilization (06.01)	9.2
Protein targeting, sorting and translocation (06.04)	9.4
Assembly of protein complexes (06.10)	11.8
Proteolysis (06.13)	12.5
Transport facilitation (07)	**5.9**
Ion transporters (07.04)	4.4
Intracellular transport (08)	**9.9**
Intracellular signal transduction (10.01.01)	**6.8**

[a] Genes were assigned to functional categories automatically using the *Arabidopsis thaliana* database (MATDB) at the server of Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de/proj/thal/db).

with this assumption, by centralizing the vertex 'sulphur', a general 'cause-to-effect' directionality of information fluxes along network paths from sulphur to distant elements has been implemented (Fig. 2C).

Mining the reconstructed network 1: combination with legacy data

In addition to its general topology and characteristics, which are indicative for properties of the given biological system, a network such as that shown in Fig. 2C contains important biological information relating to individual pathways. In this network, sulphur itself is a typical hub with a close to maximal number of direct connectors—31 (Table 1), and an even higher number of second order connectors (those which connect to direct connectors), about 200. To extract indicative routes of informational fluxes, which are directed to sulphur-responding metabolites, from such an interlaced net certain branches of connectivity were selected with the following procedure. In the upper region with regard to a certain metabolite (i.e. between sulphur and the metabolite) only those links belonging to a path 'sulphur—the metabolite' were left; in the lower region all existing links were left (Fig. 3A).

Applying the same unravelling procedure, another, now hormone-related fragment was isolated from the network (Fig. 3B).

Mining the extracted fragments provided examples of where the reconstructed network confirms and extends existing data, or leads to new knowledge. For a case in point, the positions of *O*-acetyl-serine (OAS), an important indicator of sulphate deprivation in plants (Nikiforova *et al.*, 2003; Hirai *et al.*, 2003), and *N*-acetyl-serine (NAS) were investigated. NAS is derived from OAS by non-enzymatic intramolecular *O*- to *N*-acyl migration (Flavin and Slaughter, 1965). Their relative concentrations are highly interdependent, and so both metabolites have to show coupled alterations in relative concentration levels, as is indeed reflected in their close network positioning: OAS and NAS appear in the network as closely connected vertices, equally distant from sulphur (connectors of the 2nd order, marked with green, Fig. 3A).

As another illustration, the flows of information processing regarding tryptophan were examined. In a previous study, it was suggested that the increase in tryptophan observed in sulphur-deprived plants could be explained by assuming that the reduced level of cysteine leads to a surplus in serine, which is subsequently converted into tryptophan (Nikiforova *et al.*, 2003). It has to be emphasized that this hypothesis was based on legacy data regarding biochemical pathways (for an overview of amino acid biosynthesis in plants, see Coruzzi and Last, 2000). In this respect, it is remarkable that the reconstructed network which was created solely on ungoverned analysis of co-behaviour of genes and metabolites displays only one path from sulphur to tryptophan, and this path runs via serine (Fig. 3A), therefore corroborating the hypothesis raised previously. Further, it is worth mentioning that the causal directionality 'sulphur–serine–tryptophan' could only be detected due to the centralized position of sulphur, which was implemented to the reconstructed network as *a priori* knowledge of the system exciter. This example was regarded as a striking case for the usefulness of the reconstructed network as a concept tool for identifying biological paths of communication.

Mining the reconstructed network 2: finding paths from system exciter to physiological endpoints

The systemic response of plants to sulphur deprivation results in several endpoints of biochemical pathways and physiological reactions, such as increased root formation and accumulation of anthocyanins (Nikiforova *et al.*, 2003). Alhough both physiological events are typical reactions to stress (Malamy and Ryan, 2001; Steyn *et al.*,

