Max Planck Institute of Molecular Plant Physiology
Department: Molecular Physiology
Research Group: Root Metabolism

Inferring Hypotheses from Complex Profile Data

- By Means of CSB.DB, a Comprehensive Systems-Biology Database -

Dissertation

A thesis submitted to the
Mathematisch-Naturwissenschaftliche Fakultät
of the
Universität Potsdam
for the degree of

'*doctor rerum naturalium*'
(Dr. rer. nat.)

by
Dirk Steinhauser

Potsdam, October 2004

The work presented in this thesis was carried out between October 2001 and September 2004 at the Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm.

1st Examiner: Prof. Dr. Lothar Willmitzer
Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

2nd Examiner: Prof. Dr. Joachim Selbig
Institute of Biochemistry and Biology, Potsdam University, Germany

3th Examiner: Prof. Dr. Dierk Scheel
Leibnitz Institute of Plant Biochemistry, Halle, Germany

4rd Examiner: Prof. Dr. Uwe Sonnewald
Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

This Ph.D. thesis is the account of work done between October 2001 and September 2004 in the department of Prof. Willmitzer at the Max Planck Institute of Molecular Plant Physiology, Golm, Germany. It is result of my own work and has not been submitted for any degree or Ph.D. at any other university.

Eidesstattliche Erklärung

Diese Dissertation ist das Ergebnis praktischer Arbeit, welche von Oktober 2001 bis September 2004 durchgeführt wurde in der Abteilung von Prof. Willmitzer im Max-Planck-Institut für Molekulare Pflanzenphysiologie, Golm. Ich versichere, die vorliegende Arbeit selbständig und ohne unerlaubte Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben. Ich versichere ebenfalls, dass die Arbeit an keiner anderen Hochschule als der Universität Potsdam eingereicht wurde.

Potsdam-Golm, im Oktober 2004

Dirk Steinhauser

Abstract

The past decades are characterized by various efforts to provide complete sequence information of genomes regarding various organisms. The availability of full genome data triggered the development of multiplex high-throughput assays allowing simultaneous measurement of transcripts, proteins and metabolites. With genome information and profiling technologies now in hand a highly parallel experimental biology is offering opportunities to explore and discover novel principles governing biological systems. Understanding biological complexity through modelling cellular systems represents the driving force which today allows shifting from a component-centric focus to integrative and systems level investigations. The emerging field of systems biology integrates discovery and hypothesis-driven science to provide comprehensive knowledge via computational models of biological systems.

Within the context of evolving systems biology, investigations were made in large-scale computational analyses on transcript co-response data through selected prokaryotic and plant model organisms. CSB.DB - a comprehensive systems-biology database - (http://csbdb.mpimp-golm.mpg.de/) was initiated to provide public and open access to the results of biostatistical analyses in conjunction with additional biological knowledge. The database tool CSB.DB enables potential users to infer hypothesis about functional interrelation of genes of interest and may serve as future basis for more sophisticated means of elucidating gene function. The co-response concept and the CSB.DB database tool were successfully applied to predict operons in *Escherichia coli* by using the chromosomal distance and transcriptional co-responses. Moreover, examples were shown which indicate that transcriptional co-response analysis allows identification of differential promoter activities under different experimental conditions. The co-response concept was successfully transferred to complex organisms with the focus on the eukaryotic plant model organism *Arabidopsis thaliana*. The investigations made enabled the discovery of novel genes regarding particular physiological processes and beyond, allowed annotation of gene functions which cannot be accessed by sequence homology. GMD - the Golm Metabolome Database - was initiated and implemented in CSB.DB to integrated metabolite information and metabolite profiles. This novel module will allow addressing complex biological questions towards transcriptional interrelation and extent the recent systems level quest towards phenotyping.

Table of Contents

Chapter I - General Introduction:
- Genomics & Post-Genomics -

Abstract

The past decades have seen a growing number of organisms with available complete genome sequences. The accessibility of those resources triggered the development and recent maturation of high-throughput assays. Multi-parallel analyses of transcripts, proteins and metabolites are central for functional genomics. This highly parallel experimental biology is offering opportunities to explore and discover underlying governing principles of biological systems. The following sections give a brief overview of past and recent developments of high-throughput approaches and bioinformatics which take advantage from the availability of entire genome sequences and multi-parallel techniques.

Introduction

Recent biological research is characterized by a noteworthy alteration which is mainly driven by the massive increase of sequence information and the development of high-throughput assays. Consequently, new types of experiments are made possible and allow scientists discoveries and explorations of biological processes and functions on an unprecedented scale. In the past decades various multinational coordinated efforts have focused on genome sequencing and initial gene analyses. Recently, theses large investments led to a public release of more than 30 entire or partial genome sequences (see http://www.ncbi.nlm.nih.gov/Genomes/index.html). Those breakthroughs have been made for *Escherichia coli* (Blattner et al., 1997), *Saccharomyces cerevisiae* (Goffeau et al., 1996), *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), *Oryza sativa* (Yo et al.,2002), *Drosophila melangolaster* (Adams et al, 2000), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998), *Homo sapiens* (The International Human Genome Sequencing Consortium, 2001), and many other species. Further progress of genome sequencing, in both public and private efforts, will be made for at least another hundreds of organisms in the near future (see http://www.ncbi.nlm.nih.gov/Genomes/index.html), e.g. *Lycopersicon esculentum* [Solanaceae Genome Network, (http://www.sgn.cornell.edu/index.html)]. The limitations of large-scale EST (expressed sequence tags, partial cDNA [complementary DNA] sequences) sequencing projects and the desire to gain additional information about the genome structure, such as regulatory elements, forced the initiation of genome sequencing programs (Sommerville and Sommerville, 1999). Despite the availability of having an entire list of genes for an organism, the highly complex biological processes and responses of cells and organisms still remain undiscovered. Arising from this limitation functional genomic efforts are initiated to take advantage of the entire sequence information. Instead

of providing lists of genes and gene function the main objective is to understand how components work together and comprising complex cellular functions and organisms. To gain insight into complex biological processes and to take complete advantage of sequence information novel tools and technologies are required. Although the sequencing projects are mainly focusing on the discovery of genes they have triggered the development of a variety of high-throughput assays. The recent ongoing developments and maturation of the analytical technologies in conjunction with entire genome sequences will allow us to study genetic systems in its entirety. Now, multi-parallel approaches are used to study genes and gene functions at transcript (Lockhart and Winzeler, 2000; Breyne and Zabeau, 2001), protein (Pandey and Mann, 2000; Tyers and Mann, 2003) or metabolite level (Fiehn et al., 2000; Kopka et al., 2004). Moreover, these technologies open up the possibility to monitor a large amount or all elements of the cellular inventory towards these levels. Those demands have driven the need for bioinformatics tools and computational analyses to extract biological meaningful information from the vast amount of complex data obtained.

Here we give a very brief overview of the past and recent developments, implementation and application of these techniques with the main emphasis on plant field.

Molecular Technologies

Since the discovery of the DNA structure (Watson and Crick, 1953) rapid progress has been made in the development of molecular and genomics tools used to uncover biological processes on cellular levels. The description of restriction enzymes (e.g. Arber and Linn, 1969; Nathans and Smith, 1975) in conjunction with the development of cloning (e.g. Bolivar et al., 1977) and transformation technologies (e.g. Hanahan, 1983; Mattanovich et al., 1989) in the seventies and eighties allows for experimental characterization of single genes or small sets of genes on molecular as well as biochemical level. Further breakthrough technologies such as reverse transcriptase (RT) polymerase chain reaction (PCR) (e.g. Mullis et al., 1986; Mullis und Faloona, 1987) open up the possibility to detect and quantify weakly expressed genes. In conjunction with biophysical methods transcript analyses and cDNA library generation are made possible at few as well as single cell level (Dresselhaus et al., 1994; Karrer et al., 1995; Richert et al., 1996). Enormous progress has been made in the development of tools to create and characterize genetic diversity. Transgenic knock-out populations or transposon insertions lines have paved the way to broader bases of diversity (Aarts et al., 1993; Strepp et al., 1998; Cho et al., 1999; Zhu et al., 1999). Recently, putative loss of function lines for favourite genes can be easily ordered from resource centres, which harbours background information and provide access regarding transgenic organisms (e.g. Alonso et al., 2003). The development of DNA sequencing technologies enabled analysis of the nucleic acid composition of genes. Sequence storage and public access [e.g. NCBI, (Wheeler et al., 2004); TIGR, (Quackenbush et al., 2000)] in conjunction with computational tools for sequence comparison [e.g. BLAST, (Altschul et

al., 1990)] enabled scientist to assign a putative function for novel discovered genes as well as to infer hypotheses. With the arising of automated sequencers, sequencing was scaled-up to entire genomes of various organisms (see http://www.ncbi.nlm.nih.gov/Genomes/index.html).

Functional Annotation

Assigning a basic function to the multiplicity of novel genes discovered by gene or entire genome sequencing project are, and will continue to be, an important goal in biology both now and in the future. Functions of genes can be researched from a multitude of different scientific perspectives and therefore can be implemented by a variety of technologies developed (Vukmirovic and Tilghman, 2000). Since the development of protein (Edman, 1950) and efficient DNA sequencing methods (Maxam and Gilbert, 1977; Sanger et al., 1977) protein as well as nucleic acid sequences were collected (Dayhoff, 1972; Erdman, 1978). With the gaining of sequence information various computational programs and algorithms have been developed which enabled improved experimental-driven research, for instance algorithm to translate DNA into protein sequences, to detect restriction enzyme recognition sites and promoter sites (see Roberts, 2000; Stormo, 2000). Pairwise sequence comparisons and multiple alignments by various algorithms (see Hodgman, 2002) in conjunction with amino acid relatedness matrices (see Hodgman, 2002) allowed functional prediction for genes or gene products by annotation transfer from homologous sequences (McGeoch and Davidson, 1986; Bork and Gibson, 1996). Sequence searches by conserved sequence signatures e.g. PROSITE (Bairoch, 1991) and PFAM (Sonnhammer et al., 1997) libraries enabled the identification of common motifs in novel discovered genes or gene products. Further algorithm developments open up the possibility of protein structure and functional segment predictions, e.g. based on biophysical characteristics such as charge or hydrophobicity (Hodgman, 2002). With the arising of entire genome sequences computational biology and bioinformatics were challenged with the handling of large information. Despite this it opened the way to extend and improve analyses regarding to repetetive elements, regulatory regions as well as gene prediction from genomic DNA. Furthermore, investigations have been directed to pathway reconstruction or protein interactions which are reviewed i.e. by Bork et al. (1998).

Despite the continually improved basic functional assignments and prediction of genes and gene functions the term ‘function’ is loosely defined. In the past years, it arises that ‘function’ makes only sense in a context. The emerging of improved functional assignments at higher order processes required additional information from experimental research. The lack of comprehensive technology platforms in the past did not enable such bioinformatics approaches.

Transcript Analysis

The public availability of complete genome sequence information inspired and facilitates the development of novel technologies to take the full advantage of the gained sequence information. Recently, the most prominent amongst these new technologies is transcript profiling, which allow in-parallel measurement of transcript levels of large portion or entire genomes. At this data, there are numerous analytical approaches available used for global transcript profiling (Meyers et al., 2004). These approaches can be primarily grouped in (Breyne and Zabeau, 2001; Lockhart and Winzeler, 2000):

(i) hybridization-based approaches,

(ii) sequence-based approaches,

(iii) fragment-based approaches (PCR-based approches).

Hybridization-based approaches have been used in biological science for many years and comprise high-density arrays of oligonucleotides or complementary DNAs (Schena et al., 1995; Lockhart et al., 1996). The basic principle they are based on is in-parallel hybridization of labelled (c)RNA or (c)DNA in solution to specific localized, surface-bound nucleic acid molecules (normally DNA). In general, it based on the same principle of Watson-Crick base pairing as other traditional techniques such as Northern and Southern blotting (Southern, 1975; Alwine et al., 1977). First arrays consisted of spotted DNA fragments (e.g. from genomic DNA, cDNA, plasmid libraries) on porous membrane, normally nylon, and were hybridized with radioactive labelled material (e.g. Thimm et al., 2001). Recent microarrays use glass as surface and the hybridized material is labelled with at least one fluorescent dye (Lockhart and Winzeler, 2000). The maturity of technologies for synthesizing and deposition of nucleic acids allowed miniaturization by increasing information content. Currently used array technologies enabled the deposition of more than 250,000 oligonucleotide probes or 10,000 cDNAs per square centimetre (Lockhart and Winzeler, 2000). Arrays with more than 800,000 oligonucleotides have been successfully used for whole-genome expression analysis in plants (Yamada et al., 2003). Beyond it, various modifications of the hybridization procedure and surface material used are described in the literature (Lockhart and Winzeler, 2000).

Recently, microarrays are probably the most popular and widely used technique for genome-wide high-throughput transcript profiling. In contrast, with the arising of these techniques the results of those are critically discussed (Breyne and Zabeau, 2001, Ding and Canter, 2004). The criticism is mainly targeted to possible cross hybridizations of genes with similar sequences, which is not completely resolvable but can be minimized by careful probe design. Despite these critics microarrays are successfully used for various biological applications including e.g. cancer classification (e.g. Golub et al., 1999), signal pathways discovery (Lee et al., 1999), gene function prediction (Wu et al., 2002). Initial experiments in plants were performed for expression comparison of e.g. light- and dark-grown seedlings (Desprez et al., 1998) and different tissues (Ruan et al., 1998) and different biotic and

abiotic treatments (e.g. Reymond et al., 2000). Recently, the multinational coordinated AtGenExpress consortium has generated approximately 1200 full-genome transcript profiles covering a broad range of conditions (http://www.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm). Despite the main usage of microarrays for steady-state transcript profiling this technique allows hybridization of genomic DNA (Ding and Canter, 2004). Such approach were used to analyses the replication fork movement in *Escherichia coli* (Khodursky et al., 2000), applied in medical research and in genetic diagnostics (Ding and Canter, 2004).

Sequencing-based approaches, like serial analyses of gene expression [SAGE, (Velculescu et al., 1995)] and massively parallel signature sequencing [MPSS, (Brenner et al., 2000)], represent another main technology group to measure transcript levels. SAGE is based on counting sequence tags of 14-15 bases from cDNA libraries and gives an absolute measure of gene expression. This technology is based on double-stranded cDNA synthesis by using biotinylated oligo(dT) primers and various digestion and ligation steps to obtain concatenated sequences which will be sequenced. Similar to SAGE MPSS is a parallel sequencing methods which generates short sequence signature in the range of 16-20 bases. The sequence information is obtained by a complex ligation - cleavage work flow involving hybridization of specific phycoerythrin-decoder probes. For each step image(s) are taken for base identification (Brenner et al., 2000). SAGE has been widely used for expression analyses in clinical and mammal research, but was only sporadically applied to plants (Breyne and Zabeau, 2001). MPSS application was initially described for yeast (Brenner et al., 2000). Despite the advantage of absolute quantification and generation of SAGE libraries (as well as MPSS libraries) the high amount of input RNA required restricts application to large tissues. Moreover, the obtained short sequence tags may not be unambiguous and identification requirs large EST libraries. MPSS may be more accurate due to longer sequence tags.

Fragment (PCR) -based approaches based mainly on differential display techniques and covering approaches such as arbitrarily primed (AP) PCR (Welsh et al., 1992) or cDNA-amplified fragment length polymorphism (AFLP) (Bachem et al., 1996). Basically, these approaches have been successfully applied to unravel differential gene expression of various particular biological processes applied to a broad range of organisms (Ding and Canter, 2004). Despite the time-consuming procedure the possibility of genome-wide expression analysis without the demand of a-priory sequence knowledge enabled broad application.

Another, often called PCR-based approaches, cover real-time PCR and real-competitive PCR (rcPCR) (Ding and Canter, 2004). Whereas real-time PCR is a kinetics-based quantification technique rcPCR based on competitive PCR of a ‘true’ and a mimic fragment with single base exchange followed by matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS). Real-time PCR approaches have been successfully applied to analyze the expression of transcription factors or for weakly expressed genes (Czechowski et al., 2004) as well as were commonly used for

independent examination of changed expression levels or co-responses derived from microarray experiments (Lisso et al., 2004).

Protein Analysis

Beyond the recent advances in genome-wide gene expression technologies, ongoing developments and substantial progress has been made in protein analysis and developments are still ongoing (Tyers and Mann, 2003). These developments are mainly triggered by genome sequencing and functional annotation efforts. Proteomics was first coined to describe the set of proteins encoded by the genome (Wilkins et al., 1996). Today, proteomics is a versatile, rapidly developing and open-ended effort with expanded views to e.g. protein interactions, protein isoforms and modifications as well as higher-order complexes (Schwikowski et al., 2000; Eisenberg et al., 2000; Pandey and Mann, 2000). Arising from the broader field of activity novel technologies have been introduced for qualitative and quantitative proteomics (Tyers and Mann, 2003). Traditionally, descriptions of the protein complement tend to rely on gel electrophoreses prior to subsequent identification via mass spectrometry based protocols. For efficient separation of proteins, two-dimensional gel electrophoresis which utilises protein properties, e.g. isoelectric points and apparent molecular mass, were applied and successfully used for plant proteome research (Thiellement et al., 1999). High sensitive staining protocols enabled the detection of less abundant proteins and in combination with mass spectrometry (e.g. MALDI) allowed the identification of individual spots. The drawbacks of these long-standing biochemical techniques are limited quantitative estimations, relative labour intensive and time-consuming work, partial inaccuracy as well as the difficulty to automate (Mann, 1999). Despite the possibility of relative comparison of protein extracts from treated and untreated cells by different fluorescent markers (Unlu et al., 1997) new separation and identification protocols have been developed (Tyers and Mann, 2003). The novel techniques can be predominantly assigned to (i) mass spectrometry-based and (ii) array-based proteomics technologies. Mass spectrometry-based proteomics technologies rely on electrospray ionisation in conjunction with liquid-based (e.g. chromatographic, electrophoretic) separation tools (Opiteck et al., 1997; Link et al., 1999; Aebersold and Mann, 2003). Despite a significant increase in sensitivity, speed, and the enabling of large-scale identification of proteins from complex mixtures accurate quantification is still limited. Recently, approaches for stable-isotope protein labelling for quantitative proteomics have been introduced which allow accurate measurements of small changes in protein levels when comparing two different cell types (Aebersold and Mann, 2003). One of these methods used a class of reagents termed isotope-coded affinity tags (ICAT) fused to the proteins (peptides) of the protein samples (Gygi et al., 1999). Array-based proteomics technologies complement the mass spectrometry approaches and allows to test for protein interaction partners or activities (e.g. enzyme) (Pandey and Mann, 2000). Kersten et al. (2003) demonstrate the feasibility of the protein chip technology for plants by a small set of heterologously expressed proteins. The

application of mass spectrometry methods and proteome arrays for plant proteomic research will be a growing field in plant *post-genomics* and will increase the number of reports (Thiellement et al., 1999; Rose et al., 2004).

Metabolite Analysis

The availability of entire genome sequences of organisms opens up the possibility to introduce genetic diversity or altered gene expression within theses organisms. To take full advantage of the developed resources powerful phenotyping platforms are required including approaches for systematic analyses of metabolites. As metabolites can be regarded to mirror ultimate responses of biological systems to environmental or genetic perturbations metabolome analyses may allow the linkage of genotypes to their respective phenotypes (Fiehn, 2002). Whereas this field was mostly oversight across all biological disciplines and only few approaches were successfully applied in medical research (e.g. Duez et al., 1996) stronger attempts have been made in the past few years (Fiehn et al., 2000; Kopka et al., 2004; Fernie et al., 2004). These ongoing developments and the recent maturation of technology platforms have enabled the in-parallel analysis of hundreds or up to thousands of known or unknown metabolites (Kopka et al., 2004). The metabolome of an organism is the complete set of metabolites produced by the organism during life under all possible conditions. Whereas the genome is rather constant and nucleic acids and proteins can be described by a relative simple chemistry the metabolome is the result of a highly complex chemistry conducted by noteworthy highly interlinked biochemical reactions. The highly complex nature and the enormous chemical diversity of compounds do not technically enable us to measure all metabolites in an organism by a single analytical platform. Currently, no comprehensive platform can be envisioned which can measure all metabolite in a selective and sensitive manner (Weckwerth, 2003). Unravelling the entire metabolome of an organism may not be computable and probably requires combination of different analytical technologies. This will include separation systems like gas chromatography (GC) and liquid chromatography (LC), and numerous detection systems, including mass spectrometry (MS), nuclear magnetic resonance spectroscopy (NMR), UV, and visible light spectroscopy, and enzyme based assays (Tretheway et al., 1999; Fiehn et al., 2000; Weckwerth, 2003; Kopka et al., 2004). Recently, GC-MS (Roessner et al., 2000) is the most advanced and widespread technology platform for metabolite analyses and enabled relatively broad analysis of metabolites. Metabolite profiling with GC-MS involves six main steps, namely (i) extraction, (ii) derivatization, (iii) separation, (iv) ionization, (v) detection, and (vi) evaluation which are described in more detail by Kopka et al. (2004).

High-throughput metabolite profiling by GC-MS has been initially demonstrated in the context of plant functional genomics (Fiehn et al., 2000). Recently, GC-MS profiling of plant metabolites has mainly focused on hydrophilic compounds, which were recovered in the methanol-water phase of methanol-water/chloroform extractions. GC-MS based metabolite profiling of potato (Roessner et al.,

2000; Roessner et al., 2001a), tomato (Roessner-Tunali et al., 2003), *Arabidopsis* (Taylor et al., 2002), *Lotus* (Colebatch et al., 2004) and various others have provided insights into the effects of genetic manipulation on plants, highlighting metabolic diversity amongst natural populations, and metabolic differentiation during nodulation and symbiotic nitrogen fixation.

Bioinformatics: Data Management and Analysis

The vast amounts of data obtained by recently developed high-throughput assays in conjunction with the flood of information generated by genome sequencing projects inundate researchers with data. Despite the date flood obtained in past and recently information will be rapidly accumulating by further experiments or novel technologies developed. Whereas publications now often provide data overviews publicly accessible web resources and data archives stored and gives access to these data. Main attempt has been made in the establishment of databases harbouring data derived from transcript profile measurements [SMD, (Gollub et al., 2003); NCBI-GEO, (Edgar et al., 2002)]. Recently, comprehensive resources allowing access to other profiling date are still limited but will be yielding in the next years. Despite the successfully data integration and public access standardized formats are required for cross-experiment comparison and data exchange of the structurally complex profiling information. In the past years various initiatives suggested minimal information for profiling experiments, e.g. the MIAME or MIAMET standard (Brazma et al., 2001; Bino et al., 2004), and developed mark-up languages for data exchange, such as SBML (http://sbml.org) or XML (http://www.xml.org).

Beside data storage and exchange the biggest challenge is directed towards data analysis. Hence, bioinformatics and computational analyses will be playing a more and more significant role in modern biology. Recently, the design of research project requires querying of bioinformatics databases. On the other hand, proper handling and automated analyses of high-throughput data are necessary for efficient data analyses and hypotheses extraction. The main goal is to uncover biological knowledge underneath the experimental data. In the recent years rapid progress has been made in development and application of computational algorithms to extract biological meaningful information (i.e. Yu et al., 2003), which cannot be briefly reviewed here. In general, computational data analyses encompass data normalization procedures allowing comparisons to correct for technical and biological variations. The detailed procedures are depended on profiling technology and their resulting outputs, technical design and what scientists want to know. Normalization will be followed by specific data analysis methods which are manifold but can mainly be grouped in (i) projection methods, e.g. principal [PCA, (e.g. Yeung and Ruzzo, 2001; Roessner et al., 2001b)] and independent component analyses [ICA, (e.g. Scholz et al., 2004)], (ii) classification methods, such as support vector machines (e.g. Brown et al., 2000) or discriminate analyses (e.g. Simonis et al., 2004), and (iii) clustering procedures, for instance hierarchical [HCA, (e.g. Roessner et al., 2001b)] or K-means clustering (e.g. Tavazoie et al., 1999). In

recent publication cluster procedures have been successfully applied to group genes or experiments according functional context, in conjunction with identification of regulatory networks (e.g. Pilpel et al., 2001) or used by co-clustering (Hanisch et al., 2002). In metabolomics PCA became a commonly used tool to visualize metabolite profiling data sets and for extracting of relevant information (Ward et al., 2003; Urbanczyk-Wochniak et al., 2003). Scholz et al. (2004) successfully applied ICA to metabolite profiling data. Beside these few examples there are vast amounts of reports which have applied the aforementioned methods, used combinations of them or different approaches, such as Bayesian networks (Bockhorst et al., 2003).

