Wolfgang Kopp

Master thesis

Correlation Matrices for the Analysis and Integration of Transcriptome and Metabolome Data Sets

Institute for Genomics and Bioinformatics,
Graz University of Technology
Petersgasse 14, 8010 Graz, Austria

Supervisor: Dr. Gerhard Thallinger

Graz, May 2012

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ...
(date) (signature)

Abstract

German

Um die komplexe Vernetzung zellulärer Systeme besser zu verstehen, sind in the letzten Jahren diverse bioinformatische Methoden entwickelt worden, welche Transkriptom- und Metabolomdaten sowohl separat als auch parallel analysieren. In dieser Diplomarbeit werden Korrelationsmatrizen zur Analyse und Integration von Zeitserien-Messungen von Transkriptom- und Metabolomdaten von *Arabidopsis thaliana* verwendet. Die Datensätze umfassen Kälteakklimatisierungs-, CO_2-Erhöhungs- und Sulfur-Defizienz-Experimente. Das Ziel dieser Diplomarbeit ist es, (1) globale Zusammenhänge zwischen Korrelationen und Molekülfunction zu analysieren und (2) regulatorische Zusammenhänge zwischen und innerhalb dieser zellulären Levels zu entdecken. Um generelle Zusammenhänge zwischen Korrelationswerten und funktioneller Verwandtschaft von Genen/Metaboliten zu bestimmen, wurden globale Korrelationsverteilungen untersucht. Allerdings haben diese Resultate nur eine geringe Aussagekraft zur Bestimmung von spezifischen biologischen Prozessen, welche unter den experimentellen Gegebenheiten dereguliert sind. Daher wurde eine neue Methode entwickelt, welche die Überrepräsentation von hohen Korrelationswerten innerhalb bzw. zwischen Gruppen von Genen/Metaboliten berechnet. Die Gruppierung von Genen und Metaboliten erfolgt durch die Einbindung von Annotationsbibliotheken. Dadurch werden funktionell verwandte Moleküle, die potenziell co- oder entgegengesetzt reguliert werden, identifiziert. Die Resultate der paarweisen Überrepräsentationsanalyse wurden in Form von Netzwerken dargestellt, in denen Knoten mit Annotationsbezeichnungen korrespondieren und Kanten die signifikante Anreicherung von hohen Korrelationswerten indizieren. Die Methode wurde auf Robustheit gegenüber Parametervariationen und Plausibilität der generierten Resultate durch Vergleich mit vorangegangenen Studien überprüft. Im Einklang mit vorhergehenden Studien konnten in allen Datensätzen Änderungen der Aminosäuren-Konzentrationen über die Zeit nachgewiesen werden. Für alle Datensätze wurden weiters Unterschiede in der Genexpression von Photosynthese, im primären Stoffwechsel sowie in der globalen Proteinzusammensetzung (z. B. durch induzierte Wachstumsprozesse) berichtet. Weiters wurden für die CO_2- und Sulfur-Defizienz-Datensätze Deregulationen, die im Zusammenhang mit pflanzlichen Abwehrprozessen stehen, entdeckt. Insbesondere führte Sulfur-Defizienz zu hohen Korrelationen zwischen Transkripten und Metaboliten in der Biosynthese für Glucosinolat. Die Methode konnte viele dieser zellulären Adaptationsprozesse identifizieren, wodurch die prinzipielle Anwendbarkeit und Nutzbarkeit gezeigt werden konnte.

Stichwörter: Korrelation, Transkriptom, Metabolom, Überrepräsentationsanalyse;

English

To understand more thoroughly the cellular intertwinedness, several computational approaches have recently been proposed in order to analyse and integrate transcriptome and metabolome data sets separately as well as in parallel. In this thesis, correlation matrices were utilized for the analysis and integration of transcriptome and metabolome time-series data sets of *Arabidopsis thaliana*. The data sets comprise of cold acclimation, CO_2 elevation and sulphur starvation experiments. The goals of the thesis are (1) to analyse whether correlation values are in general connected to the underlaying function of genes/metabolites and (2) to uncover putative regulatory dependencies among and within the cellular levels. Global correlation distributions were explored to identify a general relationship between correlation value observations and functional dependencies. However, these distributions are inappropriate to reveal particular, deregulated, biological processes under certain experimental settings. In the thesis, a novel method, which determines within or among annotation label enrichment of high correlation values, is proposed. Thereby, functionally related groups of gene/metabolites that are potentially co- or counter-regulated are identified. The results of the pairwise annotation label enrichment analysis are visualized in terms of networks, with nodes corresponding to annotation labels and edges corresponding to significantly enriched of high correlation values. The method was examined for robustness to parameter variations and plausibility of the generated results by drawing comparisons with previous studies. In concordance with previous studies, for all data sets tight correlations among amino acids were detected. Further, for all data sets, transcriptional deregulations of photosynthesis, primary metabolism as well as shifts in the global protein content were in agreement with previous reports. For CO_2 elevation and sulphur deficiency, annotation labels associated with plant defence processes were revealed in line with previous findings. In particular, for sulphur deficiency, tight correlations between metabolite and transcript levels of glucosinolate biosynthesis were revealed. These findings underline the general applicability and usefulness of the method.

Keywords: Correlation, Transcriptome, Metabolome, Enrichment Analysis;

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the era of high-throughput biotechnological methods, more and more large scale data sets are acquired which apparently open the door for a thorough understanding of molecular biological principles. However, unraveling cellular intertwinedness and systems responses as a whole is considered as a highly complex task which requires the development of sophisticated computational tools.

A vast array of methods has in recent history been developed which incorporate transcriptome data sets. For instance, to unravel gene function [1, 2] or to reverse engineer genetic regulatory networks [3, 4, 5] of coexpressed transcripts.

A more involved task is to analyse and interpret correlations among metabolites within a metabolic network. Steuer *et al.* [6] have analysed the emergence of metabolite-metabolite correlations by simulating stochastic differential equations of the *Saccharomyces cerevisiae* glycolysis pathway identified by Hynne *et al.* [7]. They argued that some neighbouring metabolites (connected by only a single reaction) might be well correlated whereas others are not. On the contrary, some distant metabolites in the network show a very strong correlation due to indirect effects. Furthermore, they described that it is not possible to reverse engineer the biochemical reaction network solely on the basis of the correlation network, since this problem is underdetermined.

In another theoretical analysis, Camacho *et al.* [8] employed the framework of metabolic control analysis to demonstrate cases for biochemical reaction which would cause high/moderate/low negative/positive correlation values. They have also pointed out the emergence of indirect effects. In addition, they performed simulations of yeast glycolysis pathway by adopting the model of Teusink *et al.* [9] to study random fluctuations in the steady state of the system.

In Krumsiek *et al.* [10], a probabilistic modelling approach was proposed which is based on partial correlation coefficient to reconstruct the conditional dependency structure on the metabolome scale. The group was able to show that many high absolute partial correlation coefficients correspond to known biochemical reactions. They have applied their method to blood serum samples of a large human cohort as well as to synthetically simulated reaction networks.

Specifically, integration of transcriptome and metabolome data has recently gained attention, since this offers the potential to uncover important regulatory relations between mRNA and metabolite abundance.

Urbanczyk-Wochniak *et al.* [11] analysed transcript profiles complemented by metabolic profiles of potato tuber systems by means of PCA and pairwise co-responses of transcripts and metabolites. In their study, they were able to show that phenotypic differences between cell lines could be discriminated by measuring metabolite levels. In contrast, changes in transcript abundance did not convey that information. Furthermore,

they concluded that pairwise high absolute transcript-metabolite correlations partially agree with the functional relationships between those pairs. On the other hand, a substantial number of high correlation values could not be linked directly to the underlaying function. To specifically draw conclusions on the emergence of these correlations one would have to conduct additional experiments. Nevertheless, the generated pieces of information could provide testable hypotheses for further research [11].

Bradley *et al.* [12] investigated the global stress response of the metabolome in comparison to the transcriptome of *Saccharomyces cerevisiae* using SVD. The research group showed that the first right singular vectors of both transcriptome and metabolome were significantly correlated to each other. As the observed correlations depend on the environmental conditions, they probabilistically modeled the relation between functional annotations (e.g. common KEGG pathway), observed correlation coefficients and the environmental condition. Furthermore, they tested for enrichment of gene ontology terms among genes which are highly correlated to some common metabolite. The group found that in general functionally related genes were well correlated to common metabolites. In particular, the group was able to identify new regulatory interactions between metabolites and genes (e.g. they predicted a relationship between FBP and VID24).

Takahashi *et al.* [13] employed linear dynamical systems (LDS) to analyse the transition timings of the cellular states by time-series experiments of transcript and metabolite levels of *Escherichia coli*. They identified concerted changes of transcriptome and metabolome in terms of timing with respect to the experimental stimulus. Furthermore, they used correlation networks to examine the enrichment of GO terms for biological processes among genes which are connected to some common metabolite.

Pir *et al.* [14] applied partial least squares (PLS) to a *Saccharomyces cerevisiae* data set to address three aspects: (1) to model metabolic levels as a function of transcript abundance for several experimental conditions (e.g. different medium composition, growth rate and gene knockouts), (2) to discriminate experimental conditions based on measured data and (3) to identify relations between open reading frames (ORFs) and metabolic data.

Bylesjö *et al.* [15] investigated wild-type hybrid aspen (*Populus tremula x Populus tremuloides*) data by an extension of PLS, namely (O2PLS), which is a combination of projections to latent structures (OPLS) with orthogonal signal correction (OSC), to decompose the variability of transcriptome and metabolome measurements into the joint variability, the unique variability within each data set and the residual variability. Their approach enabled the identification of strong correlations among transcripts (e.g. sharing a common GO term) and particular types of metabolites (e.g. carbonhydrates, lipids, etc.)

Dutta *et al.* [16] investigated the cellular response to elevated CO_2 concentration with *Arabidopsis thaliana* liquid cultures. They have performed a systems biology approach to study changes in transcriptional and metabolic levels. They reported shifts in photosynthetic processes and primary metabolism as well as effects on ethylene dependent signaling, which are important in biotic and abiotic response processes.

Kaplan *et al.* [17] analysed cold acclimation of *Arabidopsis thaliana* using time-series measurements of mRNA and metabolite abundance. Kaplan's group revealed new important regulatory changes in amino acid biosynthesis, GABA biosynthesis, sucrose metabolism, raffinose metabolism and carbonhydrate levels in response to the environmental stress.

Hirai *et al.* [18, 19] conducted sulphur starvation experiments with *Arabidopsis thaliana*. The group has employed a co-clustering analysis (BL-SOM) of transcript and metabolite levels. They reported similar response patterns for genes and metabolites of

glucosinolate biosynthesis. Further, they identified changes gene expression and metabolite abundance in several metabolic process i.e. photosynthetic processes and primary metabolism.

Redestig *et al.* [20] proposed a novel correlation measure based on hidden Markov models (HMM), which especially accounts for time shifts in time-series measurements, motivated by the fact that Pearson's correlation coefficients is sensitive to time-shifted profiles and noise. They applied this new correlation measure to four *Arabidopsis thaliana* data sets to integrated metabolite and transcript profiles and showed that it performed better than the lagged Pearson's correlation coefficient. In case of high noise and high shifts in time, the HMM-based correlation performed even better than ordinary Pearson's correlation. However, for a low noise level and no time shifts among the profiles, Pearson's correlation was still better than any other correlation measure. Redestig *et al.* [20] have shown that pairwise correlations performed poorly for the purpose of predicting functional connectivity (e.g. common reaction pathway) of gene-metabolite pairs. Therefore, they employed OPLS discriminant analysis (OPLS-DA) [21] to discriminate the common variability of metabolites and transcripts within KEGG pathways in contrast to uncorrelated variability.

Other methods which elucidate the relation between stochastic signals are i.e. non-parametric correlation measures (e.g. Spearman's rank correlation coefficient or Kendall's tau coefficient), the mutual information or similarity measures like the euclidean distance. Non-parametric correlation coefficients apply to data sets which are relate to each other in a non-linear way and are more robust to noise and outliers than Pearson's correlation. On the other hand, their statistical power is considerably lower than Pearson's correlation. The strength of mutual information relies on capturing non-linear relationships between the two signals, which was thoroughly analysed in Steuer *et al.* [22]. However, mutual information suffers from an even lower statistical power than non-parametric methods. Hence, both of the latter approaches require a much larger amount of data to allow for reliable significance assertions.

Popularity of network-based analysis approaches have also grown in recent history. For instance, Maere *et al.* [23] proposed the Cytoscape plug-in BiNGO which allows the estimation of overrepresented GO terms in a set of genes (e.g. subgraph of a biological network). The tool provides the opportunity to map significantly overrepresented GO terms onto the hierarchical structure of the ontology.

In a different approach, Bindea *et al.* [24] proposed the Cytoscape plug-in ClueGO to visualize functionally related terms as networks. ClueGO maps genes onto the annotation labels (e.g. KEGG pathways or GO terms) which in turn are linked to one another by utilizing kappa statistics. Furthermore, ClueGO can be used to analyse the annotation term composition of co-clustering of genes (e.g. based on the gene-expression profiles).

1.1 Goals and Outline

The major goal of the thesis was to use correlation matrices for the analysis and integration of several transcriptome and metabolome data sets in the light of functional annotation libraries. The data sets comprise of parallel microarray and mass spectrometry time-series measurements of mRNA and metabolite abundancies, respectively, for the (1) sulphur deficiency in roots [19], (2) sulphur deficiency in leaves [19], (3) CO_2 elevation [16] and cold acclimation [17].

Specifically, the following should be investigated:

- Is there a general connection between functional relationship and the observed correlation value distribution? Therefore, the global correlation value distribution is analysed in the light of functional annotation libraries.
- Which groups of molecules (e.g. transcripts or metabolites with a common function) are deregulated and how do the deregulated processes related to each other? To address this question, a novel enrichment analysis approach shall be introduced.
- Different variations of the introduced enrichment analysis approach shall be compared.
- The quality and plausability of the enrichment analysis results should be assessed on the four data sets by conducting a literature survey.

In the first part of the thesis, global correlation value distributions of three types of correlation matrices (transcript-transcript, metabolite-transcript and metabolite-metabolite) were analysed by integrating various annotation libraries (protein-protein interactions, protein domains, AraCyc pathways, KEGG pathways and GO terms for biological processes) of four different data sets. In more detail, conditional probabilities of observing a correlation value given that the gene/metabolite pairs share an annotation label were constructed. The goal of this analysis was to elucidate the global connection between functional relationship and the observed correlation values. Further, the predictive power of correlation values to infer functional similarity between a pair of molecules were examined.

In the second part, to address the question of which functionally related molecules are deregulated and how these processes related to each other, a novel enrichment strategy is introduced which was developed throughout the thesis. The algorithm takes a full correlation matrix and a set of annotation labels as an input and detects (pairs of) biological annotation categories (e.g. GO terms or KEGG pathways) for which highly positive or negative correlation values are statistically enriched. The enrichment method relies on the determination of the overrepresentation of bin counts in the histogram of observed correlation values. Of particular importance were the histogram bins at the tails of the correlation value range, since they are thought to be most valuable of conveying information of the underlaying biological system. Hence, the algorithm putatively reveals co- or counter-regulated functionally related groups of molecules. Several alternative statistical tests such as bin-specific p-value estimation as well as estimating the alternative distribution via Jackknife or bootstrap resampling were compared. Furthermore, the robustness of the algorithm's output with respect to different histogram bin sizes was investigated, which revealed high agreement of the top ranked results of the algorithm variants.

In the last part of the thesis, the results of the application of the novel enrichment algorithm to four transcriptome and metabolome time-series experiments are presented (sulphur deficiency in roots and leaves [19], cold acclimation [17] and CO_2 elevation [16]). For each data set, a range of combinations of correlation matrices (e.g. transcript-transcript or metabolite-transcript) and annotation libraries (e.g. KEGG pathways or GO terms) were analysed. Additionally, the enrichment of single annotation labels as well as pairs of annotation labels was determined. Importantly, to ease the interpretation process, the results of the pairwise enrichment approach were visualized as networks with nodes corresponding to the annotation labels and links corresponding to the statistically significant enrichment of high absolute correlation values.

Finally, for each data set, a literature survey was conducted to assess the quality and plausability of the produced results, which proved the general applicability and usefulness of the proposed approach.

Chapter 2

Methods

2.1 Dataset

Throughout the thesis four data sets were used, each consisting of time-series profiles of microarray experiments and measurements of metabolite levels using mass spectrometry. All data sets stem from measurements in *Arabidopsis thaliana* and have already been subject to filtering and normalization. The number of time steps and sampling points varied among the data sets. The data sets were downloaded from the web[1].

The cold stress response data set was generated by Kaplan *et al.* [17] on Affymetrix ATH1 microarrays and GC-TOF mass spectrometry in order to study regulatory features of the metabolome and transcriptome profiles due to cold acclimation.

The sulfur deficiency experiments in leaves and roots were conducted by Hirai *et al.* [19] utilizing custom Agilent arrays and FT-mass spectrometry complemented by HPLC and capillary electrophoresis.

The CO_2 response data were acquired using TIGR microarrays and GC-TOF mass spectrometry by Dutta *et al.* [16]. Table 2.1 briefly summarizes the experimental settings. For a more detailed description, consult the stated references.

Table 2.1: The summary of the data sets lists the number of transcripts and metabolites after filtering. The number within the parenthesis correspond to known metabolites, as opposed to the total number of measured metabolites to the left. For each experiment the time points are listed.

Description	No. Trans.	No. Metab. (Tot./Annot.)	Timepoints [h]	Source
Cold	6680	302 (87)	0/1/4/12/24/48/96	[17]
CO_2	7138	284 (76)	0/1/3/6/9/12/18/24/30	[16]
S-def. root	7342	43 (43)	3/5/12/24/48/168	[19]
S-def. leaf	7342	28 (28)	3/5/12/24/48/168	[19]

The experimental data sets are integrated with several annotation databases, which were partly obtained from TAIR[2] [25] and partly from bioconductor annotation packages [26]. Firstly, a set of protein-protein interactions (PPIs) was downloaded from AtPIN [27], which is available at TAIR. Secondly, a collection of protein domains (PDs) consisting of entries from various databases and protein domain extraction applications was downloaded from TAIR. As a large part of the protein domains from different databases

[1] `http://www.cin.ufpe.br/~igcf/Metabolites`
[2] `http://www.arabidopsis.org`

are redundant, a particular emphasis shall be placed on InterPro annotations [28], since those where used for the enrichment approach, which is introduced later in this thesis.

The third annotation resource consists of AraCyc reaction pathways [29] (available at TAIR) which assigns genes to their biochemical reaction pathway focusing on Arabidopsis species. In addition, a hierarchical organized component classification system was downloaded from KEGG BRITE using the web interface ('components with biological roles')[3].

2.2 Framework

All programming was carried out using the statistical programming environment R, Version 2.12.2 [30].

Apart from the resource at TAIR, KEGG pathways were incorporated into the analysis from the bioconductor annotation package KEGG.db, Version 2.4.5 [31] and gene ontology (GO) terms were incorporated from GO.db, Version 2.4.5 [32]. From GO, biologcial processes (GO-BP) was used. Venn diagrams were drawn using the R package VennDiagram, Version 1.0.1 [33].

The results of the enrichment of correlation values among pairs of annotation labels were visualized using Cytoscape, Version 2.8.1 [34]. Cytoscape is a network visualization tool, which one the one hand is easy to use and, on the other hand, provides a broad range of visualization features (e.g. node and edge annotations).

2.3 Notation & Nomenclature

Throughout the thesis Pearson's correlation measure will be used to measure the correlation between pairs of molecules. As described in the introduction section, one reason for this choice is that other method (e.g. non-parametric correlation measures or the mutual information) have a lower statistical power, which is in particular problematic for small sample sizes (6-9 biological replicates) as is the case for the data sets herein. Note however, in general other correlation measures would also be applicable. Pearson's correlation coefficient between the i^{th} and j^{th} molecule (e.g. transcript or metabolite levels) over n time points is given by following formula

$$(R)_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_{i.})(x_{jk} - \overline{x}_{j.})}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \overline{x}_{i.})^2 \cdot \sum_{k=1}^{n}(x_{jk} - \overline{x}_{j.})^2}} \tag{2.1}$$

In this expression, x_{ik} denotes the measurement of molecule i at time point k and $\overline{x}_{i.}$ denotes the estimated mean across all time points for the same molecule.

In particular, microarray measurements for the transcript levels are summarized by the matrix $T \in \mathbb{R}^{m \times n}$, where m corresponds to the number of measured transcripts and n corresponds to the number of experimental conditions. Similarly, the mass spectrometry measurements of the metabolites shall be denoted by $M \in \mathbb{R}^{p \times n}$ with p denoting the number of metabolites.

Based on these measurement matrices, the correlation matrices are constructed by applying Formula 2.1 for all molecule pairs of the measurement matrices. Thus, for each biological condition a transcript-transcript correlation matrix denoted by $R_{TT} \in [-1, 1]^{m \times m}$, a metabolite-metabolite correlation matrix denoted by $R_{MM} \in [-1, 1]^{p \times p}$ and a metabolite-transcript correlation matrix denoted by $R_{MT} \in [-1, 1]^{p \times m}$ are computed.

[3] `http://www.genome.jp/kegg/brite.html`

The proposed method integrates correlation matrices with biological annotation databases (e.g. KEGG pathways, GO terms, InterPro domains, etc.) to identify single annotations and pairs of annotations for which high correlation values are enriched within this subpart of the correlation matrix. The entirety of significantly enriched annotation pairs is visualized as a graph structure, which shall for convenience be referred to as **enrichment network** herein. In the enrichment network, nodes represent the annotation label and edges indicating enriched high correlation values between the nodes.

2.4 Global Correlation Structure and Functional Relationship

To evaluate the correlation structure on a global scale, knowledge about functional relationship was integrated with the correlation matrices for this analysis. For all experimental conditions, R_{TT}, R_{MM} and R_{MT} was calculated using Pearson's correlation measure. The range of possible correlation values $[-1, 1]$ was discretized to 10 equally sized bins. Subsequently, the conditional probabilities of the form

$$P(Corr = c|common\ functional\ property) \tag{2.2}$$

were computed for R_{TT} and the following annotation sources:

- Protein Interactome (Set of all PPIs)
- Protein Domains (PDs)
- AraCyc pathway
- KEGG pathway
- Gene Ontology terms for biological processes (GO-BP)

Note, that for this analysis, all available protein domain were used (e.g. InterPro, Pfam, HMMTigr). Further, the conditional probabilities for R_{MM} and R_{MT} were integrated with KEGG pathways.

For example, to integrate gene ontology terms, the conditional probabilities were computed by the following expression

$$P(Corr = c|common\ GO\ term) = \frac{count(Corr = c, any\ common\ GO\ label)}{count(any\ common\ GO\ label)} \tag{2.3}$$

and

$$P(Corr = c|no\ common\ GO\ term) = \frac{count(Corr = c, no\ common\ GO\ label)}{count(no\ common\ GO\ label)} \tag{2.4}$$

Where $count(.)$ denotes the number of joint or marginal counts over all pairs of genes / metabolites for the data sets.

2.5 Enrichment Analysis

The algorithm (see Section 3.2) is applied to the combinations of correlation matrices and annotation library shown in Table 2.2 for all data sets. Note that for the enrichment analysis with PDs only InterPro domains were used, firstly, because of the high redundancy among the PD databases and, secondly, InterPro is the most comprehensive PD library in terms of the number of PD entries. For the within and among annotation label enrichment analysis on R_{TT} only annotation labels with at least 10 gene assignments and at most 100 gene assignments were used. For the within and among annotation label enrichment analysis on R_{MM} only KEGG pathways with at least 5 metabolite assignments were used. For the enrichment analysis on R_{MT} only pathways with at least 5 metabolites/genes and at most 100 metabolites/genes were used.

Table 2.2: Enrichment within/among functional groups. For all data sets, the combination of annotation library and correlation matrix used for the analysis is indicated by a tick. For the intergration of KEGG BRITE classes (compounds with biological roles), only within annotation label enrichment was performed. Moreover, only InterPro domains were considered for the enrichment analysis based on protein domains, on the one hand, because of the high redundancy compared to other protein domain databases and, on the other hand, since InterPro is the most comprehensive protein data base available at TAIR.

	AraCyc	KEGG	PDs	GO-BP	KEGG BRITE
R_{TT}	✓	✓	✓	✓	
R_{MT}		✓			
R_{MM}		✓			✓

The results of the within and between annotation label enrichment analysis were stored in terms of tables with the associated p-values or $P(erroneous\ decision)$ for each bin. For all cases, only the bins at the tail of the histogram were considered for the enrichment analysis (e.g. $[-1, -0.8)$ and $[0.8, 1]$ or $[-1, -0.4)$ and $[0.4, 1]$). The significance level was chosen to be $\alpha = 5\%$ for the p-values as well as for $P(erroneous\ decision)$ in all cases. P-values were adjusted for multiple testing using FDR [35]. It is important to note that different variants of the algorithm were used and presented throughout the results chapter in order to keep the density of the network at a level which is easy to interpret (see Tables 2.3, 2.4, 2.5 and 2.6 for the chosen algorithm variants). In the case of pairwise enrichment analysis, the ensemble of significantly enriched pairs of annotation labels were visualized in terms of a network with nodes corresponding to annotation labels and edges indicating the significant enrichment of high absolute correlation values, with green and red edges corresponding to enrichment of positive and negative correlation values, respectively. The networks offer great benefits for the systems-wide interpretation process. Significant pairwise interactions are stored in a SIF (simple interaction file) and which are subsequently loaded into Cytoscape [34] for the purpose of visualization. Instead of the annotation label identifiers, the annotation label descriptions were shown to the user (e.g. 'Glycolysis / Gluconeogenesis' instead of KEGG ID '00010'). Hence, the enrichment networks were augmented by node description attributes (*.NA files (node attribute file)) from the annotation databases (e.g. GO description in GO.db or pathway

Table 2.3: Analysis of the S-def root data set: The different algorithm variants are listed with the used parameter settings for the S-def in roots data set. The CV intervals indicate the interval which was tested for enrichment. The chosen statistical test is given in the fifth's row. In case of truncation, the 120 smallest p-values (60 for the positive and negative interval, respectively) were visualized.

Corr. Matr.	Annot.	within / between	CV intervals	Test
R_{MM}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MM}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MT}	KEGG	both	$[-1,-0.8)$, $[0.8,1]$	Jackknife
R_{TT}	AraCyc	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	between	$[-1,-0.8)$, $[0.8,1]$	Jackknife
R_{TT}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	Bootstrap
R_{TT}	GO	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	GO	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation
R_{TT}	PD	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	PD	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation

Table 2.4: Analysis of the S-def leaf data set: The different algorithm variants are listed with the used parameter settings for the S-def in leaves data set. The CV intervals indicate the interval which was tested for enrichment. The chosen statistical test is given in the fifth's row. In case of truncation, the 120 smallest p-values (60 for the positive and negative interval, respectively) were visualized.

Corr. Matr.	Annot.	within / between	CV intervals	Test
R_{MM}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MM}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MT}	KEGG	both	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	between	$[-1,-0.8)$, $[0.8,1]$	Jackknife
R_{TT}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	Bootstrap
R_{TT}	GO	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	GO	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation
R_{TT}	PD	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	PD	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation

Table 2.5: Analysis of the cold acclimation data set: The different algorithm variants are listed with the used parameter settings for the cold acclimation data set. The CV intervals indicate the interval which was tested for enrichment. The chosen statistical test is given in the fifth's row. In case of truncation, the 120 smallest p-values (60 for the positive and negative interval, respectively) were visualized.