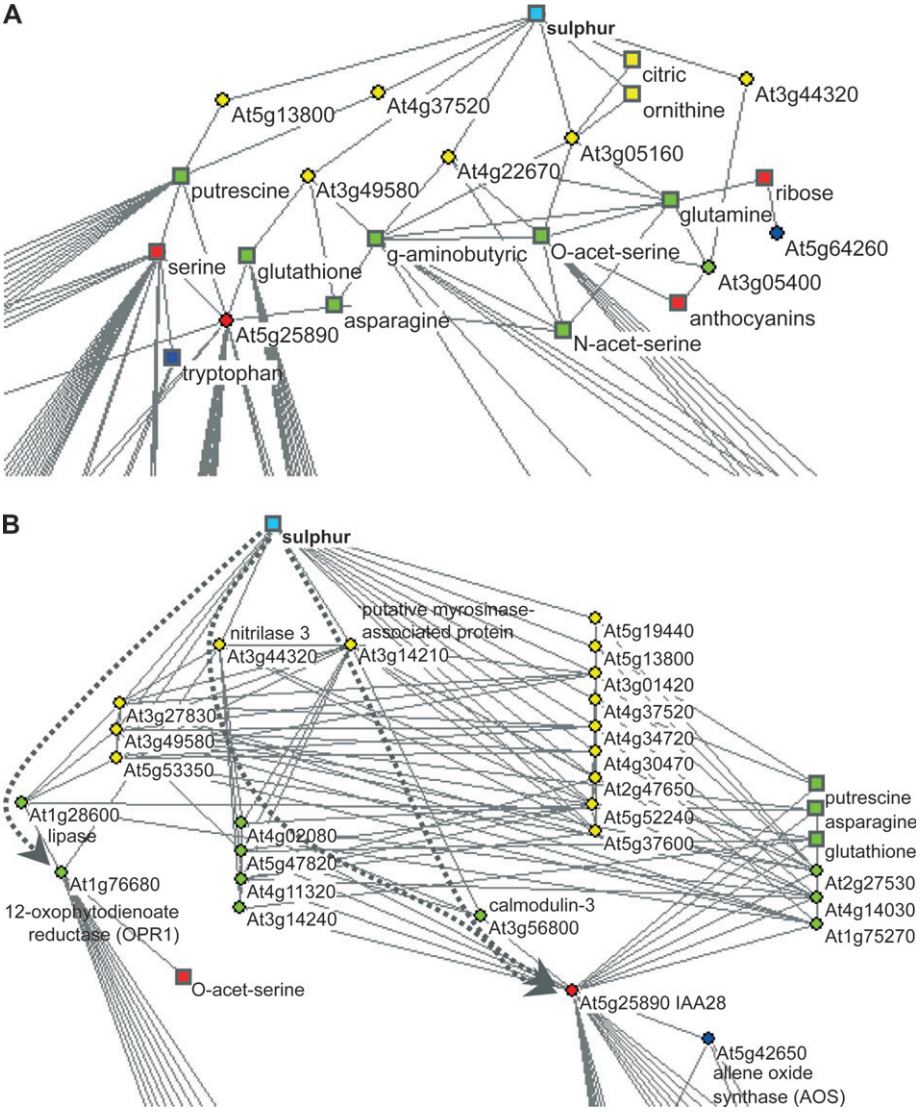

Fig. 3. Extracted fragments of the gene-metabolite response network, visualized with the Pajek computer program for large network analysis (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). Gene vertices are depicted as circles, metabolite vertices as squares. Vertices of different distance from sulphur (cyan) are marked with different colours: direct connectors are yellow, the connectors of the next orders are lime-green, then red, then blue. (A) A network fragment containing paths from sulphur to sulphur-responding metabolites. (B) A hormone-related network fragment. The jasmonate-related path and the discussed auxin-related spindle are shown with dotted arrows.

2002; Lopez-Bucio *et al.*, 2003), both do develop in sulphur-starved plants and thus they were traced in the reconstructed network.

As shown in Fig. 3A, vertex 'anthocyanins' displays only two links, and both are upstream-positioned. Among all the metabolites, this was the only one that did not possess any link to parallel or downstream vertices. Such positioning is considered as a path leading from the system exciter to a communication endpoint, in agreement with the identification of anthocyanins as a physiological endpoint of stress response.