Outlook

With the ever growing sequence and profile data it was more and more evident that understanding biological processes will not be possible with compendia of cellular elements (Oltvai and Barabási, 2002). Furthermore, the distinctness of the cellular levels, namely the genome, the transcriptome, the proteome and the metabolome, were critical discussed (Bhalla and Iyengar, 1999). An all-in-one access regarding data from all the aforementioned level is required to decipher the highly complex interactions of heterogeneous cellular components (Ideker et al., 2001). Recently, the cellular complexity is accounted as expression of a large number of functionally diverse, differently active and frequently multifunctional sets of elements which interact selectively and (non-)linearly to execute cellular function (Stephanopoulos et al., 2004; Alberghina et al., 2004). The cellular responses seem to be coherent rather than complex (Kitano, 2002). The improved multiplex high-throughput assays in conjunction with a changing perspective to approach and understand biological systems a ‘new’ field emerged - Systems Biology was reborn (see Chapter VII).

Objectives and Outline of this Thesis

With the evolving of systems level approaches in conjunction with rapidly increasing data obtained from high-throughput assays new challenges emerge for biological science (see above, and Chapter VII). The vast amount of data generated and the demands to create holistic views of the interplay between cellular elements and organization levels make cross-disciplinary research activities inevitable and highly attractive. Especially the maturity of transcript profiling technologies which allow simultaneous measurement of transcripts at full genome level enables scientists to investigate in comparative gene expression analyses. Unravelling gene-to-gene interactions will represent the basis to decipher functional modules and regulatory networks underlying biological processes. The main focus of this thesis is directed towards uncovering gene-to-gene interaction taking phylogenetic relationships into consideration. The five chapters II-VI describe the results obtained during this PhD work. Chapter II introduces the methodology as well as the implementation and development of an

open and worldwide accessible platform to the methods and results obtained (http://csbdb.mpimp-golm.mpg.de). Chapter III shows the proof the concept for investigations of gene-to-gene interactions by using the biological facile prokaryote *Escherichia coli*. With the experience made in a bacterium the concept of gene-to-gene interaction was applied to a more complex eukaryotic organism, namely *Arabidopsis thaliana*. Chapter IV outlines the attempts made to identify brassinosteroid-related genes by means of transcriptional co-responses. Chapter V describes the investigations to possible assign gene functions regarding the subtilase family which has hitherto no known function. Finally, chapter VI focuses on integration of metabolite profiling data and the provided prerequisite tools.

References

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185-2195.

Alberghina,L., Chiaradonna,F. and Vanoni,M. (2004) Systems Biology and the Molecular Circuits of Cancer. *ChemBioChem*, **5**, 1322-1333.

Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653-657.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990). Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403-410.

Alwine,J.C., Kemp,D.J. and Stark,G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA*, **74**, 5350-5354.

Arber,W and Linn,S (1969) DNA modification and restriction. *Annu. Rev. Biochem.*, **38**, 467-500.

Aarts,M.G.M., Dirkse,W.G., Stiekama,W.J. and Pereira,A. (1993) Transposon tagging of a male sterility gene in *Arabidopsis*. *Nature*, **36,** 715-717.

Bachem,C.W., van der Hoeven,R.S., de Bruijn,S.M., Vreugdenhil,D., Zabeau,M. and Visser,R.G. (1996) Visualization of differential gene expression using novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J.*, **9**, 745-753.

Bhalla,U.S. and Iyengar,R. (1999) Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381-387.

Bairoch,A. (1991) PROSITE: a dictionary of sites and pattern in proteins. *Nucleic Acids Res.*, 19, 2241-2245.

Bino,R.J, Hall,R.D, Fiehn,O., Kopka,J., Saito,K., Draper,J., Nikolau,B.J., Mendes,P., Roessner-Tunali,U., Beale,M.H. et al. (2004) Potential of metabolomics as a functional genomics tool. *Trend Plant Sci.*, (in press).

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1462.

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003b) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227-1235.

Bolivar,F., Rodriguez,R.L., Greene,P.J., Betlach,M.C., Heynecker,H.L., Boyer,H.W., Crosa,J.H. and Falkow, S. (1977) Construction and characterization of new cloning vesicles. *Gene*, **2**, 95-113.

Bork,P and Gibson,T.J. (1996) Applying motif and profile searches. *Methods Enzymol.*, **266**, 162-184.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting Function: From Genes to Genomes and Back. *J. Mol. Biol.*, **282**, 707-725.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. et al. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.*, **29**, 365-371.

Brenner,S., Johnson,M., Bridgahm,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630-634.

Breyne,P. and Zabeau,M. (2001) Genome-wide expression analysis of plant cell cycle modulated genes. *Curr. Op. Plant Biol.*, **4**, 136-142.

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA.*, *97*, 262-267.

Czechowski,T., Bari,R.P., Stitt,M., Scheible,W.R. and Udvardi,M.K. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.*, **38**, 366-379.

Cho,R.J., Mindrinos,M., Richards,D.R., Sapolsky,R.J., Anderson,M., Drenkard,E., Dewdney,J., Reuber,T.L., Stammers,M., Federspiel,N. et al. (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.*, **23**, 203-207.

Colebatch,G., Desbrosses,G., Ott,T., Krusell,L., Kloska,S., Kopka,J. and Udvardi,M.K. (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant J.*, **39**, 487-512.

Dayhoff,M.O. (1972) Atlas of Protein Sequences ans Structure. National Biochemical Research Foundation, Washington, DC.

Desprez,T., Amselem,J., Caboche,M. and Höfte,H. (1998) Differential expression in *Arabidopsis* monitored using cDNA arrays. *Plant J.*, **15**, 821-833.

Ding,C. and Cantor,C.R. (2004) Quantitative analysis of nucleic acids – the last few years of progress. *J Biochem Mol Biol.*, **37**, 1-10.

Dresselhaus,T., Lörz,H. and Kranz,E. (1994) Representative cDNA libraries from few plant cells. *Plant J.*, **5**, 605-610.

Duez,P., Kumps,A. and Mardens,Y. (1996) GC-MS profiling of urinary organic acids evaluated as a quantitative method. *Clinical Chem.*, **42**, 1609-1615.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207-210.

Edman,P. (1950) Method of amino acid sequences in peptides. *Acta Chem. Scand.*, **4**, 283–293.

Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823-826.

Erdmann,V.A. (1978) Collection of pusblished 5S and 5.8S ribosomal RNA sequences. *Nucleic Acids Res.*, **5**, r1-r13

Fernie,A.R., Trethewey,R.N., Krotzky,A.J., Willmitzer,L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, **5**, 763-769.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.*, **18**, 1157-1161.

Fiehn,O. (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155-171.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. et al. (1996) Life with 6000 Genes. *Science*, **274**, 546-567.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. et al. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94-96.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Gygi,S., Rist,B., Gerber,S.A., Turecek,F., Gelb,M.H. and Aebersold,R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.*, **17**, 994-999.

Hanahan,D. (1983) Studies on transformation of *Eschericia coli* with plasmids. *J. Mol. Biol.*, **166**, 557-580.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145-154.

Hodgman,T.C. (2000) A historical perspective on gene/protein functional assignment. *Bioinformatics*, **16**, 10-15.

Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decoding life: systems biology. *Ann. Rev. Genomics Hum. Genet.*, **2**, 343-372.

Karrer,E.E., Lincoln,J.E., Hogenhout,S., Bennett,A.B., Bostock,R.M., Martineau,B., Lucas,W.J., Gilchrist,D.G. and Alexander,D. (1995) In situ isolation of mRNA from individual plant cells: Creation of cell-specific cDNA libraries. *Proc. Natl. Acad. Sci. USA*, **92**, 3814-3818.

Kersten,B., Feilner,T., Kramer,A., Wehrmeyer,S., Possling,A., Witt,I., Zanor,M.I., Stracke,R., Lueking,A., Kreutzberger,J. et al. (2003) Generation of *Arabidopsis* protein chips for antibody and serum screening. *Plant Mol. Biol.*, **52**, 999-1010.

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

Kopka,J., Fernie,A., Weckwerth,W., Gibon,Y. and Stitt,M. (2004). Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

Khodursky,A.B., Peter,B.J., Schmid,M.B., DeRisi,J., Botstein,D., Brown,P.O. and Cozzarelli,N.R. (2000) Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proc. Natl. Acad. Sci. USA*, **97**, 9419-9424.

Link,A.J., Eng,J., Schieltz,D.M., Carmack,E., Mize,G.J., Morris,D.R., Garvik,B.M. and Yates,J.R. (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.*, **17**, 676-682

Lisso,J., Steinhauser,D., Altmann,T., Kopka,J. and Müssig,C. (2004) Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Plant Physiol.*, (submitted).

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T. Gallo,M.V, Chee,M.S, Mittmann,M., Wang,C., Kobayashi,M., Horton,H. et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.*, **2**, 108-116.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.

Mann,M. (1999) Quantitative proteomics? *Nature Biotech.*, **17**, 954-955.

McGeoch,D.J. and Davison,A.J. (1986) Alphaherpesviruses posses a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. *Nucleic Acids Res.*, **14**, 1765-1777.

Mattanovich,D., Ruker,F., Machado,A.C., Laimer,M., Regner,F., Steinkellner,H., Himmler,G. and Katinger,H. (1989) Efficient transformation of *Agrobacterium spp.* by electroporation. *Nucleic Acids Res.*, **17**, 6747.

Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, **74**, 560-564.

Meyers,B.C., Galbraith,D.W., Nelson,T. and Agrawal,V. (2004) Methods for transcriptional profiling in Plants. Be fruitful and replicate. *Plant Physiol.*, **135**, 637-652.

Mullis,K.B., Fallona,F.A., Scharf,S., Randall,S., Horn,G. und Erlich,H. (1986): Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biol.*, **51**, 263-273.

Mullis,K.B. und Faloona,F.A. (1987) Specific synthesis of DNA in vitro via a polymerase catalysed reaction. *Meth. Enzymol.*, **155**, 335-351

Nathans,D. and Smith,H.O. (1975) Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu. Rev. Biochem.*, **44**, 273-293.

Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837-846.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153-159.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763-764.

Opiteck,G.J., Lewis,K.C., Jorgenson,J.W. and Anderegg,R.J. (1997) Comprehensive on-line LC/LC/MS of proteins. *Anal. Chem.*, **69**, 1518-1524.

Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.

Reymond,P., Keizer,L.C., Bouwmeester,H.J, Sun,Z., Alvarez-Huerta,M., Verhoeven,H.A., Blass,J., van Houwelingen,A.M, de Vos, R.C., van der Voet, H. et al. (2000) Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell*, **12**, 707-719.

Richert,J., Kranz,E., Lörz,H. und Dresselhaus,T. (1996) A reverse transcriptase-polymerase chain reaction assay for gene expression studies at the single cell level. *Plant Science*, **114**, 93-99.

Roberts,R.J. (2000) The early days of bioinformatics publishing. *Bioinformatics*, **16**, 2-4.

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

Roessner,U., Willmitzer,L., and Fernie,A.R. (2001a). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.*, **127**, 749-764.

Roessner,U., Luedemann,A., Brust,D., Fiehn,O., Linke,T., Willmitzer,L. and Fernie,A. (2001b) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11-29.

Roessner-Tunali,U., Hegemann,B., Lytovchenko,A., Carrari,F., Bruedigam,C., Granot,D., and Fernie,A.R. (2003). Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol*., **133**, 84-99.

Rose,J.K., Bashir,S., Giovannoni,J.J., Jahn,M.M. and Saravanan,R.S. (2004) Tackling the plant proteome: practical approaches, hurdles and experimental tools. *Plant J.*, **39**, 715-733.

Ruan,Y., Gilmore,J. and Conner,T. (1998) Towards *Arabidopsis* genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. *Plant J.*, **15**, 821-833.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

Scholz,M., Gatzek,S., Sterling,A., Fiehn,O. and Selbig,J. (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics*, **20**, 2447-2454.

Schwikowski,B., Uetz,P and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnol.*, **18**, 1257-1261.

Simonis,N., Wodak,S.J., Cohen,G.N. and van Helden,J. (2004) Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, **20**, 2370-2379.

Somerville,C. and Sommerville,S. (1999) Plant functional genomics. *Science*, **285**, 380-383.

Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Protein Struct. Funct. Genet.*, **28**, 405-420.

Stephanopoulos,G., Alper,H. and Moxley,J. (2004) Exploiting biological complexity for strain improvements through systems biology. *Nature Biotechnol.*, **22**, 1261-1267.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16-23.

Southern,E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J.Mol.Biol.*, **98**, 503-517.

Strepp,R., Scholz,S., Kruse,S., Speth,V. and Reski,R. (1998) Plant nuclear gene knockout reveals a role in plastid division for the homolog of the bacterial cell division protein FtsZ, an ancesteral tubulin. *Proc. Natl. Acad. Sci. USA*, **95**, 4368-4376.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 218-285.

Taylor,J., King,R.D., Altmann,T. and Fiehn,O. (2002). Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, **18**, S241-248.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012-2018.

The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.

Thiellement,H., Bahrman,N. Damerval,C., Plomion,C., Rossignol,M. Santoni,V., de Vienne,D. and Zivy,M. (1999) Proteomics for genetic and physiological studies in plants. *Electrophoresis*, **20**, 2013-2026.

Thimm,O., Essigmann,B., Kloska,S., Altmann,T. and Buckhout,T.J. (2001) Response of Arabidopsis to iron deficiency stress as revealed by microarray analysis. *Plant Physiol.*, **127**, 1030-1043.

Trethewey,R.N., Krotzky,A.J., and Willmitzer,L. (1999) Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.*, **2**, 83-85.

Tyers,M. and Mann,M. (2003) From genomics to proteomics. *Nature*, **422**, 193-197.

Unlu,M., Morgan,M.E. and Minden,J.S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

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

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484-487.

Vukmirovic, O.G. and Tilghman, S.M. (2000). Exploring genome space. *Nature*, **405**, 820-822.

Ward,J.L., Harris,C., Lewis,J. and Beale,M.H. (2003) Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*, **62**, 949–957.

Watson,J.D. and Crick,F.H.D (1953) Molecular structure of Nucleic Acids. *Nature*, **171**, 737-738.

Weckwerth,W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Biol.*, **54**, 669-689.

Welsh,J., Chada,K., Dalal,S.S., Cheng,R. and McClelland,M. (1992) Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res.*, **20**, 4965-4979.

Wilkins,M.R., Pasquali,C., Appel,R.D., Ou,K., Golaz,O., Sanchez,J.C., Yan,J.X., Gooley,A.A., Hughes,G., Humphery-Smith,I. et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology*, **14**, 61-65

Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35-40.

Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255-265.

Yamada,K., Lim,J., Dale,J.M, Chen,H., Shinn,P., Palm,C.J., Southwick,A.M., Wu,H.C., Kim,C., Nguyen,M. et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **303**, 842-846.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763-774.

Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. et al. (2002) A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science*, **296**, 79-92.

Yu,U., Kee,S.H., Kim,Y.J. and Kim,S. (2004) Bioinformatics in the post-genome era. *Bioinformatics*, **37**, 75-82.

Zhu,T., Peterson,D.J. Tagliani,L., St. Clair,G., Baszczynski,C.L. and Bowen,B. (1999) Targetted manipulation of maize genes in vivo using chimeric RNA/DNA oligonucleotides. *Proc. Natl. Acad. Sci. USA*, **96**, 8768-8773.

Chapter II - Implementation & Development:

- CSB.DB: a comprehensive systems-biology database -

Dirk Steinhauser[*†], Björn Usadel[†], Alexander Luedemann, Oliver Thimm and Joachim Kopka

Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany

Received on June 5, 2004; accepted on June 30, 2004

Advanced Access publication July 9, 2004

Abstract

Summary: The open access comprehensive systems-biology database (CSB.DB) presents the results of bio-statistical analyses on gene expression data in association with additional biochemical and physiological knowledge. The main aim of this database platform is to provide tools that support insight into life's complexity pyramid with a special focus on the integration of data from transcript and metabolite profiling experiments. The central part of CSB.DB, which we describe in this application note, is a set of co-response databases, which currently focus on the three key model organisms, *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. CSB.DB gives easy access to the results of large-scale co-response analyses, which are currently based exclusively on the publicly available compendia of transcript profiles. By scanning for the best co-responses among changing transcript levels, CSB.DB allows to infer hypotheses on the functional interaction of genes. These hypotheses are novel and not accessible through analysis of sequence homology. The database enables the search for pairs of genes and larger units of genes, which are under common transcriptional control. In addition statistical tools are offered to the user, which allow validation and comparison of those co-responses that were discovered by gene queries performed on the currently available set of pre-selectable datasets.

Availability: All co-response databases can be accessed through the CSB.DB web server (http://csbdb.mpimp-golm.mpg.de/).

Contact: Steinhauser@mpimp-golm.mpg.de

[*] To whom correspondence should be addressed.

[†] The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Introduction

The availability of full genomic information (Goffeau et al., 1996; Blattner et al., 1997; Arabidopsis Genome Initiative, 2000; Lander et al., 2001) facilitated the development and spurred application of multi-parallel techniques to monitor the cellular inventory. Functional assignment of novel and partially characterized genes will continue to be, the most important goal in biological science (Wu et al., 2002; Shen-Orr et al., 2002). Modern functional genomics encompasses technologies designed for the systematic investigation of gene function at all levels of a living cell, namely the genome, the transcriptome, the proteome and the metabolome (Fiehn et al., 2000; Lockhard and Winzeler, 2000; Corbin et al., 2003). The combined and multi-parallel analyses allow the investigation of complex biological processes at full systems level (Kitano, 2002) and may become the empirical basis of understanding the paradigm of life's complexity pyramid (Oltvai and Barabási, 2002). A future task will be the discovery of functional interaction within and among the levels of the cellular inventory, e.g. among metabolome and transcriptome (Urbanczyk-Wochniak et al., 2003), and to extend knowledge from an organism-specific level towards general, organism-independent principles (Oltvai and Barabási, 2002). Hypotheses on units of genes with common function need to be associated with the currently available public knowledge of the complete cellular inventory. This information is made available in highly frequented but separate biological databases, which harbour genomic data (Mewes et al., 2004), gene expression data (Sherlock et al., 2001), information on protein properties (Schomburg et al., 2004), metabolites, and metabolic pathways (Kanehisa et al., 2004). To gain insight into the functional organization of biological networks specialized databases are required that are designed to store, handle, analyse and display the data derived from multi-parallel measurements. The comprehensive systems-biology database (CSB.DB) was developed to integrate biostatistical analyses on multi-parallel measurements of different organisms, such as *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The present goal of CSB.DB is to present a publicly accessible resource for large-scale computational analyses on transcript co-response data, which mirror the large functional network of the cellular inventory and may serve as the basis for more sophisticated means of elucidating gene function.

Project Overview

The main focus of the CSB project is the generation of easily accessible knowledge about apparent gene-to-gene interactions in sets and subsets of publicly available transcript profiling data. We implicitly make the assumption that common transcriptional control of genes is reflected in co-responding, synchronous changes in transcript levels (Steinhauser et al., 2004). For future implementations we will extent this concept to the interaction of genes with other elements of the cellular inventory, such as metabolites (Urbanczyk-Wochniak et al., 2003). Currently, no public

convention exists as to which numerical approach is best applied to detect and validate the co-response of changing transcript levels. For this reason we integrated a range of different statistical and computational algorithms, which are routinely applied in various research areas, such as Pearson's correlation, Kendall's correlation, Spearman's correlation, Euclidian distance and mutual information. Furthermore we selected a range of different datasets, which comprise three organisms and which were generated by different microarray technology platforms. Thus, the user of our co-response calculations is free to test results on different datasets and species.

The basic aim of the CSB project is to supply researchers in the field of systems biology, molecular and applied biology with statistical tools to access transcriptional co-response. We concentrate on the validation of gene co-response without requirement for the user to have a-priori knowledge about statistic methods and computational algorithms. We decided to preferentially facilitate access for those biologists who are interested in a specific gene of interest or small sets of genes. In this sense our approach is similar to simple BLAST searches (Altschul et al., 1990) of single or small number of genes. However, our approach towards the generation of novel functional hypotheses is based exclusively on simultaneous changes in transcript levels and does not require structural or sequence information.

Implementation and Structure

CSB.DB is accessible via the internet without the need to download special software to the client computer. The only system requirements are a JavaScript enabled recent web-browser and the ability to display PDF files. Only some advanced features require the JAVA extension. CSB.DB operates on a multiprocessor SuSE Linux (http://www.suse.de/de/index.html) system under an Apache web server (http://www.apache.org), and uses SAPDB (http://www.sapdb.org) as the database management system that stores the results of co-response analyses. CGI scripts, which connect the user queries with the database, are implemented in the PERL language (http://www.perl.com). The dynamic validation of discovered co-responses, graphical visualization as well as statistic algorithms, such as bootstrap and jack-knife analyses, are implemented as R (http://www.r-project.org) scripts, and can be invoked upon user selection to generate a PDF output. These files can be optionally downloaded by the user for further reference and documentation.

CSB.DB currently contains only co-response analyses, which are derived from publicly available expression profiling data. The calculated co-response data comprise pair-wise gene correlations of three model organisms, namely *E.coli* (Steinhauser et al., 2004), *S.cerevisiae*, and *A.thaliana* (Fig. 1A).

Databases and Queries

Co-response calculations based on changes in mRNA levels are the basis of functional annotation in CSB.DB and extend conventional predictions of gene function by analysis of gene homology (Wu et al., 2002). Publicly available expression profiles of various organisms represent a rich resource for cross-experiment co-response analysis of genes, but need to be critically appraised. We used transcript profiles that were quality checked according to the recommendations of the respective technology platform. Furthermore, we included only accurately measured gene spots for the assembly into multi-conditional expression data matrices. For example, our data matrices comprise approximately 20 - 50 independent transcript profiling experiments and contain only 5% missing values per gene. Besides quality checking and reduction of missing data we chose two general strategies for combining transcript datasets prior to correlation analysis. (1) We selected representative transcript profiles of as many different experimental conditions as possible. This approach allowed the search for general, constitutive gene-to-gene correlations in each organism. (2) If available we selected subsets of only those profiles, which were generated in a single set of biological experiments or under common biological conditions. These datasets allowed investigations of conditional changes in gene-to-gene co-responses as compared to constitutive co-responses. Correlations were computed with the cCoRv1.0 software (Steinhauser et al. unpublished data) and stored in organism specific co-response databases (Fig. 1A).

Rank ordered tables of pairwise gene correlations according to the selected correlation measure can be obtained using the single gene query option and using a selection of pre-defined ranking strategies (sGQ; Fig. 1B). Similar to typical BLAST queries, sGQ allows to define a gene of interest and to retrieve all genes associated by co-response, if the gene of interest is represented among the set of quality checked genes. Moreover, the variant of sGQ made available for the *Arabidopsis* co-response databases allows to select filtering according functional categories, which were reported previously together with the visualization tool MapMan (Thimm et al., 2004). The sGQ output (Fig. 1C) is presented as a HTML table, which contains the rank, the gene identifier of the co-responding gene, the correlation measure, the gene description, the number of pairs (n), the covariance (cov), the probability (*P*-value), the confidence interval (CI), the power, the mutual information [d(M), converted into distance range], and the normalized Euclidean distance [d(E)]. These statistical parameters are dynamically calculated based on the underlying test distribution of the respective pre-selected correlation coefficient (Sokal and Rohlf, 1995; Bonett and Wright, 2000). Graphical summaries of the set of co-responding genes are based on various external functional classification efforts (Thimm et al. 2004; Peterson et al., 2001; Christie et al., 2004) and/or the text search of the returned gene annotations (Fig. 1C). This survey of gene categories present in the hit list is presented below the sGQ table.