Corr. Matr.	Annot.	within / between	CV intervals	Test
R_{MM}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MM}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{MT}	KEGG	both	$[-1,-0.8)$, $[0.8,1]$	Bootstrap
R_{TT}	AraCyc	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	between	$[-1,-0.8)$, $[0.8,1]$	Jackknife
R_{TT}	KEGG	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	KEGG	between	$[-1,-0.8)$, $[0.8,1]$	Bootstrap
R_{TT}	GO	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	GO	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation
R_{TT}	PD	within	$[-1,-0.8)$, $[0.8,1]$	bin-specific p-values
R_{TT}	PD	between	$[-1,-0.4)$, $[0.4,1]$	bin-specific p-values + truncation

Table 2.6: Analysis of the CO_2 elevation data set: The different algorithm variants are listed with the used parameter settings for the carbon dioxide data set. The CV intervals indicate the interval which was tested for enrichment. The chosen statistical test is given in the fifth's row. In case of truncation, the 120 smallest p-values (60 for the positive and negative interval, respectively) were visualized.

Corr. Matr.	Annot.	within / between	CV intervals	Test
R_{MM}	KEGG	within	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{MM}	KEGG	between	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{MT}	KEGG	both	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	within	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{TT}	AraCyc	between	$[-1,-0.4)$,$[0.4,1]$	bin-specific p-values
R_{TT}	KEGG	within	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{TT}	KEGG	between	$[-1,-0.8)$,$[0.8,1]$	Bootstrap
R_{TT}	GO	within	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{TT}	GO	between	$[-1,-0.4)$,$[0.4,1]$	bin-specific p-values + truncation to 120 top ranked pairs
R_{TT}	PD	within	$[-1,-0.8)$,$[0.8,1]$	bin-specific p-values
R_{TT}	PD	between	$[-1,-0.4)$,$[0.4,1]$	bin-specific p-values + truncation to 120 top ranked pairs

description in KEGG.db) which were automatically generated by an R-script.

2.6 Analysis of Related Metabolites

Biologically related chemical compounds were grouped according to KEGG BRITE database 'Compounds with biological roles', which is available on the web[4] (see Table 2.2). This database hierarchically groups KEGG compound into biologically related compound classes (e.g. amino acids, monosaccharides, etc.). Within compound class enrichment analysis for high correlation values was conducted by means of the proposed algorithm in Section 3.2.

[4] `http://www.genome.jp/kegg/brite.html`

Chapter 3

Results

3.1 Analysis of the Distribution of Correlation Values

This section covers the examination of the distribution of correlation values for each of the four data sets. The main question, which shall be addressed here, is whether the distribution of correlation values depends on the functional relationship of two molecules (e.g. genes tend to be positively correlated if they participate in a similar function). The analysis is performed on the correlation matrices R_{TT}, R_{MM} and R_{TM} for each data set separately. For R_{TT}, protein-protein interactions, protein domains, GO terms, KEGG pathways and AraCyc pathways are integrated into the analysis. R_{MM} and R_{TM} are examined using KEGG pathways.

Figures 3.1, 3.3 and 3.2 illustrate the conditional and marginal probabilities of the correlation values given the particular biological categories. Tables 3.1, 3.2 and 3.3 summarize the numbers of pair instances which belong to a particular group.

The analysis drawn on R_{TT} (see Figure 3.1) suggests that some of the annotation libraries are particularly useful for explaining a certain fraction of the high correlation values. As can be seen, in all experimental conditions, the conditional distributions of correlation values given the existence of protein-protein interactions, a common protein domain or a common AraCyc pathways differ markedly from the marginal distribution (shown in black). Surprisingly, gene pairs with common GO terms or common KEGG pathways yield conditional distributions which are essentially equal to the marginal distribution. Similarly, the conditional distribution of correlation values given 'no common biological annotation' is very similar to the marginal distribution. Based on these observations the distribution of correlation values is (almost) conditionally independent of common GO terms and KEGG pathways. More formally,

$$
\begin{aligned}
P(Corr|\textit{no common annotation}) &\approx P(Corr|\textit{common GO term}) && (3.1)\\
&\approx P(Corr|\textit{common KEGG ID}) && (3.2)\\
&\approx P(Corr). && (3.3)
\end{aligned}
$$

Unfortunately, although there is apparently a shift towards positive correlation in the conditional distributions for an observed PPI, common protein domain or common AraCyc pathways (as opposed to KEGG pathways and GO terms), it is not possible to infer i.e. a common biological function or physical interaction based on the correlation values according to the above mentioned probability distribution. For instance, consider the problem if inferring $P(\textit{existing PPI}|Corr > 0.9)$, which might be rephrased

as 'How probable is it that the pair of proteins, associated with the transcripts, physically interact with each other if the correlation value > 0.9 was observed?'. This can be addressed by Bayes theorem in a straight forward manner. However, the issue which arises here is that the prior probability of observing a pair sharing a functional annotation $P(\textit{common annotation label})$ is much lower than the prior probability of observing no common annotation label $P(\textit{no common annotation label})$. For instance, observing a pair of genes with known or putative protein-protein interactions on R_{TT} is given by

$$\frac{P(\textit{existing PPI})}{P(\textit{no existing PPI}) + P(\textit{existing PPI})} = \frac{4235}{24765123} \approx 10^{-4} \tag{3.4}$$

(compare Table 3.1 for cold data set). Hence, this renders the inference task impossible.

Nevertheless, this analysis confirms a connection between annotation labels (e.g. for AraCyc) and observed correlation values, which underlines the potential of integrating correlation matrices with biological annotations. A noteworthy point is the observation of a huge fraction of highly positive and negative correlation values in the cold acclimation data set, which result from the fact that a large fraction of genes are differentially expressed by the cold acclimation (see Figure 3.1a). This is particularly problematic for the interpretation of specific pairwise correlations, since there are overwhelmingly many high correlation values resulting from indirect effects.

Table 3.1: Counts of transcript-transcript pairs which share a particular annotation

	Cold stress	CO_2 stress	S-def. leaf	S-def. root
# Genes	6680	7138	7342	7342
# Pairs	24765123	23328734	24925810	24925688
# Existing PPIs	4235	2539	4789	4789
# Common PDs	124620	16553	8665	8665
# Common AraCyc pathways	4923	2600	10652	10652
# Common GO terms	6914944	6201557	6053876	6053834
# Common KEGG IDs	43098	38539	42588	42588

Table 3.2: Counts of metabolite-metabolite pairs which share a particular annotation

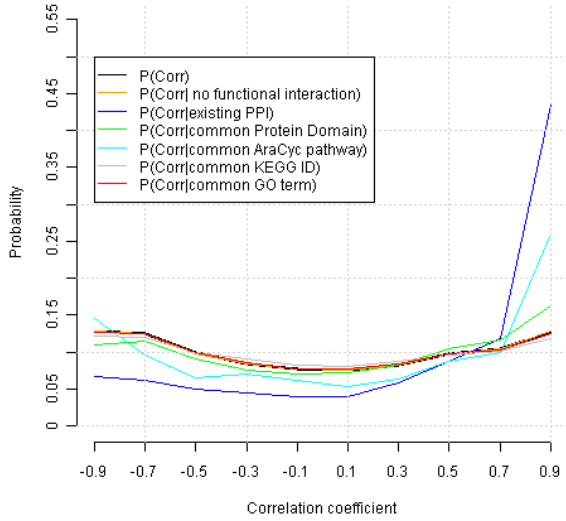
	Cold stress	CO_2 stress	S-def. leaf	S-def. root
# Metabolites	302	284	43	28
# Pairs	45451	40186	378	903
# Common KEGG IDs	1232	842	114	234

Table 3.3: Counts of transcript-metabolite pairs which share a particular annotation

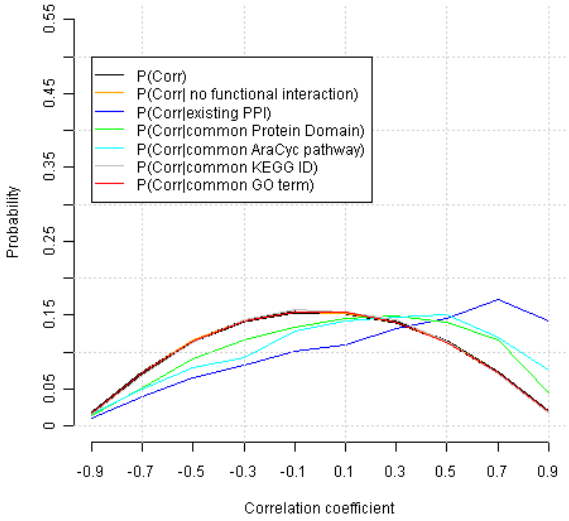
	Cold stress	CO_2 stress	S-def. leaf	S-def. root
# Metabolites	302	284	43	28
# Genes	6680	7138	7342	7342
# Pairs	1703400	2027192	205576	315706
# Common KEGG IDs	9706	9807	3453	4335

Next, the probability distributions over correlation values on R_{MM} are examined (see Figure 3.2). In all four conditions, the conditional probability of observing a high correlation value (e.g. $C >= 0.8$) is higher given that they share a KEGG pathway relative

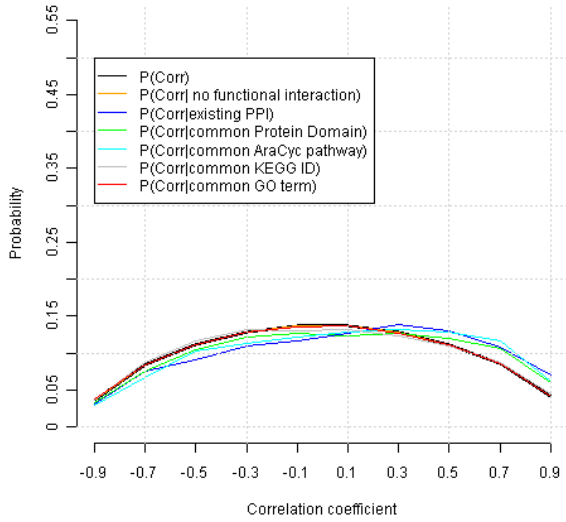
(a)



(b)



(c)



(d)

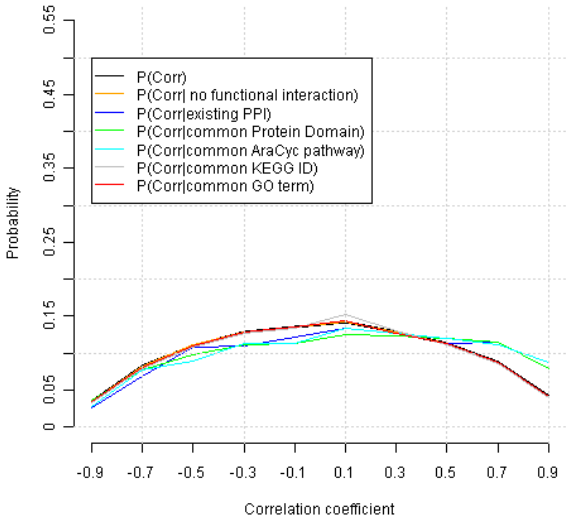
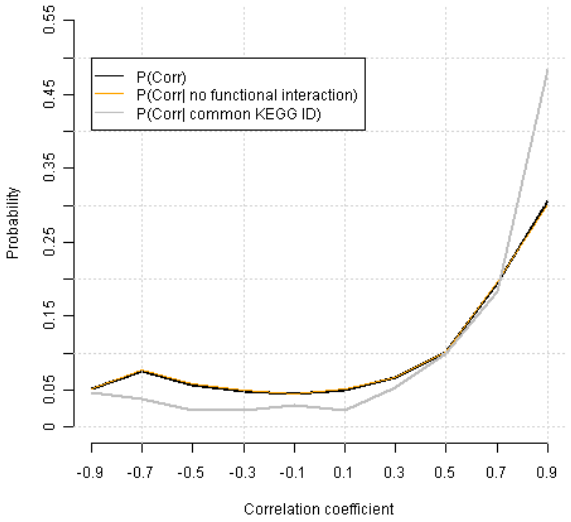
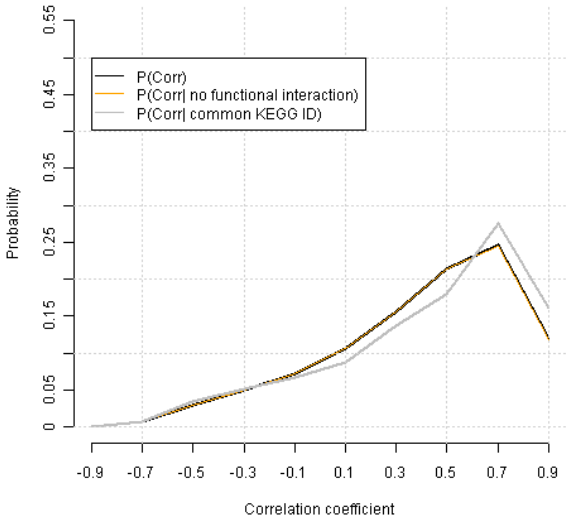


Figure 3.1: Probabilities of transcript-transcript correlations given various forms of functional dependence for each experimental condition. (a), (b), (c) and (d) show the correlation value distributions on the cold data set, the CO_2 data set and the sulphur deficiency in leaves and roots, respectively. The marginal distribution is shown in black. The conditional probabilities of observing some correlation value given known or putative protein-protein interaction, shared protein domains, shared AraCyc pathways and shared GO terms are shown in blue, green, turquoise and red, respectively. The conditional probability of observing a correlation value given that there is no functional relationship is shown in orange.

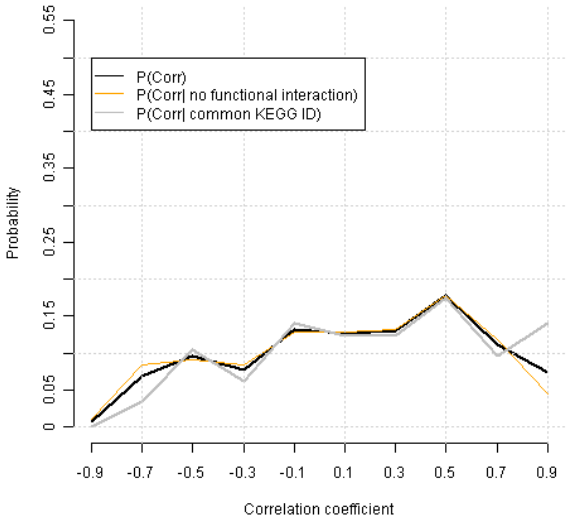
(a)



(b)



(c)



(d)

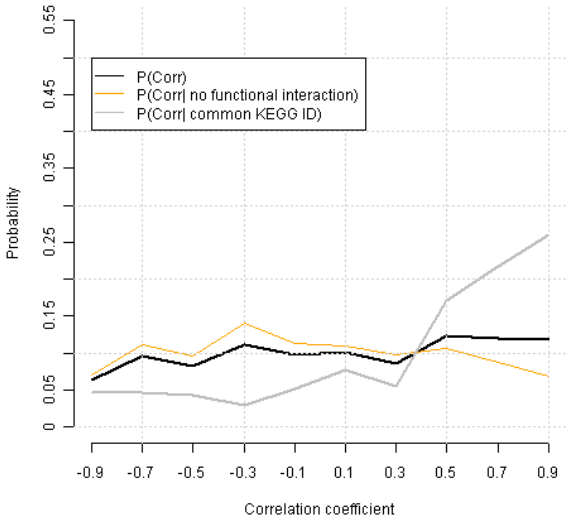
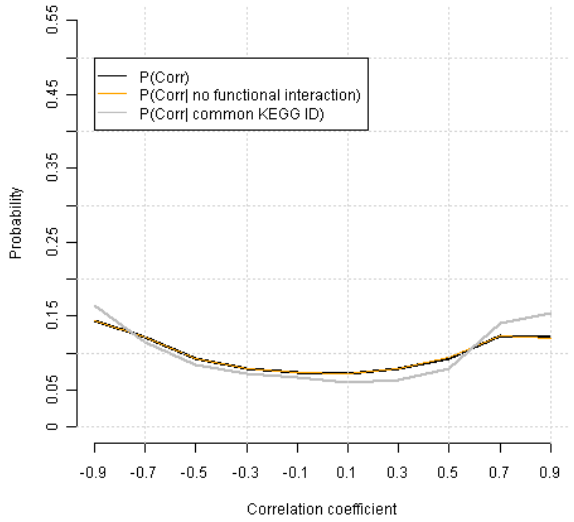


Figure 3.2: Probabilities of metabolite-metabolite correlations. (a), (b), (c) and (d) show the correlation value distributions on the cold data set, the CO_2 data set and the sulphur deficiency in leaves and roots, respectively. The marginal distribution of observing some correlation value is shown in black. The conditional probability of observing some correlation value given a common KEGG pathway for the metabolites is shown in light gray. The conditional probability of observing a correlation value given that there is no functional relationship corresponds to the orange curve.

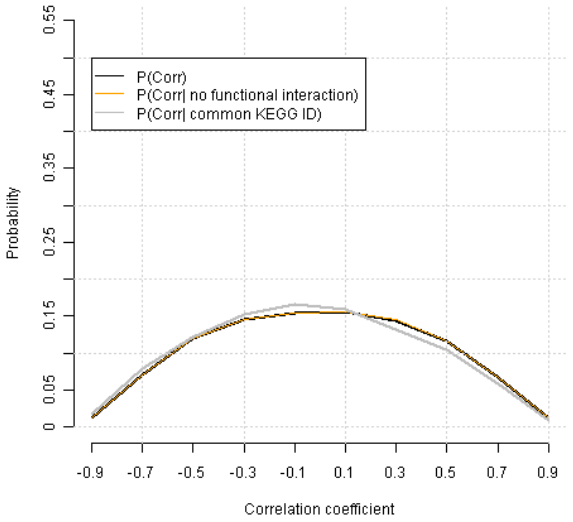
to the marginal distribution. In particular, for the cold acclimation and the sulphur starvation measured in roots a considerable difference between the probabilities is present, whereas, for the remaining two data sets, the probabilities differ only slightly. For the cold acclimation and the CO_2 elevation the marginal distribution of correlation values is markedly shifted towards positive values. This observation is far less present in sulphur starvation data sets.

The last part of this section turns to the analysis of the probability distribution of correlation values on R_{MT} (see Figure 3.3). The conditional probability of observing a correlation value for a metabolite-transcript pair with shared KEGG pathway is very similar to the marginal distribution of correlation values for all data sets. This suggests that in general there is no clear connection between the observed metabolite-transcript correlation values and the pathway co-occurance of these pairs. Hence, metabolite-transcript correlation values are much harder to interpret.

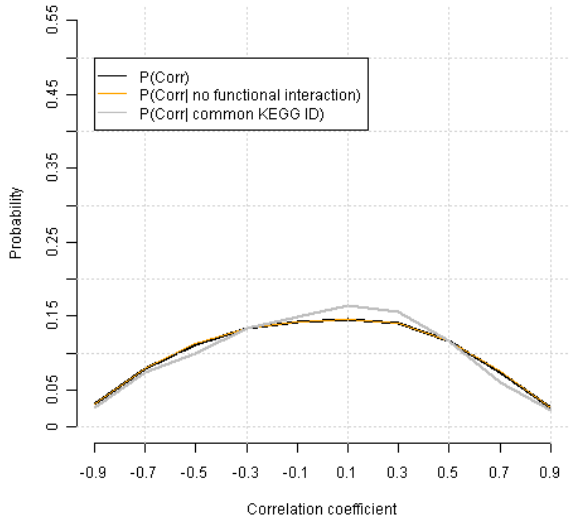
(a)



(b)



(c)



(d)

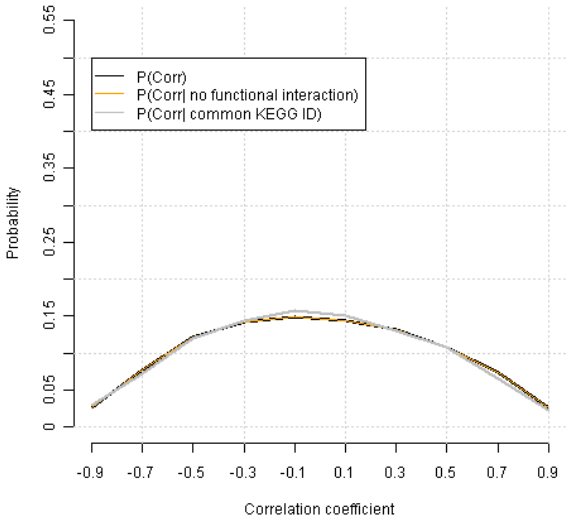


Figure 3.3: Probabilities of metabolite-transcript correlations. (a), (b), (c) and (d) show the correlation value distributions on the cold data set, the CO_2 data set and the sulphur deficiency in leaves and roots, respectively. The marginal correlation value distribution is shown in black. The conditional correlation value distribution given a common KEGG pathway for the metabolites/genes is shown in light gray. The conditional probability of observing a correlation value given that there is no functional relationship corresponds to the orange curve.

3.2 Algorithm - Enrichment of Correlation Values in Specific Regions of the Correlation Matrix

In the previous section, the global properties of correlation profiles conditioned on the functional categories was analysed in detail. However, this does not provide any information on which biological processes are subject to regulatory activity in a specific experimental context. Thus, the goal of this section is to give the outline for a method which discovers biological annotations or pairs of annotations that exhibit a closer statistical relation than would be expected by chance. The method uses the correlation values between groups of molecules (e.g. with the same annotation label) to test for statistical overrepresentation of highly positive or negative correlation values.

The algorithm requires a full correlation matrix (e.g. R_{TT}, R_{MM} or R_{MT}) and an annotation library. Basically, the method can be structured in four phases. (1) A sub-correlation matrix is extracted for the current annotation label; (2) a histogram of the correlation values in the submatrix is generated; (3) permute the molecule labels for 1000 times, extract the sub-correlation matrix and draw a histogram for each permuted instance; (4) use the counts for each bin over all permuted histogram to compute a p-value for each bin. Figure 3.4 illustrates the algorithm for within annotation label enrichment analysis in terms of a flow chart. Similarly, Figure 3.5 schematically depicts the procedure for the enrichment analysis of pairs of annotation labels. In addition, Figure 3.5 also shows a fifth's phase, which was performed manually. In this phase the resulting files (in simple interaction file (SIF)), which are generated by the enrichment procedure, were loaded into the Cytoscape environment followed by adapting the visualization. Throughout the thesis, several variants of the algorithm were considered, which shall be explained in more detail in the subsequent sections.

3.2.1 Phase 1: Generation of the Correlation Matrix

The algorithm requires a full correlation matrix (e.g. R_{MM}, R_{MT} or R_{TT}) as well as an annotation library. The generation of the entire correlation matrix is described in the methods section. This section discusses the properties of the submatrices which are extracted from the full matrix.

R_{TT} and R_{MM} are used for both within and among annotation label enrichment analysis. For the within annotation label version, a submatrix of the full correlation matrix is used, which corresponds to the currently tested annotation label (e.g. all gene which correspond to a particular KEGG pathway). This submatrix is quadratic and symmetric, hence, only the upper-right (or equivalently in the lower-left) triangular part of the matrix is used for the analysis. This is illustrated schematically in Figure 3.6a.

For the among annotation label enrichment, the submatrix consists of all cross correlations between the molecules of the two annotation labels (see Figure 3.6b). In general this submatrix is neither symmetric nor quadratic, therefore, all pairs are used for the analysis. Importantly, molecules which are assigned to both annotation labels are removed in this step, since those are deterministically set to one and would therefore bias the analysis.

In addition to the previously discussed correlation matrices, the algorithm is also used for R_{MT}. In this case, within and among annotation label enrichment analysis is performed in the same run. In general, R_{MT} is neither symmetric nor quadratic. Similarly as above, for the current annotation label or pair of labels all metabolite-transcript correlation are extracted (e.g. all metabolites and all transcripts of glycolysis). Since the entire matrix is not symmetric, the submatrix is also not symmetric, hence, all correlation values are used for the enrichment analysis (see Figure 3.6c). If the annotation label is the

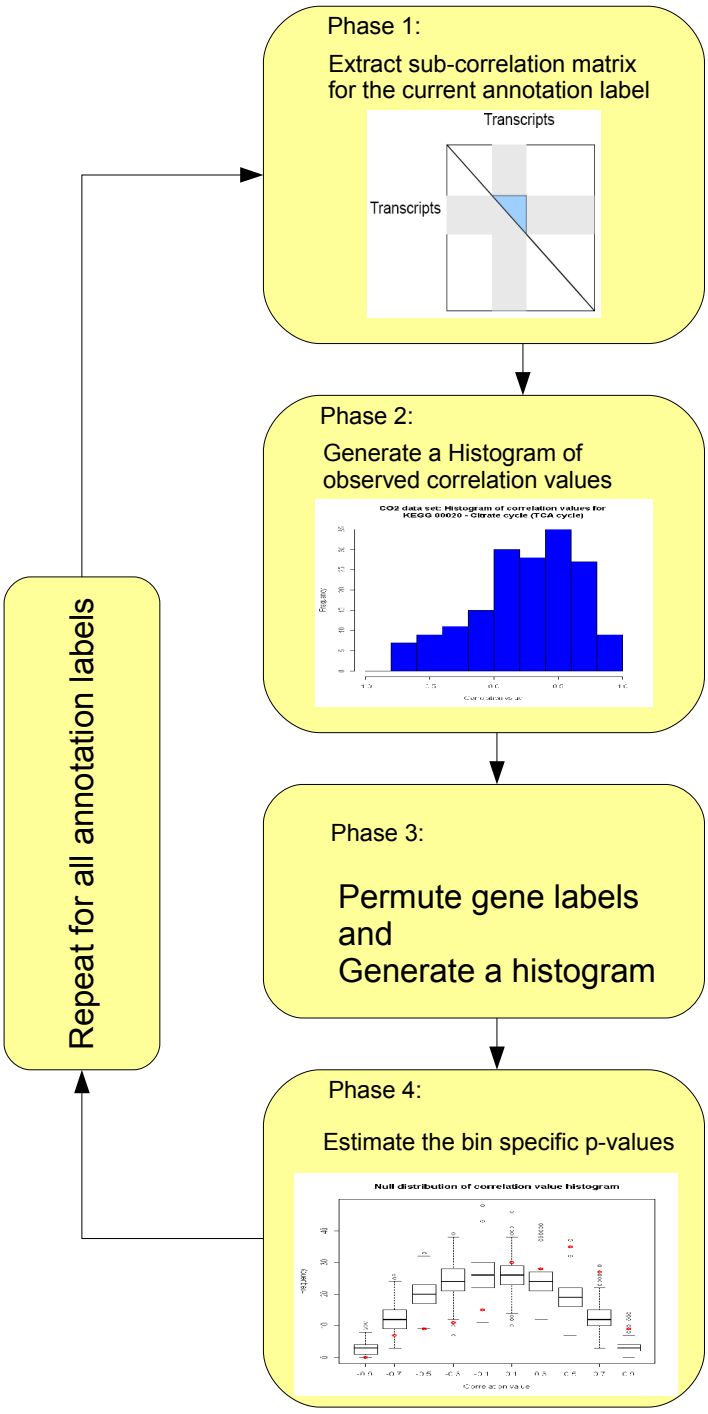


Figure 3.4: Schematic outline of the algorithm for the evaluation of enriched correlation values within functional categories. The algorithm requires an annotation library (e.g. KEGG pathways) and a correlation matrix (e.g. R_{TT}). It repeatedly extracts a submatrix (light blue, phase 1) which corresponds to the current annotation label (light gray, phase 1) and generates a histogram of observed correlation values. Subsequently, it generates a null distribution for each bin of the histogram by a permutation approach, which in turn is used to compute the p-values.