As an example of new biological information resulting from the reconstructed network analysis, a path to another physiological endpoint is presented, i.e. enhanced root formation, represented by an auxin spindle (Fig. 3B). Starting from sulphur, several interlacing redundant paths pass nitrilase 3 and a putative myrosinase-associated protein, both involved in auxin biosynthesis, and then lead via the auxin signal transduction factor calmodulin 3 to the hub-forming node IAA28, an auxin-related transcriptional factor, known to be mainly expressed in roots (Rogg *et al.*, 2001). Analysis of IAA28 expression history in the Stanford Microarray Database (http://www.stanford.edu/microarray; Gollub *et al.*, 2003) provided experimental evidence that IAA28 is, first, involved in the auxin signal transduction cascade, and second, in a conditional manner: the IAA28 gene was significantly over-expressed after exogenous auxin treatment only in the background of the *Arabidopsis* IAA24 mutant (described in Przemeck *et al.*, 1996; Hardtke and Berleth, 1998; Berleth *et al.*, 2000)

(experiment IDs 26716-26717, experimenter T Berleth). It should be stressed that this distinction has not been articulated before, even though the array data have long been publicly available. With this example, it was demonstrated that a hypothesis created by reconstructed networks can seed specific searches for verification in array databases. In this case, the data specifically derived from the IAA24 mutant profiles are in full agreement with the supposed role of IAA28 in auxin signalling. It is important to emphasize that, with the increasing growth of publicly available databases, such experimental verification of network-derived hypotheses, arising from, for example, mutants, will become of increasing importance.

The detected auxin spindle extends previous observations regarding auxin-related signalling (Fig. 4). Sulphur-deficient homeostasis causes a surplus metabolic flux via auxin and the activation of auxin-induced genes (Hirai *et al.*, 2003; Nikiforova *et al.*, 2003). The altered auxin content triggers changes in free calcium levels in plant cells (Felle, 1988; Gehring *et al.*, 1990; Kalra and Bhatla, 1999), which are sensed by calmodulin. The interaction between auxin and calmodulin has been confirmed experimentally (Naren *et al.*, 1995; Okamoto *et al.*, 1995; Choi *et al.*, 1996; Yang and Poovaiah, 2000), thus proving the relevance of this part of the path derived from the network. As a novel finding extending this path it is proposed that activated calmodulin influences the expression of the IAA28 gene, based on their relative positions within the reconstructed sulphur response network. In turn, IAA28 represses the transcription of auxin-induced genes, as shown by a gain-of-function mutation in IAA28 (Rogg *et al.*, 2001). Thus, by incorporating the network-derived interaction between calmodulin and IAA28 the auxin signalling path has been extended to a closed circuit (Fig. 4), thus providing a new example of the feedback control of multiple informational fluxes, and new evidence for the fundamental principle of a living system as a network of inferior negative feedback subordinate to a superior positive feedback.

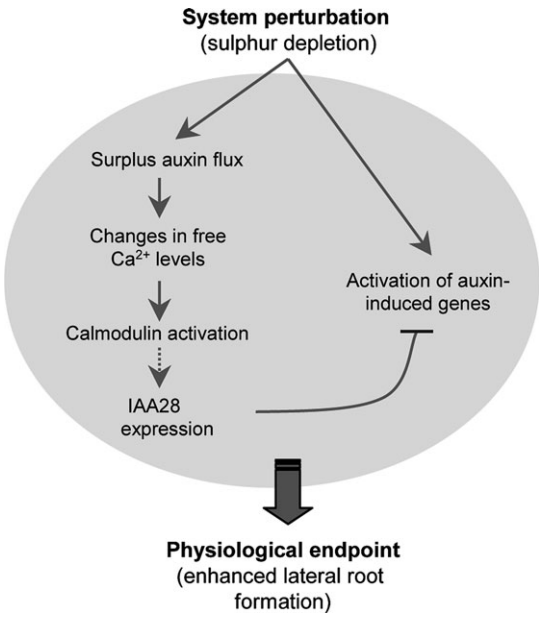

Fig. 4. The model depicts the uncovered auxin signalling circuit on the route from sulphur as system exciter to the physiological endpoint of the response development. Parts shown with plain arrows are based on legacy data, while the dashed segment of the, now closed, circuit resulted from the network analysis.