Upon user request a detailed statistical analysis may be obtained for a selected gene pair of interest. This additional validation on demand supports the detection of experimental outliers, which may be associated with technical errors or with the specific nature of a biological experiment. For this purpose a variety of graphical plots are offered (Fig. 1D).

The multiple gene query option (mGQ) allows pre-definition of up to 15 genes of interest and returns the complete set of available correlations among these genes. This option may be used to discover interdependencies of genes, which are known to contribute to a common function or pathway. To visualize data, the interrelationship is also displayed as a co-response network with extensive filtering and layout options in JAVA enabled browsers (Fig. 1C).

Finally an intersection gene query tool (isGQ) extracts those genes, which exhibit common correlations to at least two pre-defined genes of interest. The threshold settings, which are available for sGQ, may also be used for isGQ. The isGQ query may be used, if a few genes with a common function are already known. Using the intersection mode that allows finding of novel genes, which may be involved in this function, but can not be discovered based on sequence homology.

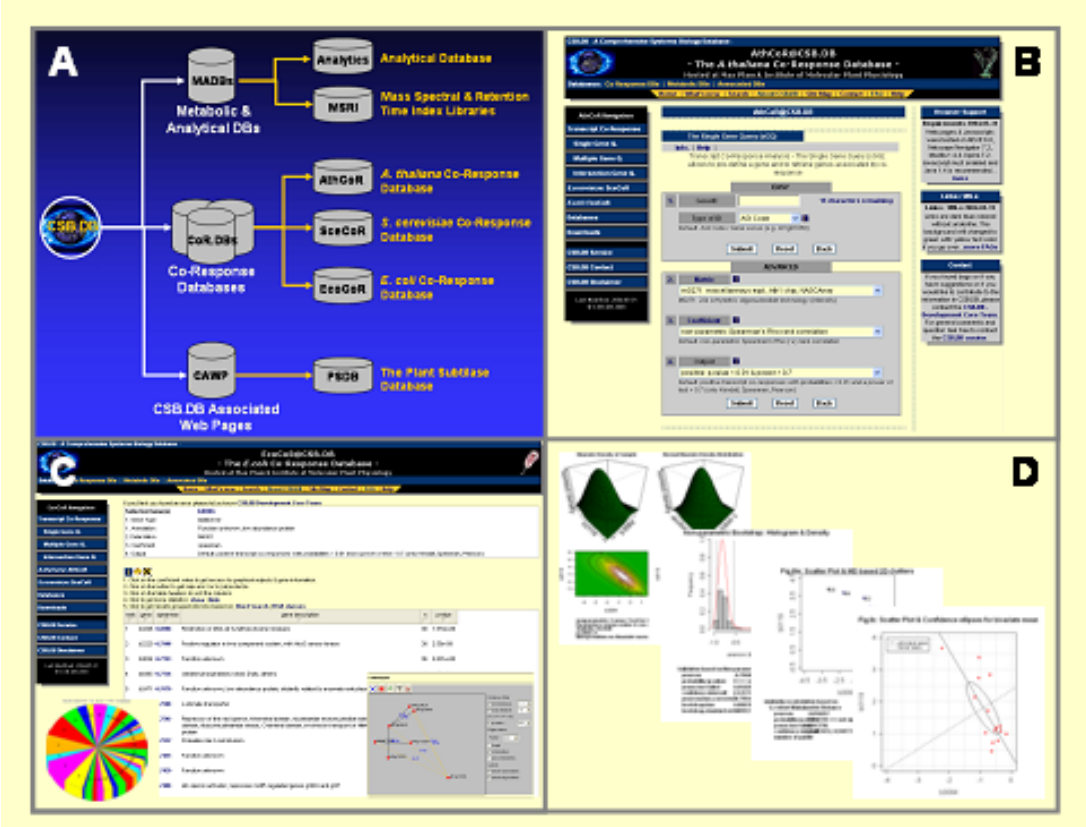

Fig. 1. Summarized overview of the current structure of CSB.DB (A) and selected examples of the available functionalities of the organism-specific co-response databases (B-D). (B) Represents one of the three possible query types, e.g. the single gene query sGQ in its HTML layout. The output of the queries is a HTML table, which contains in the case of *A.thaliana* a pie-chart summary on functional categories of the retrieved best ranking genes (C). (D) Shows examples of available gene-to-gene (bi-) plots, which can be invoked upon user request.

Outlook

We named CSB.DB 'A Comprehensive Systems-Biology Database', because we are convinced that the interpretation of gene co-response, which we currently make available to potential users, will in future require the integration of additional public resources on the present knowledge of the cellular inventory. Upon starting to use external functional classifications of genes, which among others include pathway and enzyme information, we implemented first access in our database to functional gene annotations. Thus, we laid the ground to retrieve biochemical reactions from publicly accessible metabolite databases starting from result lists of highly correlated genes.

In addition, we previously described that the combined correlation analysis of changes in metabolite and mRNA levels may be highly informative and provide novel information (Urbanczyk-Wochniak et al., 2003). Therefore, we will proceed to integrate profiling experiments and datasets into our database, which comprise measurements of changes in metabolite and transcript levels. Starting to use the same principles, which we apply to discover co-response in transcript datasets, we hope to unravel novel interactions between transcripts and metabolites. Thus, we are convinced that CSB.DB will develop into a highly useful and informative public resource.

Acknowledgements

We thank the staff of the SMD database (Sherlock et al., 2001), the ASAP database (Glasner et al., 2003) and the NASC Affymetrix Facility (Craigon et al., 2004) for the establishment of public accessible resources of microarray data. We appreciate the work of all scientists, who submitted transcript profile data to these databases and thereby made comparative investigations possible. Furthermore, we thank the staff of the Free Software Foundation (FSF) for access to software under the terms of the GNU general public license. We are grateful to Prof. Lothar Willmitzer, Prof. Mark Stitt and the Max-Planck-Institute of Molecular Plant Physiology for support of the CSB project. Furthermore, the comments from Dr. Leonard Krall, Dr. Dirk Buessis and Stefan Kempa are gratefully acknowledged. The work of B.U. is partially financed by the GABI project 0312277D.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403-410.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1462.

Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. et al. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acid Res.*, **32**, D311-D314.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.,Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S., Soupene,E. and Hunt,D.F. (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. *Proc. Natl. Acad. Sci. USA*, **100**, 9232-9237.

Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC`s transcriptomics service. *Nucleic Acid Res.*, **32**, D575-D577.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.*, **18**, 1157-1161.

Glasner,J.D., Liss,P., Plunkett,G.,III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. and Perna,J.T. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acid Res.*, **31**, 147-151.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. et al. (1996) Life with 6000 Genes. *Science*, **274**, 546-567.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acid Res.*, **32**, D277-280.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662-1664.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzttugh,W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.

Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V., Warfsmann,J. and Ruepp,A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acid Res.*, **32**, D41-44.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763-764.

Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acid Res.*, **29**, 123-125.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acid Res.*, **32**, D431-433.

Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64-68.

Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H, Ball,C.A. et al. (2001) The Stanford Microarray Database. *Nucleic Acid Res.*, **29**, 152-155.

Sokal,R.R. and Rohlf,F.J. (1995) Biometry: The principles and practice of statistics in biological research. 3rd ed. W.H. Freeman and Company New York.

Steinhauser,D., Junker,B.H., Luedemann,A., Selbig,J. and Kopka,J. (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, **20**, 1928-1939.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914-939.

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

Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255-265.

Chapter III - Proof of Concept:
- Hypothesis-driven approach to predict transcriptional units from gene expression data -

Dirk Steinhauser*, Björn H. Junker, Alexander Luedemann, Joachim Selbig and Joachim Kopka

Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany

Received on August 5, 2003; revised on February 24, 2004; accepted on February 25, 2004

Advanced Access publication March 25, 2004

Abstract

Summary: A major issue in computational biology is the reconstruction of functional relationships among genes, for example the definition of regulatory or biochemical pathways. One step towards this aim is the elucidation of transcriptional units, which are characterised by co-responding changes in mRNA expression levels. These units of genes will allow the generation of hypotheses about respective functional interrelationships. Thus, the focus of analysis currently moves from well-established functional assignment through comparison of protein and DNA sequences towards analysis of transcriptional co-response. Tools that allow deducing common control of gene expression have the potential to complement and extend routine BLAST comparisons, because gene function may be inferred from common transcriptional control.

Results: We present a co-clustering strategy of genome sequence information and gene expression data, which was applied to identify transcriptional units within diverse compendia of expression profiles. The phenomenon of prokaryotic operons was selected as an ideal test case to generate well founded hypotheses about transcriptional units. The existence of overlapping and ambiguous operon definitions allowed investigating in constitutive and conditional expression of transcriptional units in independent gene expression experiments of *Escherichia coli*. Our approach allowed identification of operons with high accuracy. Furthermore, both constitutive mRNA co-response as well as conditional differences became apparent. Thus, we were able to generate insight into possible biological relevance of gene co-response. We conclude that the suggested strategy will be amenable in general to the identification of transcriptional units beyond the chosen example of *E.coli* operons.

Availability: The analyses of *E.coli* transcript data presented here are available upon request or at http://csbdb.mpimp-golm.mpg.de/.

Contact: Steinhauser@mpimp-golm.mpg.de

* To whom correspondence should be addressed.

Introduction

Public availability of complete genome sequence information (Perna et al., 2001; Blattner et al., 1997) inspired and facilitated the development and utilization of multi-parallel techniques for monitoring the complete cellular inventory. Recent results of these technologies are made available in biological databases that harbour genomic data, gene expression data, and information about proteins, metabolites, and metabolic pathways. This information will become an empirical basis of understanding the paradigm of life's complexity pyramid (Oltvai and Barabási, 2002). Functional assignment of novel genes, which were discovered by genome sequencing projects, will continue to be the most important goal of the genomic area (Vukmirovic and Tilghman, 2000). One of the central challenges in computational biology is the discovery of regulatory networks which control gene transcription in biological model systems.

Accumulation of publicly available microarray data led to the development of a range of computational approaches to retrieve biologically meaningful information from co-responding changes of mRNA expression. A variety of computational approaches were previously applied to predict operons from full genome and transcriptome information (Zeng et al., 2002; Moreno-Hagelsieb and Collado-Vides, 2002; Ermolaeva et al., 2001; Yada et al., 1999). Tjaden et al. (2002) utilized *Escherichia coli* microarrays to monitor expression of both coding and non-coding intergenic regions. Hidden Markov models were applied to estimate gene boundaries. However, the lack of intergenic probes in routine microarray experiments currently restricts general application of this approach. Yamanishi et al. (2003) applied a generalized kernel canonical correlation analysis to group genes, which share similarities with respect to position within the genome and gene expression. However, this method was restricted to subsets of *E.coli* genes which comprised known metabolic pathways. Bockhorst et al. (2003a,b) successfully predicted operons by applying models of transcriptional units to gene sequence and expression data (Bockhorst et al., 2003a) and reported an approach based on Bayesian networks (Bockhorst et al., 2003b). Sabatti et al. (2002) re-addressed operon prediction by Bayesian classification and described required features.

Till today, no attempt has been made to assign transcriptional units by hierarchical clustering and co-clustering. Here we present a strategy (Fig. 1), which was designed to monitor occurrence of constitutive and conditional usage of transcription units in independent gene expression profiling experiments. Co-clustering was demonstrated to be a versatile tool to investigate how prokaryotic genome organization is reflected within compendia of gene expression data. Moreover, we show effects of additional, currently unknown mechanisms on gene co-response, which will be targets of further experimental verification.

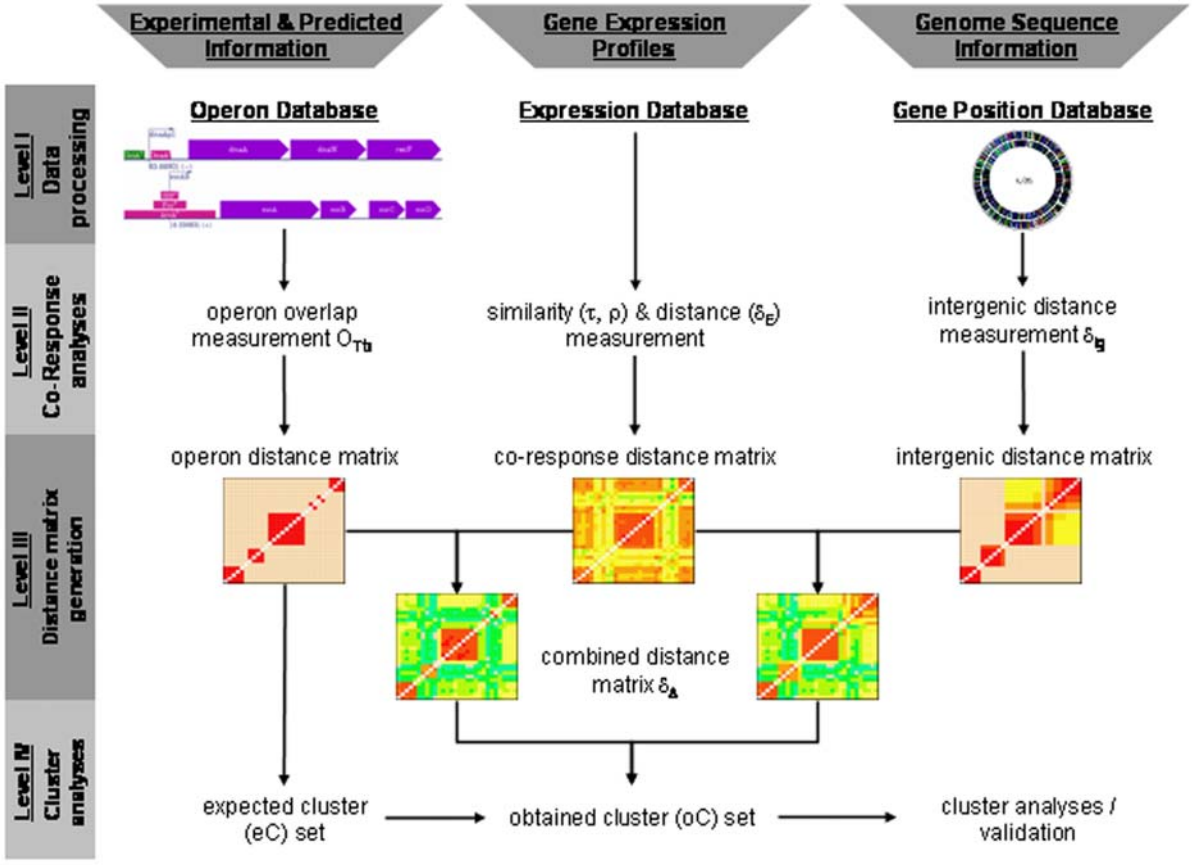

Fig. 1. Flow scheme of adopted data management and processing strategy. Initial data (level I) were converted into pairwise gene distance measures (level II), which were subsequently normalized, arranged into distance matrices and matrix combinations (level III). Finally matrices were subject to hierarchical cluster analyses and resulting cluster memberships of genes compared (level IV).

Methods

Data management and processing

Data management, subsequent data integration, and processing is summarized within a flow scheme (Fig.1). Primary data were: operon annotations, gene expression data, and gene positions within the *E.coli* genome (Fig.1, level I). These primary data were converted into pairwise gene distances: (1) common operon membership of gene pairs, (2) Pearson's linear correlation coefficient (ρ), Kendall's correlation coefficient (τ), or Euclidean distance (δ_E) of gene pairs as detected within transcript profiles, and (3) pairwise intergenic nucleotide distance of genes (Fig.1, level II). Pairwise gene distances were normalized and combined into distance matrices (Fig.1, level III). Finally, different combined and non-combined distance matrices were subject to hierarchical (co-)clustering and subsequent comparison of cluster memberships of genes. (Fig.1, level IV). The procedures were as follows:

Data source and pre-processing

Two data sources of *E.coli* gene expression profiles were used (Table 1). The first dataset, called M45 in the following, was derived from the Stanford Microarray Database using ratio 2 values only [SMD, (Sherlock et al., 2001)]. M45 contained 74 expression profiles encompassing 4264 genes, which were analyzed by colour-coded cDNA hybridization technology. M45 comprised experiments mainly related to aminoacid metabolism (Khodursky et al., 2000), DNA metabolism (Courcelle et al., 2001), and RNA decay (Bernstein et al., 2002). M45 was quality checked as recommended by the SMD tutorial (http://genome-www5.stanford.edu/help/index.html) and consequently had 43% missing data. In order to reduce the number of missing data two subsets were generated, each of which each had only 6% of missing data. M4501 was designed to maximize the number of experiments and comprised 43 experiments, which were described by 929 genes. M4502, designed to maximize the number of genes, comprised 34 experiments, which were described by 1845 genes. Data were normalized and log-transformed as suggested (Sherlock et al., 2001).

Table 1. Overview of the expression datasets and subsets.

Dataset		M45[a]	M4501[a]	M4502[a]	M96[b]	M96A[b]	M96B[b]
microarray platform[c]		cc	cc	cc	cc, on	on	cc
number of experiments		74	43	34	66	16	50
number of genes		4264	929	1845	4241	4345	4290
% of missing data		43	6	6	0	0	0
Experiment categories	subcategories						
Miscellaneous:	Miscellaneous	4	2				
Amino acid metabolism:	Tryptophan	29	13	9			
DNA metabolism:	UV radiation	15	10	8			
RNA decay:	RNA	10	10	9			
	Rifampicin	12	8	8			
Stress / Antibiotics:	Acid shock				8	1	7
	Cipro				8	1	7
	Cold shock				4		4
	Heat shock				1	1	
Growth curve:	Growth				10	4	6
Media comparison:	Media	4			8		8
Strain / mutant analyses:	Strains				27	9	18

[a]Stanford Microarray database (SMD).
[b]ASAP database.
[c]Technology platforms: cc (colour-coded EST hybridizations), on (oligonucleotide hybridizations).

The second dataset, M96, was from the ASAP collection (Glasner et al., 2003) and originated from colour-coded cDNA hybridization and oligonucleotide microarray technology (Affymetrix, Santa Clara, CA, USA). M96 consists of 66 experiments comprising 4241 genes, was log transformed, and lacked missing data. The transcript profiling experiments of M96 covered miscellaneous stress

treatments, application of antibiotics, comparison of media and growth conditions, as well as characterization of mutants (Allen et al., 2003). Two subsets of M96 were formed to separate profiles from different technology platforms, M96A and M96B (Table 1).

Topological overlap matrices of operon annotations

The operon annotations and predictions were retrieved from Regulon (Salgado et al., 2001; http://www.cifn.unam.mx/Computational_Genomics/regulondb/), EcoCyc (Karp et al., 2002; http://biocyc.org/ecoli/), and KEGG (http://www.genome.ad.jp./kegg/) databases as well as from Moreno-Hagelsieb and Collado-Vides (2002). All operon assignments were combined into two matrices. First, overlapping and conflicting operon annotations from different sources were maintained within a topological operon overlap matrix, O_{Ttu}, according to Ravasz et al. (2002). O_{Ttu} harbours the combined hypotheses of the maximum possible number and size of operons in *E.coli*. Second, an intersection operon matrix, IS_{tu}, was constructed to comprise the hypotheses of the non-ambiguous and commonly accepted minimum set of *E.coli* operons.

Genome information and weighted intergenic distance (δ_{ig}^{ω})

The EcoCyc database (Karp et al., 2002) was accessed to retrieve gene position and intergenic distances. Intergenic distance of any two genes was defined as follows: genes were separated into the two co-linear groups positioned on the two opposite circular genomic strands of *E.coli*. Within each group, the smaller of the two possible sums of all non-coding nucleotides (nt) in between any two genes was calculated. The distance of overlapping genes on the same strand was set at zero. A weighted intergenic distance (δ_{ig}^{ω}) was generated from nt distances by normalization to $0 < \delta_{ig}^{\omega} < 1$. Above an nt threshold δ_{ig}^{ω} was set to 1, below this threshold δ_{ig}^{ω} was calculated by dividing non-coding intergenic nt by the respective threshold value. Four threshold values were chosen, 2 × 2250 nt, 2 × 7250 nt, 2 × 70 000 nt, and 2 × 655 596 nt. The rationale for choosing theses thresholds reported below.

Co-response matrices

Pearson's product moment linear correlation (ρ), non-parametric Kendall's coefficient of rank correlation (τ) without correction for ties (Sokal and Rohlf, 1995), and Euclidean distance coefficients (δ_E) (Mirkin, 1996) were applied to log transformed gene expression data. Significance of correlation was tested as recommended by Sokal and Rohlf (1995). In order to generate normalized distance matrices correlation coefficients were converted according to Mirkin (1996) and Sokal and Rohlf (1995). Largest distance was assigned to negative Pearson's or Kendall's correlation coefficients. The converted distances were marked with the index of used correlation coefficient (e.g. δ_τ). Then all distance measures were normalized to the maximum distance value. Thus, all resulting normalized distances (δ^v) were in the range of $0 \leq \delta^v \leq 1$.

Joining function (λ_ψ) and combined distance (δ_Δ)

Normalized distance matrices were combined as suggested by Hanisch et al. (2002) applying a modified function, (λ_ψ), Equation (1), extended to *n* dimensions. The resulting combined distance function δ_Δ of each gene pair, g_i, g_j, with $\psi \in \{\delta_\rho^{\nu}, \delta_\tau^{\nu}, \delta_E^{\nu}, \delta_{OTtu}^{\nu}, \delta_{ig}^{\omega}, \ldots\}$ was defined as follows:

$$(1) \quad \delta_\Delta(g_1, g_2) = \frac{1}{n}\sum_{\psi=1}^{n}\left[\lambda_\psi(g_i, g_j)\right] \qquad where \quad \lambda_\psi(g_i, g_j) = \frac{1}{1 + e^{-\left[\frac{\delta_\psi(g_i, g_j) - \nu_\psi}{s_\psi}\right]}}$$

For co-response distances the control parameters of the shape of the logistic curve (ν_ψ, s_ψ) were adjusted to the median of distance distribution (ν_ψ) and to a moderate slope of $s_\psi = \nu_\psi/6$. The control parameters of the δ_{ig}^{ω}s were adjusted to $\nu_\psi = \nu_{\psi weighting} - \nu_{\psi correction} = 0.5 - 0.17578 = 0.32422$ and to $s_\psi = (\nu_{\psi weighting}/6) = 0.08$. The correction term $\nu_{\psi correction}$ was empirically determined by fitting $\nu_{\psi weighting}$ to the formula described above and the setting of parameters to $\lambda_{ig} = 0.9$ and $\delta_{ig}^{\omega} = 0.5$. For generation of O_{Ttu} the parameter were set to $\nu_\psi = 0.5$ and $s_\psi = \nu_\psi/6$. An a priori weighting of 50% (n=2) was assigned to combine, O_{Ttu} or δ_{ig}^{ω}, with distance matrices describing transcriptional co-response. Two transcriptional co-response matrices of each expression dataset were combined, i.e. 25% weight was given to either a normalized distance matrix based on Pearson's or Kendall's coefficients and residual 25% weight was assigned to the normalized Euclidian distance matrix (see above).