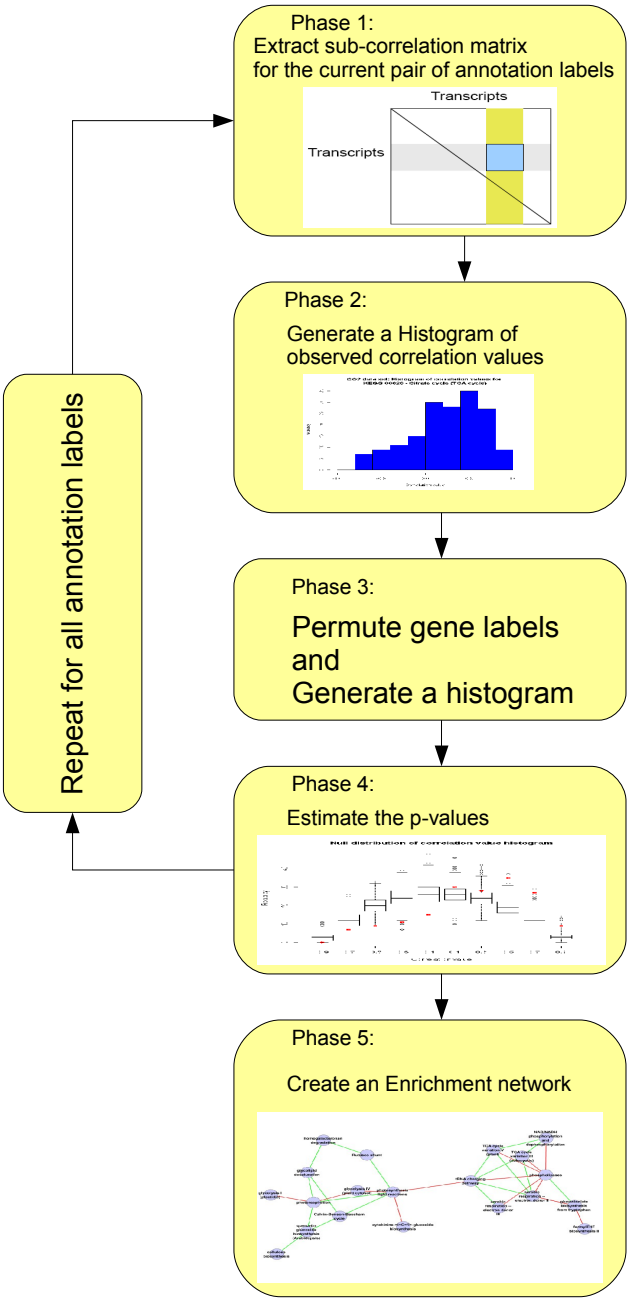


Figure 3.5: Schematic outline of the algorithm for the evaluation of enriched correlation values among functional categories. The algorithm requires an annotation library and a correlation matrix. It repeatedly extracts a submatrix (light blue, phase 1) which corresponds to the current pair of annotation labels (light gray and beige, phase 1) and generates a histogram of observed correlation values. Subsequently, it generates a null distribution for each bin of the histogram by employing a permutation approach, which in turn is used to compute the p-values. In the enrichment network, edges correspond to significantly enriched high correlation value between the adjacent nodes (annotation labels). It is important to note that only the tail bins of the histogram are considered to build the enrichment network.

same for metabolites and transcripts, there is only one possible submatrix configuration to be tested. In contrast, for the between annotation label enrichment there are two submatrices to be tested: (1) Metabolites of label A correlated with transcripts of label B and (2) metabolites of label B correlated with transcripts of label A. Moreover, for between annotation label enrichment, all metabolites and transcripts that occur in both pathways are removed from the analysis.

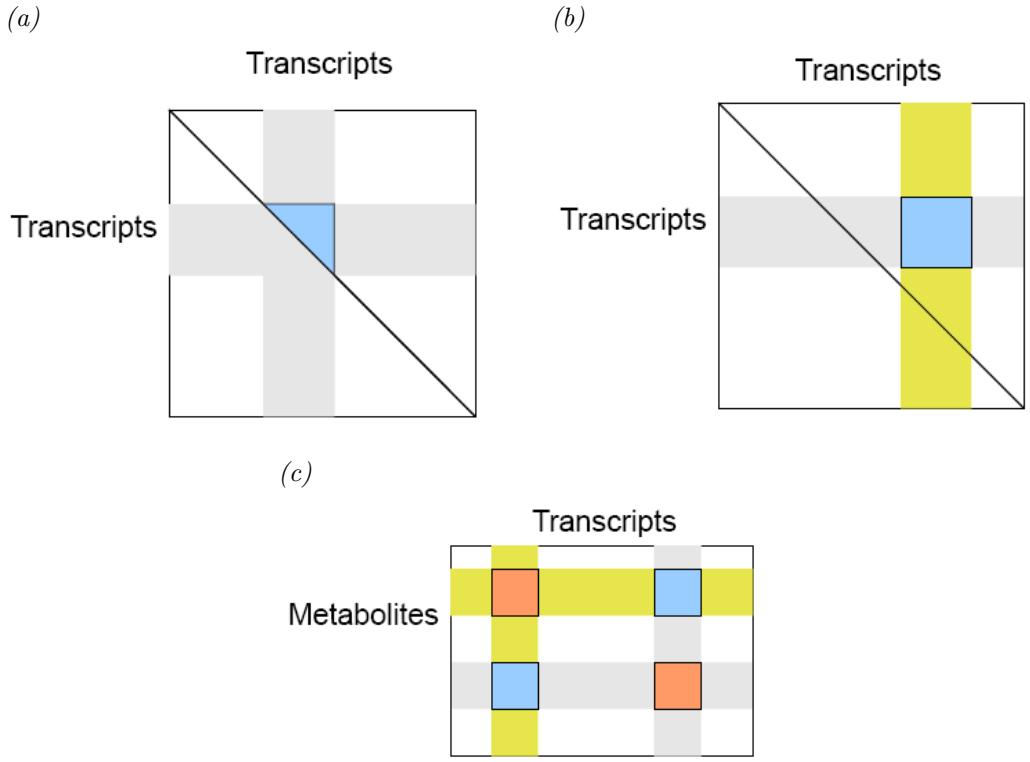


Figure 3.6: Schematic representation of the submatrix extraction. (a) and (b) are squared and symmetric matrices (e.g R_{MM} or R_{TT}), which is indicated by the diagonal line. (c) in general is non-squared and non-symmetric (e.g. R_{MT}). The shaded regions correspond to correlations of molecules that correspond to a particular pathway. In (a), the interest lays on correlation values within a particular functional category (light blue region), since the submatrix is symmetric, the analysis is drawn on the upper triangular part of the matrix. In (b), the interest lays on the distribution of correlation values between molecules of two distinct categories. This submatrix is in general non-squared and non-symmetric and hence, the whole submatrix needs to be taken into consideration for the further analysis. In (c), different data sources are integrated by correlation matrices, thus, the data types are map onto the same functional annotations (e.g. biochemical pathway for transcripts and metabolites). The entire matrix is non-symmetric and non-squared, thus, the four submatrices (orange submatrices and light blue submatices) are again non-symmetric and non-squared. Note that although the figures exemplify transcript-transcript correlations, the same applies i.e. to metabolite-metabolite correlation matrices.

3.2.2 Phase 2: Generation of an Histogram of Observed Correlation Values

With the submatrix (see discussion above), a histogram over the correlation values is computed. Therefore, the number of bins over the correlation value interval needs to be

specified. Importantly, this thesis claims that the histogram profile is indicative of co-regulatory processes for a prespecified groups of molecules. That is, i.e. for co-responding molecules the frequency or probability mass of correlation values is shifted towards positive values. Conversely, reciprocal influence manifests in a shift towards negative correlation values (e.g. pathway A is induced, while pathway B is silenced).

To address the question of which bins are most informative to draw a statistical test on, two variants are compared. Further, the robustness of the algorithm with respect to this parameter is assessed. (1) The bin intervals corresponding to $[-1, -0.8)$ and $[0.8, 1]$ were used for the enrichment analysis and (2) bin intervals ranging from $[-1, -0.4)$ and $[0.4, 1]$ were used. The results of this analysis are presented in Section 3.3.

3.2.3 Phase 3: Permutation of the Molecule Labels to Generate a Null Distribution of Correlation Value Histograms

To decide whether an annotation label or a pair of annotation labels represents a biologically relevant association, a statistical test was developed relying on permutation sampling. To construct the null distribution, gene or metabolite labels are randomly permuted for 1000 times. For each permutation step the histogram is computed. The estimated count distribution for each bin is subsequently fitted to a gamma distribution. The parameters of the gamma distribution are stored and reused for annotation labels with the same set cardinality $N_L = \left|\{m \in M : isAttributedWith(m, L)\}\right|$ with M representing the set of all molecules (e.g. all spotted microarray probes). That is, i.e. the number of genes assigned to KEGG pathway L. Consequently, permutation sampling only needs to be done once for each annotation label with the same cardinality N_L, which yields computational efficiency, if taken into account.

3.2.4 Phase 4: Estimation of the P-value or P(Erroneous decision)

The null distribution is estimated for each bin separately implying bin-specific p-value estimation. Note that only the bins representing the tails of the correlation value range are considered for the enrichment analysis, rather than those in the middle of the range, since the former once are thought to be much more valuable for elucidating changes in biological processes (e.g. $[0.8, 1]$ is biologically more relevant than $[0, 0.2)$). In this section four variants to assess the statistical significance of the observed bin counts are presented.

The first way of testing for overrepresentation of CVs is by means of estimating the bin-specific p-values using the gamma distribution which is given by

$$\text{p-value} \quad = \int_{C_{observed}}^{\infty} \frac{\beta^{\alpha}}{\gamma(\alpha)} x^{\alpha-1} \exp^{-\beta x} \, dx \tag{3.5}$$

with $C_{observed}$ denoting the observed number of correlation values in the bin of interest. α denotes the shape parameter and β the rate, which, as mentioned above, depend on the cardinality of the annotation label N_L. The p-values are adjusted for multiple testing using FDR [35].

In the second approach, in addition to the estimated null distribution the alternative distribution is estimated, again by utilizing a sampling strategy, namely bootstrapping. Based on the observed submatrix entries for a particular annotation label, the correlation values are resampled with replacement for 100 times. Each time, a histogram is generated

which in the end gives rise to the alternative distribution. Analog to the null distribution, the alternative distribution is represented in terms of a fitted gamma distribution. Hence, the following quantity is estimated using the null and alternative distribution

$$P(erroneous\ decision) = P(false\ negative) + P(false\ positive) \tag{3.6}$$

$$= \int_0^t p(x|\alpha_{bs\_a}, \beta_{bs\_a})dx + \int_t^\infty p(x|\alpha_{null}, \beta_{null})dx \tag{3.7}$$

With α_{bs_a} and β_{bs_a} representing the parameters of gamma distribution of the bootstrapped alternative distribution, whereas, α_{null} and β_{null} represent the parameters of the null distribution. The parameter t corresponds to the significance level which is found by minimizing the above expression with respect to t. For the optimization the *optimize* method of R is utilized. Note however that bootstrapping assumes independently and identically distributed (i.i.d.) samples which is obviously not the case for correlation values. In other words, the CVs are inherently related with each other. In this respect the approach violates the statistical independence assumption. Nevertheless, as discussed in the case studies, the results for this test approach are comparable to the other significance tests.

Similar to the second approach, the third test again makes use of a sampling strategy to estimate the alternative distribution. In this case, however, 100 Jackknife samples are drawn such that each time one row and one column are left out from the original submatrix, which correspond to leaving out one or two molecules for the within or between annotation label enrichment analysis, respectively. This approach corrects for the violated independence assertion of the bootstrap version. The resulting alternative distribution is again fitted to a gamma distribution. Significant instances are computed analogously to the bootstrapped version by estimating $P(erroneous\ decision)$. Note, that since Jackknife reduces the originally observed submatrix, the null distribution is also estimated for the reduced submatrix.

The fourth method, though only of minor importance for the rest of the thesis, is a weighted χ^2-test. As opposed to the previous approaches which estimate the significance for each bin separately, this variant seeks to summarize the information over the entire histogram to generate one p-value. The quantile is computed by

$$\chi^2_{test} = \sum_{i=1}^{N} \frac{(w_i \cdot O_i - w_i \cdot E_i)^2}{w_i \cdot E_i} \tag{3.8}$$

with N denoting the number of histogram bins, O_i denoting the observed count for bin i and E_i representing the expected number of correlation values in bin i. E_i results from the null distribution (see above). The weights were arbitrarily chosen to be the squared values of the correlation values for the bin means (e.g. bin 1 corresponds to the interval $[-1, -0.8)$ for which the weight is set $w_1 = 0.9^2$). The intention for the weighted χ^2-test was to place more emphasis on the tails of the histogram (e.g. high absolute correlation values), while at the same time down-weighting small correlation values. The degree of freedom was set to $\#bins - 1$. The resulting p-value was adjusted for multiple testing using FDR [35]. Note that this approach also suffers from the violated i.i.d. assumption.

3.3 Analysis of the Algorithm's Output

3.3.1 Comparison of Outcomes for Different Statistical Tests

Four different statistical tests have been designated to test for significant enrichment of correlation values: (1) Bin-specific p-value estimation, (2) bootstrapped and (3) Jackknife $P(\textit{erroneous decision})$ estimation as well as (4) p-value estimation according to weighted χ^2-test. Thus, the aim of this section is to determine the overlap between the test results.

The comparison was drawn on the CO_2 data set for within annotation label enrichment on R_{TT} using all annotation databases.

Table 3.4: Numbers of statistically significant categories on the CO_2 data set. The significance level was set to 5% for all tests. The comparison of the bin-specific statistical test variants was drawn on the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[-0.6, -0.4)$ as well as , $[0.4, 0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$. Furthermore, weighted χ^2 used the entire histogram, consisting of 10 equally sized bins, for the test.

Functional category	AraCyc	KEGG	GO	InterPro
Number of Tests	62	81	283	84
Gamma dist. + adj. p-value	11	23	33	44
Bootstrapped P(erroneous decision)	4	9	12	21
Jackknife P(erroneous decision)	8	23	39	38
Weighted χ^2	12	20	21	40

The significance assertions of the bin-specific tests were compared for agreement for the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[-0.6, -0.4)$ as well as , $[0.4, 0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$, separately. Furthermore, the weighted χ^2-test used the entire histogram. The bin-specific tests were defined to be in concordance with the weighted χ^2 variant if (1) weighted χ^2-test indicated significance and the bin-specific variant results in significance for at least one bin and (2) neither weighted χ^2-test nor the bin-specific variant (for any of the bins) indicate significance.

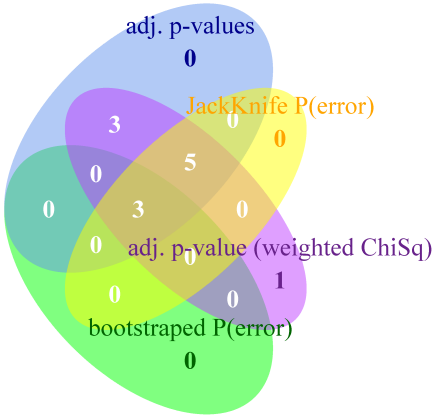
Using the p-value estimation for single bins, the p-value estimation relying on the weighted χ^2-test and $P(\textit{erroneous decision})$ estimation according to the Jackknife procedure, a comparable number of significant categories were found (see Table 3.4). Generally, bootstrapped $P(\textit{erroneous decision})$ seems to be the most stringent among the four proposed test, because it yields the smallest number of significant instances.

Next, the inspection of the intersection of the produced results reveals strong concordance with varying levels of stringency among the methods (see Figure 3.7). For all annotation libraries, virtually all significance assertions of bootstrapped $P(\textit{erroneous decision})$ were also indicated by all other tests. Furthermore, the fraction of significant annotation labels is approximately in the range of the defined significance threshold for the bootstrapped $P(\textit{erroneous decision})$. Based on the Venn diagrams, a significant overlap among results can be observed. For instance, the fractions of significant instances found by at least three of the tests are 0.53 (AraCyc), 0.79 (KEGG), 0.7 (InterPro domains) and 0.44 (GO) (compare Figure 3.7).

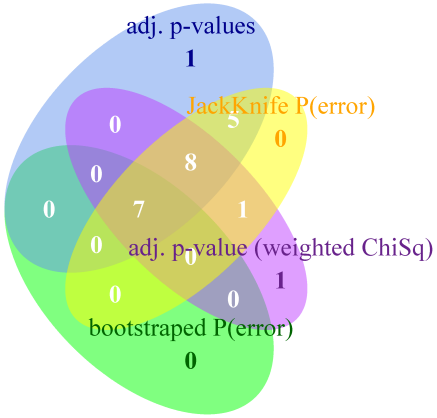
3.3.2 Analysis of the Null Distribution

This section explores the appropriateness of the gamma distribution for fitting the null distribution. In particular, two annotation labels shall be compared with markedly dif-

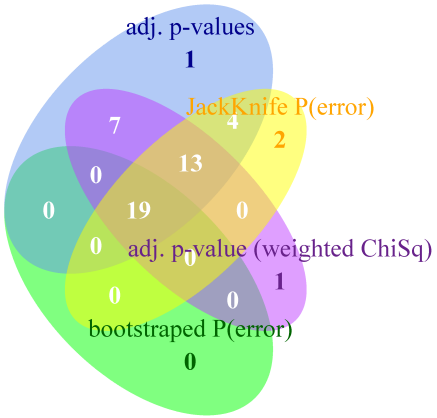
(a)



(b)



(c)



(d)

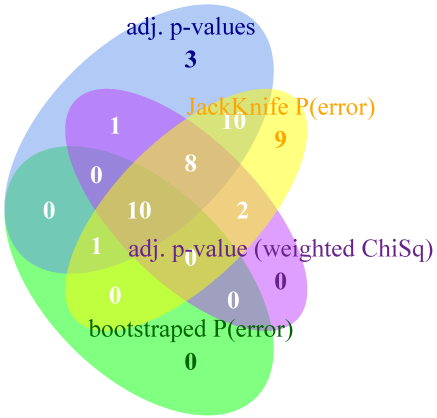


Figure 3.7: Venn diagrams showing the concordance between the results of the four discussed statistical tests for (a) AraCyc pathways, (b) KEGG pathways, (d) GO terms and (c) InterPro domains. All tests were performed for within annotation label enrichment for the CO_2 data set.

ferent numbers of molecule assignments N_L to determine any biases due to variation of the set sizes.

The analysis is performed for the AraCyc pathways 'Gluconeogenesis' and 'Pentose phosphate pathway' on for the sulphur deficiency in roots data set on R_{TT}. For both, 'Gluconeogenesis' and 'Pentose phosphate pathway', the comparison of the bin-specific null distributions with their corresponding gamma-fitted null distributions reveals that the shapes of the distributions are very similar (see Figures 3.8 and 3.9). There were no adverse effect on the shape similarity caused by the particular bin choice (compare different bins in Figures 3.8 and 3.9) as well as due to different annotation set cardinalities (e.g. pentose phosphate pathway has set cardinality $N_{PPP} = 11$, whereas, gluconeogenesis has set cardinality $N_{Gluc} = 30$). Furthermore, the histograms of correlation values are markedly different for the two AraCyc pathways (compare Figures 3.8a and 3.9a) which underlines the appropriateness of the choice of the gamma distribution.

Another way of comparing the similarity of distributions is to generate Q-Q plots. Figures 3.8 and 3.9 shows the corresponding Q-Q plots for the CV intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$ of AraCyc 'Gluconeogenesis' and 'Pentose phosphate pathway', respectively. The plots underscore the similarity of the distributions, though for high quantiles the distributions seem to disagree slightly.

The last point in this section illustrates the p-value estimation for the examples of within pathway enrichment analysis of AraCyc pathway 'Gluconeogenesis' and 'Pentose phosphate pathway' for sulphur deficiency, R_{TT} and CV interval $[0.8, 1]$ (see Figures 3.8b and 3.9b). P-value estimation is based on the fitted gamma distribution (red). The green area indicates the 5% significance level and the blue vertical line marks the observed number of correlation values for the correlation value interval $[0.8, 1]$. Additionally, the histogram displays the number of observed correlation values due to permutation sampling for this bin. According to the statistical test, 'Gluconeogenesis' is statistically enriched for highly positive correlation values, whereas, 'Pentose phosphate pathway' is not enriched for highly positive correlation values.

3.3.3 Analysis of the Bootstrap- and Jackknife-based Statistical Test

This section examines the statistical tests which are based on resampling of the observed correlation values to estimate the alternative distribution. Firstly, the general adequacy of these resampling strategies shall be illuminated. Secondly, the representation of the resampled alternative distributions in terms of gamma distributions is inspected.

As described in Section 3.2.4, the bootstrapped estimation of $P(erroneous\ decision) = P(false\ negative) + P(false\ positive)$ is performed by sampling correlation values with replacement from the observed submatrix of the correlation matrix. This corresponds to independently picking samples of pairs of molecules from the submatrix. The analysis revealed that the bootstrapped alternative distribution is adequately represented by a fitted gamma distribution (see Figure 3.10b). An illustration of a statistical test using bootstrap resampling of the alternative distribution is shown in Figure 3.10a for the KEGG pathway 00010, CV interval $[0.8, 1]$, CO_2 elevation and R_{TT}. The bootstrapped alternative distribution (blue) is shown in contrast to the null distribution (red + histogram). The significance threshold t was chosen such that $P(erroneous\ decision)$ (green area) is minimized.

Jackknife sampling seems to be a more natural choice to introduce variability to the observed correlation matrix, because it preserves relationships among correlation values within the matrix. In contrast, bootstrap sampling may result in correlation value his-

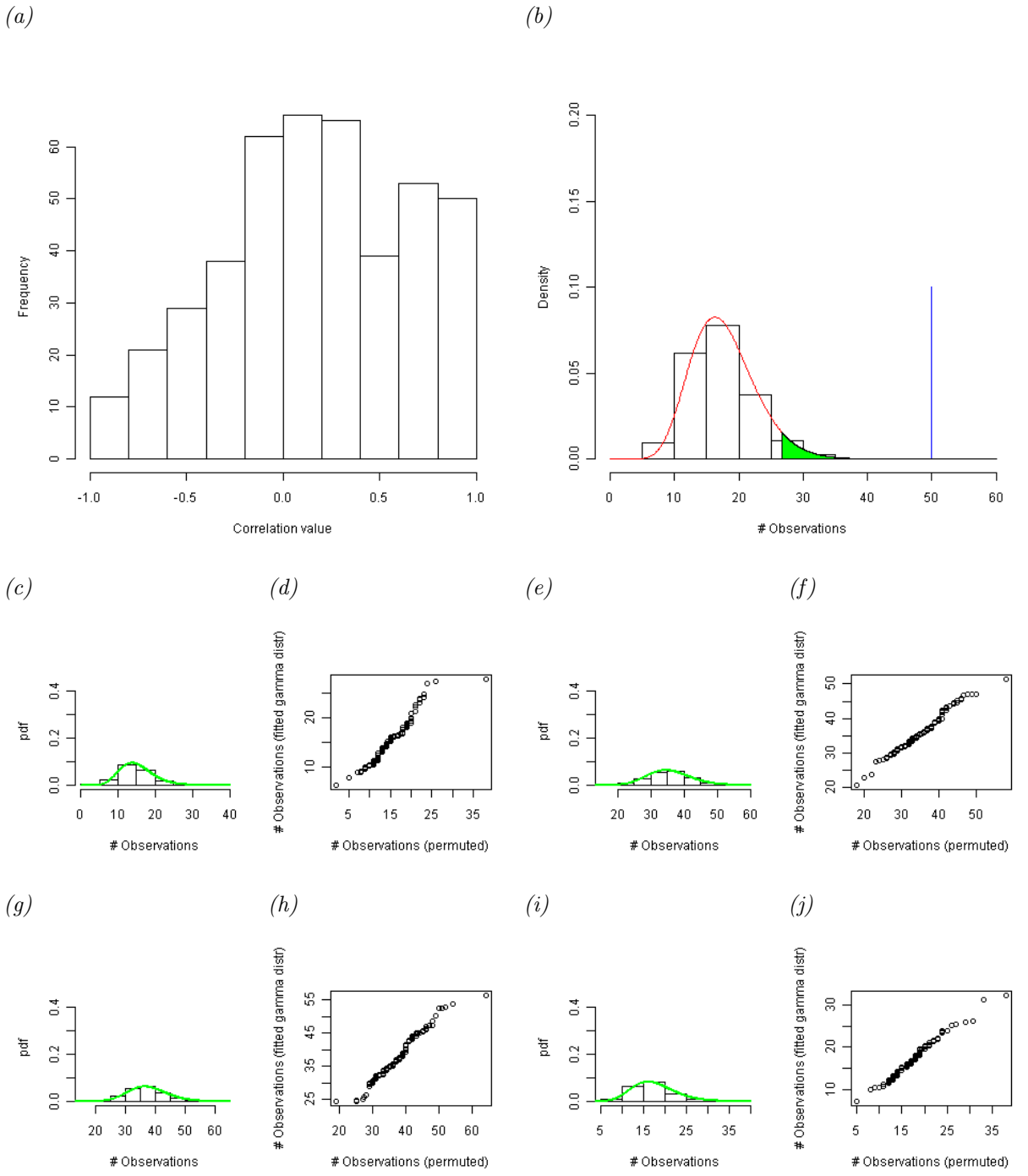


Figure 3.8: Comparison of null distribution histograms with fitted gamma distribution for 'Gluconeogenesis' of AraCyc for the S-def roots data set. (a) shows the within pathway correlation value distribution of AraCyc 'Gluconeogenesis'. The cardinality of this pathway for the S-def roots data set is $N_{Gluc} = 30$. (b) Illustrates the estimation of the p-value. The null distribution is shown in red and as a histogram. The green area corresponding to the 5% significance level and the blue vertical line marks the observed number of CVs in the bin $[0.8, 1]$. (c), (e), (g) and (i) show the histograms of the null distributions along with the fitted gamma distribution (green) for the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$, respectively. Furthermore, (d), (f), (h) and (j) show the Q-Q plots of the fitted gamma distribution against the originally sampled null distributions for the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$, respectively.