At first sight, transcriptional repression of auxin-induced genes under conditions of depleted nutrient supply, which fosters exploitation of new soil layers by enhanced root growth, seems difficult to reconcile. However, the metabolic overflow to auxin caused by sulphur deficiency via the path 'sulphur–serine–tryptophan' (proposed by Nikiforova *et al.*, 2003, and detected as an information flux in the reconstructed hypo-sulphur response network, present study) may well require negative feedback control over the expression of auxin-induced genes, which can be provided via the activity of IAA28. For the overall survival strategy of a plant as a system, feedback auxin regulation can serve as a way to save resources under conditions in which further root enforcement leads to fatal resource expenditure and, as a result, the inability to set seeds for a new generation.

Mining the reconstructed network 3: check on reliability

The implemented causal directionality allows hub input (upstream positioned) and output (downstream positioned) signals to be distinguished. If a hub is formed by a gene involved in transcriptional regulation, then in comparative promoter analysis hub output genes are expected to contain common motifs in their promoter regions. This expectation comes solely from the network topology analysis and therefore can serve as a check for reliability in network reconstruction. As an example, statistical motif analysis was undertaken in promoter sequences of IAA28 input and output genes, as hub-forming IAA28 possesses the highest number of the direct connectors among the genes with annotated transcription factor activities (Table 1). Using the tool for motif analysis available at the TAIR web page (*http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp*), twice as many common motifs were found in the queried output, than input, promoters. This result fits the expectation from our network topology, and thus supports causal directionality of the reconstructed network.

Hub-containing networks support the plait concept

The reconstructed response network also highlights the distinctions between specific and common parts of response routes. According to the plait concept first proposed by Nikiforova *et al.* (2003), the transduction of specificity during stress responses within a non-specific signalling stream is provided by the specific interlacing of

biosynthetic pathways. In the example of the auxin-related spindle (Fig. 3B), one can follow the specific interlacing of correlation paths, some of which pass calcium-sensing calmodulin. Calcium-related signalling systems are well known to be redundant (Bowler and Fluhr, 2000; Yang and Poovaiah, 2002), and involvement of calcium as an intermediate signalling molecule in responses to diverse stresses necessitates its interactions in a combinatorial fashion, resulting in a dense network of interactions. Thus, calcium among other intermediate signalling molecules can be considered as an element of the main non-specific signalling stream. Such an intermediate positioning in a signal transduction cascade may lead to cross-tolerance against multiple stresses (Bowler and Fluhr, 2000). However, a specific signal should first reach the general non-specific stream. As strands of hair interlace into a plait, biological specificity at the beginning of the signalling path connects with multiple (environmental) signals at the end, thus providing the most effective, specific reaction. End responses are connected by a non-specific main signalling stream, which contains common signalling molecules, like calcium or activated oxygen species (discussed by Bowler and Fluhr, 2000), and which probably provides multiple crosstalk connections between pathways, as described in plants (Genoud and Metraux, 1999). The analysis of the reconstructed response network showed that, besides biosynthetic pathways, other correlation paths of either known, or of as yet unknown, natures provide a flow of information from initial excitement to physiological endpoints. As an extension of the plait concept, those sulphur-specific elements belonging to the upper non-interlaced part of the plait can now be dissected. In the network, these are probably the vertices between sulphur and calmodulin, and among them nitrilase 3 was indeed shown to be highly specific for sulphur stress response (Kutz *et al.*, 2002; Hirai *et al.*, 2003; Nikiforova *et al.*, 2003).

Concluding remarks

The described approach for reconstructing gene-metabolite communication networks with implemented causality opens many research opportunities on systems behaviour. Indispensability of the hubs for information processing suggests examining alternative networks in which hub genes are knocked out. In the search for response specificity, the comparison of response networks of, for example, other nutritional stresses, will allow the identification of nutrient-specific response elements.

High-precision molecular tracking of resolved informational fluxes provides potential gains in the understanding of many responses, including stress, disease, and therapeutic effects. The advantages of this innovative approach can be demonstrated as applied to stress biology or biomedicine. When a stress/disease agent is known, tracing of signalling pathways to response endpoints through the reconstructed network as a whole, and the identification of control hubs through the analysis of network topology will highlight potential control points and lead to target drug discovery. If the stress/disease agent is unknown, causality can be implemented from downstream effects (endpoint reactions) and may lead to the identification of the agent. Taken together, reconstruction of the gene/metabolite network with implemented causal directionality provides an extension towards a consistent development of systems biology.