Hierarchical cluster analysis and cluster validation

For the classification of genes, the unweighted average linkage clustering algorithm (UPGMA) was applied to normalized distance matrices (Mirkin, 1996). Expected operon clusters (eC) were generated by use of δ_{OTtu} of the operon overlap matrix (see above). Cluster validation was performed by measuring the degree of correspondence between the expected cluster (eC) and the obtained cluster (oC). In detail, the cluster specific match coefficient (CMC) reflects the ratio of elements, i.e. genes, from eC which are observed to occur within oC. For example, if eC = oC, obtained clusters perfectly match expectations. The combined match coefficient (MC) was defined at selected clustering heights as the sum of all CMCs divided by the number of expected clusters. MC represents the portion of all genes which were found to belong to expected operons, for example, if MC = 1.0, all genes were found to group into respective expected clusters. The cluster specific reassignment coefficient (CRC) is the ratio of those genes that are not expected to occur in eC as compared to the genes that are correctly grouped into oC. CRC is indicative of the portion of novel genes which were unexpectedly assigned to any given cluster. The sum of all CRCs at a specific clustering height divided by the number of expected clusters yields the reassignment coefficient (RC). For example, if RC = 1.0, the number of mis-assigned genes is equal to the number of correctly assigned genes.

Statistical analysis and software

The Mantel test and respective analysis of variance, the non-parametric Kruskal-Wallis and two-sample Wilcoxon rank sum tests, tests of homogeneity as well as parametric three-way factorial ANOVA were computed as described by Sokal and Rohlf (1995). The Cramer-test was performed according to Baringhaus and Franz (2004). All statistical tests were applied to iterated random selections of data subsets.

Computations were performed using the statistical software environment R (http://www.r-project.org) version 1.6.1. and 1.6.2 with the libraries 'mva', 'exactRankTests', 'vegan', 'e1071', 'tseries', 'ctest' and 'cramer'. Calculations were executed with PERL scripts.

Results

The goal of this work was to investigate how the prokaryotic genome organisation, namely polycistronic operon structure, is reflected within different compendia of gene expression profiles from *E.coli* and whether functional linkage of genes can be detected by clustering technologies. We selected the prokaryotic operon structure because genes that are co-regulated in physical units of common polycistronic messenger RNA (mRNA) can be expected to high correlation in transcriptome analyses independent of the nature of underlying biological experiments. This strong co-regulation should allow precise classification by clustering technologies irrespective of the nature of distance measure applied. However, initial attempts to retrieve clusters of genes that constitute operons within combined sets or subsets of M45 and M96 failed.

Operon classification by clustering

The co-response matrices of four data subsets M4501, M4502, M96A and M96B were generated using, Kendall's τ, Pearson's ρ, and Euclidean distance δ_E. These matrices were subject to HCA and subsequently cluster membership of genes was compared to expected clusters as represented by O_{Ttu}. Criterion for clustering quality was MC (cluster match coefficient), criterion for gene mis-assignment was RC (reassignment coefficient; refer to the earlier section 'hierarchical cluster analysis and cluster validation'). Only results of M4501, which were representative of all other data subsets, are shown (Fig. 2). Irrespective of the applied distance measure, only a minority of genes was correctly assigned to expected operons, e.g. MC < 0.08, when accepting 50% mis-assignments (RC = 1.0) (Fig. 2).

The analysis of best pairwise gene associations according to Kendall's τ, Pearson's ρ (data not shown), or Euclidean distance δ_E demonstrated, that only 50% of those gene associations, which were expected to result of genes belonging to the same operons, ranked among the top 5 – 10% (Fig. 3). We tested whether this observation was caused by applying O_{Ttu}, which included ambiguous as well as predicted operon definitions. For this purpose we applied IS_{tu}, i.e. we restricted our analysis to the minimum intersection of available operon annotations (refer to 'topological overlap matrices of operon annotations'). MC however, only increased approximately 2-fold at RC = 1.0 and frequency distributions of relative rank positions were independent of choice of either O_{Ttu} or IS_{tu} as was supported by a non-parametric Cramer test. Therefore, we had to assume other factors than precision of current operon annotation, which either cause absence of expected pairwise gene associations or which are caused by other regulatory mechanisms of coordinated gene expression, such as transcription factors or mRNA processing. We ruled out an artefact caused by the choice of transcript profiling technology because datasets M96A, and M96B, did not show fundamental differences (Fig. 3). In the following we first characterize the nature of the datasets, unravel properties, which obscure operon units, and demonstrate that co-clustering and the use of data subsets allowed overcoming this inherent problem of transcriptome analyses.

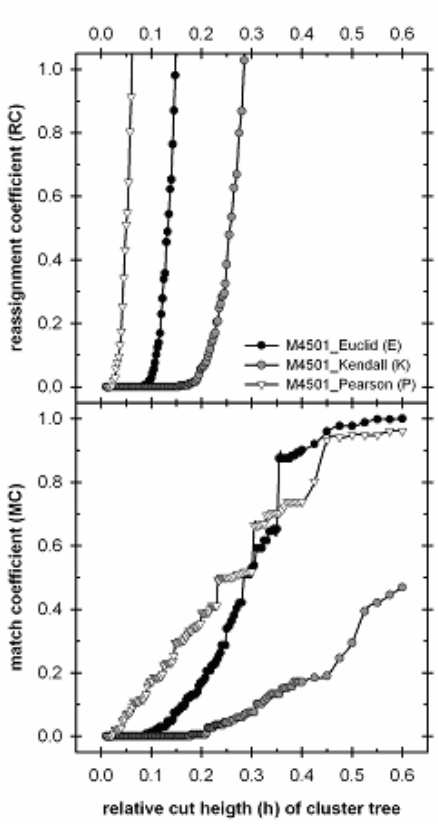

Fig. 2. Plot of match coefficient (MC) (bottom) and the reassignment coefficient (RC) (top) resulting from HCA of distance matrices from dataset M4501 at increasing relative heights (h). RC is shown in the range of 0 to 1.0 with 1.0 representing 50% novel gene assignments. Inset shows applied distance measures.

Properties of transcript datasets

We performed a comparison of the datasets M45 and M96 by applying hierarchical cluster analysis to a δ_E association matrix of the respective compendium experiments (Fig. 4). In dataset M4501 the majority of nodes are in the heterogeneity range of $0 \leq h_e \leq 0.2$ and experiment grouping strongly reflects the nature of underlying biological experiments (Fig. 4a). Similar results were obtained from either M45 or subset M4502 (data not shown). M96 exhibited higher inherent heterogeneity, $0.2 \leq h_e \leq 0.4$. Biological experiments were partially reflected by clustering, but experiments from different technology platforms were clearly separated (Fig. 4b). Non-parametric analysis of variance by Mantel

testing revealed a highly significant ($\tau = 0.5668$, $P << 0.001$) difference of variance between the experiments of different technology platforms within M96, as measured by median Kendall's τ association; in detail, among experiments of Affymetrix technology $\tau_{median} = 0.558$, among experiments of cDNA technology $\tau_{median} = 0.453$, and in between experiments of different technology platforms $\tau_{median} = 0.222$. Therefore, dividing M96 into M96A and M96B according to technology platform was justified.

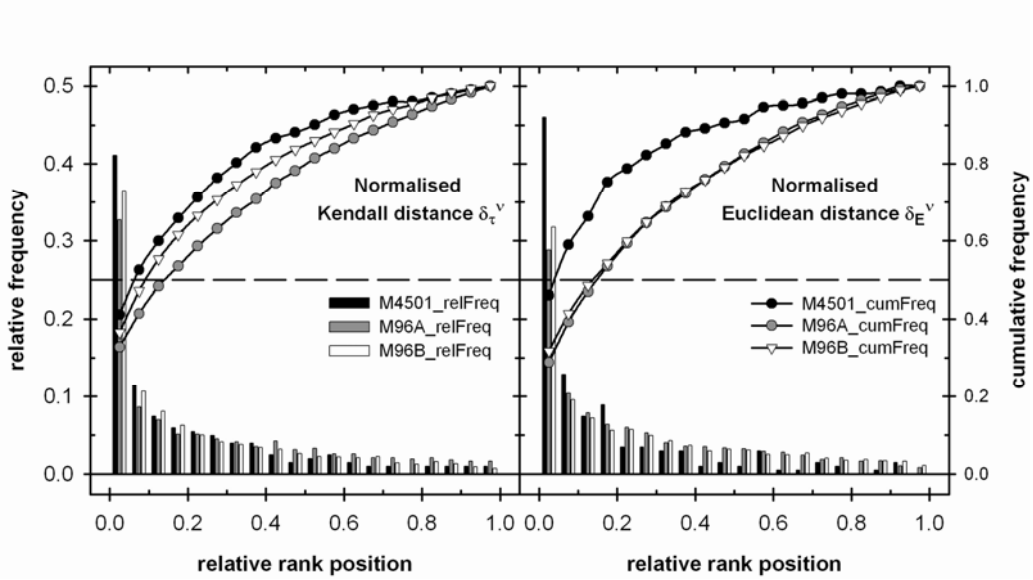

Fig. 3. Histogram plots and cumulative frequency of the relative rank distributions of all gene pairs, which belong to same operons. Best ranking was according to normalized Kendall distance δ_τ^v and normalized Euclidean distance δ_E^v. Dashed lines mark 50% cumulative frequency. Relative rank was rank of gene divided by total number of genes available in respective dataset, M4501, M96A, and M96B.

Analysis of gene pair association within the different datasets revealed significantly different data structures. The Kendall's τ distribution of gene pair association in M4501 ($\tau_{median} = 0.452$) and M4502 ($\tau_{median} = 0.537$) exhibited strong shifts to positive values, whereas M96A ($\tau_{median} = 0.075$) and M96B ($\tau_{median} = 0.068$) were centered approximately to zero (Fig. 5a). In all datasets we observed more significant positive gene associations than significant negative associations. Furthermore, even though the datasets appeared to be of highly diverse structure, the datasets tested positive for the presence of common gene pair associations, when the Mantel test was applied to compare the gene co-response matrices. In addition, test of homogeneity applied to gene pair associations from the above matrices revealed homogeneity levels of 47.8% – 90.1%. Thus, all datasets contained portions of similar information on pairwise gene associations. This observation was incentive of subsequent comparative analyses.

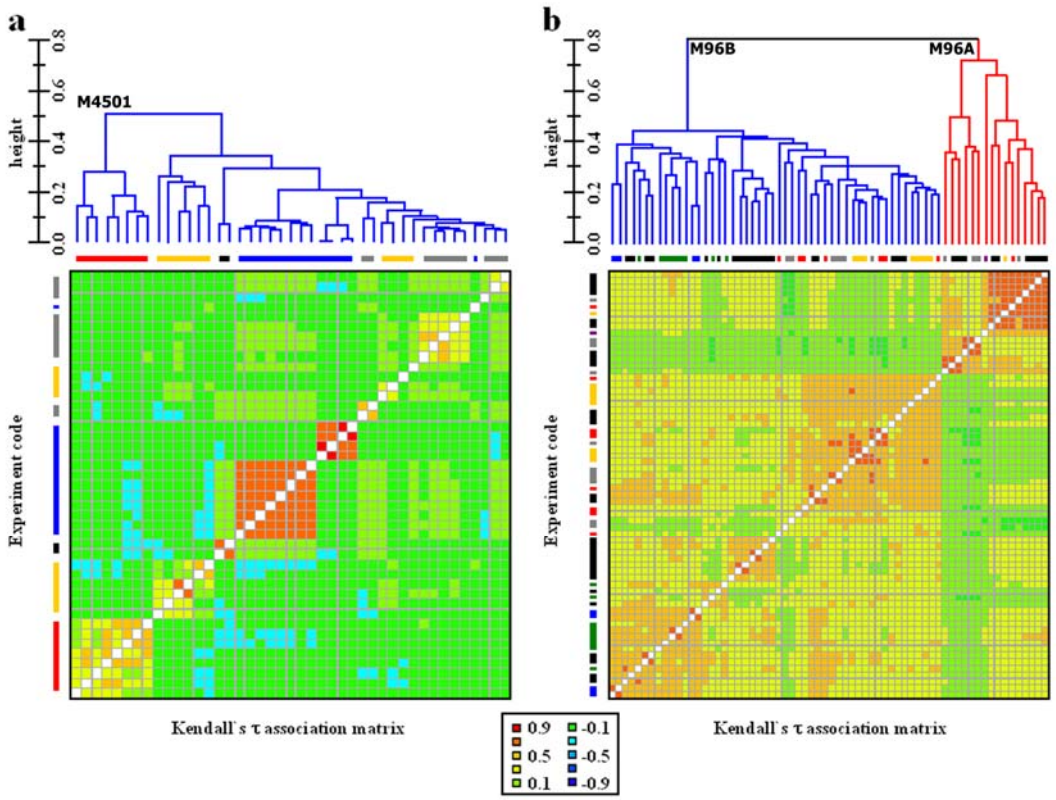

Fig. 4. Comparison of dataset M4501 and M96 by hierarchical clustering of experiments applying Euclidean distance δ_E (top) to a Kendall's τ association matrix (bottom). The experiment categories are colour-coded to the left: M4501 (a), RNA decay – Rifampicin (red), RNA decay – RNA (yellow), miscellaneous (black), amino acid metabolism – tryptophan (blue), and DNA metabolism – UV radiation (grey); M96 (b) cold-shock (blue), various strain/mutant characterization (black), media comparison (green), acid-shock (yellow), antibiotic (red), heat-shock (purple), and growth curve (grey).

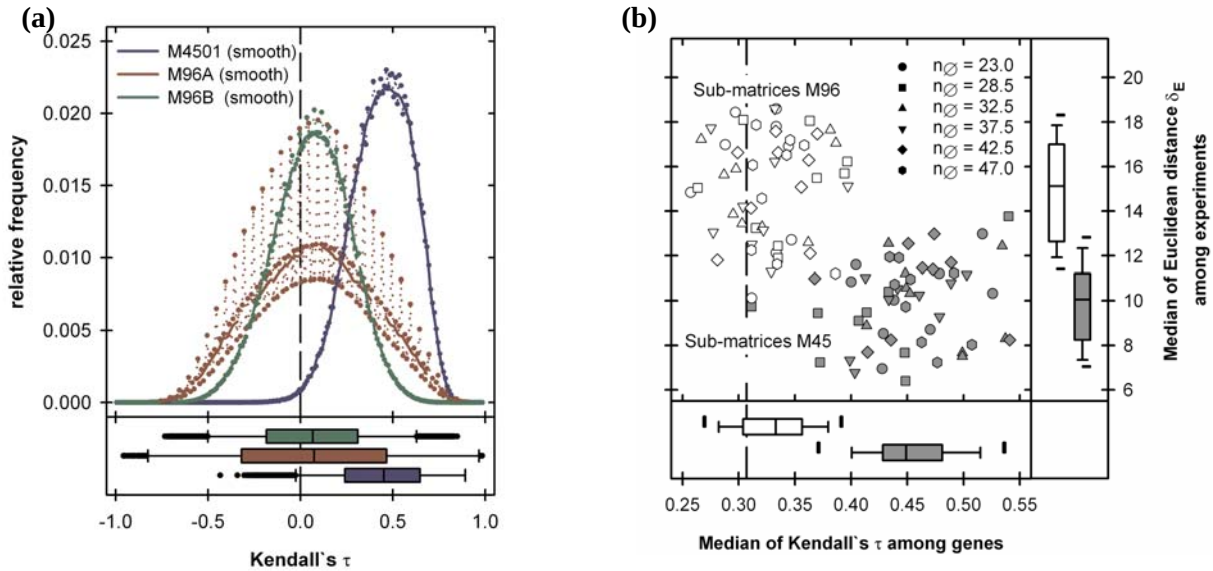

Fig. 5. (a) Histograms (top) and box-plots (bottom) of all gene-pair association from three Kendall's τ correlation matrices, M4501, M96A, and M96B. The bold lines represent smoothed frequency distributions. **(b)**

Scatter-plots and box-plots of median Euclidean distances δ_E representing heterogeneity among experiments and median Kendall's τ of pair wise gene associations from subsets of gene expression data, marked according to source, M45 (grey), M96 (white). Subsets were created at random, comprised approximately 200 genes each, and were characterised by average numbers of gene pairs ($n_{\varnothing}$, inset). The dashed vertical line represents the critical significance threshold of gene associations for n = 22 at probability $\alpha = 0.05$.

Distribution of gene associations

A major concern of our investigation was the positive shift of the distribution of Kendall's τ gene pair associations evident within M4501 (Fig. 5a) and M4502 (data not shown). We investigated whether the choice of gene expression experiments and experimental diversity (Fig. 5a) had an impact on the shift of Kendall's τ distribution of gene pair association (Fig. 5a). In other terms we asked, whether prevailing positive associations in dataset M4501 might reflect higher suitability of this dataset to investigate prokaryotic operon organization as compared to M96.

For this purpose, we created 10 random gene sets, comprising approximately 200 genes each. Genes were chosen only once and had to be present in both datasets. Transcript data of each gene set were extracted separately from M96 and M45. Experiments of each dataset were chosen at random to comprise data subsets that had average numbers of gene pairs as follows: 47.0, 42.5, 37.5, 32.5, 28.5, and 23.0. The numbers of gene pairs were smaller than the numbers of experiments, because of missing data in M45. Heterogeneity among experiments of each selection was determined by median δ_E (Fig. 5b). Our analyses revealed a relation of median δ_E and median Kendall's τ (τ_{median}), e.g. reduced heterogeneity of experiments coincided with a positive shift of τ_{median} of gene pair associations. The portion of significant positive gene pair associations was increased in M45 (Fig. 5b) as compared to M96. Subsequently, by application of parametric ANOVA on τ_{median} distributions we tested the factor that might influence the above observation, namely choice of gene subsets, number of gene pairs, i.e. 23.0 - 47.0, or data source, i.e. M96 or M45. No first- and second-order interactions among these factors were found. Only the data source had a significant influence, ANOVA: $F_s > 293$, $P = 2.50e^{-15}$. A subsequent non-parametric Kruskal-Wallis test ($P < 7.00e^{-06}$) substantiated this finding. In conclusion, the shift of τ_{median} distributions of pair wise gene associations was inherent to datasets and not biased by different numbers of experiments or choice of technology (see above).

However, on comparing Kendall's τ distribution of all genes from M96, either M96A or M96B (Fig. 5a), with the distributions of gene subsets from M96 (Fig. 5b), we observed that the τ_{median} shifted from 0.053 (M96, 4241 genes) to approximately 0.25 - 0.4 (M96, approx. 200 genes each). Thus, shifts of τ_{median} can be caused by the choice of gene subsets. In order to investigate this observation, we subdivided gene associations into one group that represents gene associations by operon structures (type I associations) and a second much larger group of all gene associations which do not describe operon structures (type II associations). Comparative analysis of Kendall`s τ distributions indicated

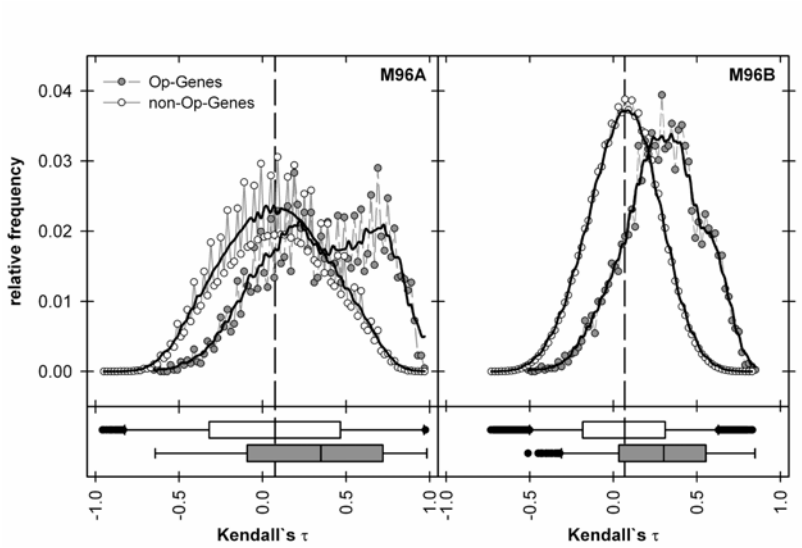

Fig. 6. Kendall's τ distribution of type I associations, i.e. related to operon structure, and type II associations, i.e. not related to operon structure apparent in data subsets M96A and M96B. Dashed lines indicate median of Kendall's τ for the combined type I and type II associations. The bold lines represent smoothed frequency distributions.

fundamentally different properties (Fig. 6). While type II associations were approximately normal distributed and exhibited almost equal numbers of high positive and negative Kendall's τ, type I associations were predominantly positive as was supported by Wilcox rank sum testing and shown previously with an independent dataset (Sabatti et al., 2002). A portion of 75 % of type I associations were overlapping with significant and positive type II associations and thus, the high numbers of type II associations were obscured (Fig. 7). In addition to previous observations we found evidence of bi-modal type I associations (Fig. 6; M96A), indicative of a mechanism, which apparently uncouples associations based on polycistronic mRNA. We define these associations as type III, i.e. those associations that according to operon annotations were expected to be significantly positive, but were found to be non-significant or had even negative and significant Kendall's τ. The type III associations were variable in numbers, as was exemplified by M96A and M96B (Fig. 6), highly operon specific, and dependent on the choice of experiments (Fig. 7).

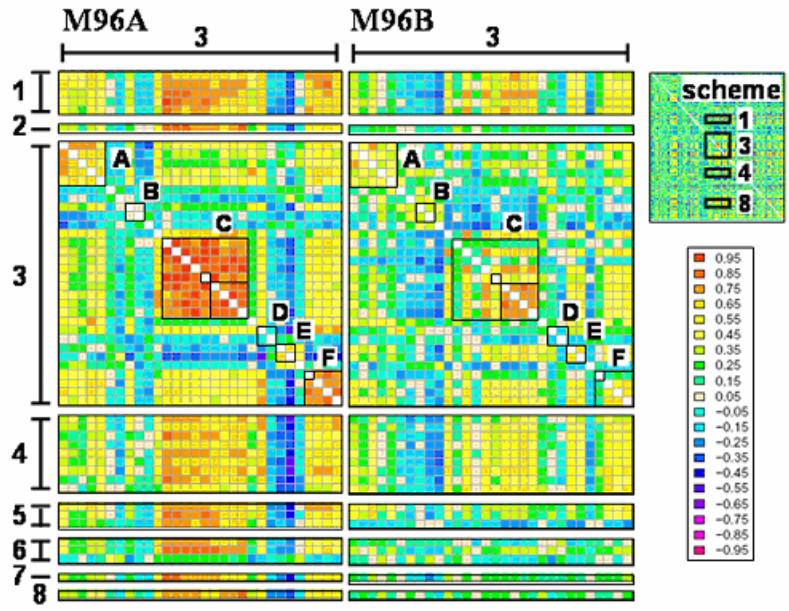

Fig. 7. Colour-coded partial visualization of Kendall's τ association matrix of genes from the chromosomal section nt 15.995 - 16.718. Nearest neighbour associations (section 3) and selected distant associations (sections 1, 2, 4-8) from data subsets M96A (left) and M96B (right) are shown. Sections were: (1) operon *cyoABCDE*, (2) nmpC, (3) region *ybgI* to *tolA*, (4) region *flgB* to *flgJ*, (5) *trpD*, *trpE* and *trpL*, (6) operon *ivbL-ilvBN*, (7) *metH*, and (8) *malE*. Section 1 represents operons: A - *ybgIJKL-nei*, B - *ybgPQ*, C - *sdhCDAB-b0725-sucABCD*, D - *hrsA-ybgG*, E - *cydAB*, and F - *yvgC-tolQRA*. Alternative or ambiguous operon annotations are marked by boxes.