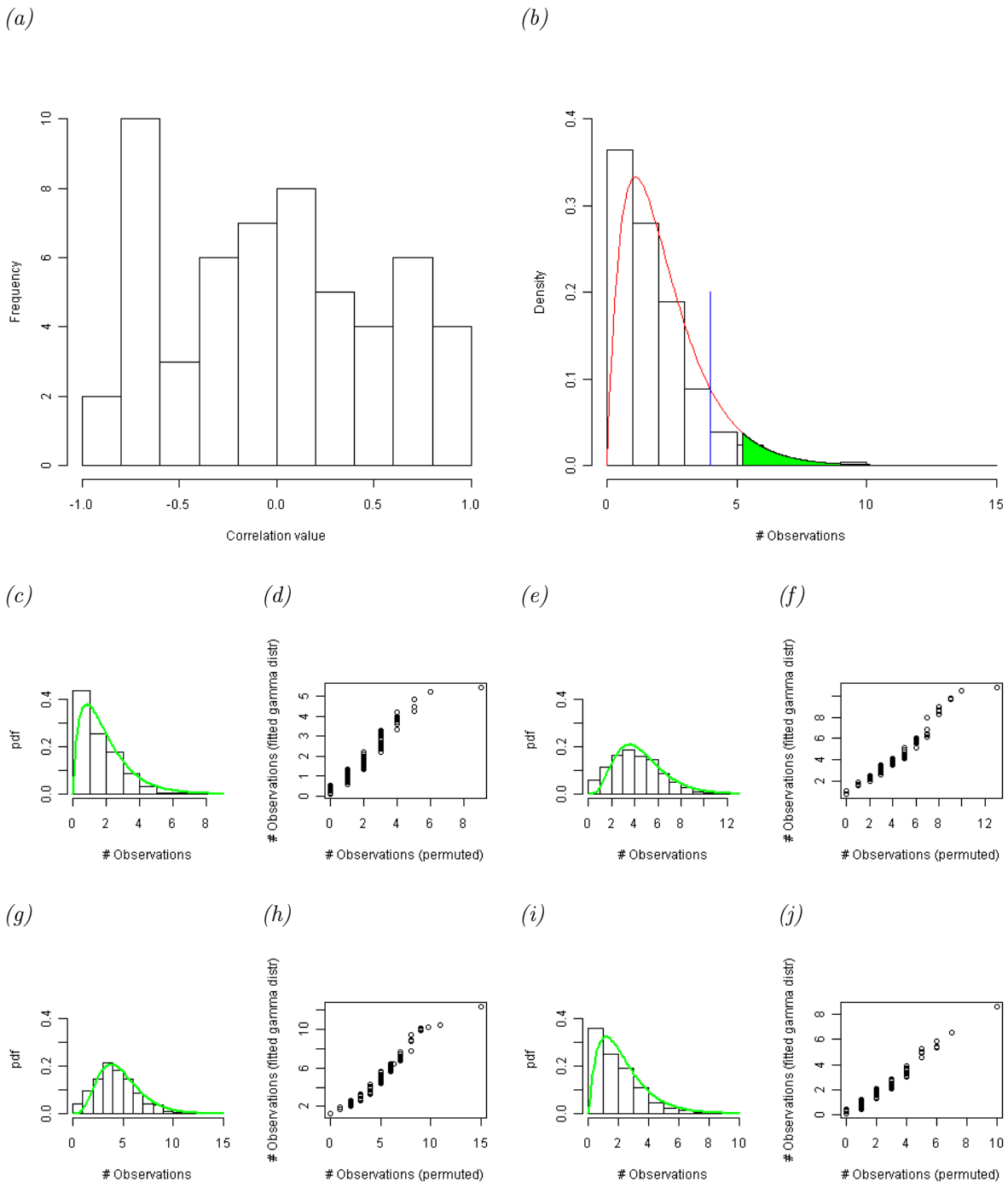


Figure 3.9: Comparison of null distribution histograms with fitted gamma distribution for 'Pentose phosphate pathway' of AraCyc for the S-def roots data set. (a) shows the within pathway correlation value distribution of AraCyc 'Pentose phosphate pathway'. The cardinality of this pathway for the S-def roots data set is $N_{PPP} = 11$. (b) illustrates the estimation of the p-value. The null distribution is shown in red and as a histogram, with the green area corresponding to the 5% significance level and the blue vertical line marking the observed number of CVs in the bin $[0.8, 1]$. The histogram represents to original null histogram, whereas the fitted gamma distribution is shown in red. (c), (e), (g) and (i) show the histograms of the null distributions along with the fitted gamma distribution (green) for the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$, respectively. Furthermore, (d), (f), (h) and (j) show the Q-Q plots of the fitted gamma distribution against the originally sampled null distributions for the correlation value intervals $[-1, -0.8)$, $[-0.8, -0.6)$, $[0.6, 0.8)$ and $[0.8, 1]$, respectively.

tograms which are very unlikely caused by a real correlation matrix. For the purpose of testing the appropriateness of the Jackknife variant, the KEGG pathways 00071, 00010 and 00020 are tested for within pathway enrichment for CO_2 elevation and R_{TT}. The analysis revealed that the resampled alternative distributions are relatively dissimilar compared to the corresponding gamma distributions (see Figures 3.11b, 3.11d and 3.11f). The actual shape of the sampled alternative distribution depends on the set cardinality and the overall shape of the histogram. Despite the limitation of the gamma distribution to represent the Jackknife alternative distribution adequately, the test based on Jackknife sampling performs comparably to the other test variants (see Section 3.3.1).

This can be explained by the fact that the optimized parameter t (which corresponds to the significance threshold) is relatively robust against shape variations of the alternative distribution. Figures 3.11a, 3.11c and 3.11e presents the performance of Jackknife sampling.

(a) *(b)*

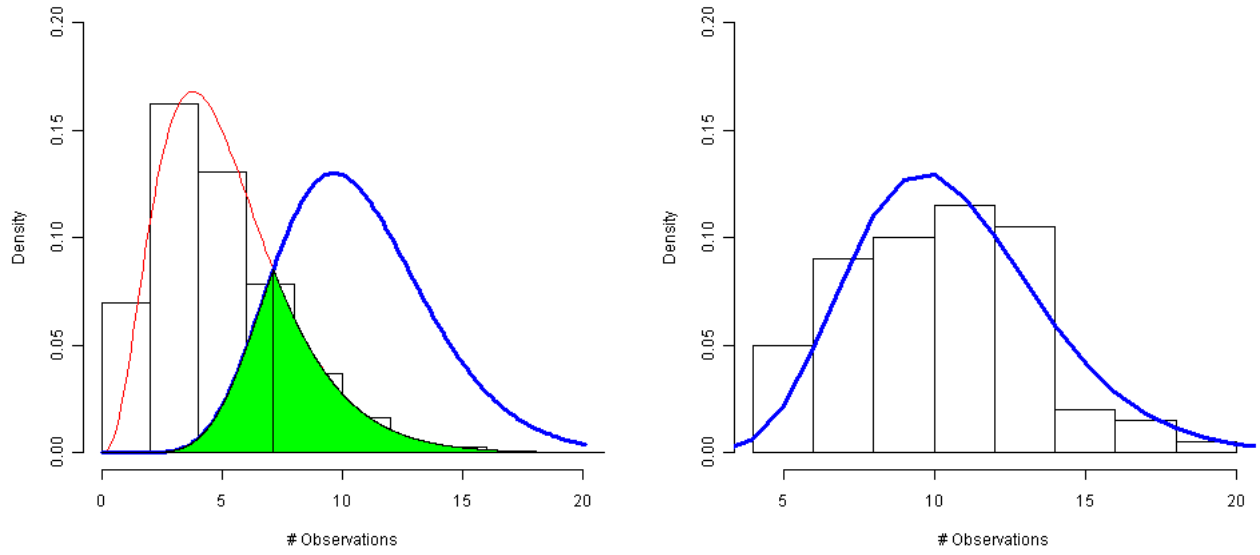


Figure 3.10: Histogram of null distribution and bootstrapped alternative distributions for R_{TT}, KEGG pathway '00010 - Glycolysis / Gluconeogenesis' and CV interval $[0.8, 1]$. (a) shows the null distribution (histogram + fitted gamma distribution in red) along with the bootstrapped gamma fitted alternative distribution (blue). The green region corresponds to $P(\text{erroneous decision})$. (b) shows the histogram along with the fitted gamma distribution of the bootstrapped alternative distribution. The gamma distribution is also appropriate for fitting the bootstrapped alternative distribution.

3.3.4 Variation of the Histogram Bin Sizes

Up to this section, the p-values and $P(\textit{erroneous decision})$ were examined primarily for the bin intervals $[-1, -0.8)$ and $[0.8, 1]$. However, apparently, a broader range of correlation value might be informative to determine biological association of a group of molecules. That is, i.e. is there a shift towards positive correlation values in general, in contrast to the question whether only high correlation values are observed. Figure 3.12 illustrates the motivation for using increased correlation value intervals. In this section,

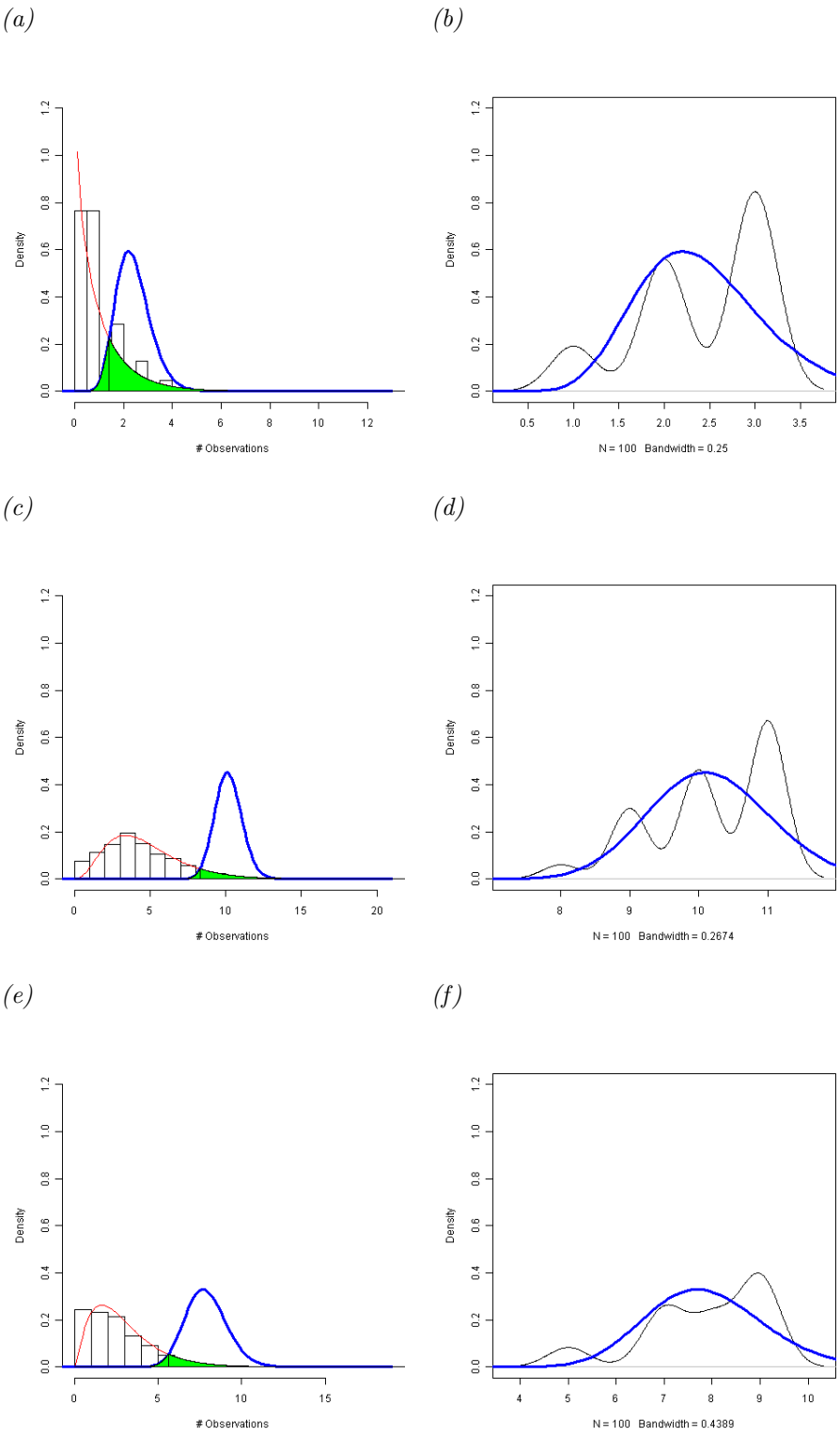


Figure 3.11: Histogram of null distribution and Jackknife alternative distributions for the CO_2 data set and R_{TT} using the CV interval $[0.8, 1]$. (a), (c) and (e) show three examples of estimating $P(\text{erroneous decision})$ according to the Jackknife alternative distribution for KEGG pathway 00071, 00010 and 00030, respectively. The null distribution is shown in red and as a histogram along with the alternative distribution in blue. The green region corresponds to $P(\text{erroneous decision})$. (b), (d) and (f) show the non-parametric densities of the Jackknife alternative distribution in black along with the fitted gamma distribution in blue.

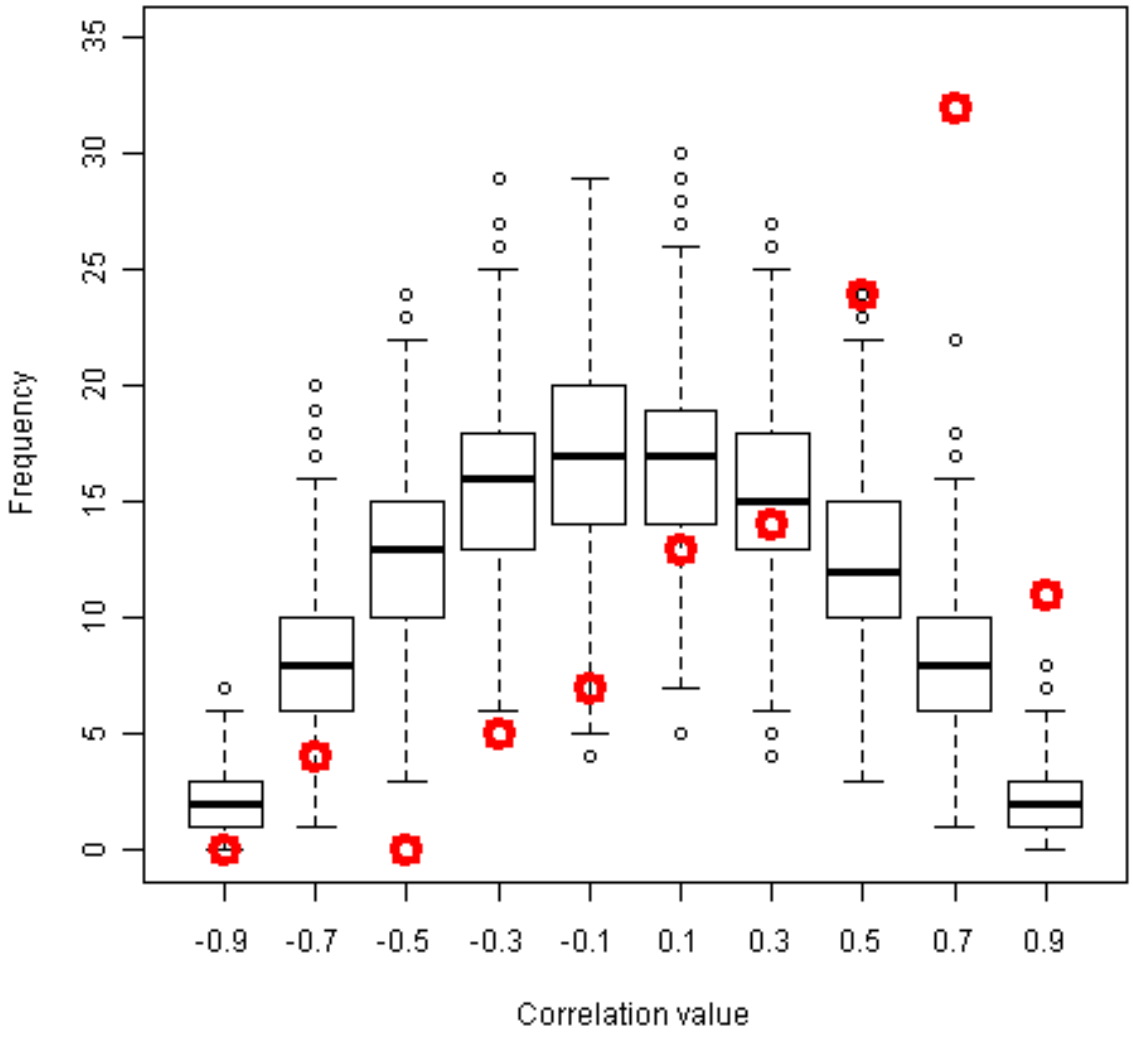


Figure 3.12: Boxplots of the bin-specific null distributions with observed correlation value counts for the example of CO_2 Elevation, R_{TT} and the pair of AraCyc annotations 'aerobic respiration – electron donor II' and 'TCA cycle variation III (eukaryotic)'. For all ten bins over the correlation value range a boxplot is drawn, which illustrates the estimated null distribution for that bin. Furthermore, the observed number of correlation values in each bin is marked (red circle). The majority of correlation values is shifted towards positive correlation values. Hence, using a broader range of correlation values (e.g. $[0.4, 1]$*) might further increase statistical significance.*

(a) (b)

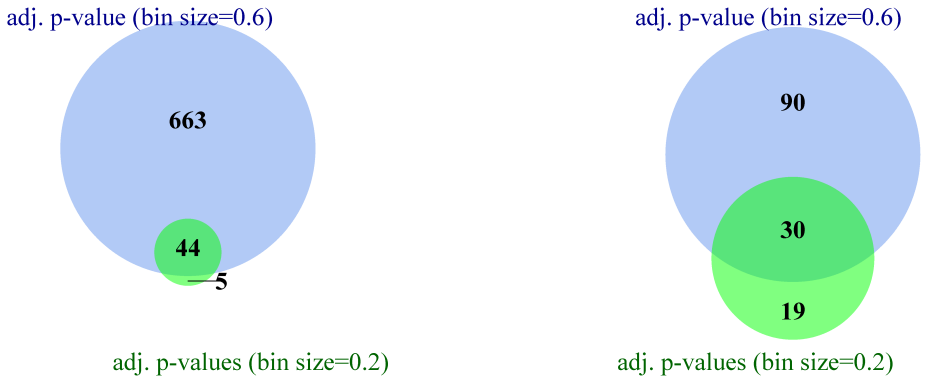


Figure 3.13: The Venn diagram illustrates the number of statistically significant pairs of KEGG pathways for CO_2 elevation and R_{TT}. (a) shows the concordance between two test variants of varying bin size and a fixed significance level of 5%. Bin size 0.6 corresponds to using the CV intervals $[-1, -0.4)$ and $[0.4, 1]$, whereas, bin size 0.2 corresponds to the intervals $[-1, -0.8)$ and $[0.8, 1]$. Increasing the bin interval resulted in a massive increase of significant results. (b) shows the overlap after truncation of the results from the tests on $[-1, -0.4)$ and $[0.4, 1]$ to the 120 top smallest p-values.

the two significance test variants are compared. Firstly, CV enrichment is estimated using bin-specific p-values for the correlation value intervals $[-1, -0.8)$ and $[0.8, 1]$, as performed previously. Secondly, CV enrichment is estimated using bin-specific p-values for the correlation value intervals $[-1, -0.4)$ and $[0.4, 1]$.

The enrichment analysis was performed for the CO_2 elevation data set and pairs of KEGG pathway on R_{TT}. Out of 118 KEGG pathways a total of 2059 pairwise pathway enrichment tests were performed.

The results of the two enrichment test variants were surprisingly divergent. CV enrichment for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ generated an extraordinary high number of significance assertions (707 significant KEGG pairs). In comparison, using the CV intervals $[-1, -0.8)$ and $[0.8, 1]$ results in only 49 significant KEGG pairs (see Figure 3.13). Using the kappa statistics to estimate the concordance between the tests results in $\kappa = 0.075$. These observations lead to the suggestion that the CV enrichment for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ might have produced many false positives.

Hence, as the excessive number of significant pairwise interactions is implausibly high, it was investigated whether the overlapping annotation label pairs exhibit smaller p-values compared to the non-overlapping ones for the test variant on the interval $[-1, -0.4)$ and $[0.4, 1]$ (compare Figure 3.13a).

Therefore, the set of significant KEGG pairs for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was truncated to the 120 top ranked pairs (60 for the positive correlations and 60 for the negative) according to the p-values. The comparison between the truncated list of significant pairs on $[-1, -0.4)$ and $[0.4, 1]$ and the results on $[-1, -0.8)$ and $[0.8, 1]$ revealed that 30 significant pairs still show overlap, which is a significant concordance (see

Table 3.13b; kappa statistics = 0.276).

Table 3.5: Concordance between the enrichment tests of variant 1 (using the CV intervals $[-1, -0.8)$ *and* $[0.8, 1]$*) and variant 2 (using the CV intervals* $[-1, -0.4)$ *and* $[0.4, 1]$*). For fixed significance level* α *=5% for both variants, variant 2 yields an excessive number of significance assertions. After truncation of the result list of variant 2 to the 120 smallest p-values (60 for the positive negative intervals, respectively), a substantial overlap of 30 significance assertions was observed. Kappa statistics was used to estimate the agreement between the test variants.*

	Variant 1	Variant 2	Overlap	kappa statistics
fixed 5% sig. level	49	707	44	0.075
truncated Variant 2	49	120	30	0.276

3.4 Integration of Metabolite Profiles with Biological Component Classes

In this section we analyse the correlation profile of functionally related metabolites as classified by KEGG BRITE 'Compounds with biological roles' [36]. The classification system is characterized by three hierarchical levels. The aim of this section is to reveal compound classes for which the assigned metabolites are enriched in positive or negative correlation values. In other words, strongly correlated metabolite classes shall be identified (e.g. high correlation of several amino acid species). The analysis was performed by utilizing the introduced enrichment analysis approach with bin-specific p-value estimation for the CV intervals $[-1, -0.8)$ and $[0.8, 1]$.

In 3 out of 4 experimental conditions, 'Peptides' and its subclasses (Amino acids in level 2 and Common/Other amino acids in level 3) were significantly enriched for high positive CVs (see Tables 3.6, 3.7 and 3.8). For the cold stress data set, 'Monosaccharides' were significantly enriched for highly negative correlation values. For the sulphur starvation in leaves, no statistically significant compound classes were found (see Table 3.6). However, the examination of the histograms of CVs for 'Peptides' suggests highly positive correlation of amino acids in all experimental conditions (see Figure 3.14).

Table 3.6: Correlated classes of metabolites for the COLD stress data set. According to the hierarchical classification of metabolite groups, several amino acid levels were found to be tightly correlated. 'Monosaccharides' were found to be enriched for negative correlation values. Note that 'Amino acids' and 'Common/Other amino acids' are subclasses of 'Peptides'.

Metabolite class	Hierarchy level	p-value for $C \leq -0.8$	p-value for $C \geq 0.8$
Peptides	1	–	3.07e-6
Amino acids	2	–	2.03e-6
Common amino acids	3	–	1.29e-5
Other amino acids	3	–	6.38e-6
Carbonhydrates	1	0.107	–
Monosaccharides	2	0.031	–

Table 3.7: Correlated classes of metabolites for the CO_2 stress data set. According to the hierarchical classification of metabolite groups, several amino acid levels were found to be tightly correlated. Note that 'Amino acids' and 'Common/Other amino acids' are subclasses of 'Peptides'.

Metabolite class	Hierarchy level	p-value for $C \leq -0.8$	p-value for $C \geq 0.8$
Peptides	1	–	0.0145
Amino acids	2	–	0.0086
Common amino acids	3	–	0.026
Other amino acids	3	–	0.024

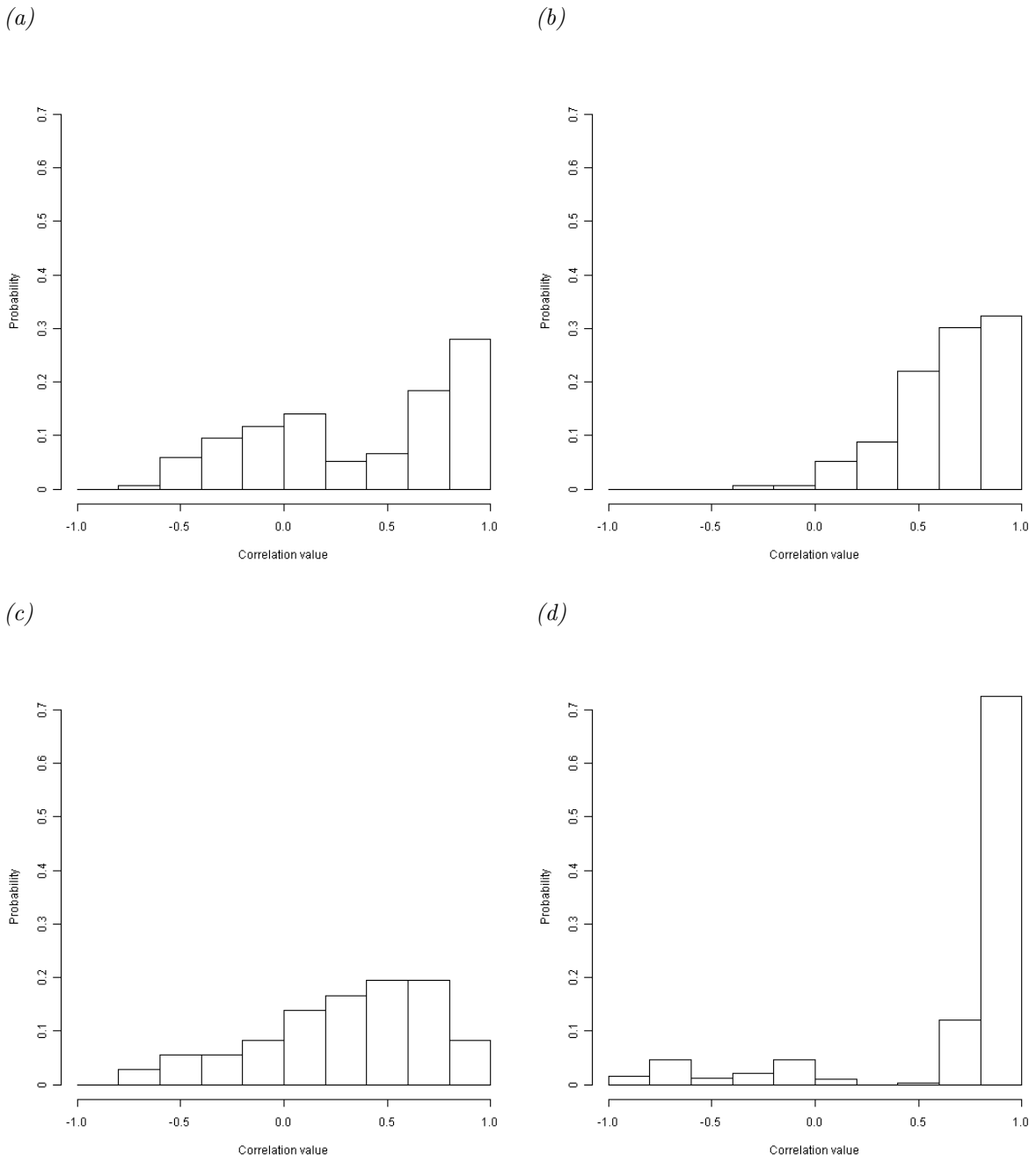


Figure 3.14: Histograms over correlation values for KEGG BRITE 'Peptides'. (a), (b), (c) and (d) show the histograms of the CV frequency within metabolites of 'Peptides' for the CO_2, S-def in roots and leaves and cold data set, respectively. All histograms suggest strong correlation of metabolites in the KEGG BRITE class of 'Peptides' indicated by a marked shift of the CVs towards positive values. Note that for S-def in leaves also a shift towards positive CVs is present, though not for the interval $[0.8, 1]$*, in particular.*

Table 3.8: Correlated classes of metabolites for the sulphur deficiency stress in roots. According to the hierarchical classification of metabolite groups, several amino acid levels were found to be tightly correlated. Note that 'Amino acids' and 'Common/Other amino acids' are subclasses of 'Peptides'.

Metabolite class	Hierarchy level	p-value for $C \leq -0.8$	p-value for $C \geq 0.8$
Peptides	1	–	0.00076
Amino acids	2	–	0.00077
Common amino acids	3	–	0.00036
Other amino acids	3	–	0.00084

3.5 Analysis of Sulphur Starvation Data Sets

This section focuses on the analysis of the sulphur deficiency data sets [18, 19]. Both data sets comprise of microarray and mass spectrometry time-series measurements for the time points 3h, 5h, 12h, 24h, 48h and 168h of mRNA and metabolite level levels, respectively. The data set was generated to reveal important regulatory interactions among transcripts and metabolites which are affected by sulphur limitation, as sulphur is known play important roles in sulphur metabolism, plant defence processes, stress response etc. [37].

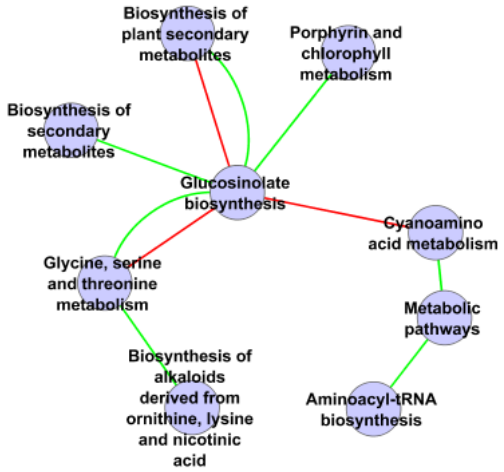


Figure 3.15: Enrichment network derived from S-def roots, R_{MM} and KEGG pathways. Bin-specific p-value estimation was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

Results drawn from $\mathbf{R_{MM}}$: There were no significant within KEGG pathway enrichments for high CVs for both two data sets.