Acknowledgements

We thank Megan McKenzie for manuscript proofreading. This work was supported by the EU commission through funding of FP5 project QLRT-2000-00103 and by the Max Planck Society.

References

Albert R, Jeong H, Barabasi AL. 2000. Error and attack tolerance of complex networks. *Nature* **406,** 378–382.

Askenazi M, Driggers EM, Holtzman DA, *et al*. 2003. Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nature Biotechnology* **21,** 150–156.

Barabasi AL, Albert R. 1999. Emergence of scaling in random networks. *Science* **286,** 509–512.

Batagelj V, Mrvar A. 1998. Pajek—Program for Large Network Analysis. *Connections* **21,** 47–57.

Berleth T, Mattsson J, Hardtke CS. 2000. Vascular continuity, cell axialization and auxin. *Plant Growth Regulation* **32,** 173–185.

Bowler C, Fluhr R. 2000. The role of calcium and activated oxygens as signals for controlling cross-tolerance. *Trends in Plant Science* **5,** 241–246.

Butte AJ, Kohane IS. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing* **5,** 415–426.

Choi YJ, Cho EK, Lee SI, Lim CO, Gal SW, Cho MJ, An GH. 1996. Developmentally regulated expression of the rice calmodulin promoter in transgenic tobacco plants. *Molecular Cell* **6,** 541–546.

Clipsham R, Zhang YH, Huang BL, McCabe ERB. 2002. Genetic network identification by high density, multiplexed reversed transcriptional (HD-MRT) analysis in steroidogenic axis model cell lines. *Molecular Genetics and Metabolism* **77,** 159–178.

Coruzzi G, Last R. 2000. Amino acids. In: Buchanan BB, Gruissem W, Jones RL, eds. *Biochemistry and molecular biology of plants.* Rockville, Maryland: American Society of Plant Physiologists, 358–410.

Daub CO, Kloska S, Selbig J. 2003. MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics* **19,** 2332–2333.

Daub CO, Steuer R, Selbig J, Kloska S. 2004. Estimating mutual information using B-spline functions: an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* **5,** 118.

Edwards JS, Palsson BO. 2000. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences, USA* **97,** 5528–5533.

Fauchon M, Lagniel G, Aude JC, Lombardia L, Soularue P, Petat C, Marguerie G, Sentenac A, Werner M, Labarre J.

2002. Sulfur sparing in the yeast proteome in response to sulfur demand. *Molecular Cell* **9,** 713–723.

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays* **24,** 267–274.

Felle H. 1988. Auxin causes oscillations of cytosolic free calcium and pH in *Zea mays* coleoptiles. *Planta* **174,** 495–499.

Fiehn O, Weckwerth W. 2003. Deciphering metabolic networks. *European Journal of Biochemistry* **270,** 579–588.

Flavin M, Slaughter C. 1965. Synthesis of the succinic ester of homoserine, a new intermediate in the bacterial biosynthesis of methionine. *Biochemistry* **4,** 1370–1375.

Forst CV. 2002. Network genomics: a novel approach for the analysis of biological systems in the post-genomic era. *Molecular Biology Reporter* **29,** 265–280.

Gehring CA, Irving HR, Parish RW. 1990. Effects of auxin and abscisic acid on cytosolic calcium and pH in plant cells. *Proceedings of the National Academy of Sciences, USA* **87,** 9645–9649.

Genoud T, Metraux JP. 1999. Crosstalk in plant cell signalling: structure and function of the genetic network. *Trends in Plant Science* **4,** 503–507.

Gollub J, Ball CA, Binkley G, *et al*. 2003. The Stanford microarray database: data access and quality assessment tools. *Nucleic Acids Research* **31,** 94–96.

Grosu P, Townsend JP, Hartl DL, Cavalieri D. 2002. Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Research* **12,** 1121–1126.

Hardtke CS, Berleth T. 1998. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO Journal* **17,** 1405–1411.

Hirai MY, Fujiwara T, Awazuhara M, Kimura T, Noji M, Saito K. 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. *The Plant Journal* **33,** 651–663.