Operon classification by co-clustering

We demonstrated above that type I and type II associations cannot be differentiated without utilization additional knowledge. Co-clustering was suggested to integrate multiple information sources for cluster analyses (Hanisch *et al.,* 2002). We modified this technology to allow overlay of operon annotation, intergenic distance and transcriptional co-response data into combined matrices and subsequent HCA (Fig. 1). We first adjusted co-clustering to enforce classification of genes belonging to the same operons (Fig. 8; OpOVLP). The choice of joining function and weighting were as described earlier (refer to the section 'joining function and combined distance'). Consequently MC approximated 1.0 and RC was 0.0 at clustering height 0.5. The stringency of co-clustering was set to merge even those genes that exhibited negative correlation to other members of these operons into correct operon clusters. Representative results from M4501 are shown (Fig. 8). Instead of adjusting joining function and weighting of co-clustering to allow novel association to operon clusters or rule out previous annotations, we maintained settings but substituted O_{Ttu} for matrices, which described physical proximity and co-linearity of genes on chromosomal strands. The maximum nt allowed applied to construct matrices of weighted intergenic distances (δ_{ig}^{ω}) was optimized to approximate the co-clustering results obtained by O_{Ttu} (Fig. 8). Use of the maximum nt distance threshold, namely 655 596 nt representing the average of non-coding nt divided by 2 of both DNA strands, improved results markedly, MC > 0.3 at RC 1.0 as compared to direct clustering (Fig. 2). Applying a 70 000 nt threshold doubled MC at RC = 1.0. The number of 70 000 nt is equal to the maximum of non-coding nt observed in between any set of 16 co-linear and adjacent genes. The choice of 16 genes was motivated by the largest known operon of *E. coli,* which has 15 genes. The choice of further thresholds, 2250 and 7250 nt, was equal to the maximum length of 95% or of all *E. coli* genes, respectively. Application of these thresholds mimics MC traces obtained with O_{Ttu} (Fig. 8), whereas

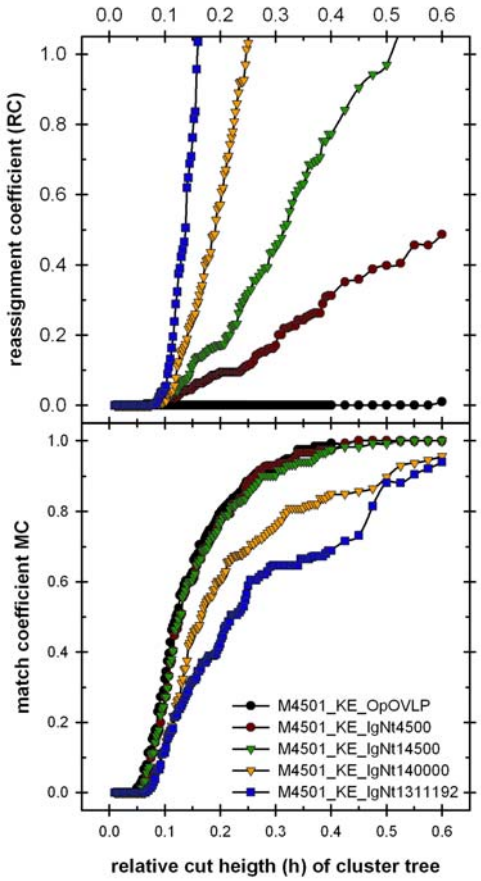

Fig. 8. Plot of match coefficient (MC) (bottom) and the reassignment coefficient (RC) (top) resulting from HCA of distance matrices from dataset M4501 at increasing clustering heights (h). RC is shown in the range of 0.0 to 1.0 with 1.0 representing 50% mis-assigned genes. Kendall's τ distance measure was applied. Inset shows different co-clustering: OpOVLP forced co-clustering by O_{Ttu}.

RC, i.e. the percentage of novel reassignments, can be fine-tuned by selecting the threshold number of nt (data not shown). MC traces obtained with equal settings were mostly independent of datasets. For example, when we applied the 2250 nt threshold, MC only ranged from 0.87 to 1.0 at RC = 1.0. In contrast, RC traces varied widely. RC traces of M96A and M96B had steeper slopes as compared to M4501 and similarly M96B had a steeper RC slope as compared to M96A (data not shown). This observation was indicative of RC slopes increasing with heterogeneity of experiments, $h_{M4501} < h_{M96A} < h_{M96B}$.

Analysis of operon structures

The clustering results can now be used to investigate operon annotations as well as compare and validate transcriptional co-response in different datasets, e.g. under different experimental conditions. In the following, we analyse two exemplary *E. coli* operons in detail using combined matrices (δ_Δ) of δ_{ig}^{ω} at 2250nt threshold with Kendall's τ matrices of the different transcript datasets. Gene clusters were created with clustering heights set at RC = 0.5, i.e. allowing a portion of 33.3% novel gene assignments. Because M96 had a more complete representation of the full genome and most of the genes discussed below were only partially represented in M45 we focussed our investigations on M96.

The complex operon *sdhCDAB-(b0725)-sucABCD* (Fig. 7, section 3C) codes for succinate dehydrogenase (*sdhCDAB*), components of the 2-oxoglutarate dehydrogenase complex (*sucAB*), and parts of succinyl-CoA synthetase (*sucCD*) (Cunningham and Guest, 1998). This operon allows control of a major constituent of the tricarboxylic acid cycle in central metabolism. Prevoius experimental characterization identified two promotor elements, P_{sdh} and the internal P_{suc}, as well as the co-transcription of the entire operon into a 'nine-cistron' mRNA (Cunningham and Guest, 1998). RegulonDB assigned three possible transcripts: (1) the entire *sdhCDAB-b0725-sucABCD* mRNA, (2) a *sucABCD* mRNA and (3) a single-gene *b0725* transcript. In contrast, EcoCyc annotates only the first two mRNAs. Co-clustering of δ_Δ of M96A supports activity of P_{sdh} and expression of the full 'nine-cistron' transcript (b0721 - b0729). All genes of this operon were merged into one cluster (Table 2) and respective type I gene associations were all significant ($P << 0.001$). Dataset M96B and underlying experiments revealed differential regulation of transcriptional co-response (Fig. 7; Table 2). This dataset supports transcriptional activity of P_{suc}, as is evident through significant type I associations of the *sucABCD* operon (b0726 - b0729). All other expected associations were type III, e.g. absent or strongly reduced. Indeed we can propose one possible underlying mechanism of type III associations: differential use of stacked promoters in *E.coli*. These promoters control overlapping subsets of genes, which can be differentially controlled under varying experimental conditions, as was demonstrated.

As a second example, we selected operon *nlpB-dapA,* which according to RegulonDB may be controlled by P_{dapA} (Salgado *et al.*, 2001). *NlpB* (b2477) encodes lipoprotein-34 and *dapA* (b2477) code for dihydrodipicolinate synthase (EC 4.2.1.52). Co-cluster analyses of M96A and M96B merged

both genes into one cluster (Table 3). In addition, adjacent *purC* (b2476), which encodes a subunit of phosphoribosylaminoimidazole-succinocarboxamide synthase, was assigned to this cluster irrespective of choice of dataset (Table 3). All associations of *purC* with *nlpB-dapA* were significant. Based on the proximity of *purC* and *nlpB-dapA* we can propose a transcriptional unit *purC-nlpB-dapA*. Two mechanisms may be the cause of this unit, either operon structure or a strong transcriptional control of two adjacent genes by a common transcription factor. The mechanism remains to be investigated by experimental analyses of transcript length.

Table 2. Kendall's τ association of operon *sdhCDAB-(b0725)-sucABCD* and co-clustering using combined matrices (δ_Δ) of δ_{ig}^{ω} at 2250 nt threshold with Kendall's τ matrices derived from M96A and M96B.

	b0721	b0722	b0723	b0724	b0725	b0726	b0727	b0728	b0729	Type	RIC[a]	rel h[b]	RC[c]
M96A													
b0721			avg.	0.84				operon avg.			2	0.00	0.07
b0722	0.88							0.80			2	0.00	0.07
b0723	0.83	0.92									4	0.03	0.09
b0724	0.75	0.80	0.88								3	0.02	0.08
b0725	0.88	0.90	0.88	0.80						I	1	0.00	0.06
b0726	0.80	0.85	0.83	0.78	0.85			avg.	0.81		1	0.00	0.06
b0727	0.63	0.68	0.77	0.85	0.72	0.73					4	0.03	0.09
b0728	0.72	0.77	0.85	0.83	0.77	0.82	0.92				5	0.15	0.13
b0729	0.72	0.73	0.82	0.87	0.73	0.78	0.78	0.80			4	0.03	0.09
M96B													
b0721			avg.	0.81				operon avg.			5	0.69	0.26
b0722	0.21							0.41			3	0.05	0.10
b0723	0.17	0.09									5	0.69	0.26
b0724	0.16	0.72	0.13								3	0.05	0.10
b0725	0.21	0.47	-0.02	0.45						I	4	0.17	0.13
b0726	0.11	0.64	0.09	0.66	0.43			avg.	0.73		2	0.01	0.08
b0727	0.06	0.57	0.19	0.65	0.34	0.74					1	0.00	0.07
b0728	0.14	0.62	0.03	0.63	0.38	0.82	0.72				2	0.01	0.08
b0729	0.13	0.61	0.19	0.71	0.41	0.70	0.74	0.68			1	0.00	0.07

[a]Rank of merger into operon cluster. [b]Relative height at merger. [c]Reassignment coefficient of data set at merging height.

Table 3. Kendall's τ association of transcriptional unit *purC-nlpB-dapA* (operon *nlpB-dapA*) and co-clustering of combined matrices (δ_Δ) of δ_{ig}^{ω} at 2250nt threshold with Kendall's τ matrices derived from M96A and M96B.

	b2476	b2477	b2478	Type	RIC[a]	rel h[b]	RC[c]
M96A							
b2476		avg.	0.81		2	0.06	0.10
b2477	0.88			I	1	0.02	0.08
b2478	0.83	0.72			1	0.02	0.08
M96B							
b2476		avg.	0.59		1	0.02	0.08
b2477	0.52			I	2	0.16	0.13
b2478	0.66	0.59			1	0.02	0.08

[a]Rank of merger into operon cluster.
[b]Relative height at merger.
[c]Reassignment coefficient of data set at merging height.

Discussion

In this paper we choose a hypothesis-driven co-clustering approach for the identification of transcriptional units. As ideal test cases of co-transcribed genes we used operon structures which result in physically linked co-transcription. Furthermore, we applied our approach to three independent sets of biological experiments using an overlap matrix O_{Ttu}, which represents the combination of all available annotations of polycistronic *E.coli* operons. We clearly demonstrate the failure of the assumption, that polycistronic mRNA inevitably results in high gene-to-gene correlation of transcript measurements. We unravel two major mechanisms that contribute to obscure operon structures within transcript profiles. First, presence of type II associations, which include synergistic (positive Kendall's τ) and antagonistic (negative Kendall's τ) control of distant genes by common transcription factors. These associations dominate in numbers any correlation matrix and overlap with type I associations. Second, type III associations exist. Expected high transcriptional co-response due to operon structures may indeed be conditional, because of known or still undiscovered stacked promoters. An example of the differential use of stacked promoters under different experimental conditions is shown. Moreover, additional mechanisms contributing to type III associations might be functional, such as post transcriptional mRNA processing and degradation.

We conclude that only recruiting additional information will allow extraction of operon structures from gene expression data. We successfully applied co-clustering technology to include gene distance information and demonstrated that gene distance as suggested earlier by Sabatti *et al.* (2002) can effectively substitute information about known operon annotations (Fig. 8). Furthermore, we show evidence, that comparative analyses on data subsets, which describe defined experimental interventions, will be highly informative as compared to global analyses of compendium datasets. The presence of binding sites for multiple transcription factors within many promoter regions as well as occurrence of stacked promoters driving different gene subsets of operons indicate that many overlapping transcription units may exist and can be used in response to varying stimuli. Our analyses demonstrate differential as well as constitutive use of exemplary transcriptional units. Transcription units were shown to be highly dependent on experimental conditions (Fig. 7, Table 1). We envision that analyses of constitutive activity or conditional use of operons and transcriptional units controlled by transcription factors will be imminent task of transcriptome analyses and will lead to further experimental investigations.

Acknowledgements

We thank the staff of the SMD and the ASAP database for the establishment of public accessible sources for microarray data as well as all scientists who submitted transcript profile data to these databases and thereby enabled comparative investigations. Furthermore, we thank the Free Software

Foundation (FSF) for access to software under the terms of the GNU general public license. We acknowledge L. Krall, A. Fernie and J. Kehr for critical reading of this manuscript. Furthermore the comments from the two anonymous referees are gratefully acknowledged.

References

Allen,T.E., Herrgård,M.J., Liu,M., Qiu,Y., Glasner,J.D., Blattner,F.R. and Palsson,B.Ø. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: a model-driven analysis of heterogeneous datasets. *J. Bacteriol.*, **185**, 6392-639.

Baringhaus,L. and Franz,C. (2004) On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190-206..

Bernstein,A.J., Khodursky,A.B., Lin.,P.-H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarray. *Proc. Natl. Acad. Sci. USA*, **99**, 9697-9702.

Blattner,F.R., Plunkett III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis.,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1462.

Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F.R. and Craven,M. (2003a) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34-i43.

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003b) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227-1235.

Courcelle,J., Khodursky,A., Peter,B., Brown,P.O. and Hanawalt,P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, **158**, 41-64.

Cunningham,L. and Guest,J.R. (1998) Transcription and transcript processing in the *sdhCDAB-sucABCD* operon of *Escherichia coli*. *Microbiology*, **144**, 2113-2123.

Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucl. Acids. Res.*, **29**, 1216-1221.

Glasner,J.D., Liss,P., Plunkett III, G., Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. and Perna,N.T. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acid Res.*, **31**, 147-151.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145-154.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2003) The EcoCyc Database. *Nucl. Acid. Res.*, **30**, 56-58.

Khodursky,A.B., Peter,B.J., Cozzarelli,N.R., Botstein,D., Brown,P.O. and Yanofsky,C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **97**, 12170-12175.

Mirkin,B. (1996) Mathematical Classification and Clustering. (Book Series: Nonconvex Optimisation and Its Application: Volume 11). Kluwer Academic Publishers.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329-S336.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763-764.

Perna,N.T., Plunkett III,G., Burland,V., Mau,B., Glasner,J.D., Rose,D.J, Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A., Pósfai,G., Hackett,J., Klink,S., Boutin,A., Shao,Y., Miller,L., Grotbeck,E.J., Davis,N.W., Lim.,A., Dimalanta,E.T., Potamousis,K.D., Apodaca,J., Anantharaman,T.S., Lin,J., Yen,G., Schwartz,D.C., Welch,R.A. and Blattner,F.R. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* 0157:H7. *Nature*, **409**, 529-533.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabási,A.-L. (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science*, **297**, 1551-1555.

Sabatti,C., Rohlin,L., Oh,M.-K. and Liao,C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucl. Acids. Res.*, **30**, 2886-2893.

Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Diaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucl. Acids. Res.*, **29**, 72-74.

Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Wenig,S., Jin,H., Ball.,C.A., Eisen,M.B., Spellman,P.T., Brown,P.O., Botstein,D. and Cherry,J.M. (2001) The Stanford Microarray Database. *Nucleic Acid Res.*, **29**, 152-155.

Sokal,R.R. and Rohlf,F.J. (1995) Biometry: The principles and practice of statistics in biological research. 3rd ed. W.H. Freeman and Company New York.

Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, **18**, 337-344.

Vukmirovic,O.G. and Tilghman,S.M. (2000) Exploring genome space. *Nature*, **405**, 820-822.

Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987-993.

Yamanishi,Y., Vert,J.-P., Nakaya,A. and Kanehisa,M. (2003) Extraction of correlated gene clusters from multiple genomic data by generalised kernel canonical correlation analysis. *Bioinformatics*, **19**, i323-i330.

Zheng,Y., Szustakowski,J., Fortnow,L., Roberts,R. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221-1230.

Chapter IV - Application to *A.thaliana*: - Identification of brassinosteroid-related genes by means of transcript co-response analyses -

Janina Lisso[a], Dirk Steinhauser[a], Thomas Altmann[b], Joachim Kopka[a] and Carsten Müssig[*b]

[a]Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

[b]Universität Potsdam, Institut für Biochemie und Biologie, Genetik, c/o MPI für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

Abstract

Brassinosteroid (BR) effects are largely mediated at the level of transcriptional regulation. Previous expression profiling experiments identified several BR-responsive genes in Arabidopsis. However, BR-responsiveness of the vast majority of identified genes depends on the genotype and specific choice of experimental conditions, and for numerous genes likely remained uncovered hitherto. In this study, the comprehensive systems-biology database (CSB.DB) was used to reveal BR-related genes from 51 gene expression profiles exploiting the concept of co-response analysis – genes exhibiting simultaneous changes in transcript levels are assumed candidates of common transcriptional regulation. In the present study CSB.DB was queried for transcriptional co-responses with the BR-signalling components *BRI1* and *BAK1*. A total of 301 genes out of 9694 genes represented by transcript measurements in the database showed co-responses with both genes, including several known BR-up-regulated genes. The identified genes point towards roles of BR in cell wall modification, water transport, dehydration/cold response, transcriptional control, light signalling, and protein degradation, and imply interactions with auxin and ethylene signalling. Affymetrix expression profiling and real-time RT-PCR analysis of BR-mutants, BR-, and brassinazole treated plants confirmed the BR-dependent expression of 72 genes. Analysis of GA- and paclobutrazol-treated plants demonstrated that the identified genes were virtually unaffected by GA and secondary effects such as altered growth. Our results demonstrate that transcript co-response analysis presents a valuable public resource to uncover common regulatory patterns of genes.

[*] To whom correspondence should be addressed.

Introduction

BR-deficient and BR-insensitive mutants in *Arabidopsis*, pea, tomato, barley, and rice show dwarfism (Bishop, 2003). The growth-promoting effect of BR was assigned to changes in transcript levels of genes involved in cell wall modifications such as xyloglucan endotransglucosylase/hydrolases and expansins. Other BR-regulated genes point to further mechanisms contributing to growth. Enhanced resistance of BR-treated plants to temperature, salt, water, phytopathogens, and other environmental stresses was reported (Khripach et al., 2000; Nakashita et al., 2003; Sasse, 1999). However, the underlying molecular basis is unknown. The growth effect of exogenous BR is light-dependent. For instance, exogenous BR stimulates growth of *Arabidopsis* hypocotyls in the light, but inhibits growth in the dark (Choe et al., 2001; Wang et al., 2002). *Arabidopsis* mutants such as *det2*, *cpd*, and *bri1* display short hypocotyls, opened cotyledons, and emergence of primary leaves in darkness. These findings suggest a crosstalk between photomorphogenesis and steroid signal transduction (Turk et al., 2003).

More than 50 BR-responsive genes have hitherto been identified in *Arabidopsis*. However, the majority of these genes do not show consistent BR-dependent expression in different BR-mutants, under different environmental conditions, and upon BR-treatment (Müssig et al., 2002). These inconsistencies suggest additional regulatory effects, secondary events in BR-deficient dwarf mutants, and not physiological responses upon BR-treatment, for instance of tissues which normally contain low levels of BR. Thus, the usually applied approaches are limited. On the other hand, the identified genes most likely present only a subset of genes involved in the mediation of BR-effects such as growth promotion. One major reason for incomplete uncovering of genomic effects is the limited number of experiments, since gene expression can vary even under highly controlled conditions and many genes probably fail to meet the stringent selection criteria routinely applied in expression profiling experiments.

Cross-experiment co-response analysis provides an alternative approach which is based on the assumption that common transcriptional control of genes is reflected in co-responding, synchronous changes in transcript levels. Thus co-response analysis describes common changes of transcript levels among gene pairs. Publicly available expression profiles represent a rich resource for such cross-experiment investigations. In this study, the CSB.DB (http://csbdb.mpimp-golm.mpg.de/; Steinhauser et al., 2004) was used to identify BR-responsive genes. The CSB.DB provides access to co-response analysis based on 51 expression profiles representing numerous independent experiments which were generated with Affymetrix ATH1 arrays.

In order to identify BR-responsive genes co-response analyses could be performed with known BR-responsive genes or genes involved in BR-signalling. The use of known BR-responsive genes for co-response analyses may result in the identification of further BR-responsive genes. However, BR-responsive genes are positioned downstream within the signalling cascade. The use of BR-signalling

components presents a superior alternative. BR responses depend on signalling components such as BRI1, BAK1, BIN2, BZR1, and BES1. BRI1 is an essential receptor (component) for BR-responses. The BR-insensitivity of the *bri1* mutant (Clouse et al., 1996; Kauschmann et al., 1996; Li and Chory, 1997) indicates that major BR-responses depend on BRI1. BAK1 is a receptor-like kinase which forms a heterodimer with BRI1 (Li et al., 2002; Nam and Li, 2002). BAK1 was identified independently by a yeast two-hybrid screen for BRI1-interacting proteins (Nam and Li, 2002) and as suppressor of a weak *bri1* allele (Li et al., 2002). In the presence of BR a phosphorylation cascade is initiated which receives additional input of proteins such as BSU1 (Mora-Garcia et al., 2004), and finally results in the regulation of BR-responsive genes. Components such as BZR1 and BES1 regulate subsets of BR-responsive genes (Yin et al., 2002). However, these downstream components could mediate responses to other stimuli as well, since the complex phosphorylation cascade likely receives additional input and certainly serves to modulate BR-responses with respect to tissue specificity, environmental conditions, and developmental stages. The use of upstream signalling components for transcript co-response analyses can be expected to result in a more robust identification of BR-related genes.

In this study, both the *BRI1* gene and the *BAK1* gene were used to identify associated genes by transcript co-responses. Only the intersection of genes showing co-responses with both *BRI1* and *BAK1* were considered because the parallel use of two genes allows removing BR-independent transcriptional co-responses associated with either gene. As a pre-requisite for co-response analysis both genes showed changes in transcript levels throughout the set of expression profiles. The *BRI1* and *BAK1* genes show variable transcript levels in different organs (Li and Chory, 1997; Li et al., 2002; Nam and Li, 2002), however, further parameters affecting *BRI1* and *BAK1* transcript levels as well as the molecular basis of the regulation of transcript levels are barely known.

In the following we describe discovery of common co-responsive genes through CSB.DB. We demonstrate supportive crosschecking with publicly available Affymetrix expression profiles provided by the AtGenExpress consortium (http://www.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm). Pairwise comparisons were performed with transcript profiles of BR-, GA-, brassinazole (BRZ) - and paclobutrazol (PAC) - treated plants. Furthermore, 44 cell wall and growth-related genes were selected for experimental validation by means of real-time RT-PCR.

Results

Identification of potential BR-induced genes by means of transcript co-response analyses

The central part of CSB.DB is a set of co-response databases (CoR.DBs) which are based on publicly available transcript profiles. By scanning for the best co-responses among changing transcript levels

CSB.DB allows to infer hypothesis on common regulation of gene expression. In this study, CSB.DB was used to identify BR-regulated genes. Our analysis was restricted to positive transcript co-responses. As mentioned above, from a biological point of view the use of genes encoding upstream signalling components (such as *BRI1* and *BAK1*) appears superior in comparison to genes encoding downstream signalling components (such as *BZR1*). In fact, the *BRI1* and *BAK1* genes showed consistent co-responses with known BR-responsive genes such as *AGP4*, *BEE1*, *GTL1*, *KCS1*, *PRO1*, and *TIP2.1* (*δ-TIP*) (Friedrichsen et al., 2002; Goda et al., 2002; Müssig et al., 2002), whereas *BZR1* barely did. Transcript patterns of *BRI1* and *BAK1* were similar (Fig. 1).