The analysis of KEGG pathway pairs of the root data set revealed metabolites of glucosinolate biosynthesis as highly enriched for positive and/or negative correlation values with i.e. 'Glycine, serine, and threonine', 'Biosynthesis of plant secondary metabolites', 'Porphyrin and chlorophyll metabolism' and 'Cyanoamino acid metabolism', underlining its importance in cellular response to sulphur starvation (see Figure 3.15). There were no significant pairs of KEGG pathways found for the leaf data set.

Results drawn from $\mathbf{R_{MT}}$: In the leaf data set, metabolites of 'Glucosinolate biosynthesis' are enriched for negative correlations with genes from 'Ribosome' and 'Porphyrin and chlorophyll metabolism'. Moreover, metabolites from 'Glucosinolate biosynthesis' are enriched for positive correlations paired with genes of 'Limonene and pinene degradation'. Ribosomal genes are enriched for negative correlation values with metabolites of 'Aminoacyl-tRNA biosynthesis', 'Cyanoamino acid metabolism', 'Metabolic pathways and 'Biosynthesis of secondary metabolites' (see Figure 3.16a).

For the root data set, the approach revealed genes of 'Glucosinolate biosynthesis' as enriched for highly negative correlation values with metabolites of 'Glycine, serine and threonine metabolism', 'Porphyrin and chlorophyll metabolism' and 'Cyanoamino acid metabolism'. Interestingly, genes of 'Glucosinolate biosynthesis' are enriched for high positive and negative CVs with metabolites of the same pathway (see Figure 3.16b) . Moreover, the enrichment network suggests an association of genes of 'Photosynthesis - antenna proteins' with metabolites of 'Nitrogen metabolism', 'Cyanoamino acid metabolism', 'Biosynthesis of secondary metabolites' and 'Aminoacyl-tRNA biosynthesis',

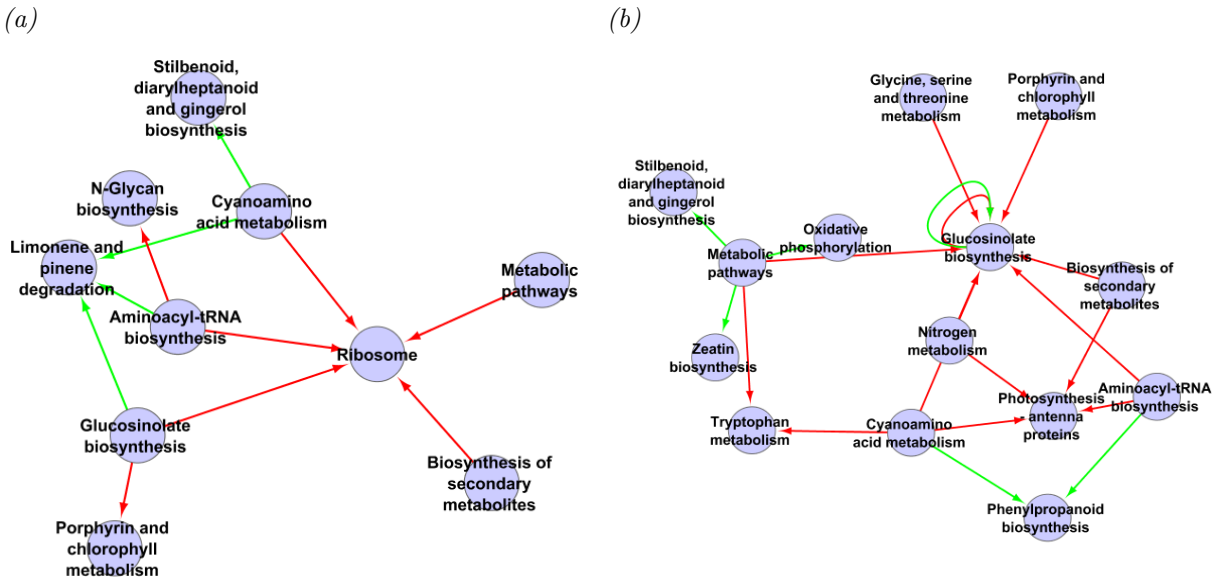


Figure 3.16: Enrichment networks derived from S-def (a) leaves and (b) roots, R_{MT} and KEGG pathways. (a), bin-specific p-value estimation was used for the CV intervals $[-1, -0.8)$ *and* $[0.8, 1]$*. (b), Jackknife sampling was used for the CV intervals* $[-1, -0.8)$ *and* $[0.8, 1]$*. Red and green links correspond to significantly enriched correlation values in the ranges* $[-1, -0.8)$ *and* $[0.8, 1]$*, respectively. The arrows point from metabolites of some category to the transcripts of another category.*

while metabolites of 'Cyanoamino acid metabolism' and 'Aminoacyl-tRNA biosynthesis' are both enriched for positive CVs with 'Phenylpropanoid biosynthesis' (see Figure 3.16b).

Results drawn from $\mathbf{R_{TT}}$: For the leaf data set, photosynthesis related pathways and 'Glucosinolate biosynthesis' were found to be highly enriched for positive CVs for AraCyc and KEGG annotations. Similarly, for the enrichment analysis with GO terms, genes of 'Glucosinolate biosynthetic processes' were found to be highly correlated. Furthermore, 'Response to chitin' was among the top ranked GO terms with respect to the p-value for positive correlation values. Using the InterPro domains for within annotation label enrichment, 'DNA-binding, integrase-type', which is a DNA binding domain that occurs i.e. in transcription factors and 'Heat shock protein Hsp20' domain were among the top ranking InterPro domains.

For the root data set, the proposed method revealed 'Photosynthesis - antenna proteins' and 'Glucosinolate biosynthesis' to be enriched for high positive CVs using the AraCyc library. The KEGG annotations resulted in similar findings. Using GO terms, 'Ribosome biosynthesis', 'Glucosinolate biosynthesis' and 'Photosynthesis' were among the most highly enriched terms for positive CVs with respect to the p-values. Integrating the InterPro domains revealed 'Chlorophyll A-B binding protein, plant' and 'DNA-binding, integrase-type' were among the top ranked domains for positive CVs.

According to the pairwise enrichment analysis for the root data set, 'Glucosinolate biosynthesis' is densely connected to other AraCyc pathways (see Figure 3.17a). There are links from 'Photorespiration' and 'Photosynthesis light reaction' to several amino acid degradation or biosynthesis pathways (e.g. 'Valine degradation', 'Leucine biosynthesis' and 'Leucine degradation I'). Additionally, several pathways which can be assigned to primary metabolism (various amino acid degradation/biosynthesis pathways, carbonhydrate pathways, citrate cycle and photosynthesis related pathways etc.) are present in the network (see Figure 3.17a).

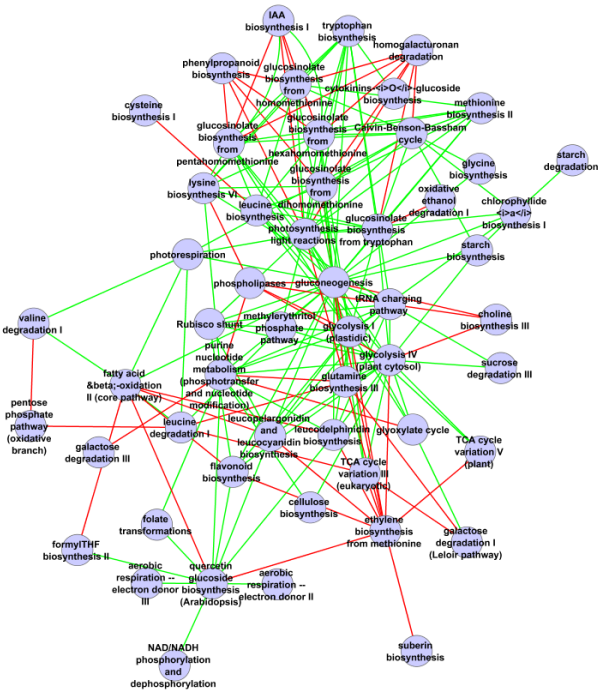
For the leaf data set, the most tightly connected annotation labels are again 'Glucosinolate biosynthesis' , pathways associated with amino acid and carbonhydrate metabolism, citrate cycle and pathways associated with photosynthesis (see Figure 3.17b). A comparison of the generated networks between both tissue types i.e. for AraCyc pathways (see Figure 3.18) shows considerable overlap between the results. For instance, computation of the intersection of the two graphs reveals that 'Photosynthesis light reaction' to be enriched for high CVs with 'Glycolysis', 'Gluconeogenesis', 'Glucosinolate biosynthesis' and 'Calvin-Benson-Bassham cycle' in both tissue types. Further, edges from several amino acid biosynthesis pathways (lysine, leucine, methionine) to glucosinolate biosynthesis are preserved among the tissue types. For the other annotation libraries, the results also show a considerable overlap.

Integrating the KEGG pathway annotations yields largely similar information compared to the AraCyc pathways. For instance, 'Glucosinolate biosynthesis' is again linked with 'Photosynthesis'. Moreover, KEGG pathways such as 'RNA polymerase', 'RNA degradation', 'Proteasome', 'Ubiquinone and aminoacyl-tRNA biosynthesis', which are not present in AraCyc, suggest general changes of protein abundance, induced by the experimental conditions (see Figure 3.18).

In concordance with the previous enrichment networks, integrating GO terms yields high node degrees for 'Glucosinolate biosynthetic process' and 'Ribosome biogenesis' for the root data set (see Figure 3.19a). In both, root and leaf data set, 'Photosynthesis', 'Photorespiration' and/or 'Photosynthesis light reaction' are enriched for positive correlations with 'Glucosinolate biosynthetic process' (see Figure 3.19). Some of the GO terms which are connected to 'Ribosome biogenesis' (e.g. 'DNA replication', 'DNA repair') suggest global adaption processes in response to sulphur deficiency (see Figure 3.19a). Several GO terms, which are associated with primary metabolic processes can be found in both networks (see Figure 3.19). For instance, 'Glycolysis', 'Leucine biosynthetic process', 'Arginine biosynthetic process' or 'Malate metabolic process'. 'Cellular response to sulphate starvation' is enriched for negative correlation values compared to 'response to chitin' for the leaf data set, which itself is enriched for positive correlation values with 'Response to other organisms' and 'Jasmonic mediated signaling pathway' (see Figure 3.19b).

Using the InterPro domains for the pairwise enrichment analysis revealed several protein domains that are specific for transcription factors or proteins with function in signal transduction. For instance, 'DNA-binding integrase-type', 'Pathogenesis-related transcriptional factor/ERF', 'Toll/interleukin-1 receptor homology (TIR) domain' for 'Protein kinase, catalytic domain' for both, the root and leaf data set (see Figure 3.20). Additionally, glutathione-S-transferase domains, which are constituent parts of glutatione-S-transferases, can be found in both networks.

(a)



(b)

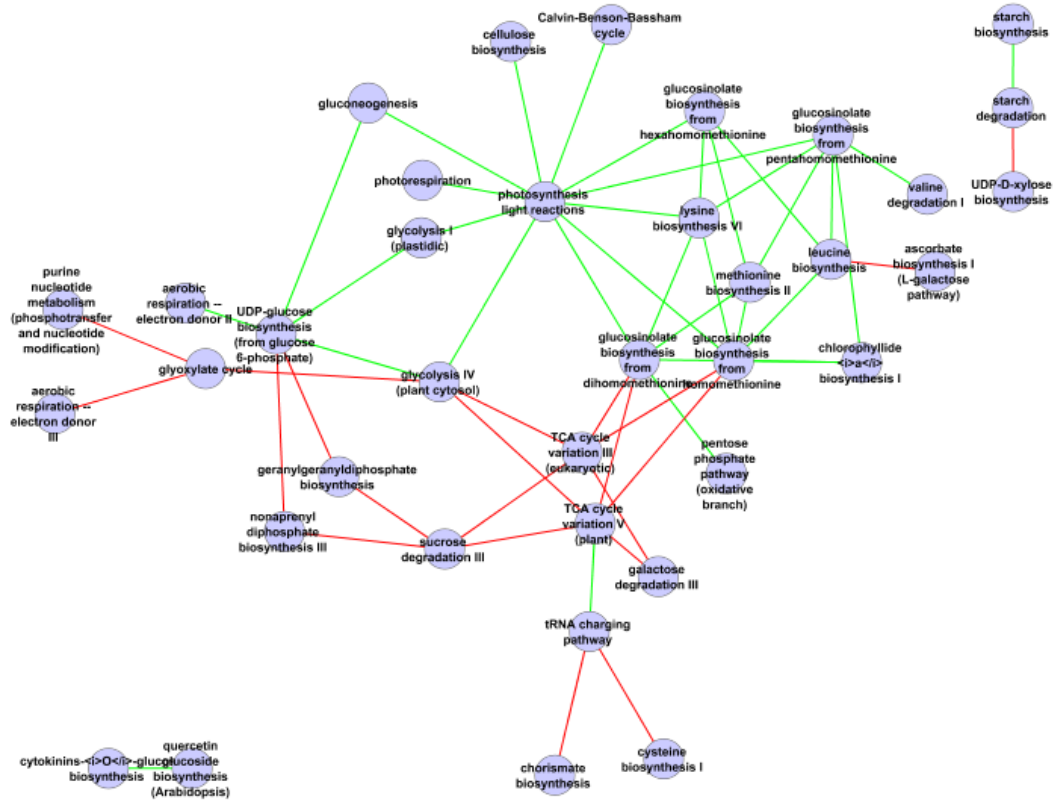
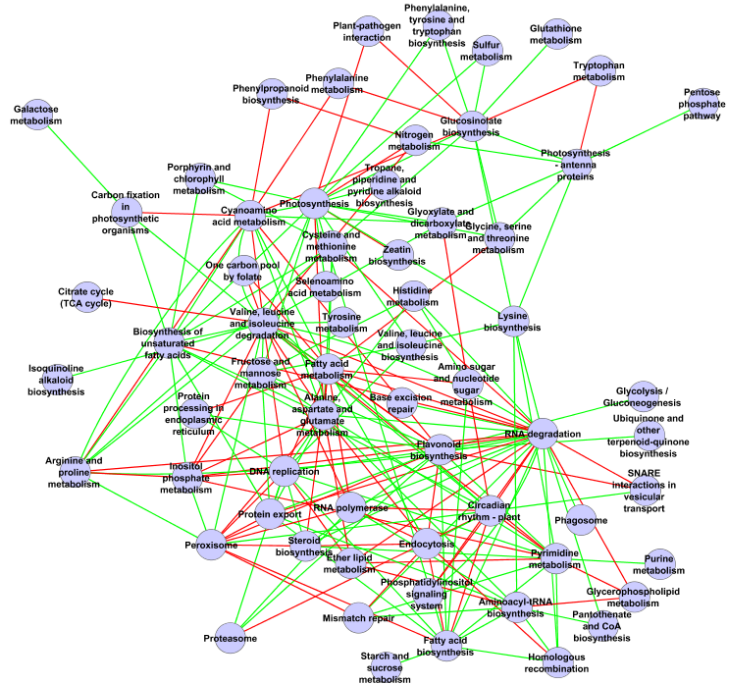


Figure 3.17: Enrichment networks derived from S-def (a) roots and (b) leaves, R_{TT} and AraCyc pathways. Jackknife sampling for the CV intervals $[-1, -0.8)$ and $[0.8, 1]$ was performed in both cases. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

(a)



(b)

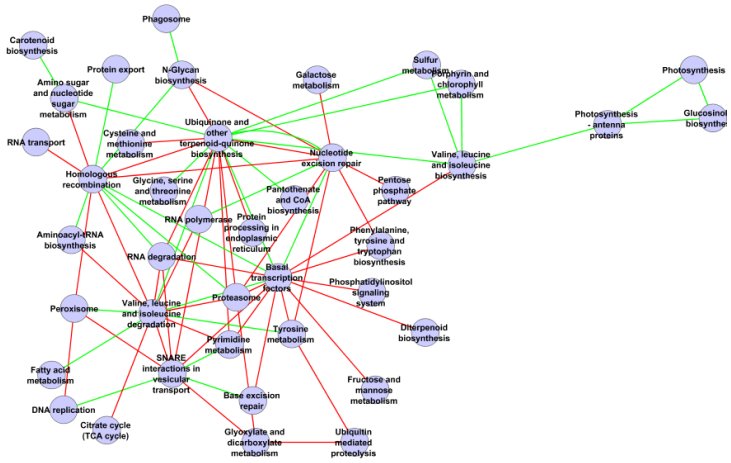
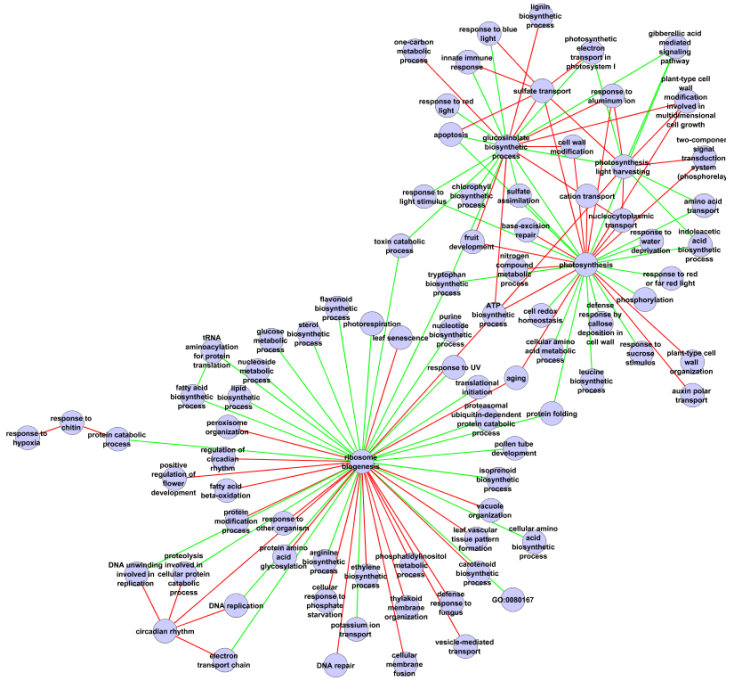


Figure 3.18: Enrichment networks derived from S-def (a) leaves and (b) roots, R_{TT} and KEGG pathways. Bootstrap sampling for the CV intervals $[-1, -0.8)$ and $[0.8, 1]$ was performed in both cases. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

(a)



(b)

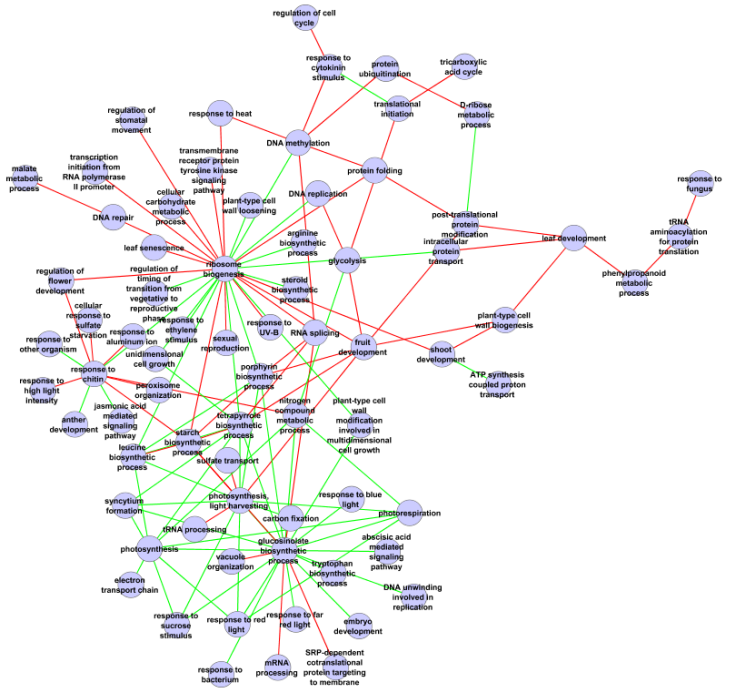
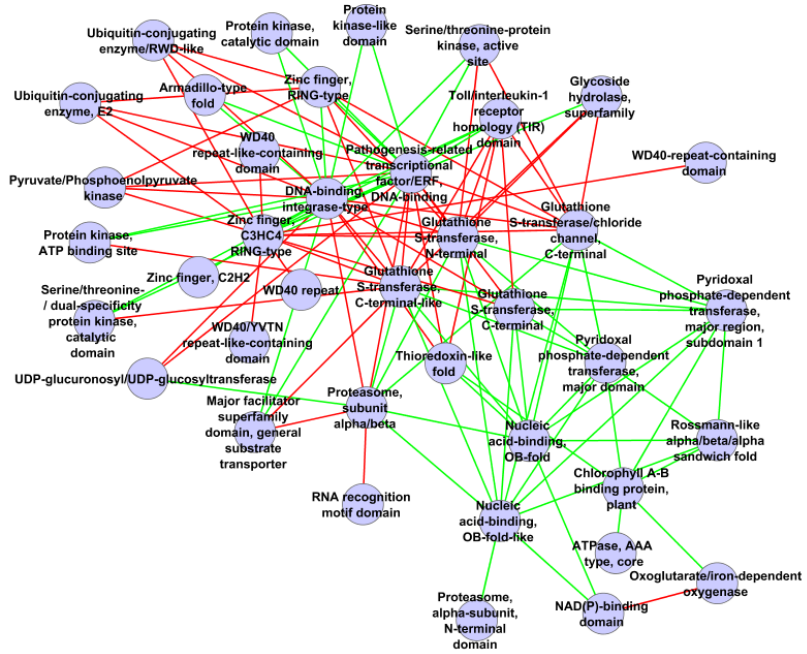


Figure 3.19: Enrichment networks derived from S-def (a) leaves and (b) roots, R_{TT} and GO terms (biological processes). Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used in both cases. Links were drawn for the 120 top ranked GO term pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

(a)



(b)

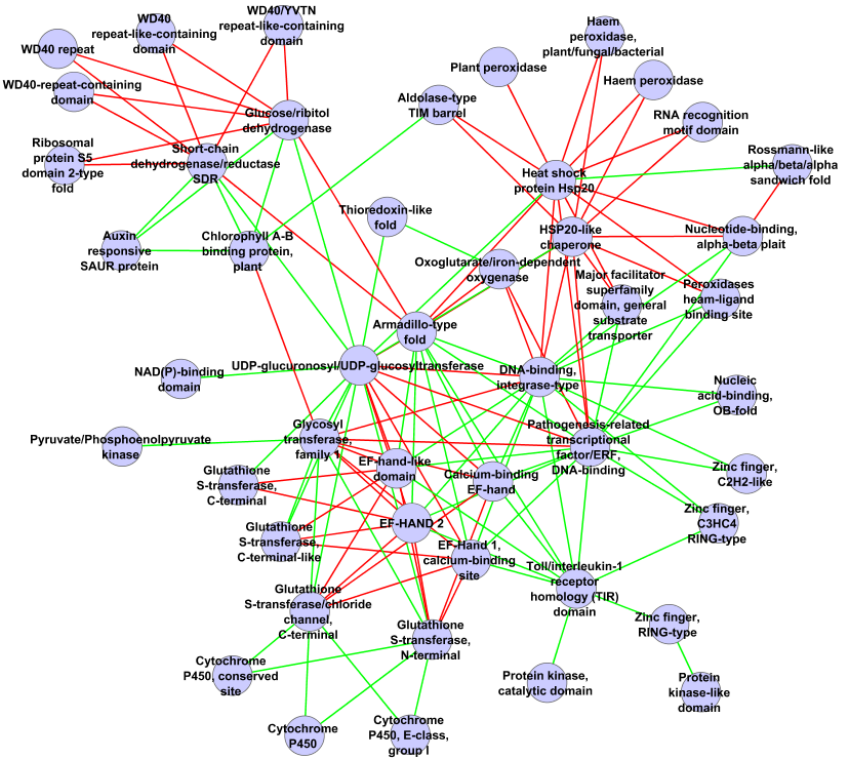


Figure 3.20: Enrichment networks derived from S-def (a) leaves and (b) roots, R_{TT} and InterPro PDs. Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used in both cases. Links were drawn for the 120 top ranked InterPro domains pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

3.6 Analysis of Cold Acclimation Data Sets

This section focuses on the analysis of the data set published by Kaplan *et al.* [17]. The data set comprises of microarray and mass spectrometry time-series measurements for the time points 0h, 1h, 4h, 12h, 24h, 48h and 96h of mRNA and metabolite levels, respectively. The data set was aimed at identifying regulatory interactions among transcripts and metabolites as well as deregulated processes, which are affected by cold acclimation. Cold acclimation induces complex changes of metabolite and transcript abundances, to prevent the plant from being injured by ice formation under freezing temperatures [38].

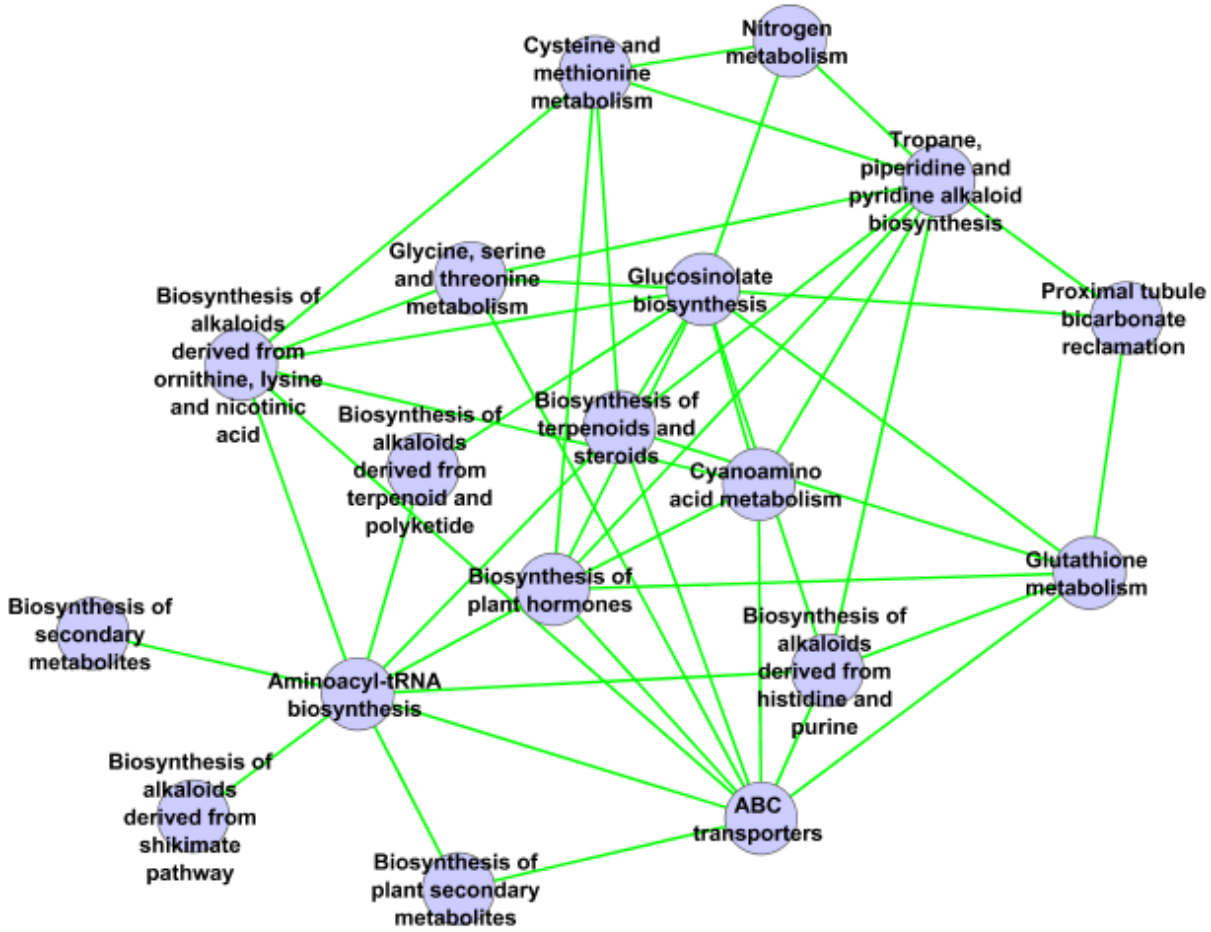


Figure 3.21: Enrichment network derived from cold acclimation, R_{MM} and KEGG pathways. Bin-specific p-value estimation was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively. Note that for this data set no significant enrichment for negative CVs was detected.