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K. 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **101,** 10205–10210.

Jeong H, Tombor B, Albert R, Oltval ZN, Barabasi AL. 2000. The large-scale organization of metabolic networks. *Nature* **407,** 651–654.

Kalra G, Bhatla SC. 1999. Distribution of membrane-bound calcium and activated calmodulin in cultured protoplasts of sunflower (*Helianthus annuus* L.). *Current Science* **76,** 1580–1584.

Korzeniewski B. 2001. Cybernetic formulation of the definition of life. *Journal of Theoretical Biology* **209,** 275–286.

Kutz A, Müller A, Hennig P, Kaiser WM, Piotrowski M, Weiler EW. 2002. A role for nitrilase 3 in the regulation of root morphology in sulphur-starving *Arabidopsis thaliana*. *The Plant Journal* **30,** 95–106.

Lopez-Bucio J, Cruz-Ramirez A, Herrera-Estrella L. 2003. The role of nutrient availability in regulating root architecture. *Current Opinion in Plant Biology* **6,** 280–287.

Malamy JE, Ryan KS. 2001. Environmental regulation of lateral root initiation in *Arabidopsis*. *Plant Physiology* **127,** 899–909.

McCabe ERB. 2002. Vulnerability within a robust complex system-DAX-1 mutations and steroidogenic axis development. *Journal of Clinical Endocrinology and Metabolism* **87,** 41–43.

Mueller LA, Zhang PF, Rhee SY. 2003. AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiology* **132,** 453–460.

Naren A, Prasad TG, Sashidhar VR, Kumar MU. 1995. Involvement of calcium in brassinolide and auxin-induced cell elongation. *Current Science* **69,** 777–780.

Nikiforova V, Freitag J, Kempa S, Adamik M, Hesse H, Hoefgen R. 2003. Transcriptome analysis of sulfur depletion in *Arabidopsis thaliana*: interlacing of biosynthetic pathways provides response specificity. *The Plant Journal* **33,** 633–650.

Nikiforova VJ, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R. 2005. Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of *Arabidopsis* plants. *Plant Physiology* **138,** 304–318.

Okamoto H, Tanaka Y, Sakai S. 1995. Molecular cloning and analysis of the cDNA for an auxin-regulated calmodulin gene. *Plant Cell Physiology* **36,** 1531–1539.

Przemeck GKH, Mattsson J, Hardtke CS, Sung ZR, Berleth T. 1996. Studies on the role of the *Arabidopsis* gene *MONOPTEROS* in vascular development and plant cell axialization. *Planta* **200,** 229–237.

Rogg LE, Lasswell J, Bartel B. 2001. A gain-of-function mutation in IAA28 suppresses lateral root development. *The Plant Cell* **13,** 465–480.

Rosenfeld N, Elowitz MB, Alon U. 2002. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology* **323,** 785–793.

Schuster S, Klamt S, Weckwerth W, Moldenhauer F, Pfeiffer T. 2002. Use of network analysis of metabolic systems in bioengineering. *Bioprocess and Biosystems Engineering* **24,** 363–372.

Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31,** 64–68.

Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420,** 190–193.

Steuer R, Kurths J, Daub CO, Weise J, Selbig J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18,** S231–S240.

Steyn WJ, Wand SJE, Holcroft DM, Jacobs G. 2002. Anthocyanins in vegetative tissues: a proposed unified function in photoprotection. *New Phytologist* **155,** 349–361.

Thieffry D, Huerta AM, Perezrueda E, Colladovides J. 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Eschherichia coli*. *Bioessays* **20,** 433–440.

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. MapMan: a user-driven tool to display genomics data sets onto digrams of metabolic pathways and other biological processes. *The Plant Journal* **37,** 914–939.

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR. 2003. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports* **4,** 989–993.

Yang TB, Poovaiah BW. 2000. Molecular and biochemical evidence for the involvement of calcium/calmodulin in auxin action. *Journal of Biological Chemistry* **275,** 3137–3143.

Yang TB, Poovaiah BW. 2002. A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signalling pathways in plants. *Journal of Biological Chemistry* **277,** 45049–45058.