BRI1 and *BAK1* were used to screen for co-responses with all 9694 genes represented in the data matrix. The data matrix (termed nasc0271) comprised 51 manually selected expression profiles (http://csbdb.mpimp-golm.mpg.de/csbdb/home/matrices/ath_nasc0271.html; Experimental procedures). It contained a wide range of experimental conditions and minimal overlap of identical experiments. Importantly, expression profiles of BR-mutants or BR-treated plants were not included. Since a gene had to be measured with high quality, i.e. a detection call of Marginal or Present (according to standard parameters of the MAS 5.0 software) in at least 85% of the experiments, the matrix contained information on 9694 genes rather than on the total of > 22.000 genes represented on the ATH1 array. Several statistical parameters were applied to retrieve genes displaying co-responses. In particular, the Spearman's non-parametric rank correlation coefficient (r_s), the p-value, and the power were used. The exclusion criteria for these parameters were >0.35, <0.01, and >0.7, respectively. 720 and 1179 genes showed co-responses with *BAK1* and *BRI1*, respectively, whereas 301 genes showed co-responses with both genes. Public databases contained

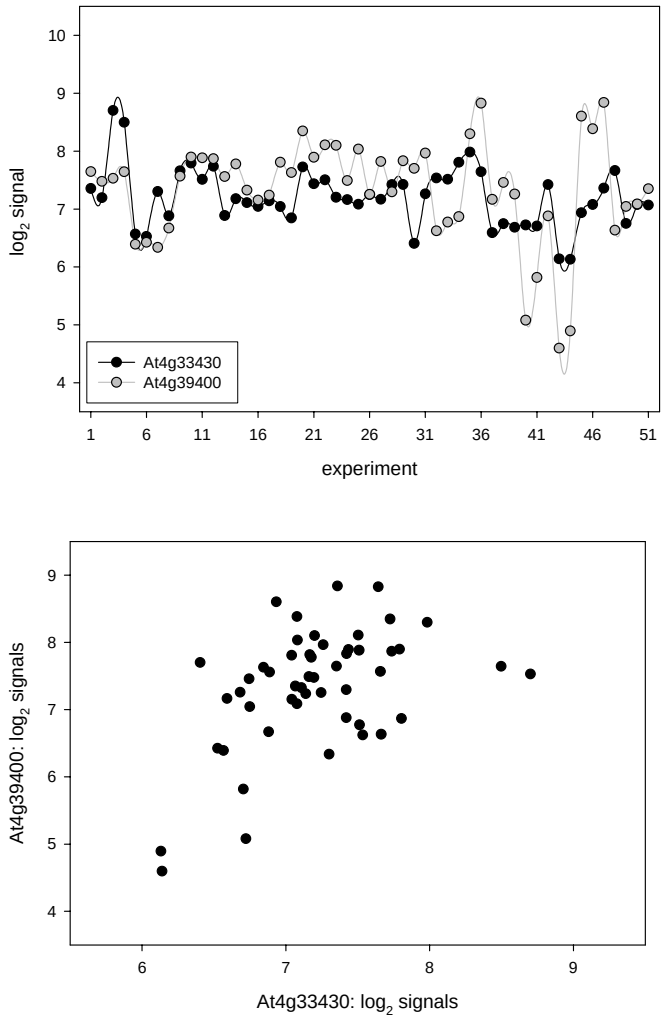

Fig. 1. *BRI1* (At4g39400) and *BAK1* (At4g33430) transcript levels denoted as $\log_2$ signals in 51 expression profiles (incorporated in the CSB.DB data matrix nasc0271) (top). Lines are drawn to aid interpretation. Scatter plot of the $\log_2$ signals (bottom).

information that allowed functional categorisation of more than 50 % of these genes (Fig. 2 and supplementary EXCEL file sheet 1 [301 genes]).

Non-parametric bootstrap analysis with 2.000 bootstrap samples was performed with 135 functionally classified genes (including *BRI1* and *BAK1*). The resulting Spearman correlation coefficients (r_s), p-values, and power values confirmed the correlated behaviour of both the *BAK1* gene and the *BRI1* gene with the respective other 134 genes, and also resulted in a complete matrix of all pairwise gene correlations (supplementary EXCEL file sheet 2 [Spearman rs], sheet 3 [pvalue], and sheet 4 [power]).

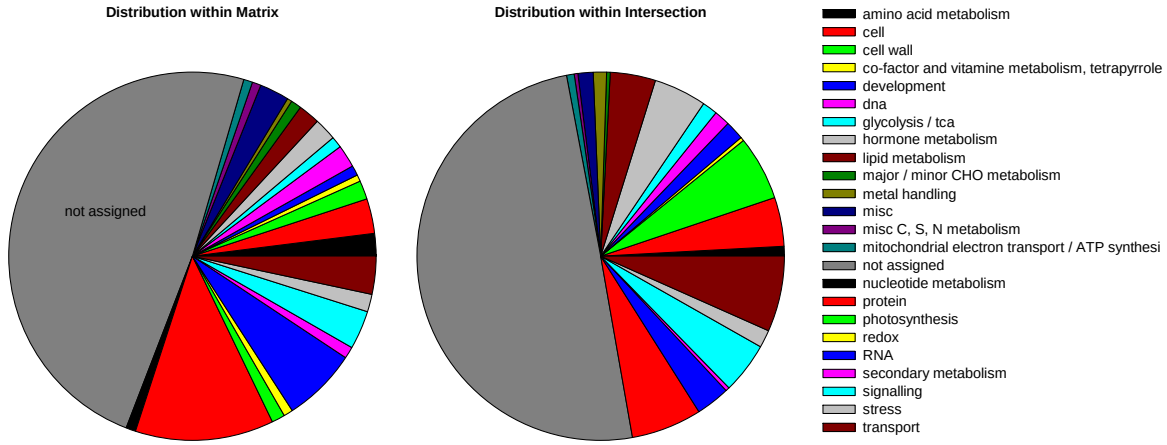

Fig. 2. Functional classification of 301 genes showing co-responses with the BRI1 and BAK1 genes. Categories were deduced from the MapMan software (Thimm et al., 2004) but slightly modified (supplementary information Table S1).

Expression analysis of identified genes by means of Affymetrix expression profiles

AtGenExpress is a multinational coordinated effort to uncover the transcriptome of the multicellular model organism *Arabidopsis thaliana* (http://www.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm). Several groups contributed expression profiles based on the Affymetrix ATH1 full genome chip technology. For this study expression profiles from AtGenExpress, which were established by Hideki Goda (Plant Science Center and Plant Functions Laboratory, RIKEN, Japan), and 2 previously published profiles (Coll-Garcia et al., 2004) were used.

Expression profiles were analyzed using the stringent settings of the statistical algorithms applied in the GCOS software. The detection *P* value (with standard parameters) was applied to remove genes with Absent and Marginal calls. Induced genes were expected to be Present in experimental hybridisations, i.e. experiments representing higher phytohormone levels in comparison to the corresponding baseline experiments. Repressed genes were expected to be Present in baseline experiments, i.e. experiments representing lower phytohormone levels in comparison to the corresponding experimental hybridisations. Simultaneously, change P values with an exclusion limit

of <0.01 or >0.99 and signal log ratios of ≥0.8 or ≤ 0.8 were used in pairwise comparisons in order to identify changes in transcript levels with high reliablity.

30 expression profiles of brassinolide (BL)-, 24-epibrassinolide (EBL)-, castasterone (CS)- or control-treated wild-type and BR-deficient *det2* plants were used for 15 pairwise comparisons. The results were screened for intersections with the 301 genes identified by means of transcript co-responses with *BRI1* and *BAK1*. Upon BR-treatment 55 genes showed stronger expression in at least two independent situations (supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]), whereas 13 genes including the *BRI1* homolog At1g72180 revealed weaker expression in at least 2 independent experiments. Moreover, the *BRI1* gene also showed reduced transcript levels upon BR-treatment. These up- or down-regulated genes did not show any conflicting expression pattern, i.e. were not down- or up-regulated in other situations.

Exogenous BR strongly promotes growth. The observed transcript co-responses could be related to the growth response rather than to specific BR action. In this case expression may be induced by other growth-promoting compounds such as GA. To unravel this 24 expression profiles of GA_3- or control-treated wild-type and *ga1-5* plants were used for 12 pairwise comparisons. Genes with significantly increased or decreased transcript levels were compared with the 301 candidate genes. Only one gene showed stronger expression upon GA-treatment according to the criteria applied for the BR-treatments (supplementary EXCEL file sheet 8 [Affx results GA3, PAC]).

Reduced BR-levels should result in weaker expression of BR-related genes. Brassinazole (BRZ) is a specific BR biosynthesis inhibitor (Asami et al., 2000). We analysed 12 expression profiles of BRZ- or control-treated wild-type plants and performed 8 pairwise comparisons. According to the obtained results 17 of the 301 candidate genes showed reduced expression in the presence of BRZ in at least two independent situations (supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]). In order to test whether reduced GA-levels also affect transcript levels of the 301 genes 8 expression profiles of paclobutrazol (PAC)- or control-treated wild-type plants were used for 4 pairwise comparisons. The results were filtered for significantly altered transcript levels of the 301 candidate genes. Only 2 genes showed weaker expression in the presence of PAC according to the criteria used for BRZ treatments (supplementary EXCEL file sheet 8 [Affx results GA3, PAC]). Thus, transcript levels of identified genes were marginally affected by GA and GA-mediated growth processes. The obtained Affymetrix results are also given in an all-inclusive table (supplementary EXCEL file sheet 6 [complete Affx results]).

Expression analysis of 44 growth-related genes in BR-mutants by means of real-time RT-PCR

Given these findings the transcript co-response data allowed to dissect the molecular basis of BR-promoted growth. Since BR-promoted growth represents a focus of our current work (Coll-Garcia et al., 2004; Müssig et al., 2003), we subjected growth-related genes to experimental examination. The

301 candidate genes were manually screened. 44 genes with predicted or known functions in growth processes and cell wall modifications were selected for real-time RT-PCR analysis. BR-induced genes were expected to display reduced transcript levels in different BR-deficient and BR-insensitive mutants. The combination of different genotypes provides means to exclude changes in transcript patterns which are restricted to a specific genotype. In fact, previous expression profiles revealed different changes in BR-mutants (Goda et al., 2002; Müssig et al., 2002), and only a subset of genes showed consistent changes. This finding is in agreement with phenotypic differences of BR-mutants. BR-treated plants were also analysed. Wild-type and *dwf1-6* plants were treated with 300 nM EBL and plant material was harvested 5 h after treatment.

We performed 3 sets of independent experiments. Firstly, wild-type, *dwf1-6*, *cbb3*, and *cbb2* plants (all C24 background, mutants allelic to the *dim*, *cpd*, and *bri1* mutants; Kauschmann et al., 1996) were grown in half-concentrated Murashige and Skoog medium. Plant material was harvested 14 d or 19 d after sowing. Secondly, wild-type, *dwf1-6* plants, and *CPD*-antisense plants (Schlüter et al., 2002) were grown in soil. Plant material was harvested 28 d after sowing. Thirdly, wild-type and *dwf1-6* plants were grown in half-concentrated Murashige and Skoog medium and either treated with a control solution or 300 nM EBL.

Transcript levels were determined by means of real-time RT-PCR. The amplification efficiency *E* of each primer pair (Table 1) was determined from the log slope of SYBR Green fluorescence versus cycle number in the exponential phase, and was used to normalise the readout for each primer pair and run (Czechowski et al., 2004; see Experimental procedures). 8 fold change values from 3 sets of experiments were established (Table 2 and Table 3). Underlying C_T values are provided in Table 4. Pairwise comparisons were organized like the Affymetrix pairwise comparisons (i.e. wild-type versus BR-mutant and BR-treatment versus control treatment). Therefore, fold change values greater than 1.0 indicate a positive BR-effect on transcript levels.

23 genes showed reduced transcript levels in different BR-mutants (Table 2). These genes include the known BR-responsive *KCS1* and *TIP2.1* (*δ-TIP*) genes (Coll-Garcia et al., 2004; Goda et al., 2002) (encoding a β-ketoacyl-CoA synthase required for wax biosynthesis and an aquaporin, respectively), genes presumably involved in cell wall modifications (At1g27600, *AGP21* [At1g55330], At2g06850, *AGP9* [At2g14890], At3g05910, At3g24480, At3g28180, At3g57790, *FLA2* [At4g12730], At4g13340, and At4g18670), 4 aquaporins (*TIP1.1* [At2g36830], *PIP1.2* [At2g45960], *TIP2.1* [At3g16240], and *TIP1.2* [At3g26520]), and a *KCS1* homologue [At5g43760]). In general, fold changes of growth-related genes were more pronounced in 14 and 19 d old plants in comparison to 28 d old plants, likely due to reduced growth rates in older plants (Table 2). 6 (of 23) genes also displayed higher mRNA levels in BR-treated plants (Table 2). 18 genes did not exhibit BR-dependent expression or showed variable transcript levels (Table 3). 3 genes displayed stronger expression in BR-mutants or weaker expression upon BR-treatment (Table 3). The 44 genes subjected to RT-PCR analysis were highlighted in the supplementary tables to aid comparisons with Affymetrix data (supplementary

EXCEL file sheet 6 [complete Affx results], sheet 7 [Affx results BRs, BRZ], and sheet 8 [Affx results GA3, PAC]).

Table 1. Primers used for real-time RT-PCR analysis. The amplification efficiency (*E*) was determined for each primer pair (see Experimental procedures).

Gene	sense primer	antisense primer	*E* ± SD
At1g01120	ACCGAAGCTAAGGGTCGGGTTA	GTAACTCTTTCCAAACCGCACT	0.838 ± 0.007
At1g01620	CGGAACCTTCGTCCTTGTCT	ATTGGGAGTGGTGCCAAAAT	0.936 ± 0.007
At1g03870	GGTGAGATGTTCGGAGAGCA	TGACTTATCCGACGGTGACG	0.862 ± 0.003
At1g27600	CGACGTCCCTTTTCACATCC	TCCATCTCGCTTTCATCAGC	0.906 ± 0.003
At1g31310	TGTGGAGGAGGTTGGACGTA	CTCTCCCTAATCGCGTTTGC	0.857 ± 0.005
At1g55330	CAGCTCCAAGCCCAACTTCT	TGAAACCAGATGCCAAAGCA	0.896 ± 0.049
At1g70210	ACGAGAGCCCTGAGACTTGG	GGCCATCGCTTTCATCAGTC	0.861 ± 0.012
At2g06850	CTGAACAATGGCGTCGTCTC	CTGGCATAACCGGGAACCTA	0.894 ± 0.016
At2g14890	TTGGATCTGTTCTCGTCTGG	TCCCAAGCAATACAGTGGAA	0.936 ± 0.002
At2g16850	AGAAGTGCCCGTGACTCTCA	CCAAAGCTTCTGGCTGGATT	0.868 ± 0.001
At2g22840	GCGATGCGTGGATCAGATAA	GAAGCCCCTCGAGAATTTTG	0.757 ± 0.011
At2g36400	TTTTGGTGGTGGTGGTGGTA	TCTTGCTTCATCTCCGAACA	0.833 ± 0.021
At2g36830	TCGCTTGCCTCATCCTTAAA	TCGAAAACGAAAGCGTTCAA	0.981 ± 0.000
At2g45960	CCAAGAGAAACGCTCGTGAC	CCAGTAATGGGGATGGTTGC	0.898 ± 0.036
At3g05910	CGGTGGCAATAGCTGTAGGA	CACCAAGTTGTGGCAGCTCT	0.874 ± 0.011
At3g16240	CACAGTCATCACCGGAGTTT	GATCCTAGTCCAGCCGCAAC	0.934 ± 0.012
At3g24480	ACCGCCAGTCCATCACAGTT	GTGGTGGGGGAGATGCATAG	0.892 ± 0.018
At3g26520	GTTTCTGGCCGTTGGATCAT	TTGCACAAAAGCCTTCCAGA	0.991 ± 0.000
At3g28180	GCAACAACATCGTTGGCATT	CGCCTTCTTCCAAACCGATA	0.907 ± 0.005
At3g53420	CCACCAATTCGTTCTGAGAG	GTGTTTAGACGTTGGCAGCA	0.887 ± 0.014
At3g57790	TTACAGTGCAGGGCTTGTGG	AACCCGTTCACATGCTCCAT	0.917 ± 0.014
At3g61430	GACCATTCCTGGGATGACCA	TGGCTCTGATGACAACCACA	0.953 ± 0.026
At4g02500	CGGTGTTCGAAGCTGATGAT	GAATCCCCCAGTAACCGTGA	0.952 ± 0.006
At4g12420	TCGGACACCCTGACAATGTT	CGTGAATCCAATGCTCTTCG	0.854 ± 0.003
At4g12730	CGTGGCATTTACCTCGCTTT	TTCCAAATCTCCACACCAAG	0.899 ± 0.000
At4g13340	AGGACCACTACCACCGGTCA	CGATCTGTTCGTGTCACGTC	0.791 ± 0.004
At4g14130	CTGGAGACCCCAACACATCA	TGGGTTGACTCTTTGGGAAA	0.842 ± 0.013
At4g18670	GCCGTCGTCACCTAGTCCTC	GGTGGGGATGGAGGAGAGTA	0.724 ± 0.001
At4g19410	CTTGGTCCGGTGACAAAGGT	ATGCACTCCGGCCATAAAAC	0.704 ± 0.000
At4g23400	ATTAACCCGGCCAGGAGTCT	GATGGTACAGAGCAGCAAGC	0.917 ± 0.001
At4g25260	TTGAGCAATTCGCTGAAGGA	GCGGTCTCATCTGTCAGTGC	0.897 ± 0.000
At4g25620	ACGGGTCGTCAAGGCATAAT	AACCGGAACGGTTTAGCTTG	0.833 ± 0.001
At4g31590	CGGGACCTATGCAGCTTTTC	TCAGATTCGCCTTCTTCCAT	0.890 ± 0.001
At5g01210	GGGGAAAACCGTTAGCTGTG	ACTTCCAGATCCACGCTTCC	0.832 ± 0.008
At5g05170	ATTGCCAGCCGTTTGTCTCT	TATACCCGTGGCGAAAATGG	0.949 ± 0.007
At5g07830	CAGCTACGGGTTTACGCACA	CCACGTTTATGCCATTGCTG	0.870 ± 0.022
At5g12250	GGACAATGAAGCCCTTTACG	CCGGGAACCTAAGACAGCAT	0.773 ± 0.042
At5g19770	TCTCCGTCCGTCGAAGAAGT	GACCTGGATCCCAGCTTGTC	0.884 ± 0.001
At5g20250	AACTCGCGATTGTTTGTTCG	CACGCCAAGAACTCCAGTGT	0.892 ± 0.016
At5g26670	GGAGCTGTCACAATCTGGTC	ACGAATGCAGCAAAATGTCA	0.838 ± 0.007
At5g43760	AATCTTGGTGGAATGGGATG	GCGTATGAGTTTGGTTGCAC	0.872 ± 0.027
At5g55730	TAATGTCGGCTCATGGATGC	CATTGCATCATCTCCTGGAC	0.986 ± 0.002
At5g60390	TTGACAGGCGTTCTGGTAAGG	CAGCGTCACCATTCTTCAAAAA	not determined
At5g64740	CCCTGCCATCTGTCTTCTCA	AGAAGAGCGCCATGAAGAGG	0.884 ± 0.006
At5g67260	CCGTATGTTTCTCGATGTGC	ATTGTAGCCGATGGCCGATA	0.822 ± 0.009

Table 2. Real-time RT-PCR analysis of cell wall and growth-related genes. For comparison of basal transcript levels wild-type plants and different BR mutants were grown under aseptic conditions (columns 1 – 6) or in soil (columns 7 and 8) and harvested after 14 days (column 1), 19 days (columns 2 – 6), and 28 days (columns 7 and 8), respectively. For BR-treatments (columns 5 and 6) plant material was either treated with a control solution or 300 nM 24-epibrassinolide. Plant material was harvested 5 h after treatment. Control treatments were also used to compare basal transcript levels in wild-type and *dwf1-6* plants (column 4). Fold change ratios were calculated taking into account amplification efficiency (*E*) of all primer pairs (Table 1). RT-PCR data (i.e. C_T values) are given in Table 4.

		sterile culture			sterile culture			soil	
		WT vs *dwf1-6*	WT vs *cbb2*	WT vs *cbb3*	WT vs *dwf1-6* controls	WT BR vs control	*dwf1-6* BR vs control	WT vs *dwf1-6*	WT vs αCPD
Gene		14d	19d	19d	19d	19d	19d	28d	28d
Weaker expression in at least two mutants, stronger expression upon BR-application:									
KCS1	At1g01120	1.7	1.5	2.2	1.1	1.6	1.4	3.1	1.6
	At1g27600	2.0	0.9	1.6	1.9	2.3	1.4	1.4	2.0
	At3g24480	1.4	1.6	1.7	1.4	2.5	1.8	1.6	1.3
CSLC4	At3g28180	1.4	0.9	1.9	2.5	1.5	3.0	0.8	1.1
	At3g57790	1.3	1.5	1.5	7.3	2.1	3.9	2.6	1.8
	At4g13340	1.6	1.1	2.9	1.2	1.7	1.9	0.8	1.7
Weaker expression in at least two mutants:									
PIP1.3	At1g01620	3.7	1.4	1.8	2.5	0.9	1.6	0.9	0.8
AGP21	At1g55330	2.1	3.7	4.1	1.9	1.0	1.7	1.2	2.3
EXT	At2g06850	2.5	1.3	2.0	1.6	1.2	1.2	1.0	0.8
AGP9	At2g14890	1.8	1.9	1.7	1.6	0.7	1.0	1.2	2.0
PIP2.8	At2g16850	3.6	2.3	3.7	2.5	0.8	1.5	0.8	1.2
TIP1.1	At2g36830	1.4	1.6	1.2	1.7	0.7	1.2	0.7	1.2
PIP1.2	At2g45960	1.9	1.8	1.8	1.8	0.7	1.0	1.5	1.0
	At3g05910	1.2	1.2	1.5	1.6	1.1	1.6	1.7	1.4
TIP2.1	At3g16240	2.8	1.5	2.3	1.9	0.9	1.3	0.8	1.0
TIP1.2	At3g26520	1.6	2.2	1.7	1.8	1.5	1.2	1.4	1.8
PIP2.1	At3g53420	2.6	1.4	1.9	1.5	0.7	0.6	1.7	1.0
PIP1.1	At3g61430	2.7	1.5	1.7	1.9	0.9	1.0	1.6	1.1
SKU5	At4g12420	1.4	1.3	2.0	1.0	1.1	1.0	0.8	0.9
FLA2	At4g12730	1.5	2.5	2.7	0.8	1.2	1.0	1.3	1.7
	At4g18670	1.1	1.5	2.1	1.4	0.8	1.0	1.0	1.2
	At4g25260	1.3	1.9	1.1	1.3	0.7	0.9	1.0	1.9
	At5g43760	1.8	1.6	1.7	1.3	0.9	1.0	0.8	0.7

Table 3. Real-time RT-PCR analysis of cell wall- and growth-related genes. Experiments and methods as described in legend of Table 2.

		sterile culture			sterile culture			soil	
		WT vs *dwf1-6*	WT vs *cbb2*	WT vs *cbb3*	WT vs *dwf1-6* controls	WT: BR vs control	*dwf1-6*: BR vs control	WT vs *dwf1-6*	WT vs αCPD
Gene		14d	19d	19d	19d	19d	19d	28d	28d
No consistent BR-dependent or variable expression:									
FLA9	At1g03870	1,6	1.1	0.8	1.0	1.5	1.5	1.1	1.0
	At1g31310	0.9	0.5	0.6	1.1	1.8	1.1	0.6	1.0
CYCD1.1	At1g70210	2.3	0.8	1.2	2.8	0.4	0.8	0.9	1.0
	At4g02500	1.1	1.0	1.1	0.8	0.8	1.1	1.2	0.8
	At4g19410	1.1	0.9	1.1	1.1	2.3	1.1	1.0	1.5
PIP1.5	At4g23400	2.0	1.7	1.4	1.3	0.7	0.5	0.7	0.7
	At4g25620	1.0	0.7	1.1	1.0	1.1	0.8	0.6	0.6
CSLC5	At4g31590	1.2	1.0	1.5	1.8	0.5	0.7	1.5	1.0
	At5g01210	1.0	1.1	1.0	0.5	1.1	0.7	1.1	0.9
CESA3	At5g05170	1.4	1.0	1.3	1.1	1.0	1.0	1.3	1.2
	At5g07830	0.8	1.0	1.2	1.0	1.0	1.4	1.8	1.4
TUB6	At5g12250	1.2	1.2	1.4	1.3	0.9	1.0	1.5	1.7
TUA3	At5g19770	1.2	1.3	1.1	1.3	0.6	1.1	1.5	1.8
DIN10	At5g20250	1.7	1.4	2.2	0.5	2.3	1.3	0.4	1.4
	At5g26665	1.0	0.8	1.3	1.5	0.8	0.8	1.2	1.2
FLA1	At5g55730	1.0	1.2	1.0	0.8	1.1	0.6	1.3	1.8
CESA6	At5g64740	1.2	1.1	1.8	1.4	1.0	1.2	1.2	0.7
CYCD3.2	At5g67260	1.5	1.0	1.1	1.7	0.5	0.7	0.7	1.0
Stronger expression in mutants or weaker expression upon BR-application:									
GRL1	At2g22840	1.2	1.1	1.6	1.4	0.5	0.7	0.6	0.6
GRL3	At2g36400	1.2	1.0	1.3	1.2	0.6	0.6	0.4	0.5
XTR7	At4g14130	0.4	0.4	1.1	0.2	3.8	0.3	0.1	0.6