Results drawn from $\mathbf{R_{MM}}$: There are no significant within KEGG pathway enrichments. For the between KEGG pathway enrichment analysis, several pathways for alkaloid biosynthesis pathways are linked to 'Aminoacyl-tRNA biosynthesis' (see Figure 3.21). The network suggests high CVs of metabolites contained in 'Glutathione metabolism', 'Glycine, serine and threonine metabolism', 'Cyanoamino acid metabolism', 'Nitrogen metabolism', 'Alkaloid biosynthesis' and 'Biosynthesis of terpenoids and steroids' with metabolites of 'Glucosinolate biosynthesis'.

Results drawn from $\mathbf{R_{MT}}$: The analysis revealed gene expression changes of many

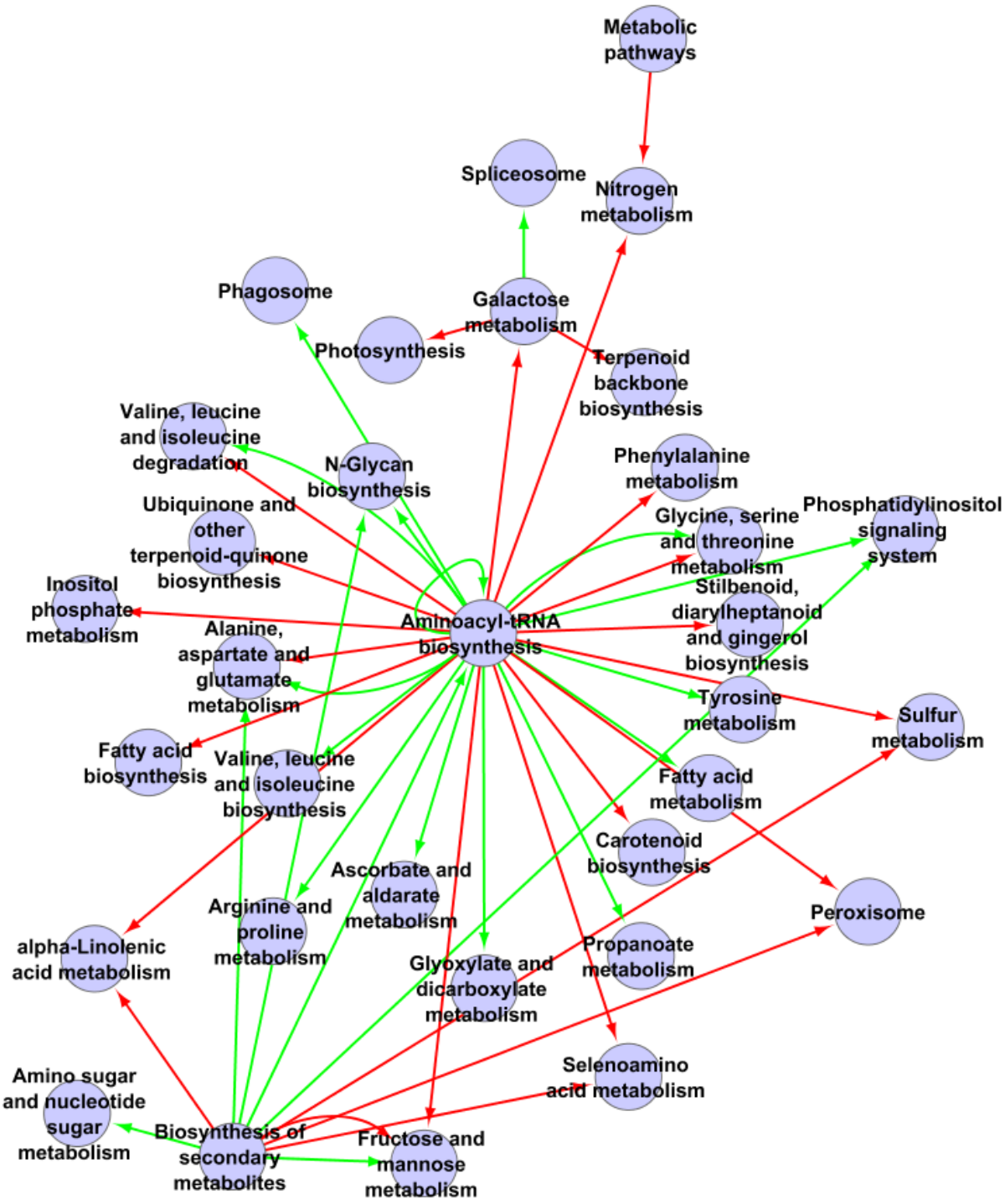


Figure 3.22: Enrichment network derived from cold acclimation, R_{MT} and KEGG pathways. Bootstrapped $P(\text{erroneous decision})$ estimation was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively. The arrows point from metabolites of some category to the transcripts of another category.

amino acid related pathways, for biosynthesis and degradation, indicating changes in transcript levels imposed by cold acclimation (see Figure 3.22). Furthermore, several genes related with functions in carbonhydrate metabolism are deregulated. Metabolites of 'Galactose metabolism' are enriched for negative correlation values with genes of 'Photosynthesis' and 'Terpenoid backbone biosynthesis' and for positive correlation values with genes of 'Splicosome'. Metabolites of 'Aminoacyl-tRNA biosynthesis' exhibit high positive and negative CVs with genes of many KEGG pathways (e.g. amino acid biosynthesis or metabolism).

Results drawn from $\mathbf{R_{TT}}$: First, the top ranking within annotation label enrichment results shall be presented.

Using the AraCyc pathways, 'Photosynthesis light reactions', 'Chlorophyllide a biosynthesis', 'Calvin-Benson-Bassham cycle', 'TCA cycle variation III & V' and 'Gluconeogenesis' are highly enriched for positive CVs. This suggests a tight transcriptional regulation of 'Photosynthesis' and primary metabolism.

Similarly, 'Photosynthesis' and 'photosynthesis - antenna proteins' are among the top ranked KEGG pathways. Moreover, using KEGG pathways 'Splicosome', 'Proteasome', 'RNA degradation' also appears among the most significant pathways.

For the GO term analysis, 'Photosynthetic electron transport in photosystem I', 'Proteolysis involved in cellular protein catabolic process', 'Chlorophyll biosynthetic process', 'Photosynthesis light harvesting' and 'Translational initiation' were detected.

Within InterPro domain enrichment revealed 'Pentatricopeptide repeat', several helicase domains (e.g. 'DEAD/DEAH box type, N-terminal, RNA helicase'), 'WD40 repeat' (and related domains) among the top ranked results.

In the following, the results of the pairwise enrichment test for all annotation label databases shall be described.

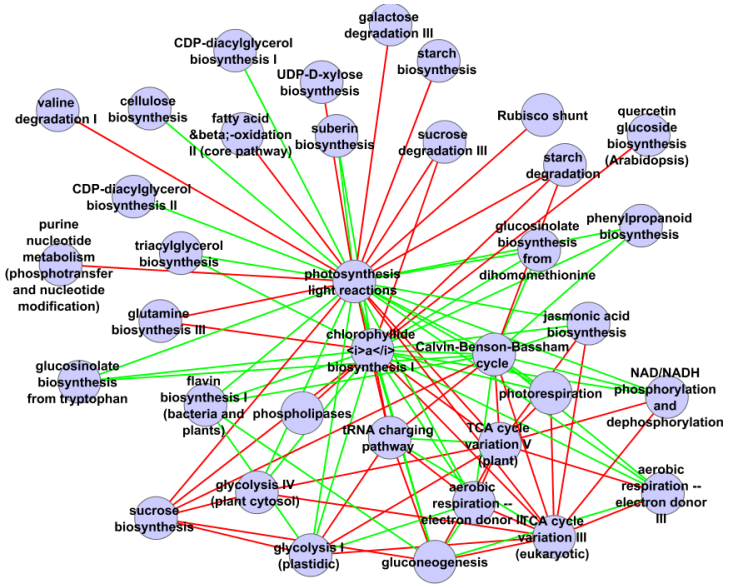
The comparison of the generated enrichment networks using bin-specific p-values estimation (Figure 3.23a) and the bootstrapping approach (see Figure 3.23b) shows strong concordance between the results, despite of variations of the algorithm parameters. In particular, highly significant pairs of annotation labels are present in both network, which underlines the robustness of the method.

The enrichment network for the AraCyc pathways yields (1) 'Starch biosynthesis' is enriched for negative CVs with 'Photorespiration', 'Calvin-Benson-Bassham cycle', 'Chlorophyllide a biosynthesis I' and 'Gluconeogenesis'; (2) 'Sucrose biosynthesis' is linked to 'Gluconeogenesis', 'Glycolysis', 'Calvin-Benson-Bassham cycle', 'Chlorophyllide a biosynthesis', 'NAD/NADH phosphorylation and dephosphorylation' and 'Glucosinolate biosynthesis from dihomomethionine'; (3) 'Phospholipases' is enriched for negative correlations with 'Chlorophyllide a biosynthesis I'; (4) 'Triacylglycerol biosynthesis', a storage form for fatty acids, is enriched for positive correlations with photosynthetic processes; (5) 'Fatty acid β-oxidation II' is enriched for negative correlations with photosynthesis; (6) glucosinolate biosynthesis and photosynthesis are enriched for positive correlation values; (7) and, 'Flavonoid biosynthesis' is enriched for negative CVs with 'Chorophyllide a biosynthesis' (see Figures 3.23a and 3.23b);

The enrichment network based on R_{TT} and KEGG pathways (see Figure 3.24) indicates a global shift of RNA and protein abundance (e.g. see 'Proteasome', 'Splicosome', 'RNA degradation', 'Protein processing in endoplasmatic reticulum'). Additionally, several enriched pairs of KEGG pathways correspond to pairs in the AraCyc networks (e.g. 'Photosynthesis', 'Glucosinolate biosynthesis', 'Glycolysis', 'Gluconeogenesis', Carbon fixation in photosynthetic organisms' or 'TCA cycle').

Several GO processes associated with photosynthesis constitute hub nodes of the network (see Figure 3.25). Furthermore, some hub nodes are related to general changes in

(a)



(b)

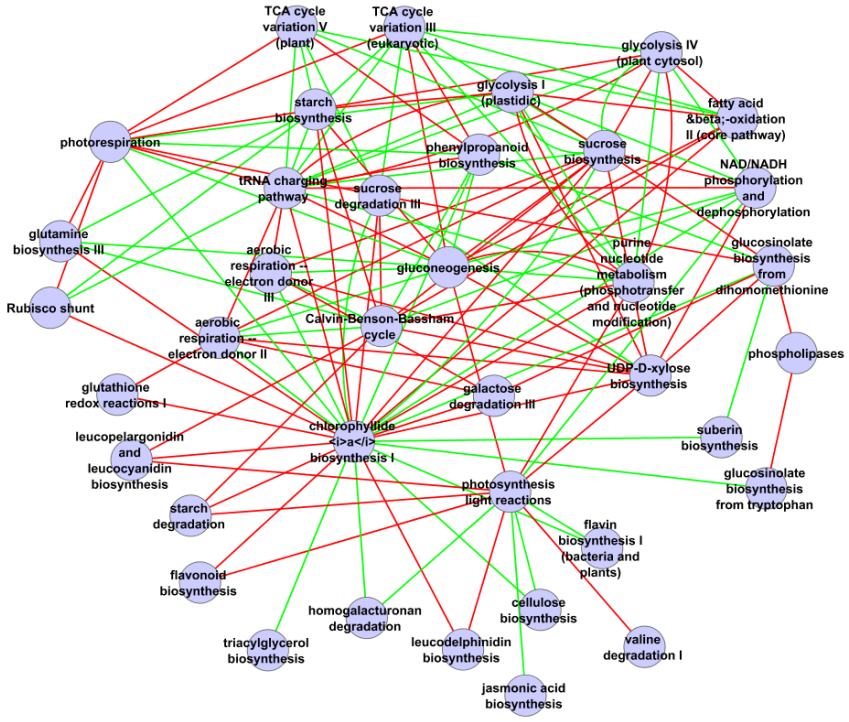


Figure 3.23: Enrichment networks derived from cold acclimation, R_{TT} and AraCyc pathways. (a), illustrates significantly indicated pairs of AraCyc pathways using ordinary p-value estimation for the CV intervals $[-1,-0.4)$ (red) and $[0.4,1]$ (green) for the top 120 ranked AraCyc pairs (60 for positive and 60 for negative CVs) according to the p-values. (b), shows the resulting network using bootstrapped $P(\text{erroneous decision})$ estimation for the CV intervals $[-1,-0.8)$ (red) and $[0.8,1]$ (green).

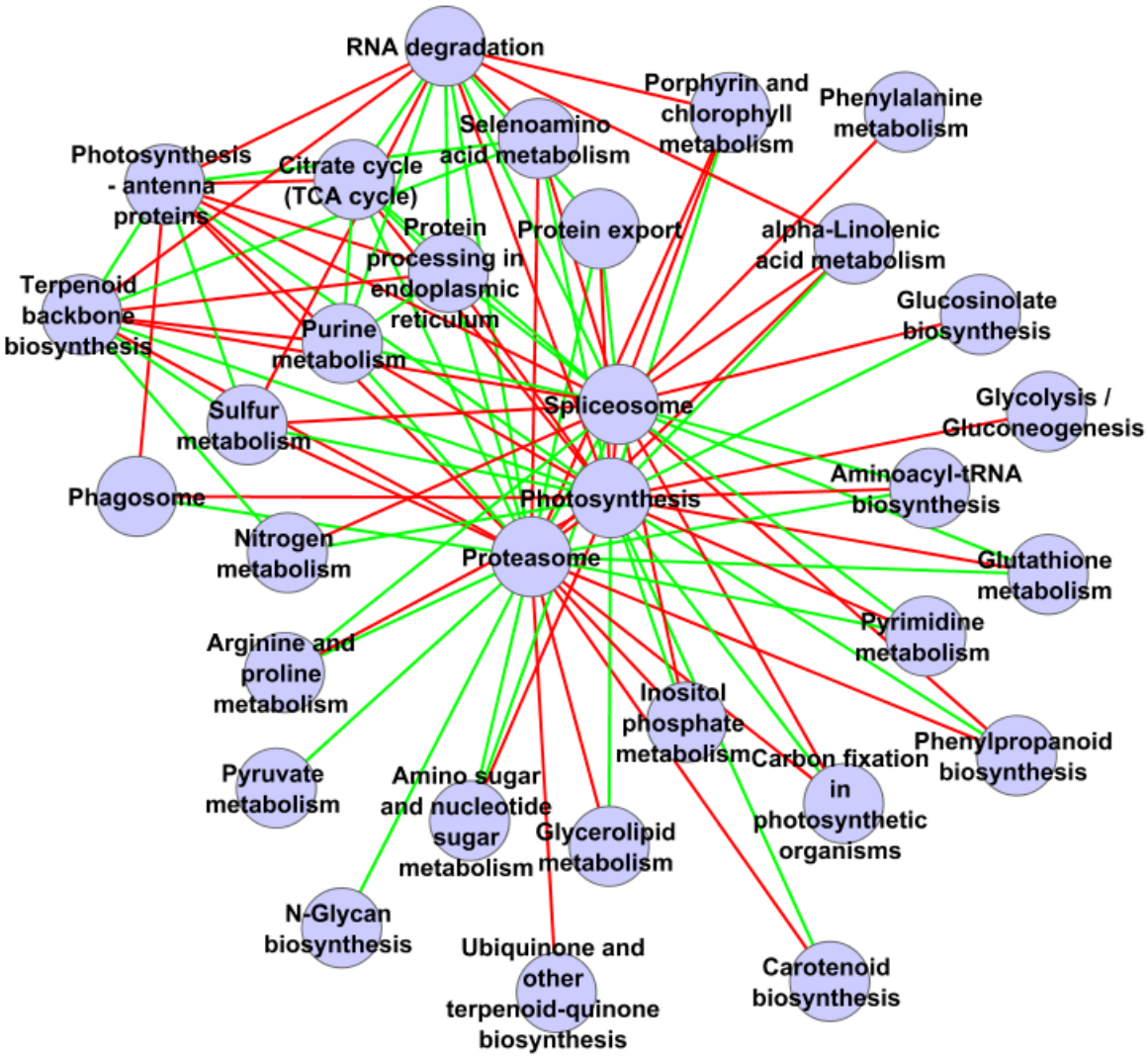


Figure 3.24: Enrichment network derived from cold acclimation, R_{TT} and KEGG pathways. Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used followed by truncation to the top ranked pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

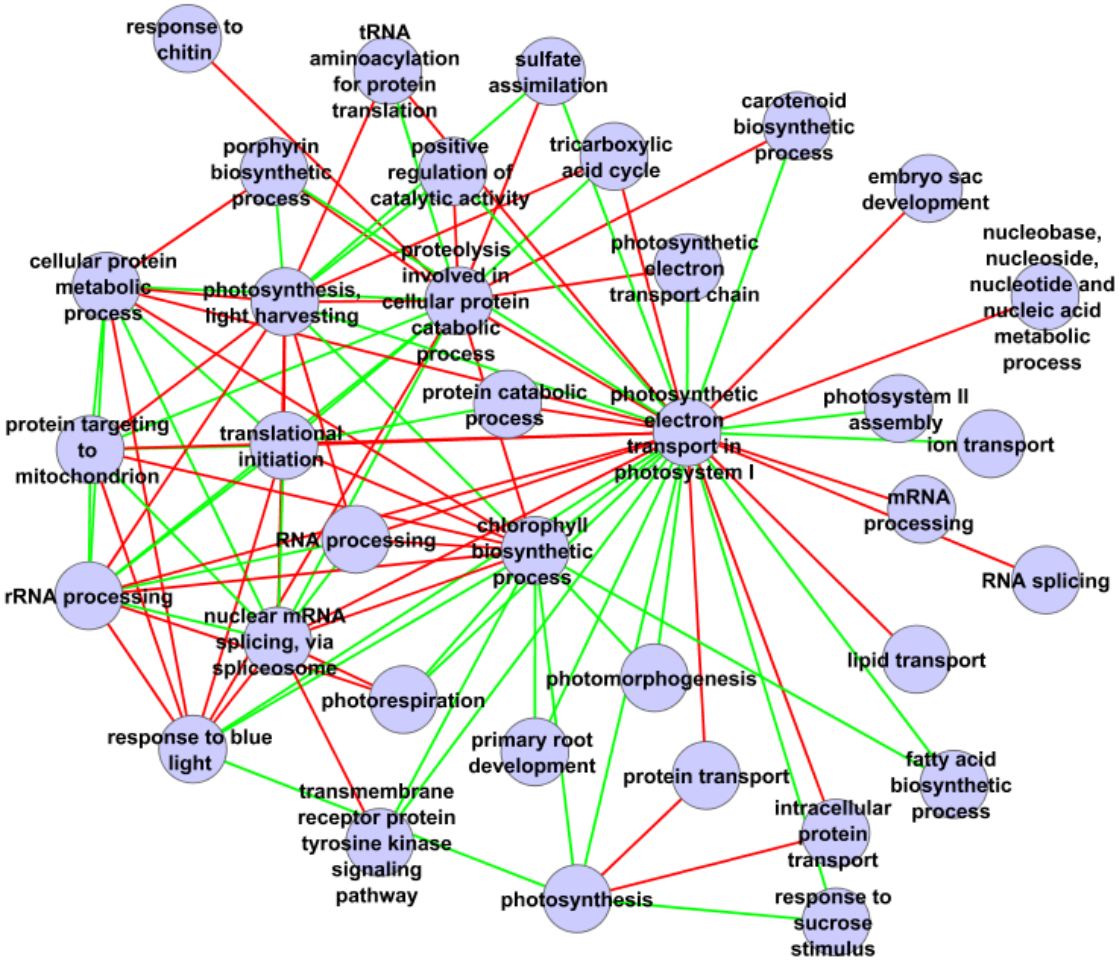


Figure 3.25: Enrichment network derived from cold acclimation, R_{TT} and GO terms (biological processes). Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used followed by truncation to the top ranked pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

RNA or protein abundance (e.g. 'RNA processing', 'tRNA aminoacylation for protein translation', 'Translational initiation', 'Protein catabolic process'). For the GO term pair 'Sulphur assimilation' and 'Photosynthesis', positive CVs are enriched (compare Figure 3.25).

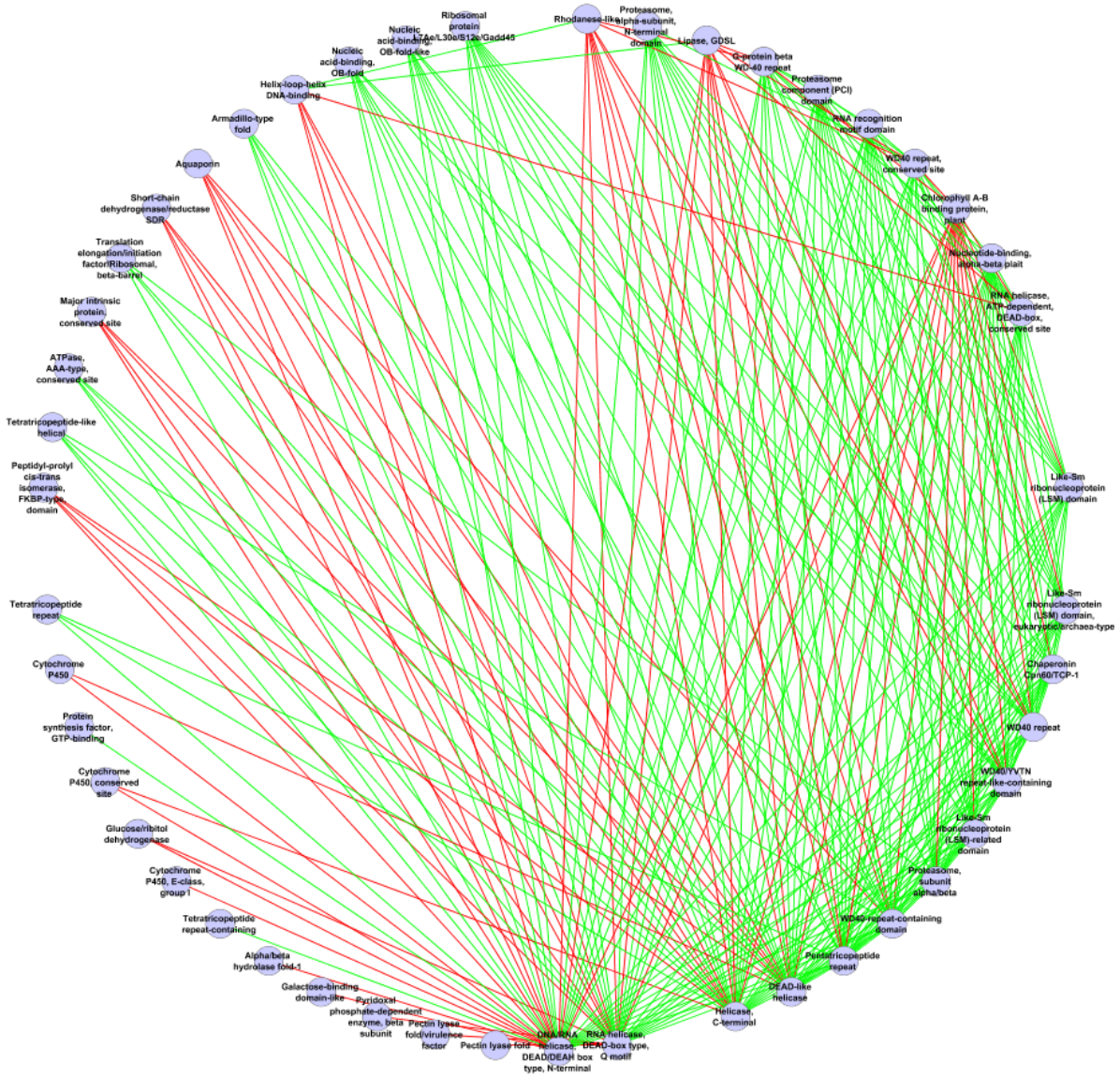


Figure 3.26: Enrichment network derived from cold acclimation, R_{TT} and InterPro PDs. Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used followed by truncation to the top ranked pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

For the enrichment network based on InterPro domains, a relatively high number of highly significant categories was found (see Figure 3.26). Protein domains with the highest node degrees in general correspond to domains which are also top ranked according to the within annotation label enrichment. This network, as the previous ones, reflects global change in protein and RNA composition due to the occurrence of i.e. proteasome-specific domains, ribosome-specific domains, DNA/RNA helicases.

3.7 Analysis of Carbon Dioxide Elevation Data Sets

This section presents the results of the analysis of the elevated CO_2 elevation data set [16]. The data set comprises of microarray and mass spectrometry time-series measurements for the time points 0h, 1h, 3h, 6h, 9h, 12h, 18h, 24h and 30h of mRNA and metabolite levels, respectively. The data set aims at a better understanding of important regulatory interactions among transcripts and metabolites due to CO_2 elevation. Of particular importance were the study of primary metabolism and processes related to photosynthesis.

Results drawn from $\mathbf{R_{MM}}$: Within annotation label enrichment analysis did not

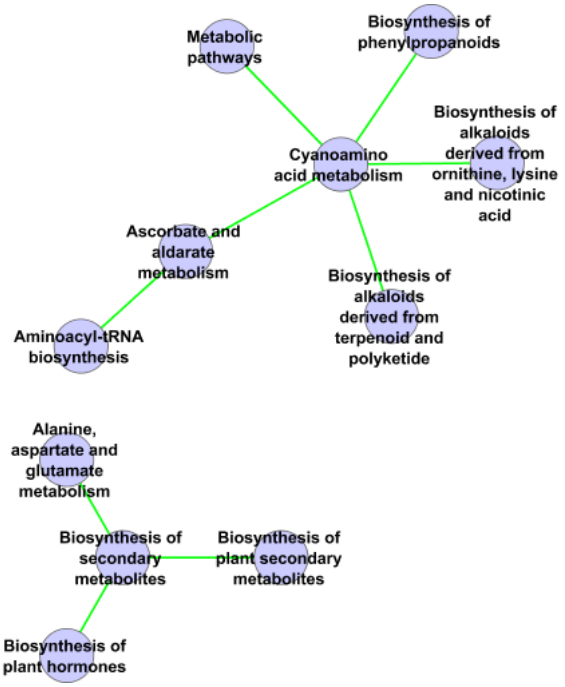


Figure 3.27: Enrichment network derived from CO_2 elevation, R_{MM} and KEGG pathways. Bin-specific p-value estimation was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

reveal any pathways for which high CVs are overrepresented among the metabolites.

For the enrichment analysis of pairs of KEGG pathways, metabolites of 'Cyanoamino acid metabolism' are enriched for positive correlation values with metabolites of 'Alkaloid biosynthesis', 'Ascorbate and aldarate metabolism' and 'Biosynthesis of phenylpropanoids' (see Figure 3.27).

Results drawn from R_{MT}: Only a very small number of metabolite-to-transcript annotation pairs were enriched for highly positive or negative correlation values. Note that the only nodes with out-degree greater than zero are 'Metabolic pathways', 'Aminoacyl-tRNA biosynthesis' and 'Biosynthesis of secondary metabolites'. In other words, the nodes which represent groups of metabolites are fairly general due to the small number of metabolites in the data set (see Figure 3.28). Nevertheless, the pathways 'Ubiquitin mediated proteolysis' and 'Proteasome', which are classified as 'Folding, Sorting and Degradation', and 'Aminoacyl-tRNA biosynthesis', which is classified as 'Translation', according to KEGG BRITE [36], indicate global changes in protein composition in parallel with shifts in metabolite concentrations (see Figure 3.28). Moreover, 'Photosynthesis' was revealed to be enriched for negative correlation values with these rearrangement processes.