Table 4. RT-PCR analysis of 44 cell wall and growth-related genes in 12 situations used for the calculation of fold change values in Tables 2 and 3. C_T values were calculated from 3 technical replicates per experiment. C_T values of the *eIF1α* control gene were subtracted from C_T values of the genes of interest to account for different cDNA amounts. The resulting nC_T values were subtracted from an arbitrary value (i.e. 30). Numbers give the difference ($30 - nC_T$) and the standard error (SE) of C_T values for each gene of interest. Higher numbers indicate stronger expression.

		sterile culture			sterile culture			sterile culture			soil		
		WT	*cbb2*	*cbb3*	WT	WT control	WT BR	*dwf1-6*	*dwf1-6* control	*dwf1-6* BR	WT	*dwf1-6*	αCPD
Gene		19d	19d	19d	14d	14d	14d	14d	14d	14d	28d	28d	28d
KCS1	At1g01120	8.8 ± 0.0	8.1 ± 0.0	7.5 ± 0.0	9.4 ± 0.0	9.0 ± 0.0	9.7 ± 0.0	8.5 ± 0.0	8.8 ± 0.0	9.4 ± 0.0	8.6 ± 0.0	6.7 ± 0.0	7.8 ± 0.0
PIP1.3	At1g01620	7.2 ± 0.0	6.6 ± 0.0	6.2 ± 0.0	8.3 ± 0.0	7.9 ± 0.0	7.8 ± 0.0	6.3 ± 0.1	6.5 ± 0.0	7.2 ± 0.0	7.3 ± 0.0	7.5 ± 0.0	7.8 ± 0.0
FLA9	At1g03870	6.1 ± 0.0	6.0 ± 0.0	6.5 ± 0.0	6.2 ± 0.0	5.9 ± 0.1	6.6 ± 0.1	5.5 ± 0.0	5.9 ± 0.0	6.5 ± 0.1	6.3 ± 0.0	6.1 ± 0.1	6.3 ± 0.0
	At1g27600	7.0 ± 0.0	7.2 ± 0.0	6.3 ± 0.0	6.3 ± 0.0	6.3 ± 0.0	7.6 ± 0.0	5.2 ± 0.0	5.4 ± 0.0	5.9 ± 0.0	6.2 ± 0.0	5.7 ± 0.0	5.1 ± 0.0
	At1g31310	3.2 ± 0.0	4.3 ± 0.0	4.0 ± 0.0	3.5 ± 0.0	3.7 ± 0.1	4.7 ± 0.0	3.7 ± 0.0	3.5 ± 0.1	3.7 ± 0.0	3.8 ± 0.1	4.7 ± 0.0	3.7 ± 0.0
AGP21	At1g55330	10.3 ± 0.0	8.3 ± 0.0	8.1 ± 0.0	11.0 ± 0.0	11.1 ± 0.0	11.2 ± 0.0	9.8 ± 0.0	10.2 ± 0.1	11.0 ± 0.0	10.5 ± 0.0	10.2 ± 0.0	9.2 ± 0.0
CYCD1.1	At1g70210	4.5 ± 0.1	4.9 ± 0.1	4.2 ± 0.0	4.8 ± 0.0	4.9 ± 0.0	3.5 ± 0.0	3.4 ± 0.0	3.2 ± 0.0	2.8 ± 0.0	4.4 ± 0.0	4.5 ± 0.0	4.4 ± 0.0
EXT	At2g06850	6.8 ± 0.0	6.4 ± 0.0	5.7 ± 0.0	8.4 ± 0.0	8.1 ± 0.0	8.3 ± 0.0	7.0 ± 0.0	7.4 ± 0.0	7.7 ± 0.0	7.6 ± 0.0	7.6 ± 0.0	8.0 ± 0.0
AGP9	At2g14890	10.2 ± 0.1	9.2 ± 0.0	9.4 ± 0.0	11.3 ± 0.4	10.8 ± 0.0	10.4 ± 0.0	10.4 ± 0.0	10.1 ± 0.0	10.1 ± 0.0	10.4 ± 0.0	10.2 ± 0.0	9.3 ± 0.0
PIP2.8	At2g16850	5.7 ± 0.0	4.3 ± 0.0	3.6 ± 0.0	5.6 ± 0.0	5.8 ± 0.1	5.4 ± 0.1	3.6 ± 0.2	4.3 ± 0.1	5.0 ± 0.0	5.4 ± 0.0	5.8 ± 0.0	5.1 ± 0.0
GRL1	At2g22840	5.9 ± 0.0	5.7 ± 0.2	5.1 ± 0.1	5.2 ± 0.0	5.3 ± 0.0	4.2 ± 0.0	4.9 ± 0.0	4.7 ± 0.0	4.0 ± 0.0	5.6 ± 0.1	6.5 ± 0.0	6.6 ± 0.1
GRL3	At2g36400	5.0 ± 0.0	5.0 ± 0.0	4.6 ± 0.0	5.9 ± 0.0	6.0 ± 0.1	5.1 ± 0.0	5.5 ± 0.0	5.7 ± 0.0	4.9 ± 0.0	5.0 ± 0.0	6.3 ± 0.1	6.2 ± 0.0
TIP1.1	At2g36830	7.5 ± 0.0	6.7 ± 0.0	7.2 ± 0.1	7.7 ± 0.0	7.9 ± 0.0	7.3 ± 0.0	7.2 ± 0.1	7.1 ± 0.0	7.5 ± 0.0	8.0 ± 0.0	8.5 ± 0.0	7.7 ± 0.0
PIP1.2	At2g45960	9.5 ± 0.0	8.6 ± 0.1	8.6 ± 0.0	9.2 ± 0.0	9.2 ± 0.1	8.6 ± 0.0	8.2 ± 0.1	8.3 ± 0.0	8.3 ± 0.0	9.7 ± 0.1	9.0 ± 0.0	9.6 ± 0.0
	At3g05910	8.4 ± 0.0	8.1 ± 0.0	7.7 ± 0.0	8.8 ± 0.0	8.9 ± 0.1	8.9 ± 0.0	8.5 ± 0.0	8.2 ± 0.1	8.9 ± 0.0	8.8 ± 0.0	7.9 ± 0.0	8.2 ± 0.1
TIP2.1	At3g16240	7.9 ± 0.0	7.2 ± 0.0	6.6 ± 0.0	9.5 ± 0.0	9.1 ± 0.1	8.9 ± 0.1	7.9 ± 0.1	8.1 ± 0.3	8.5 ± 0.0	9.4 ± 0.0	9.8 ± 0.0	9.4 ± 0.0
	At3g24480	9.8 ± 0.0	9.0 ± 0.1	9.0 ± 0.0	9.5 ± 0.0	9.5 ± 0.1	11.0 ± 0.0	9.0 ± 0.0	9.1 ± 0.0	10.0 ± 0.1	9.2 ± 0.0	8.4 ± 0.0	8.8 ± 0.4
TIP1.2	*At3g26520*	10.9 ± 0.0	9.8 ± 0.2	10.1 ± 0.1	11.9 ± 0.0	12.3 ± 0.0	11.7 ± 0.0	11.2 ± 0.0	11.5 ± 0.0	11.2 ± 0.0	11.5 ± 0.0	11.0 ± 0.0	10.7 ± 0.0
CSLC4	At3g28180	3.5 ± 0.0	3.8 ± 0.1	2.5 ± 0.0	3.8 ± 0.1	4.4 ± 0.0	5.0 ± 0.0	3.3 ± 0.1	2.9 ± 0.0	4.6 ± 0.1	5.2 ± 0.0	5.5 ± 0.0	5.1 ± 0.0
PIP2.1	At3g53420	7.7 ± 0.0	7.1 ± 0.0	6.7 ± 0.0	8.6 ± 0.0	8.3 ± 0.0	7.7 ± 0.0	7.1 ± 0.0	7.7 ± 0.0	6.8 ± 0.0	8.9 ± 0.1	8.1 ± 0.0	8.9 ± 0.0
	At3g57790	7.3 ± 0.0	6.7 ± 0.0	6.6 ± 0.0	7.1 ± 0.2	7.5 ± 0.3	8.6 ± 0.0	6.7 ± 0.2	4.4 ± 0.0	6.5 ± 0.1	7.9 ± 0.0	6.4 ± 0.0	7.0 ± 0.0
PIP1.1	At3g61430	9.6 ± 0.0	9.0 ± 0.0	8.8 ± 0.0	9.6 ± 0.0	9.4 ± 0.0	9.2 ± 0.0	8.1 ± 0.0	8.4 ± 0.0	8.4 ± 0.1	9.7 ± 0.0	9.0 ± 0.1	9.5 ± 0.0

	At4g02500	7.9 ± 0.0	8.0 ± 0.0	7.7 ± 0.2	7.9 ± 0.0	8.0 ± 0.0	7.7 ± 0.0	7.8 ± 0.0	8.3 ± 0.0	8.4 ± 0.0	8.5 ± 0.0	8.2 ± 0.1	8.7 ± 0.0
SKU5	At4g12420	9.7 ± 0.0	9.2 ± 0.0	8.5 ± 0.0	10.4 ± 0.0	10.3 ± 0.0	10.5 ± 0.0	9.9 ± 0.0	10.3 ± 0.0	10.4 ± 0.0	9.3 ± 0.0	9.6 ± 0.1	9.4 ± 0.0
FLA2	At4g12730	10.2 ± 0.0	8.8 ± 0.0	8.7 ± 0.0	10.7 ± 0.0	10.2 ± 0.0	10.4 ± 0.0	10.1 ± 0.0	10.6 ± 0.0	10.6 ± 0.0	10.0 ± 0.0	9.6 ± 0.0	9.2 ± 0.1
	At4g13340	6.0 ± 0.1	5.8 ± 0.0	4.2 ± 0.2	7.0 ± 0.0	7.0 ± 0.0	7.9 ± 0.0	6.2 ± 0.1	6.6 ± 0.1	7.7 ± 0.0	7.0 ± 0.0	7.4 ± 0.0	6.1 ± 0.0
	At4g14130	6.0 ± 0.0	7.5 ± 0.0	5.9 ± 0.0	6.3 ± 0.1	5.9 ± 0.0	8.1 ± 0.0	7.9 ± 0.0	8.9 ± 0.0	6.9 ± 0.0	5.2 ± 0.0	9.0 ± 0.0	6.0 ± 0.0
	At4g18670	4.1 ± 0.0	3.4 ± 0.0	2.8 ± 0.2	3.2 ± 0.0	3.6 ± 0.0	3.2 ± 0.0	3.0 ± 0.1	3.0 ± 0.0	3.0 ± 0.0	3.4 ± 0.0	3.5 ± 0.0	3.2 ± 0.0
	At4g19410	8.5 ± 0.0	8.7 ± 0.3	8.3 ± 0.0	7.7 ± 0.1	7.9 ± 0.1	9.5 ± 0.0	7.4 ± 0.1	7.7 ± 0.0	7.8 ± 0.0	9.4 ± 0.0	9.3 ± 0.0	8.6 ± 0.0
PIP1.5	At4g23400	8.9 ± 0.0	8.1 ± 0.0	8.3 ± 0.0	9.8 ± 0.0	9.5 ± 0.1	8.9 ± 0.0	8.8 ± 0.0	9.1 ± 0.0	8.1 ± 0.0	9.1 ± 0.0	9.7 ± 0.0	9.5 ± 0.0
	At4g25260	7.1 ± 0.0	6.1 ± 0.0	5.9 ± 0.2	8.0 ± 0.0	7.6 ± 0.0	7.1 ± 0.1	7.6 ± 0.0	7.2 ± 0.4	7.0 ± 0.1	7.0 ± 0.0	6.1 ± 0.1	7.1 ± 0.0
	At4g25620	7.3 ± 0.0	7.8 ± 0.0	7.2 ± 0.0	7.8 ± 0.0	7.9 ± 0.0	8.1 ± 0.1	7.7 ± 0.0	7.9 ± 0.0	7.6 ± 0.0	7.3 ± 0.0	8.2 ± 0.1	8.1 ± 0.0
CSLC5	At4g31590	7.0 ± 0.0	7.0 ± 0.1	6.4 ± 0.0	5.8 ± 0.0	6.3 ± 0.0	5.2 ± 0.0	5.6 ± 0.0	5.4 ± 0.0	4.9 ± 0.4	6.1 ± 0.0	5.5 ± 0.0	6.1 ± 0.1
	At5g01210	6.7 ± 0.0	6.4 ± 0.1	6.7 ± 0.0	8.6 ± 0.0	7.9 ± 0.0	8.1 ± 0.0	8.6 ± 0.0	9.0 ± 0.0	8.4 ± 0.1	7.5 ± 0.0	7.4 ± 0.1	7.7 ± 0.0
CESA3	At5g05170	10.1 ± 0.2	10.1 ± 0.1	9.7 ± 0.1	10.3 ± 0.0	10.2 ± 0.0	10.2 ± 0.0	9.8 ± 0.0	10.1 ± 0.0	10.1 ± 0.0	9.4 ± 0.0	9.1 ± 0.0	9.2 ± 0.0
	At5g07830	6.6 ± 0.1	6.6 ± 0.0	6.3 ± 0.0	6.1 ± 0.1	6.1 ± 0.0	6.1 ± 0.0	6.4 ± 0.1	6.1 ± 0.0	6.7 ± 0.0	7.5 ± 0.0	6.5 ± 0.1	7.0 ± 0.1
TUB6	At5g12250	9.3 ± 0.0	8.9 ± 0.0	8.7 ± 0.0	9.1 ± 0.0	9.4 ± 0.0	9.1 ± 0.0	8.7 ± 0.1	8.8 ± 0.0	8.9 ± 0.0	9.5 ± 0.0	8.8 ± 0.0	8.5 ± 0.1
TUA3	At5g19770	7.3 ± 0.0	7.0 ± 0.0	7.1 ± 0.0	7.0 ± 0.1	7.6 ± 0.1	7.0 ± 0.0	6.8 ± 0.0	7.2 ± 0.0	7.3 ± 0.0	8.0 ± 0.0	7.3 ± 0.0	7.1 ± 0.0
DIN10	At5g20250	7.2 ± 0.1	6.7 ± 0.0	6.0 ± 0.0	9.2 ± 0.1	8.3 ± 0.0	9.5 ± 0.1	8.4 ± 0.0	9.3 ± 0.0	9.8 ± 0.0	7.6 ± 0.0	9.2 ± 0.0	7.1 ± 0.1
	At5g26665	5.8 ± 0.0	6.1 ± 0.0	5.4 ± 0.1	5.3 ± 0.0	5.3 ± 0.0	5.0 ± 0.0	5.4 ± 0.1	4.6 ± 0.0	4.3 ± 0.0	5.8 ± 0.0	5.5 ± 0.0	5.5 ± 0.1
	At5g43760	7.2 ± 0.0	6.4 ± 0.0	6.3 ± 0.0	7.4 ± 0.0	7.4 ± 0.1	7.3 ± 0.0	6.5 ± 0.0	7.0 ± 0.0	7.1 ± 0.0	6.7 ± 0.0	7.2 ± 0.0	7.3 ± 0.1
FLA1	At5g55730	7.2 ± 0.0	6.9 ± 0.0	7.2 ± 0.0	7.2 ± 0.0	6.8 ± 0.0	6.9 ± 0.0	7.2 ± 0.0	7.2 ± 0.0	6.5 ± 0.0	7.4 ± 0.0	7.0 ± 0.1	6.6 ± 0.0
CESA6	At5g64740	8.6 ± 0.0	8.5 ± 0.1	7.7 ± 0.0	8.1 ± 0.0	8.4 ± 0.0	8.4 ± 0.0	7.9 ± 0.1	7.8 ± 0.0	8.1 ± 0.0	8.8 ± 0.0	8.5 ± 0.0	9.3 ± 0.0
CYCD3.2	At5g67260	7.1 ± 0.0	7.0 ± 0.0	6.9 ± 0.1	6.8 ± 0.0	7.0 ± 0.0	5.7 ± 0.0	6.1 ± 0.0	6.1 ± 0.0	5.4 ± 0.0	6.5 ± 0.0	7.0 ± 0.0	6.6 ± 0.0

Transcript co-responses point at additional BR functions

Transcript co-responses with *BRI1* and *BAK1* point to the molecular basis of additional BR-effects. The positive impact of BR on plant growth upon cold stress and salt stress has been reported (Kamuro and Takatsuto, 1999). The *ERD4* (*EARLY RESPONSIVE TO DEHYDRATION 4*), *ERD6*, *ERD15*, and At1g29470 (similar to *ERD3*) showed transcript co-responses with *BRI1* and *BAK1* (supplementary EXCEL file sheet 1 [301 genes]). In addition, transcript co-responses were detected for genes involved in the auxin response (*IAA7*, *IAA14*, *IAA16*, *TIR1* homologs [At3g26810 and At4g03190], and *GRH1* [At4g03190]), auxin transport (*PIN3* and *AUX1*), and ethylene response (*EIL1*, *AtER6* [At3g11930], *ERF7*, and *EIN2*), providing further evidence for phytohormone interactions. The expression of the *IAA14* and *AtER6* (At3g11930) genes was previously reported to be weaker in roots of the *dwf1-6* mutant in comparison to roots of wild-type plants (Müssig et al., 2003). Several transcription factors showed co-responses, including *BEE1*, *GTL1*, and *MYC1* which have been shown to be BR-induced (Friedrichsen et al., 2002; Müssig et al., 2002). The co-responses of genes encoding ubiquitin-conjugating enzymes (At1g63800 and At1g64230), the ubiquitin-ligase RMA1, and F-box proteins (At1g30200, At1g67480, At2g18280, At3g06380, At3g61060, At5g27920, and At5g60570) indicate a role of BR in protein degradation. In fact, protein levels of positive mediators of BR responses such as BZR1 and BES1 depend on the presence of BR (He et al., 2002; Wang et al., 2002; Yin et al., 2002).

Discussion

Transcript co-response analysis reveals BR-related genes

Genomic BR-effects depend on several parameters. A major determinant is the genotype. Different BR-mutants show different transcript patterns. BR-treatment of BR-deficient plants results in largely different changes of transcript patterns in comparison to BR-treated wild-type plants. BR-responses depend on the developmental stage, environmental conditions, and tissue (Müssig et al., 2002; Müssig et al., 2003). Transcript co-response analysis with various expression profiles allows ruling out outliers and conditional changes of transcript levels, because the matrix combines different genotypes, environmental conditions, and other factors.

The observed transcript co-responses partly differ for components of the BR-signalling pathway. 42 % (301 of 720) of genes correlating with *BAK1* also showed a co-response with *BRI1*. In case of *BRI1* 26 % (301 of 1179) of co-responding genes also showed a co-response with *BAK1*. This discrepancy could reflect differences in the signalling crosstalks in which both proteins are involved. *BRI1* is an indispensable BR-receptor component, whereas a null allele of *BAK1* results in reduced (but not abolished) sensitivity to BR (Li et al., 2002). The BRI1 protein might interact with other proteins (i.e. BAK1 homologs). BRI1 could also bind another ligand, since tomato BRI1 perceives both BR and the

peptide hormone systemin (Montoya et al., 2002; Scheer and Ryan, 2002; Wang and He, 2004). Conversely, BAK1 could interact with proteins different from BRI1. Thus, use of more than one BR-signalling component for co-response analyses is essential in order to exclude correlations not related to BR action.

301 genes showed co-responses with both *BRI1* and *BAK1*. More than 50 % of them could be functionally classified (Fig. 2 and supplementary information). The robust estimation of co-responses of 135 functionally classified genes (including *BRI1* and *BAK1*) was confirmed by means of bootstrap analyses (supplementary EXCEL file sheet 2 [Spearman r_s], sheet 3 [pvalue], and sheet 4 [power]). A subset of genes (such as *AGP4*, *BEE1*, *GTL1*, *KCS1*, *PRO1*, and *TIP2.1*) was already described as BR-induced and validates the relevance of co-response analysis. However, the majority of identified genes were hitherto not regarded as BR-related.

Analysis of 301 co-responding genes in Affymetrix expression profiles

A large set of publicly available Affymetrix expression profiles was used to address the question whether the expression of identified candidate genes are BR-dependent and not regulated by GA. 21 % of the genes (62 of 301) turned out as BR-induced in at least two experiments (criteria: Present in experimental hybridisation, signal log ratio ≥0.8, change P value <0.01; supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]). In contrast, only 3 genes (1 %) were affected by altered GA-levels (criteria as aforementioned; supplementary EXCEL file sheet 8 [Affx results GA3, PAC]). Thus, the identified genes are not regulated by GA or GA-mediated growth processes.

In-depth analysis of 44 growth-related genes by means of real-time RT-PCR

Since the observed transcript co-responses were not a consequence of altered growth but rather reflected BR-dependent effects we proceeded to analyze the molecular basis of BR-promoted growth. BR promotes growth in all plant organs at early developmental stages. BR promotes both cell division (Hu et al., 2000; Oh and Clouse, 1998) and cell elongation (Asami et al., 2000; Kauschmann et al., 1996). There is increasing evidence for the involvement of a multitude of molecular mechanisms, e.g. cell wall loosening (Wang et al., 1993), acidification of wall space (Cerana et al., 1983), carbohydrate allocation (Goetz et al., 2000), carbon assimilation (Schlüter et al., 2002), and control of aquaporin activity (Morillon et al., 2001). BR apparently coordinates and integrates diverse processes required for growth. Transcript levels of 44 growth-related genes were analyzed in 4 BR-mutants. Correlated behaviour of these genes in the nasc0271 data matrix is shown in Fig. 3. 23 genes (52 %) showed weaker expression in BR-mutants in comparison to wild-type plants and partly also showed increased mRNA levels after BR application (Table 2). 13 (of 23) genes were also BR-induced in at least 2 microarray experiments (supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]). These genes in particular point at BR-effects on cell wall metabolism and water transport across membranes. The low fold changes comply with previous findings, which consistently showed subtle BR effects on

transcript levels (Goda et al., 2002; Müssig et al., 2002). The Affymetrix data along with the real-time RT-PCR data demonstrated a positive BR-effect on mRNA levels of 72 genes.