Results drawn from $\mathbf{R_{TT}}$: The top ranking results for the within annotation label enrichment analysis according to bin-specific p-value estimation for the CV ranges

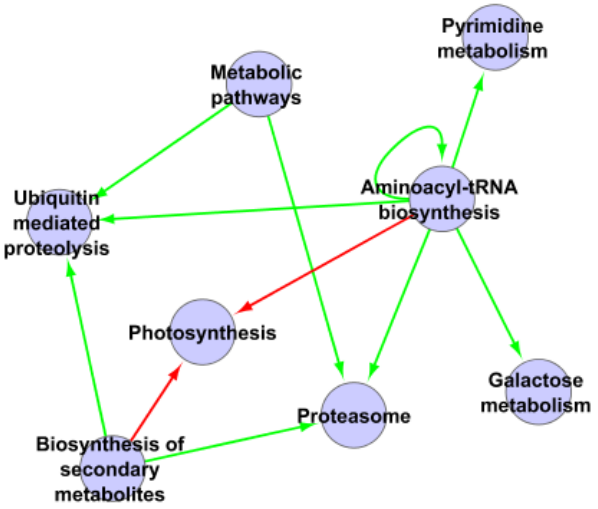


Figure 3.28: Enrichment network derived from CO_2 elevation, R_{MT} and KEGG pathways. Bin-specific p-value estimation was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively. The arrows point from metabolites of some category to the transcripts of another category.

$[-1, -0.8)$ and $[0.8, 1]$ listed below for each annotation library. Using KEGG pathways, the pathways 'Photosynthesis', 'Photosynthesis - antenna proteins', 'Purine metabolism', 'Proteasome' and 'Pyrimidine metabolism' were most pronouncedly enriched for positive CVs. Similarly, 'Calvin-Benson-Bassham cycle' (aka. carbon fixation), 'Rubisco shunt', 'Photorespiration' and 'Photosynthesis light reactions' were enriched for positive correlation values within AraCyc pathways. Additionally, 'NAD/NADH phosphorylation and dephosphorylation', 'TCA cycle variation III & V' and 'Aerobic respiration – electron donor II & III' were enriched for positive correlation values by integrating AraCyc pathways.

Within group enrichment based on GO terms resulted in the following top ranking terms: 'Apoptotic process', 'Response to biotic stimulus', 'Photosynthesis', 'Photosynthesis light harvesting', 'Response to chitin', 'Ribosome biogenesis', 'Phosphate ion transport' and 'Cellular response to phosphate starvation'.

For the InterPro within group enrichment top ranking domains were 'Pentatricopeptide repeat', 'Protein kinase-like domain', 'Protein kinase, catalytic domain', 'Serine/threonine-protein kinase, active site', 'AAA+ ATPase domain', 'DNA-binding, integrase-type' and 'AP2/ERF domain'.

'Pentatricopeptide repeats' were found in proteins which are related to mitochondria or chloroplasts [28]. The protein kinase related domains might suggest expression changes for genes associated with signal transduction, while, DNA-binding domains and AP/ERF are identified with transcription factors [28].

Inspection of the network reveals that 'Photosynthesis light reactions' is linked to 'Calvin-Benson-Bassham cycle' and 'Rubisco shunt' (see Figure 3.29). Furthermore, 'Glycolysis' is enriched for negative correlation values with 'Photosynthesis light reaction' and 'Photorespiration' and genes of 'NAD/NADH phosphorylation and dephosphorylation' are are overrepresented for positive CVs with genes of 'TCA cycle', which in turn is linked to 'Aerobic respiration – electron donor'.

For the pairwise enrichment analysis using KEGG pathways, 'Proteasome', 'Ubiquitin mediated proteolysis', 'Protein processing in endoplasmatic reticulum', 'RNA polymerase', 'Spliceosome', 'Purine metabolism', 'Pyrimidine metabolism' and 'Aminoacyl-

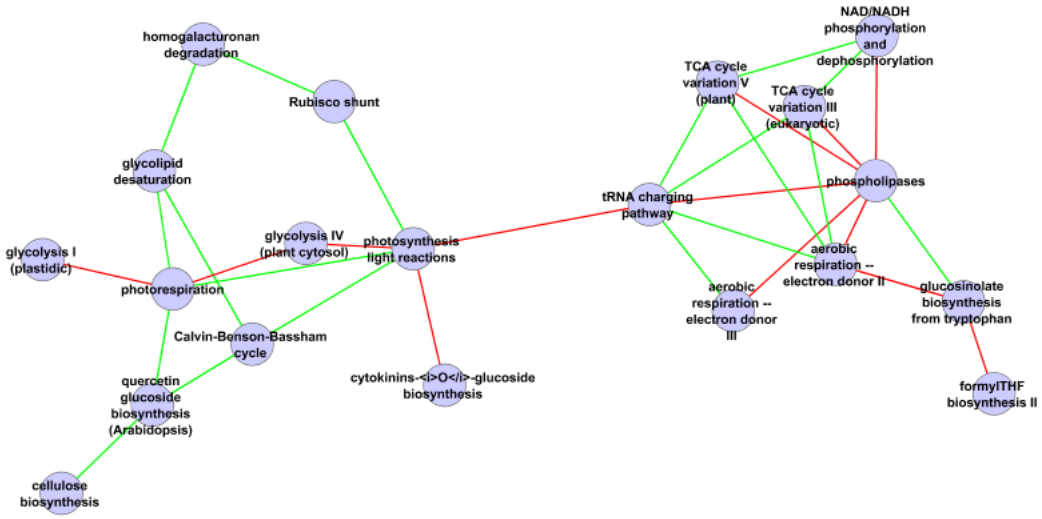


Figure 3.29: Enrichment network derived from CO_2 elevation, R_{TT} and AraCyc pathways. Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

tRNA biosynthesis' were found to be the nodes with the highest node degrees, which suggests a global change in gene expression and protein abundance (see Figure 3.30). Moreover, genes of 'Photosynthesis' were revealed to be enriched for positive CVs with 'Glyoxylate and dicarboxylate metabolism'. For the pairwise enrichment analysis using the GO terms, 'Response to biotic stimulus', 'Ribosome biogenesis', 'Response to chitin' and 'Phosphate transport' are hub nodes of the network. These pathways are also top ranked using the within annotation label enrichment analysis (see above). Furthermore, 'Nucleosome assembly', 'Protein transport', 'Protein targeting to mitochondrion' and 'Phloem or xylem histogenesis' are enriched for positive correlation values with 'Ribosome biogenesis'. 'Response to ethylene stimulus', 'Response to jasmonic acid stimulus' and 'Jasmonic acid mediated signaling pathway' are enriched for positive correlation values with 'Response to biotic stimulus' (see Figure 3.31).

According to the results by integrating InterPro domains, several transcription factors related domains (e.g. 'Pathogenesis-related transcriptional factor/ERF, DNA-binding' or 'Transcription regulator HTH, Myb-type DNA-binding') as well as to domains related with functions in signal transduction (e.g. 'Protein kinase, ATP binding site' or 'Serine-threonine/tyrosine-protein kinase catalytic domain') are linked to one another and to other domains. This suggests changes in regulatory activity on the transcription level. Some of the protein kinases specific domains are also linked to cytochrome P450-specific domains.

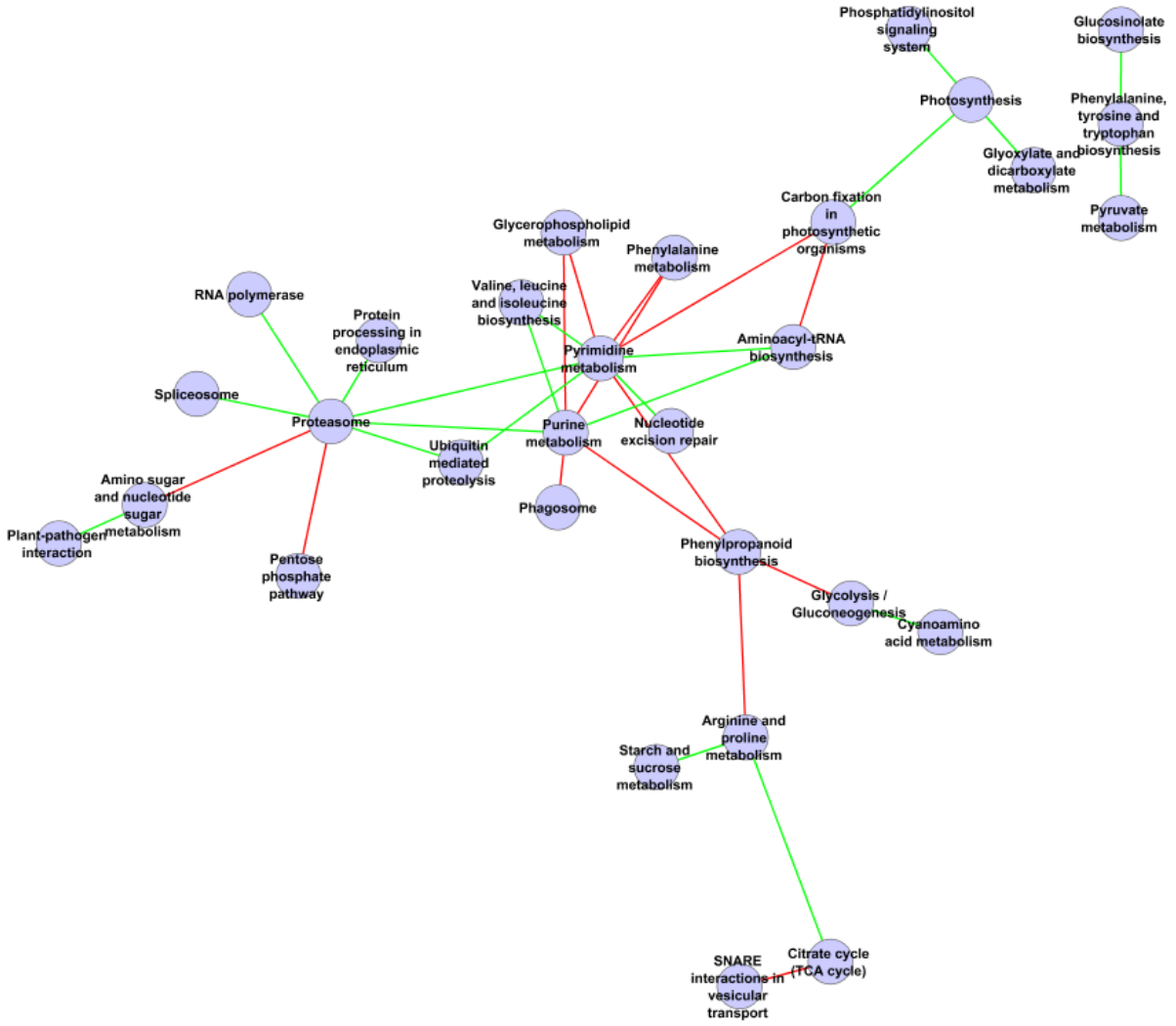


Figure 3.30: Enrichment network derived from CO_2 elevation, R_{TT} and KEGG pathways. Bootstrapped $P(\text{erroneous decision})$ estimation for the CV intervals $[-1, -0.8)$ and $[0.8, 1]$ was used. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.8)$ and $[0.8, 1]$, respectively.

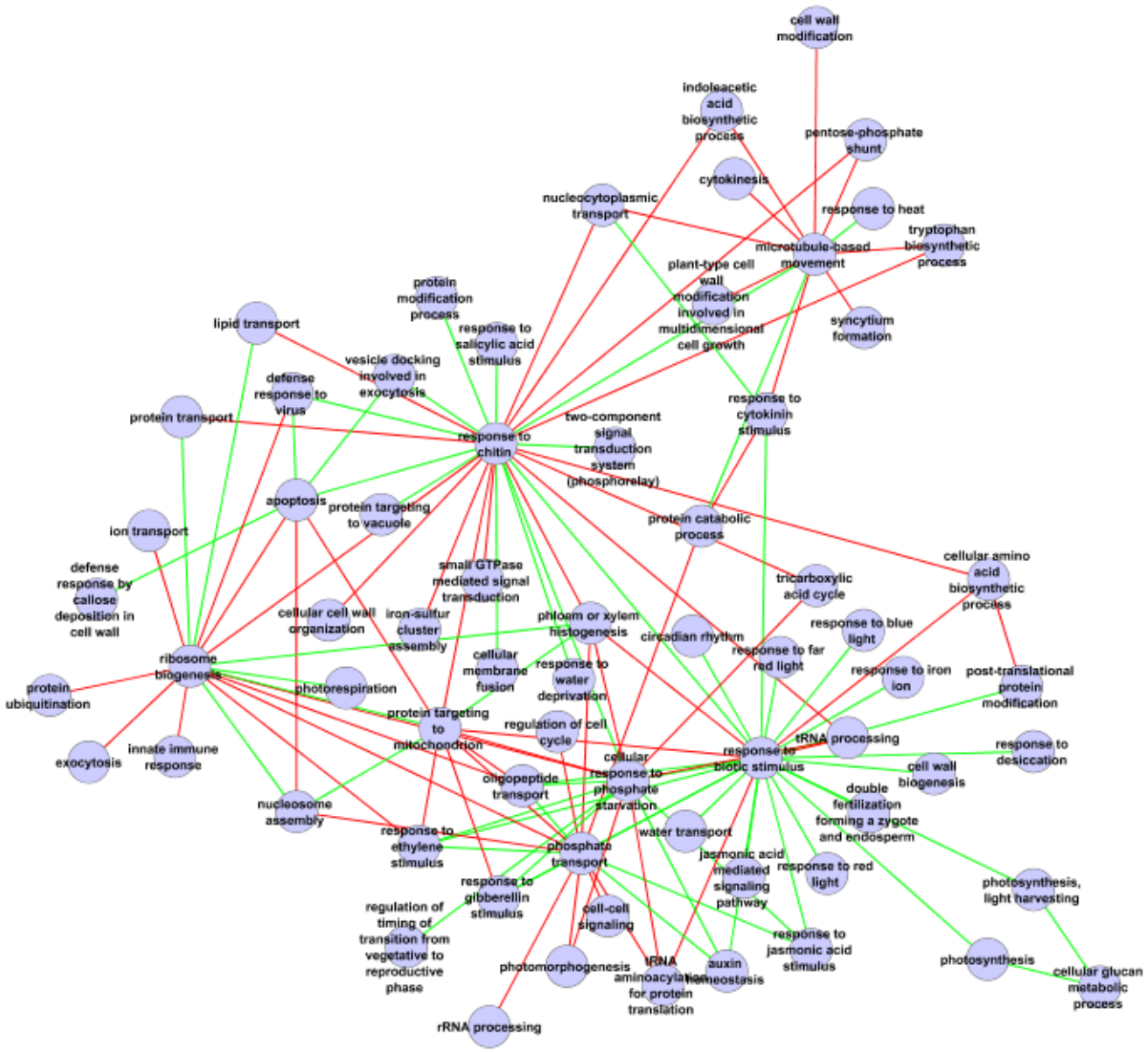


Figure 3.31: Enrichment network derived from CO_2 elevation, R_{TT} and GO terms (biological processes). Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used followed by truncation to the top ranked pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

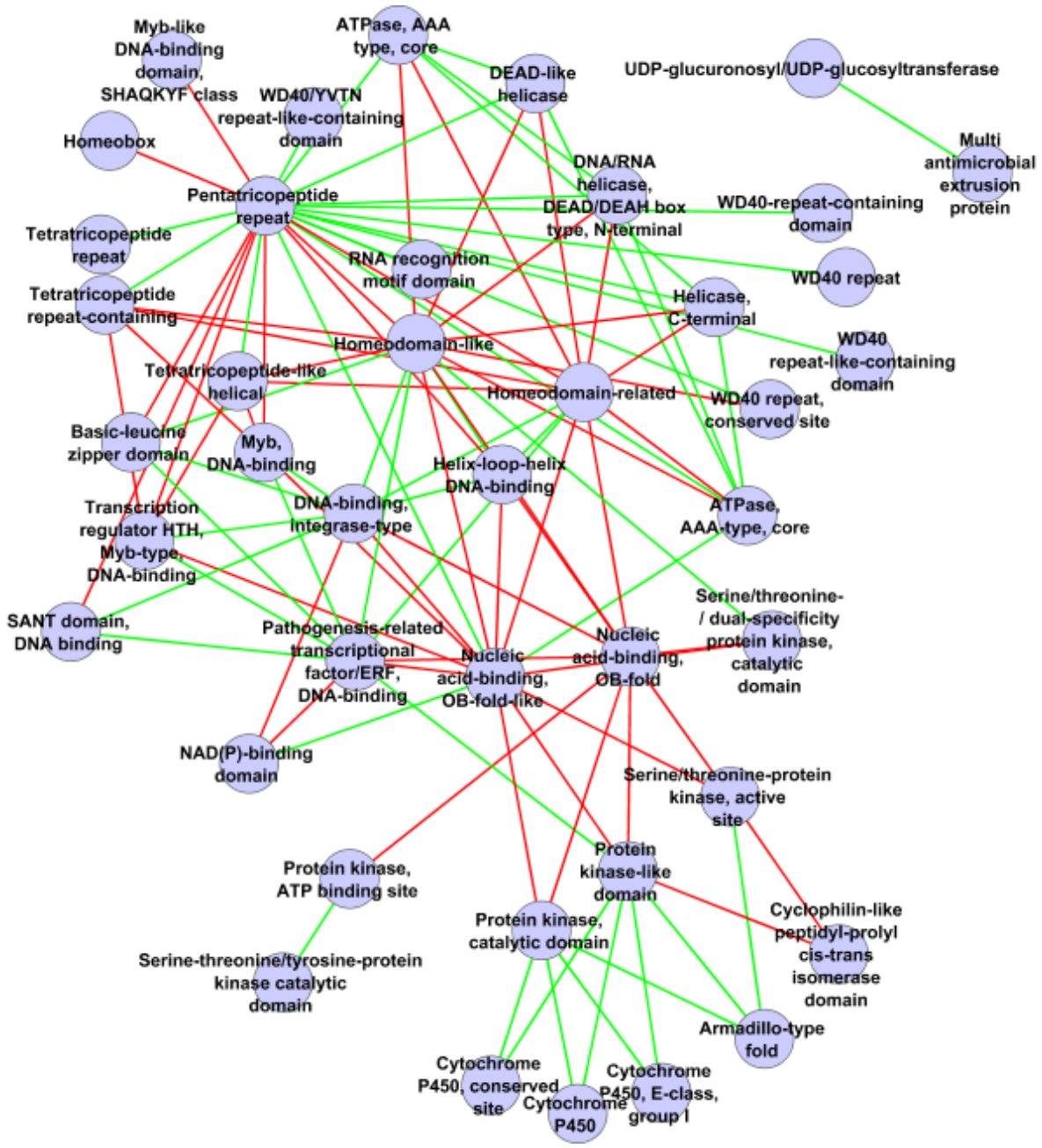


Figure 3.32: Enrichment network derived from CO_2 elevation, R_{TT} and InterPro PDs. Bin-specific p-value estimation for the CV intervals $[-1, -0.4)$ and $[0.4, 1]$ was used followed by truncation to the top ranked pairs (60 for positive and 60 for negative CVs) according to the p-values. Red and green links correspond to significantly enriched correlation values in the ranges $[-1, -0.4)$ and $[0.4, 1]$, respectively.

Chapter 4

Discussion

Similarity measures like Pearson's correlation measure have successfully been applied in order to elucidate relationships among molecule levels and putative regulatory dependencies in biological systems (e.g. [6, 2]). Recently, several research groups have proposed to take advantage of parallel analysis of several omics-data sets (e.g. transcriptome and metabolome). It has been demonstrated that parallel analysis of i.e. transcriptome and metabolome data sets potentially illuminate regulatory connectivity between the cellular levels, which in the case of separately analysing the omics-data sets would not emerge [11, 18, 19]. On the contrary, firstly, due the complicated regulatory dependencies between e.g. metabolome and transcriptome and, secondly, due to the lack of information about i.e. protein content, post-translationional modifications or enzymatic activity the interpretation of metabolite-transcript correlations is a challenging task [20].

The goal of the thesis was to analyse and integrate transcriptome and metabolome data sets using correlation matrices and various annotation libraries. In particular, the potential usefulness of correlations to reveal biologically relevant information about the underlaying system were assessed to answer two general questions: (1) Is there a general connection between functional relationship and the observed correlation value and, (2) which annotation labels (e.g. reaction pathways or GO terms) are deregulated and how are these processes related to each other?

4.1 Analysis of Global Correlation Profiles

As shown in the result section, for transcript-transcript R_{TT} and metabolite-metabolite R_{MM} correlation matrices a considerable fraction of high correlation values can be attributed to some biological function. For some annotation databases, hence, the conditional probabilities $P(Corr|\ some\ functional\ relation)$ and $P(Corr|\ no\ functional\ relation)$ differ especially for the high correlation values (e.g. high positive correlation values are more likely if some functional relationship exists), which was, however, not the case for KEGG pathways and GO terms. Unfortunately, an overwhelming fraction of high positive or negative correlation values was also found when no functional or physical relationship was present. In other words, the prior probability of 'no common function' is much higher than the prior probability of 'any functional relation'. This in turn implies that it is impossible to use this global scale information to infer the functional relationship after observing some correlation value (e.g. using Bayes' rule to query $P(any\ functional\ relation|Corr)$).

Nevertheless, it is still reasonable to assume that functional grouping of genes/metabolites might be connected to high correlation values among groups of molecules, albeit, on the global scale i.e. a functional relationship between genes/metabolites was averaged.

Put differently, correlation values in a local context might be useful to identify putative, co-regulatory processes between molecules.

For the cold and the CO_2 data set, a global shift towards positive correlation values was observed for R_{MM}. One explanation for this observation might be substantial effects on metabolic adaption processes of the plant, imposed by the experimental conditions. Note, however, that markedly different numbers and types metabolites were acquired be the research groups, which might have biased this analysis. For instance, a considerable number of metabolites are amino acids in all data sets. The problem of metabolome data acquisition was also discussed in Nikiforova *et al.* [37]. Nevertheless, the analysis showed that metabolic profiles convey relevant information about the cellular state, which might be used to address questions regarding i.e. metabolic regulation.

On a global scale metabolite-transcript correlation matrices R_{MT} do not seem to provide much information about pathway co-occurrence of the molecular species. Hence, one could argue that R_{MT} is much harder to interpret than, for instance, correlations among genes or metabolites separately. In other words, there is no easily applicable principle to reveal the cause of an observed correlation value between a metabolite-to-transcript pair. This seems also clear because of fundamentally different regulatory principles within these cellular layers and the complicated intertwinedness of the layers. Moreover, there is no information present i.e. about the protein abundances, enzyme activities or post-translational modifications, which challenges the interpretation of metabolite-transcript correlations [20, 39, 40, 41]. Despite these complications, Urbanczyk-Wochniak *et al.* [11] have demonstrated that some metabolite-transcript correlations convey information about regulatory connectivity.

4.2 Novel Enrichment Algorithm

Though the global correlation value distribution is connected with functional relationship for some correlation matrices and annotation libraries, this analysis does not provide any information about which particular functions/pathways are actually deregulated. Furthermore, interpretation of correlation values between single pairs of molecules has been demonstrated to perform poorly, because of many indirect effects and spurious high absolute correlation values [20]. One strategy to overcome this issue is by grouping genes/metabolites *a priori* according to their functional relationship and perform an enrichment or overrepresentation analysis (e.g. Gene set enrichment analysis [42] or BiNGO [23] are two popular examples, which reveal groups of statistically differentially regulated groups of genes). The success of these methods motivated the development of a novel approach which integrates correlation matrices and annotation libraries in order to identify groups of molecules for which high absolute correlation values are overrepresented. The results of the approach putatively elucidate regulatory interactions between and within groups of genes/metabolites.

In general, the algorithm seeks to filter out within or among annotation label enrichment of high correlation values. Hence, in contrast to the above section, where the correlation matrix was analysed on a global scale, this algorithm examines local (sub)-correlation matrices that are enriched for high (positive or negative) correlation values.

One major advantage of the method is its flexibility to integrate different data sources. This was demonstrated in this thesis using measurements of metabolite levels and measurements of transcript abundances. The method exploits the histogram of correlation values in order to aggregate (dis)similarity information among groups of transcripts/metabolites. Groups of molecules were formed by incorporating annotation libraries, which en-

abled the generation of a system-wide picture of pathway relationships, relationships between biological processes or relationships among protein domains.

The method was employed for within as well as for among annotation label enrichment of high correlation values. Moreover, it was proposed to visualize the entirety of significant pairs annotation label in terms of networks, referred to as enrichment networks herein, because of the eased interpretation process of the algorithm's output. The enrichment network might provide the possibility to guide further in depth analysis as well as guiding the discovery of finer-grained cellular processes (e.g. which particular set of molecules was responsible for the enrichment of a pathway interaction?), but this was beyond the scope of this thesis.

Functional grouping of genes or metabolites is thought to be crucial to uncover a high-level view of the disregulated processes. Redestig *et al.* [20] argued grouping is especially important in cases where millions of correlation values are investigated. ClueGO [24] is a bioinformatics tool which generates networks among annotation labels, which is therefore similar to the proposed method. However, ClueGO links annotation labels based on the similarity of gene assignments to the annotation labels using kappa statistics. In contrast, herein, the annotation associations emerge from the measured data sets. Furthermore, by testing for enrichment of negative CVs the proposed method not only concentrates on group similarity, but also on potentially counter-regulated groups of transcripts/metabolites, which offers a richer view on cellular processes.

In the following, the properties and limitations of the algorithm are recapitulated along with possible extensions or improvements.

Firstly, in this thesis, correlation matrices where generated using Pearson's correlation measure. Other possible correlation measures might be non-parametric correlation measures (e.g. Kendall's tau) or mutual information, these are more appropriate for capturing nonlinear relationships at the price of a lower statistical power. Thus, much more biological replicates would be required to get reliable results with those correlation measures [22]. On the other hand, Pearson's correlation measure is sensitive to shifts among the transcript or metabolite time-series profiles. Redestig *et al.* [20] have proposed a HMM-based correlation measure which copes with shifts and relatively high noise levels in time-series profiles. More elaborate similarity measures might also be utilized for this analysis, for instance, the local shape-similarity measure proposed by Balasubramaniyan *et al.* [43]. Though, this similarity measure is not appropriate for short-time series profiles as encountered in this thesis.

Secondly, obviously the quality of the generated results strongly depends on the quality of the annotation database as well as the comprehensiveness of the acquired omics-data. It has been argued that metabolite-metabolite or metabolite-transcript correlation values are more difficult to interpret than i.e. transcript-transcript correlations due to the nature of the interdependencies and regulatory mechanisms. The interpretation was also challenged by the fact that only a relatively small number of metabolites was identified (compare Table 2.1). As these correspond to only few metabolite classes (e.g. amino acids or carbonhydrates), the metabolome view is far from complete. This is also reflected by fairly general enriched metabolite annotation labels (e.g. 'Metabolic pathway' or 'Biosynthesis of secondary metabolites') considering the analysis drawn on R_{MM} or R_{MT} in the case studies. With the advancement of metabolomic analysis techniques and the improvement and extension of the annotation libraries the approach is believed gain its usefulness to integrate omic-data across different levels. A further aspect, that challenges the analysis with this approach is the nature of the data sets. For instance, for the cold stress data set, a relatively high fraction of high correlation values was observed compared with for the sulphur limitation experiments. Hence, using a single significance level (e.g. 5%) results in

an excessive number of significant results for the former example, whereas, only very few significant instances emerge for the latter one. This challenges the visual representation and inspection of the generated enrichment networks. To address this issue, different statistical tests were applied to optimize the network size. For instance, if a large number of significant instances was found only the top ranking results were presented or a more stringent test was used (e.g. bootstrapping approach).

Thirdly, annotation libraries often comprise annotation labels which are hierarchically structured or strongly overlapping. In other words, there might be several different significant annotation label pairs which are basically induced by a common set of molecules (e.g. occurrence of WD40-repeat InterPro domains in Figure 3.32). A consequence of this is that the absence of edges between two nodes in the enrichment networks might either be caused by the acceptance of the null hypothesis (no enrichment of high CVs) or due to the absence of a statistical test, which challenges the intuitive interpretation of the networks. A possible improvement would be to preprocess the annotation library so as to make the library non-redundant (e.g. remove hierarchy superclasses, compute redundancies, remove strongly overlapping annotations, introduce annotations which represent the overlap, etc.). Removal of redundancies due to the hierarchical structure of annotation libraries is also performed by ClueGO [24].

Fourth, it is difficult to define a general notion of which histogram profiles need to be considered as statistically significant. In this thesis, it was claimed that high correlation values are much more informative than low correlation values. Therefore, the method concentrated on testing the overrepresentation of histogram bin counts for high correlation values, rather than examining the entire shape of the histogram. Thus, this rises the question of which correlation value intervals to use for the test. In this thesis, two variants have been compared with each other (e.g. using the CV intervals $[-1, -0.8)$ and $[0.8, 1]$ in contrast to $[-1, -0.4)$ and $[0.4, 1]$). The enrichment analysis for correlation values on the interval $[-1, -0.4)$ and $[0.4, 1]$ results in an implausible high number of significant annotation pairs. Fortunately, however, the top ranking significant annotation labels or annotation pairs showed high concordance between the bin size variants, which underlines the robustness with respect to the ranks of the result lists.

A variant which was intended to capture the entire histogram shape was the weighted χ^2-test. However, it is difficult to define weights for the χ^2-test which capture adequately biological relevant features. Further, it suffers from the violation of the i.i.d. assumption. Therefore, this variant was not used for the case studies.

Fifth, the null distribution was computed once for each input correlation matrix and annotation label set cardinality by permuting the molecules labels. The parameters of the subsequently fitted gamma distribution were stored and reused for each annotation label with the same set size. This yields huge computational benefits.

Unfortunately, in some cases this particular sampling strategy has resulted in 'over-counting' of high correlation values, hence to overestimation of the p-values. Especially in the case of pairwise enrichment analysis where the molecules of one of the participating annotation labels are tightly correlated. Based on the same systematic biases, it was also observed that hub nodes in the enrichment networks, in general, also are annotations with high enrichment for high CVs. Consequently, links in the enrichment network which connect a hub node and a node with only one edge might likely correspond to a false positive result.

To overcome this issue, a different permutation sampling approach might be implemented. For the pairwise enrichment analysis, i.e. clamping the gene/metabolite labels of one annotation label while shuffling the annotation labels of the other annotation label might be considered. This would result in different null distributions for each correlation

matrix, annotation label and set size of the partner label, as opposed to the variant used herein, where different set sizes correspond to different null distributions. In this case, however, the efficiency gained by storing and reusing the null distribution parameters is largely lost resulting in a much larger algorithmic runtime. Despite the systematic bias due to simplified assumptions for the permutation sampling, highly significant results might still be of biological relevance.

Several strategies for assessing statistical significance (e.g. p-value estimation based on the fitted gamma distribution, bootstrapped and Jackknife alternative distribution estimation, etc.) have been implemented and compared against each other. They mostly differ in terms of their stringency, but show strong overlap among the generated results. This argues for the relative similarity of these tests.

4.3 Integration of KEGG BRITE

In line with Kaplan *et al.* [17] and Dutta *et al.* [16], the method reported significant co-response patterns for amino acids. This pattern was also discovered for the sulphur limitation data sets, although, Hirai *et al.* [19] did not explicitly mention this fact. As sulphur metabolism is directly linked to amino acids levels [37, 44], this metabolic links might be reflected by the results of the algorithm (compare Figure 3.14). For the cold acclimation data set, monosaccharides were identified to be enriched for negative correlation values, which might suggests a reciprocal metabolic regulation among the present monosaccharides. Stitt *et al.* [45] discussed thoroughly the important relation of acquired freezing tolerance and cellular sugar concentrations, which might be attributed to this observation (compare Table 3.6).

4.4 Sulphur Deficiency Data Set

Hirai *et al.* [19] mentioned marked changes and co-clustering of glucosinolates as well as genes associated with glucosinolate biosynthesis, which were in line with the results herein. Importantly, this observation was recovered from R_{TT}, R_{MM} and R_{MT} (compare Figures 3.15, 3.16, 3.17, 3.18 and 3.19).

Nikiforova *et al.* [37] revised molecular transport and related biochemical pathways, which are affected by sulphur limitation. For instance, sulphur metabolism is linked by biochemical reactions directly to glycine and serine metabolism (compare Figures 3.15, 3.18).

Hirai's clustering analysis (BL-SOM) [18] reported genes of photosynthesis and glucosinolate biosynthesis pathway to be tightly co-expressed for the leaf data set, which was also detected by the enrichment network approach (compare Figures 3.17, 3.18 and 3.19). A known phenotypic consequence of sulphur depletion is chlorosis in leaves [37], which refers to a lack of chlorophyll, which might be connected to the chlorophyll related annotations for several annotation libraries. In addition, Hirai *et al.* [18] reported major changes in primary metabolism and glucosinolate biosynthesis as well as photosynthesis in roots, which were also revealed by the proposed method (compare Figure 3.17a).

The algorithm discovered an association between 'Glucosinolate biosynthesis' and 'Sulfate assimilation', which is in line with the review of Rausch *et al.* [44] (compare Figures 3.17, 3.18 and 3.19), who mentioned that plants under sulphur limitation assimilate to this environmental condition. Several sulphur containing molecules such as glucosinolates, glutathiones play an important role in plant defence mechanisms [44]. Inspection of the enrichment networks recover many plant defence specific annotations and relationships

among the annotations, which are in line with current knowledge about the relationships. Among those, the presence 'Response to ethylene' (in the GO-based network) suggests a modulation of plant defense responses [46] (compare Figures 3.16a and 3.19).

Furthermore, the InterPro domain-based networks highlight several Glutathione-S-transferase specific domains, which are important to dispose of xenobiotics [47]. Glutathiones are also known to provide a redox buffer against reactive oxygen [48, 44]. Genes containing glutathione-S-transferase domain are enriched for negative correlations with EF-hand containing genes, according to the networks. Both, glutathione-S-transferases and EF hand proteins have been linked to plant response upon insect wounding [49] (compare Figure 3.20). Importantly, Hirai *et al.* [19] also noted regulatory changes in genes coding for glutathione-S-transferases.

Another interesting aspect is that plants upon sulphur limitation undergo morphological changes, such as enhancement of root growth [37]. Many nodes and interactions in the enrichment networks are in line with this observation (e.g. pathways and terms that are associated with transcription, DNA replication and translation; compare Figures 3.18 and 3.19). Furthermore, many transcription factor or protein kinase related protein domains are statistically interacting with each other, according to the results. It might be speculated that some of the proteins associated with the domains play a role in the realization of these morphological adaptations (compare Figure 3.20).

4.5 Cold Acclimation Data Set

During cold acclimation plants exhibit a number of cellular changes. Stitt *et al.* [45] discussed thoroughly the tight correlation between the development of freezing tolerance and sugar levels. The enrichment networks revealed several pathways sugar metabolism (e.g. starch biosynthesis and degradation, galactose degradation, sucrose degradation, gluconeogenesis in Figures 3.23, 3.24 and 3.25).

Moreover, the networks also revealed tightly linked pathways/terms related with glycolysis, citrate cycle, NAD/NADH phosphorylation and dephosphorylation, which are in a close connection to sugar levels [50].

Cold acclimation is also related to changes photosynthesis. Photosynthesis slows down as a biochemical consequence of cold acclimation [45]. Photosynthesis (and related annotation labels like chlorophyll, carbon fixation, etc.) are also tightly connected to many other categories in the corresponding networks, suggesting a putative regulatory connection between those nodes (compare Figures 3.23, 3.24 and 3.25).

In line with Kaplan *et al.* [17], enrichment of positive CVs is present between 'Sulphur assimilation' and 'Photosynthesis' (compare Figure 3.25).

Another interesting aspect of plants exhibiting freezing tolerance is that water content is reduced in the cells, while the protein content is increased [45, 17]. An inspection of the networks shows that many annotation labels indicate a general shift in protein abundance and transcriptional activity (e.g. 'Proteasome', 'Splicosome' and 'Aminoacyl-tRNA biosynthesis' in Figures 3.24 and 3.25). These findings might be linked to the reported increase in protein abundance during cold acclimation.

In agreement with Kaplan *et al.* [17], the algorithm reports a general increase of amino acid biosynthesis genes and an change in amino acid abundance (compare Figure 3.14).

As reported by Hannah *et al.* [51], cold acclimation causes downregulation of photosynthesis and upregulation of flavonoid metabolism. These reciprocally regulated processes also emerged from the enrichment network based on AraCyc pathway (compare Figure 3.23) .

The enrichment networks suggests similarities between gene expression changes of glutathione metabolism, proteasome and spliceosome (compare Figure 3.24), hence all are up regulated. Taking together that glutathione is a prominent antioxidant [48] and concentrations of antioxidants are increase upon cold acclimation [45], this might suggest an important role of glutathiones during cold acclimation.

On the same data set, Kaplan *et al.* [17] reports increased expression of genes associated the synthesis of phospholipids, starch, sugar, flavonoid, protein amino acids and terpenoid biosynthesis, while genes of photorespiration, folic acid, sulphate assimilation, ethylene, fatty acid, gluconeogenesis, amino acids, brassinosteroids and chlorophyll biosynthesis were found as downregulated. Aggregating the information of the enrichment networks based on several annotation libraries, virtually all of the previously mentioned processes where present in the generated results (compare Figures 3.23, 3.24 and 3.25).

In particular for the cold data set and the pairwise enrichment analysis drawn on InterPro domains resulted in an excessive number of significant instances, even after truncation of the results to the top ranked. In fact, all links correspond to p-value=0 in Figure 3.26. An explanation for this observation might be that, on the one hand, a much higher number of pairwise tests were performed (compare Table 3.1; higher number of gene-pairs with common InterPro domain) and, on the other hand, as the global correlation profile analysis has revealed, protein domains are informative for the explanation of observed correlation values (see Figure 3.1; in comparison with e.g. GO term, KEGG pathways).

Nevertheless, nodes with the high node degrees might be biologically relevant domains. For instance, some proteins containing 'Pentatricopeptide repeat' have been associated with post-transcriptional processes such as sequence-specific RNA-binding proteins, RNA stabilization or RNA processing [28] (compare Figure 3.26). Proteins carrying a 'DEAD-/DEAH box type, N-terminal, RNA helicase' domain are thought to be important for various processes such as RNA synthesis or degradation (e.g. splicing, transport, ribosome biosynthesis, translation or degradation) [28] and 'WD40 repeats' facilitate protein-protein interactions and occur in proteins which participate in signal transduction, transcriptional regulation to cell cycle control and apoptosis [28].

These protein domains underline general shifts in the protein level due to the acquired freezing tolerance.

4.6 Carbon Dioxide Elevation Data Set

In line with Dutta *et al.* [16], who reported an increase in amino acid abundance during the duration of the experiment, the method revealed changes of amino acid levels (compare Figure 3.14). In particular, it was reported that amino acids are accumulated.

Furthermore, transcriptional regulation of primary metabolism is tightly interacting, according to the enrichment networks (e.g. see glycolysis, citrate cycle or NAD/NADH (de)phosphorylation; compare Figure 3.29), which fits into the picture about the biological relationship of TCA and electron transport and oxidative phosphorylation ([50] chapter 16 and 17).

Moreover, genes of 'Photosynthesis' were revealed to be enriched for positive CVs with 'Glyoxylate and dicarboxylate metabolism', which is in agreement with the findings of Dutta *et al.* [16]. They noted a significant decrease of photosynthesis and carbon utilization, which is present in the networks be links between photosynthesis-related annotations and Calvin-Benson-Bassham cycle or carbon fixation (compare Figure 3.29 and 3.30). Additionally it was revealed that 'Photosynthesis light reactions' is linked to 'Rubisco shunt' (compare Figure 3.29). Dutta *et al.* [16] mentioned that carbon fixation

competes with photosynthesis for the activity of rubisco, which might be reflected by this part of the network.

Dutta *et al.* [16] reported a similar decline of mRNA levels of genes in 'Photosynthesis' and gene related with glyoxylate, glycolate and glycerate metabolism, which agrees with the result generated by the proposed method (see 'Photosynthesis' and 'Glyoxylate and dicarboxylate metabolism'; Figure 3.30).

It is known that plants exhibit enhanced growth and structural changes upon increased CO_2 levels [52, 53]. In line with this, there is strong evidence for adaption processes of the plant due to tightly interlinked KEGG annotations like 'Proteasome', 'Ubiquitin mediated proteolysis', 'Protein processing in endoplasmatic reticulum', 'RNA polymerase', 'Spliceosome', 'Purine metabolism', 'Pyrimidine metabolism' and 'Aminoacyl-tRNA biosynthesis', which suggests a global change in gene expression and protein abundance (see Figure 3.30). In other words, this suggests that some proteins are degradated while at the same time other proteins are synthesized. Moreover, i.e. 'Nucleosome assembly', 'Protein transport', 'Protein targeting to mitochondrion' or 'Phloem or xylem histogenesis' are enriched for positive correlation values with 'Ribosome biogenesis' in the GO-based network (compare Figure 3.31), which underlines the above argument.

The analysis using InterPro annotations revealed genes carrying a 'Pentatricopeptide repeat' domain to be tightly co-expressed. Proteins with such domains are related with targets in mitochondria or chloroplasts [28]. Some of these proteins might have a putative role in sequence-specific RNA-binding [28, 54].

The enrichment analysis revealed protein kinase specific domains within and among InterPro domains. They suggest changes in the expression of genes coding for signalling proteins. Moreover, 'DNA-binding integrase-type' indicates the change in gene expression of transcription factors (compare Figure 3.32). Interestingly, the AP2/ERF domain, which was studied by Pré *et al.* [46] to reveal its role in ethylene-mediated signalling in plant defense, occurs among the top ranked InterPro domains for within annotation label enrichment. As described, jasmonic acid, beside ethylene, is another important hormone playing a role in plant defense [46]. Both, ethylene and jasmonic acid are associated with AP2/ERF domain carrying transcription factors to respond to particular pathogens. In line with these results, the proposed method revealed relationships between 'Response to ethylene stimulus', 'Response to jasmonic acid stimulus', 'Jasmonic acid mediated signalling pathway' and 'Response to biotic stimulus' (compare Figure 3.31). Dutta *et al.* [16], firstly, linked CO_2 elevation to effects in ethylene biosynthesis, which can be confirmed by the approach proposed in this thesis. Furthermore, they conducted an time-point specific differential expression analysis which resulted in the enrichment of 'Defense response' and 'Response to biotic stimulus'. The latter one, also occurs among the top ranked GO terms revealed by the proposed method herein. The network based on the InterPro integration recovered 'Pathogenesis-related transcriptional factor/ERF, DNA-binding', which agrees with the knowledge about ethylene and jasmonic acid induced response to pathogens [46]. Furthermore, for the InterPro network, several other transcription factor specific domains can be found, some of which are also linked to 'Pathogenesis-related transcriptional factor/ERF, DNA-binding'. The presence of edges between protein kinases specific domains and ERF transcription factor domains might suggest adaption processes in signal transduction (compare Figure 3.32).

The presence of links between cytochrome P450 domains with protein kinase domains, suggests gene expression changes in signal transduction in parallel with expression changes in cytochrome P450. Cytochrome P450 is a family of genes which catalyse a wide range of biochemical reactions. Among the many functions, cytochrome P450 participates in the biosynthesis or degradation of hormones, signalling molecules, defense compounds and

xenobiotics [55]. In this context, it might be speculated that cytochrome P450 plays a role in coping with biotic stress or plant defence compounds, which would fit into the picture of sustained ethylene signalling [16].

4.7 Conclusion and Future directions

The proposed algorithm recovers results from recent studies [16, 17, 18, 19, 20] on R_{TT} by integrating various annotation libraries, though there is still potential to improve to method's performance. In addition, many findings, which were not explicitly reported in these papers, can be linked to evidence of similar studies (e.g. AP2/ERF domain which is linked to ethylene and jasmonic acid-related signaling, was detected as one of the top ranked results for the within enrichment analysis, CO_2 data set and R_{TT} [46, 28]). Especially, top ranked annotation labels or annotation pairs (e.g. in terms of their p-values or $P(erroneous\ decision)$) recover many previously mentioned cellular adaptation processes in response to these environmental stresses. Hence, this thesis provides a proof of concept for a novel enrichment analysis approach, which is believed to be valuable for identifying high level system-wide regulatory relationships and, additionally, might be useful to guide further experimental studies. However, in order to elucidate the true false positve and false negative rate of the method it is necessary to compare the performance of the algorithm either to a bench mark data set or to a similar bioinformatical approach (e.g. ClueGO [24]), which is part of future work.

In general, R_{MM} is much harder to interpret than R_{TT} [6, 8]. The proposed algorithm in general succeeded in finding several associations among metabolites based on their level profile. For instance, amino acids were found to be tightly correlated in all data sets. However, for R_{MM} often only fairly general annotation labels can be used for statistical tests, because of the small number of acquired metabolites and the limitation of using at least five metabolites for the enrichment test.

An even harder task is concerned with the interpretation of R_{MT}. The quality of the algorithm depends strongly on the comprehension of the annotation libraries and the number of molecules acquired. Hence, the small number of identified metabolites challenged the quality of the results drawn on metabolic data used in here. Nevertheless, in some cases (e.g. glucosinolate biosynthesis in S-def conditions) the algorithm produced plausible results, in concordance with the literature. However, the analysis of R_{MT} also suffered from the small number of measured metabolites. Again, metabolite nodes in the networks correspond to fairly general annotation labels hindering the interpretation process of the resulting networks.

A number of advancements and extensions might be useful for the algorithm in order to optimize its performance (e.g. improved permutation sampling approach) as well as to improve its usability for biologists.

Out of the developments throughout this thesis, either a bioconductor package or a Cytoscape plug-in might be developed. With particular focus on the optimization of the method, the network visualization and network annotation and the user-interactive navigation within the network. For instance, a richer set of graph annotation might be useful such as color code to represent within annotation libraries or line width might be associated with the p-value levels. Furthermore, parameters of the algorithm like significance levels, significance test variant or bin intervals in the histogram should interactively navigated. An important point is also to remove redundancies within annotation libraries, e.g. as proposed by Bindea *et al.* [24].

With the establishment of new and advanced high-throughput solutions (e.g. the

establishment of maps of molecule-kinase interactions [56] or biotechnological approaches to measure the affinity of chemical compounds to proteins [57, 58]), the method might also be considered for an adaption to integrate a broader spectrum of omics data.

Bibliography

[1] Brown PO, Botstein D, *et al.*: **Exploring the new world of the genome with DNA microarrays**. *Nature Genetics* 1999, **21**(1 Suppl):33–37.

[2] Stuart JM, Segal E, *et al.*: **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* 2003, **302**(5643):249–255.

[3] Basso K, Margolin AA, *et al.*: **Reverse engineering of regulatory networks in human B cells**. *Nature Genetics* 2005, **37**(4):382–390.

[4] Segal E, Shapira M, *et al.*: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nature Genetics* 2003, **34**(2):166–176.

[5] Bar-Joseph Z, Gerber GK, *et al.*: **Computational discovery of gene modules and regulatory networks**. *Nature Biotechnology* 2003, **21**(11):1337–1342.

[6] Steuer R, Kurths J, *et al.*: **Observing and interpreting correlations in metabolomic networks**. *Bioinformatics* 2003, **19**(8):1019–1026.

[7] Hynne F, Danø S, Sørensen PG: **Full-scale model of glycolysis in *Saccharomyces cerevisiae***. *Biophysical Chemistry* 2001, **94**:121–163.

[8] Camacho D, de la Fuente A, *et al.*: **The origin of correlations in metabolomics data**. *Metabolomics* 2005, **1**:53–63.

[9] Teusink B, Passarge J, *et al.*: **Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry**. *European Journal of Biochemistry* 2000, **267**(17):5313–5329.

[10] Krumsiek J, Suhre K, *et al.*: **Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data**. *BMC Systems Biology* 2011, **5**:21.

[11] Urbanczyk-Wochniak E, Luedemann A, *et al.*: **Parallel analysis of transcript and metabolic profiles: a new approach in systems biology**. *EMBO Reports* 2003, **4**(10):989–993.

[12] Bradley PH, Brauer MJ, *et al.*: **Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae**. *PLoS Computational Biology* 2009, **5**:e1000270.

[13] Takahashi H, Morioka R, *et al.*: **Dynamics of time-lagged gene-to-metabolite networks of Escherichia coli elucidated by integrative omics approach**. *Omics: A Journal of Integrative Biology* 2011, **15**(1-2):15–23.

[14] Pir P, Kırdar B, *et al.*: **Integrative investigation of metabolic and transcriptomic data**. *BMC Bioinformatics* 2006, **7**:203.

[15] Bylesjö M, Eriksson D, *et al.*: **Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data**. *The Plant Journal* 2007, **52**(6):1181–1191.

[16] Dutta B, Kanani H, *et al.*: **Time-series integrated "omic" analyses to elucidate short-term stress-induced responses in plant liquid cultures**. *Biotechnology and Bioengineering* 2009, **102**:264–279.

[17] Kaplan F, Kopka J, *et al.*: **Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content**. *The Plant Journal* 2007, **50**(6):967–981.

[18] Hirai MY, Yano M, *et al.*: **Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(27):10205.

[19] Hirai MY, Klein M, *et al.*: **Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics**. *Journal of Biological Chemistry* 2005, **280**(27):25590.

[20] Redestig H, Costa IG: **Detection and interpretation of metabolite–transcript coresponses using combined profiling data**. *Bioinformatics* 2011, **27**(13):i357.

[21] Bylesjö M, Rantalainen M, *et al.*: **OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification**. *Journal of Chemometrics* 2006, **20**(8-10):341–351.

[22] Steuer R, Kurths J, *et al.*: **The mutual information: detecting and evaluating dependencies between variables**. *Bioinformatics* 2002, **18**(suppl 2):S231–S240.

[23] Maere S, Heymans K, *et al.*: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448–3449.

[24] Bindea G, Mlecnik B, *et al.*: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25**(8):1091–1093.

[25] Rhee SY, Beavis W, *et al.*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community**. *Nucleic Acids Research* 2003, **31**:224–228.

[26] Gentleman RC, Carey VJ, *et al.*: **Bioconductor: Open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**:R80.

[27] Brandão M, Dantas L, *et al.*: **AtPIN: Arabidopsis thaliana protein interaction network**. *BMC Bioinformatics* 2009, **10**:454.

[28] Apweiler R, Attwood TK, *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Research* 2001, **29**:37–40.

[29] Mueller LA, Zhang P, *et al.*: **AraCyc: a biochemical pathway database for Arabidopsis**. *Plant Physiology* 2003, **132**(2):453–460.

[30] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011, `[http://www.R-project.org/]`. [ISBN 3-900051-07-0].

[31] Carlson M, Falcon S, *et al.*: *KEGG.db: A set of annotation maps for KEGG.* 2012, `[http://www.bioconductor.org/packages/2.10/data/annotation/html/KEGG.db.html]`. [R package version 2.4.5].

[32] Carlson M, Falcon S, *et al.*: *GO.db: A set of annotation maps describing the entire Gene Ontology.* 2012, `[http://www.bioconductor.org/packages/2.10/data/annotation/html/GO.db.html]`. [R package version 2.4.5].

[33] Chen H: *VennDiagram: Generate high-resolution Venn and Euler plots* 2011, `[http://CRAN.R-project.org/package=VennDiagram]`. [R package version 1.0.1].

[34] Shannon P, Markiel A, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Research* 2003, **13**(11):2498–2504.

[35] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300.

[36] Kanehisa M, Goto S, *et al.*: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Research* 2012, **40**(D1):D109–D114.

[37] Nikiforova VJ, Gakière B, *et al.*: **Towards dissecting nutrient metabolism in plants: a systems biology case study on sulphur metabolism**. *Journal of Experimental Botany* 2004, **55**(404):1861–1870.

[38] Xin Z, *et al.*: **Cold comfort farm: the acclimation of plants to freezing temperatures**. *Plant, Cell & Environment* 2000, **23**(9):893–902.

[39] Carrari F, Baxter C, *et al.*: **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 Physiology* 2006, **142**(4):1380–1396.

[40] Gibon Y, Usadel B, *et al.*: **Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes**. *Genome Biology* 2006, **7**(8):R76.

[41] Ladurner AG: **Rheostat control of gene expression by metabolites**. *Molecular Cell* 2006, **24**:1–11.

[42] Subramanian A, Tamayo P, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545.

[43] Balasubramaniyan R, Hüllermeier E, Weskamp N, Kämper J: **Clustering of gene expression data using a local shape-based similarity measure**. *Bioinformatics* 2005, **21**(7):1069–1077.

[44] Rausch T, Wachter A: **Sulfur metabolism: a versatile platform for launching defence operations**. *Trends in Plant Science* 2005, **10**(10):503–509.

[45] Stitt M, Hurry V: **A plant for all seasons: alterations in photosynthetic carbon metabolism during cold acclimation in *Arabidopsis***. *Current Opinion in Plant Biology* 2002, **5**(3):199–206.

[46] Pré M, Atallah M, *et al.*: **The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense**. *Plant Physiology* 2008, **147**(3):1347–1357.

[47] Leustek T, Martin MN, *et al.*: **Pathways and regulation of sulfur metabolism revealed through molecular and genetic studies**. *Annual Review of Plant Biology* 2000, **51**:141–165.

[48] Cnubben NHP, Rietjens IMCM, *et al.*: **The interplay of glutathione-related processes in antioxidant defense**. *Environmental Toxicology and Pharmacology* 2001, **10**(4):141–152.

[49] Stotz HU, Pittendrigh BR, *et al.*: **Induced plant defense responses against chewing insects. Ethylene signaling reduces resistance of Arabidopsis against Egyptian cotton worm but not diamondback moth**. *Plant Physiology* 2000, **124**(3):1007–1018.

[50] Voet D, Voet JG, *et al.*: *Fundamentals of Biochemistry*. New York, USA: John Wiley & Sons Inc. 2008.

[51] Hannah MA, Wiese D, *et al.*: **Natural genetic variation of freezing tolerance in Arabidopsis**. *Plant Physiology* 2006, **142**:98–112.

[52] Pritchard SHG, Rogers HOH, *et al.*: **Elevated CO2 and plant structure: a review**. *Global Change Biology* 1999, **5**(7):807–837.

[53] Ward JK, Strain BR: **Elevated CO2 studies: past, present and future**. *Tree Physiology* 1999, **19**(4-5):211–220.

[54] Delannoy E, Stanley WA, *et al.*: **Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles.** *Biochemical Society Transactions* 2007, **35**(Pt 6):1643.

[55] Werck-Reichhart D, Bak S, *et al.*: **Cytochromes P450**. *Arabidopsis Book* 2002, **1**:e0028.

[56] Fabian MA, Biggs WH, *et al.*: **A small molecule–kinase interaction map for clinical kinase inhibitors**. *Nature Biotechnology* 2005, **23**(3):329–336.

[57] Barglow KT, Cravatt BF: **Activity-based protein profiling for the functional annotation of enzymes**. *Nature Methods* 2007, **4**(10):822–827.

[58] Rix U, Superti-Furga G: **Target profiling of small molecules by chemical proteomics**. *Nature Chemical Biology* 2008, **5**(9):616–624.