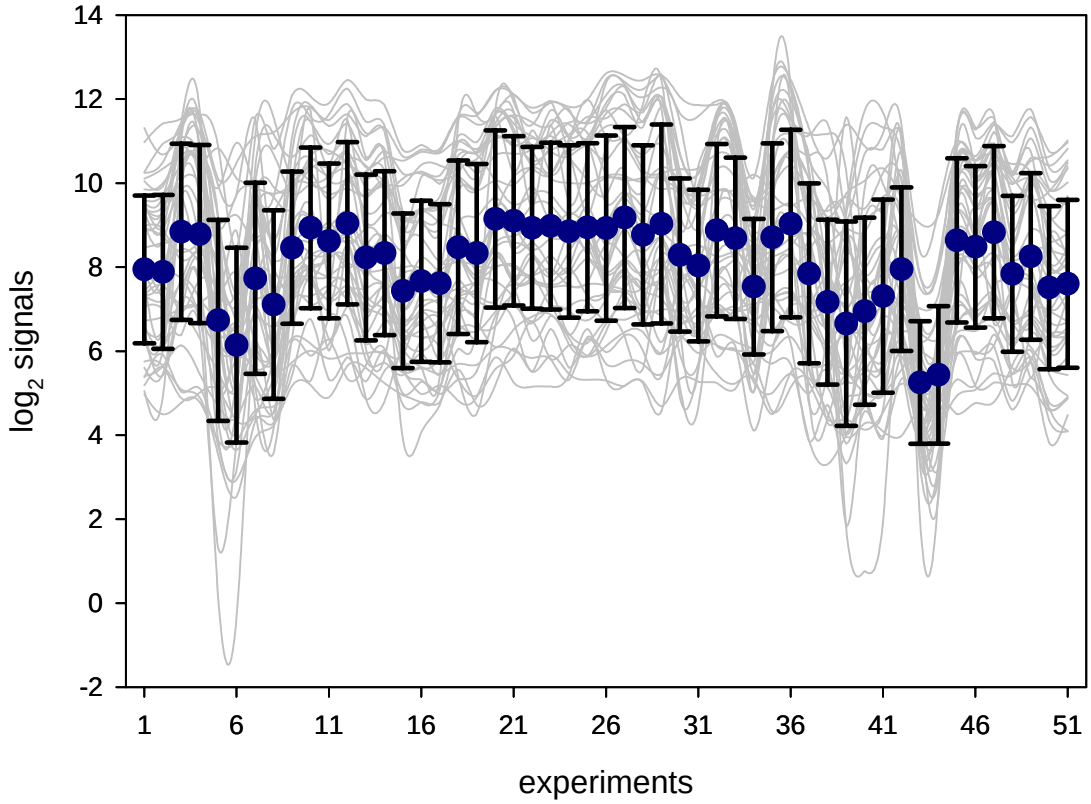

Fig. 3. Transcript levels of 44 cell wall and growth-related genes denoted as log2 signals in 51 expression profiles (incorporated in CSB.DB data matrix nasc0271). Dots and bars give average log2 signals of all genes and standard deviations, respectively. Lines are drawn to aid interpretation.

A subset of positive co-responding genes points at antagonistic signalling events

The statistical approach was based on a wide range of experiments but did not include BR-mutants or BR-treated plants. One single group of 301 genes was identified which showed positively correlated expression (supplementary EXCEL file sheet 1 [301 genes]). Experimental evaluation using BR-mutants, BR- and BRZ-treated plants divided this homogenous group into 3 groups: BR-induced genes (in total different 72 genes [24 %]), BR-repressed genes (in total 16 genes [5 %]), and genes which were not affected by altered BR-levels or showed variable expression (213 genes [71 %]) (Table 2, Table 3, supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]).

Weaker expression in plants upon BR-treatment and stronger expression in BR-mutants appears to contradict BR-induction inferred from positive transcript co-responses with *BRI1* and *BAK1*. However, it could indicate antagonistic signalling events. Antagonistic regulation could be switched off in the

presence of physiological BR-levels (i.e. in the experiments used for CSB.DB). Upon BR-treatment antagonistic signalling events may occur which serve to limit responses and counter toxic BR-effects. Remarkably, expression of *BRI1* and a *BRI1* homolog (At1g72180) is also repressed upon BR-treatment (supplementary EXCEL file sheet 7 [Affx results BRs, BRZ]). However, the BRI1 protein is essential for BR-responses. Thus, BR-repressed genes at least partly represent components required for BR-responses, and repression in the presence of elevated BR-levels does not exclude their function as positive mediators of BR-responses.

213 genes did not show significantly altered expression or transcript levels varied. Expression of these genes is not strictly BR-dependent. Positive transcript co-responses with *BRI1* and *BAK1* may reflect a context dependent functional environment. Interestingly, several genes showing high sequence similarity to BR-signalling components also showed co-responses but did not display significant changes of mRNA levels in the Affymetrix and RT-PCR experiments (genes similar to *BRI1*: At1g28440, At1g72180, and At2g01950; BAK1: At1g07650 and At3g14840; *BRZ1/BES1*: At1g78700). However, pseudo-correlations cannot be excluded.

CSB.DB: a valuable public resource to uncover genomic effects

In this study, we focused on the analysis of genes showing positive transcript co-responses. Numerous genes were identified, which point towards previously known BR-effects, such as promotion of growth and stress resistance, and expected phytohormone interactions. In addition BR appears to affect processes such as protein degradation. Besides investigating positive co-responses CSB.DB also allows screening for genes which display negative correlation coefficients. Following the presented approach 404 genes would have had negative transcript co-responses with both *BRI1* and *BAK1* (r_s <[-0.35], p-value <0.01, power >0.7). For the sake of clarity these genes were not reported in the current study. These 404 genes point towards novel BR actions. For instance, several genes encode transcriptional regulators or proteins involved in chromatin remodelling. Future work will not only address the analysis and experimental verification of negative co-responses, but will also include the analysis of other data matrices. We currently construct subsets of expression experiments which will allow screening for co-responses under specific experimental conditions, for example temperature stress or comparison between root and leaf thus facilitating the identification of tissue specific effects. These data matrices will be made publicly available through CSB.DB.

Material and Methods

Co-response analyses

Transcript co-responses were retrieved from the CSB.DB (a comprehensive systems-biology database; http://csbdb.mpimp-golm.mpg.de; Steinhauser et al., 2004) for data matrix nasc0271. 51 publicly available expression profiles were obtained from NASCarrays (http://affymetrix.arabidopsis.info/; Craigon et al., 2004) and used for the generation of the data matrix. The profiles were originated through the 22k Affymetrix Chip technology (Affymetrix, La Jolla, CA). The data matrix was established in October 2003. At that time 123 expression profiles from 22 experiments were available. The number of Present and Marginal calls (according to the MAS 5.0 algorithm) was calculated for each profile. Profiles from 21 experiments were included into the database. In the majority of cases 2 or 3 profiles per experiment with the highest numbers of Present and Marginal calls were selected. Thus nasc0271 covers 9694 genes out of > 22.000 genes represented on the ATH1 array, which are well measured in at least 85% of the underlying expression profiles. Co-responses were obtained via the intersection gene query (isGQ) tool of CSB.DB which allows the extraction of those genes exhibiting common correlation to *BAK1* and *BRI1*. The correlations are based on non-parametric Spearman's rank-order correlation (r_s). For both genes positive transcriptional co-responses with an uncorrected probability for multiple comparisons of <0.01 (99%) and a power of test of >0.7 (70%) were retrieved. Out of the initial result (which covers 301 genes) 135 genes were selected and used for statistical in-depth analyses. To test for influence of individual or small subsets of underlying expression profiles transcriptional co-responses were confirmed with bootstrap analyses. The bootstrap spearman correlation, the probability, the confidence interval and the power of the test were re-computed with the statistical software environment R (http://www.r-project.org) by non-parametric bootstrap analyses with 2,000 numbers of bootstrap samples (Efron and Tibshirani, 1993) on the log base 2 and range normalised signal intensities.

Distribution of functional categories

The manually curated assignments into bins were used from the MapMan software (Thimm et al., 2004). The obtained assignments of genes into bins were slightly modified for those bins, which cover few genes. The merging of bins and the resulting descriptions is illustrated in Table SI (see supplementary information). Genes without assignment or with unclear classification were termed 'unclassified'. The relative impact (ri) of a gene with multiple assignments (n_{assign}) onto each category was defined as: $ri = 1 / n_{assign}$. The distribution of retrieved genes into functional categories was computed by adding-up the relative assignment coefficient for each gene per category.

Plant material and growth conditions

Two growth conditions were applied. Firstly, *Arabidopsis thaliana* cv. C24 (wild-type), *dwf1-6* (*cbb1*), *cbb2*, and *cbb3* mutants (Kauschmann et al., 1996) were grown in half-concentrated MS medium supplemented with 1 % sucrose and solidified with 0.7 % agar under a 16 h day (140 µE, 22°C)/8 h (22°C) night regime. Plants were harvested 14 d or 19 d after sowing. Roots were discarded. Secondly, Arabidopsis thaliana cv. C24, the *dwf1-6* mutant (Kauschmann et al., 1996), and *CPD*-antisense plants (Schlüter et al., 2002) were grown in soil under long day conditions (16 h fluorescent light, 180 µE, 20°C, 70% relative humidity/8 h dark, 16°C, 75% relative humidity). Above ground organs were harvested 28 d after sowing. BR-treatments were conducted as described (Müssig et al., 2002).

Real-time RT-PCR analysis

Total RNA was isolated with the Invisorb Spin Plant RNA kit (Invitek, Berlin, Germany). One µg of total RNA was then reverse-transcribed with the Superscript II reverse transcriptase (Invitrogen) in a reaction volume of 28.5 µl to generate first-strand cDNA. Real time RT-PCR was performed with 1 µl of a 1:3.5 dilution of the first-strand cDNA reaction and the SYBR Green reagent (Applied Biosystems, Foster City, CA) in a 25 µl volume on a Perkin Elmer Geneamp 5700 machine. Primer sequences were given in Table 1. Data were normalized to the *eIF1a* gene (At5g60390) and then compared according to the formula (considering as example the *KCS1* gene):

$$nC_T = C_{T\ KCS1} - CT_{\ eIF1a}$$

$$\Delta C_T = \text{signal log ratio} = nC_{T\ mutant} - nC_{T\ WT}$$

Amplification efficiency (*E*) was checked for all primer pairs (Czechowski et al., 2004). In short, the *E*-values were derived from the log slope of the fluorescence versus cycle number curve for a particular primer pair, using the equation $(1 + E) = 10^{slope}$. The *E*-values for all primer pairs are summarized in Table 1 and used to calculate normalised fold change values, using the equation $(1 + E)^{\Delta CT}$. Control experiments showed that the use of different control genes (either *eIF1α* or *eIF4A1* [At3g13920 primers: ACAATGTGGTTGTCGAAGAGCTG and GCAGAGCAAACACAGCAAC AGAA]) did not bias the results with respect to the signal log ratios.

Analysis of Affymetrix expression profiles

Expression analysis was performed with the MAS 5.0 and GCOS software (Affymetrix). The output of every experiment was multiplied by a scaling factor to adjust its average intensity to a target intensity of 100. Results of Absolute and Comparison expression analysis were imported into MS Access2003 and screened for significant changes according to the criteria mentioned in the text. Table 5 specifies the AtGenExpress CEL files used for this study. Plant material was grown in liquid MS medium for 7 d at 23°C prior to the BL-, CS-, BRZ-, GA- and PAC-treatments. Two own profiles were used (7 h EBL and control treatment of wild-type [Col-0] plants grown in half-concentrated MS medium supplemented with 1 % sucrose and solidifife with 0.7 % agar; Coll-Garcia et al., 2004).

Table 5. CEL files from AtGenExpress used in this study.

Experiment	Genotype	Experiment	Baseline
10 nM BL	Wild-type (Col-0)		
30 min		BL30-2 and BL30-3	mock 30-2 and mock 30-3
1 h		BL1-2 and BL1-3	mock 1-2 and mock 1-3
3 h		BL3-2 and BL3-3	mock 3-2 and mock 3-3
10 nM BL	*det2-1*		
30 min		det2BL30-1 and det2BL30-2	det2m30-1 and det2m30-2
1 h		det2BL1-1 and det2BL1-2	det2m1-1 and det2m1-2
3 h		det2BL3-1 and det2BL3-2	det2m3-1 and det2m3-2
1 µM GA3	Wild-type (Col-0)		
30 min		GA3 30-2 and GA3 30-3	mock 30-2 and mock 30-3
1 h		GA3 1-2 and GA3 1-3	mock 1-2 and mock 1-3
3 h		GA3 3-2 and GA3 3-3	mock 3-2 and mock 3-3
1 µM GA3	*ga1-5*		
30 min		GA1-5G30-1 and GA1-5G30-2	GA1-5m30-1 and GA1-5m30-2
1 h		GA1-5G1-1 and GA1-5G1-2	GA1-5m1-1 and GA1-5m1-2
3 h		GA1-5G3-1 and GA1-5G3-2	GA1-5m3-1 and GA1-5m3-2
100 nM CS	*det2-1*	CS3h-1 and CS3h-2	m3h-1 and m3h-2
10 µM BRZ220	Wild-type (Col-0)		
3 h		mock 3h-1 and mock 3h-2	2203h-1 and 2203h
12 h		mock12h-1 and mock 12h-2	220 12h-1 and 220 12h-2
10 µM BRZ91	Wild-type (Col-0)		
3 h		mock 3h-1 and mock 3h-2	91 3h-1 and 91 3h-2
12 h		mock12h-1 and mock 12h-2	91 12h-1 91 12h-2
10 µM PAC	Wild-type (Col-0)		
3 h		mock 3h-1 and mock 3h-2	pac 3h-1 and pac 3h-2
12 h		mock12h-1 and mock 12h-2	pac 12h-1 and pac 12h-2

Acknowledgement

We acknowledge the AtGenExpress consortium (Lutz Nover, Thomas Altmann, Detlef Weigel) and Hideki Goda for supply of expression profiles. We thank Björn Usadel for assisting us in statistical analyses. We acknowledge the NASCArrays and all scientists who submitted transcript profiles. We are grateful to Tomasz Czechowski and Michael Udvardi for relaying details with regard to real-time RT-PCR analyses.

References

Asami,T., Min,Y.K., Nagata,N., Yamagishi,K., Takatsuto,S., Fujioka,S., Murofushi,N., Yamaguchi,I. and Yoshida,S. (2000) Characterization of brassinazole, a triazole-type brassinosteroid biosynthesis inhibitor. *Plant. Physiol.*, **123**, 93-99.

Bishop,G.J. (2003) Brassinosteroid mutants of crops. *J. Plant Growth Regul.*, **22**, 325-335.

Cerana,R., Bonetti,A., Marre,M.T., Romani,G., Lado,P. and Marre,E. (1983) Effects of a brassinosteroid on growth and electrogenic proton extrusion in Azuki bean epicotyls. *Physiol. Plant.*, **59**, 23-27.

Choe,S., Fujioka,S., Noguchi,T., Takatsuto,S., Yoshida,S. and Feldmann,K.A. (2001) Overexpression of *DWARF4* in the brassinosteroid biosynthetic pathway results in increased vegetative growth and seed yield in *Arabidopsis*. *Plant J.*, **26**, 573-582.

Clouse,S.D., Langford,M. and McMorris,T.C. (1996) A brassinosteroid-insensitive mutant in *Arabidopsis thaliana* exhibits multiple defects in growth and development. *Plant Physiol.*, **111**, 671-678.

Coll-Garcia,D., Mazuch,J., Altmann,T. and Müssig,C. (2004) EXORDIUM regulates brassinosteroid-responsive genes. *FEBS Let.*, **563**, 82-86.

Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acid Res.*, **32**, D575-D577.

Czechowski,T., Bari,R.P., Stitt,M., Scheible,W.-R. and Udvardi,M.K. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.*, **38**, 366-379.

Efron, B. and Tibshirani, R. (1993) An introduction to the Bootstrap. Chapman and Hall, New York, London.

Friedrichsen,D.M., Nemhauser,J., Muramitsu,T., Maloof,J.N., Alonso,J., Ecker,J.R., Furuya,M. and Chory,J. (2002) Three redundant brassinosteroid early response genes encode putative bHLH transcription factors required for normal growth. *Genetics*, **162**, 1445-1456.

Goda,H., Shimada,Y., Asami,T., Fujioka,S. and Yoshida,S. (2002) Microarray analysis of brassinosteroid-regulated genes in *Arabidopsis*. *Plant Physiol.*, **130**, 1319-1334.

Goetz,M., Godt,D.E. and Roitsch,T. (2000) Tissue-specific induction of the mRNA for an extracellular invertase isoenzyme of tomato by brassinosteroids suggests a role for steroid hormones in assimilate partitioning. *Plant J.*, **22**, 515-522.

He,J.-X, Gendron,J.M., Yang,Y., Li,J. and Wang,Z.-Y. (2002) The GSK3-like kinase BIN2 phosphorylates and destabilizes BZR1, a positive regulator of the brassinosteroid signaling pathway in Arabidopsis. *Proc. Natl. Acad. Sci. USA*, **99**, 10185-10190.

Hu,Y., Bao,F. and Li,J. (2000) Promotive effect of brassinosteroids on cell division involves a distinct CycD3-induction pathway in *Arabidopsis*. *Plant J.*, **24**, 693-701.

Kamuro,Y. and Takatsuto,S. (1999) Practical application of brassinosteroids in agricultural fields. In *Brassinosteroids: Steriodal Plant Hormones* (Sakurai, A., Yokota, T. and Clouse, S.D., eds). Tokyo: Springer Verlag, pp. 223-241.

Kauschmann,A., Jessop,A., Koncz,C., Szekeres,M., Willmitzer,L. and Altmann,T. (1996) Genetic evidence for an essential role of brassinosteroids in plant development. *Plant J.*, **9**, 701-713.

Khripach,V., Zhabinskii,V. and De Groot,A. (2000) Twenty years of brassinosteroids: Steroidal plant hormones warrant better crops for the XXI century. *Ann. Bot.*, **86**, 441-447.

Li,J. and Chory,J. (1997) A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell*, **90**, 929-938.

Li,J., Wen,J.Q., Lease,K.A., Doke,J.T., Tax,F.E. and Walker,J.C. (2002) BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell*, **110**, 213-222.

Montoya,T., Nomura,T., Farrar,K., Kaneta,T., Yokota,T. and Bishop,G.J. (2002) Cloning the tomato Curl3 gene highlights the putative dual role of the leucine-rich repeat receptor kinase tBRI1/SR160 in plant steroid hormone and peptide hormone signaling. *Plant Cell*, **14**, 3163-3176.

Mora-Garcia,S., Vert,G., Yin,Y.H., Cano-Delgado,A., Cheong,H. and Chory,J. (2004) Nuclear protein phosphatases with Kelch-repeat domains modulate the response to brassinosteroids in *Arabidopsis*. *Genes Dev.*, **18**, 448-460.

Morillon,R., Catterou,M., Sangwan,R.S., Sangwan,B.S. and Lassalles,J.P. (2001) Brassinolide may control aquaporin activities in *Arabidopsis thaliana*. *Planta*, **212**, 199-204.

Müssig,C., Fischer,S. and Altmann,T. (2002) Brassinosteroid-regulated gene expression. *Plant Physiol.*, **129**, 1241-1251.

Müssig,C., Shin,G.H. and Altmann,T. (2003) Brassinosteroids promote root growth in *Arabidopsis*. *Plant Physiol.*, **133**, 1261-1271.

Nakashita,H., Yasuda,M., Nitta,T., Asami,T., Fujioka,S., Arai,Y., Sekimata,K., Takatsuto,S., Yamaguchi,I. and Yoshida,S. (2003) Brassinosteroid functions in a broad range of disease resistance in tobacco and rice. *Plant J.*, **33**, 887-898.

Nam,K.H. and Li,J.M. (2002) BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. *Cell*, **110**, 203-212.

Oh,M.H. and Clouse,S.D. (1998) Brassinolide affects the rate of cell division in isolated leaf protoplasts of petunia hybrida. *Plant Cell Rep.*, **17**, 921-924.

Sasse,J. (1999) Physiological actions of brassinosteroids. In *Brassinosteroids: Steriodal Plant Hormones* (Sakurai, A., Yokota, T. and Clouse, S.D., eds). Tokyo: Springer Verlag, pp. 137-161.

Scheer,J.M. and Ryan,C.A. (2002) The systemin receptor SR160 from *Lycopersicon peruvianum* is a member of the LRR receptor kinase family. *Proc. Natl. Acad. Sci. USA*, **99**, 9585-9590.

Schlüter,U., Köpke,D., Altmann,T. and Müssig,C. (2002) Analysis of carbohydrate metabolism of *CPD* antisense plants and the brassinosteroid-deficient *cbb1* mutant. *Plant Cell Environ.*, **25**. 783-791.

Steinhauser,D., Usadel,B., Lüdemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, (in press).

Thimm,O., Bläsing,O., Gibon,Y., Nagel,A., Meyer,S., Krüger,P., Selbig,J., Müller,L.A., Seung,Y.R. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914-939.

Turk,E.M., Fujioka,S., Seto,H., Shimada,Y., Takatsuto,S., Yoshida,S., Denzel,M.A., Torres,Q.I. and Neff,M.M. (2003) CYP72B1 inactivates brassinosteroid hormones: an intersection between photomorphogenesis and plant steroid signal transduction. *Plant Physiol.*, **133**, 1643-1653.

Wang,T.-W., Cosgrove,D.J. and Arteca,R.N. (1993) Brassinosteroid stimulation of hypocotyl elongation and wall relaxation in pakchoi (*Brassica chinensis* cv Lei-Choi). *Plant Physiol.*, **101**, 965-968.

Wang,Z.Y. and He,J.X. (2004) Brassinosteroid signal transduction - choices of signals and receptors. *Trends Plant Sci.*, **9**, 91-96.

Wang,Z.-Y., Nakano,T., Gendron,J., He,J., Chen,M., Vafeados,D., Yang,Y., Fujioka,S., Yoshida,S., Asami,T. and Chory,J. (2002) Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis. *Dev. Cell*, **2**, 505-513.

Yin,Y.H., Wang,Z.Y., Mora-Garcia,S., Li,J.M., Yoshida,S., Asami,T. and Chory,J. (2002) BES1 accumulates in the nucleus in response to brassinosteroids to regulate gene expression and promote stem elongation. *Cell*, **109**, 181-191.

Chapter V - Application to *A.thaliana*: - Inferring Hypotheses For Gene Functions: The *Arabidopsis thaliana* Subtilase Gene Family -

Carsten Rautengarten[b*], Dirk Steinhauser[a*], Annick Stintzi[c], Dirk Büssis[b], Andreas Schaller[c], Joachim Kopka[a] and Thomas Altmann[†b]

[a]Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

[b]Universität Potsdam, Institut für Biochemie und Biologie, Genetik, Karl-Liebknecht-Str. 24-25, Haus 26, 14476 Golm, Germany

[c]Universität Hohenheim, Institut für Physiologie und Biotechnologie der Pflanzen, Emil-Wolff-Str.25, 70593 Stuttgart, Germany

Abstract

The gene family of subtilisin-like serine proteases (subtilases, AtSBTs) in *Arabidopsis thaliana* comprises 56 members, which were identified by homology and motif searches. Hence, it is among the largest multigene protease family in *Arabidopsis*, divided into six distinct subfamilies. Whereas five families are similar to pyrolysins, two genes share stronger homology to animal kexins. 31 (53%) of the subtilase genes are organized in tandem clusters, 18 (32%) are located in segmental duplicated genomic regions. Mutant screens confirmed 144 T-DNA insertion lines with knockouts for 55 out of the 56 subtilases. Apart from *sdd1*, none of the confirmed homozygous mutants revealed any obvious visible phenotypic alteration during growth under standard conditions. Computational analyses based on transcriptional co-expression and co-response pattern revealed at least two expression networks of subtilases. Network analyses suggest that functional redundancy may exist among subtilases by less homolog genes. Furthermore, two hubs were identified, which may be involved in signalling or represent higher order regulatory factors involved in responses to environmental cues. A particular enrichment of co-regulated genes with metabolic functions was observed for four subtilases possibly representing late responsive elements of environmental stress. The kexin homologs show stronger associations with genes of transcriptional regulation context. Based on the analyses presented here and in accordance with previously characterized subtilases, we propose three main functions of subtilases: Involvement in (I) control of development, (II) protein turnover, and action as (III) downstream components of signalling cascades.

[*] The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[†] To whom correspondence should be addressed.

Supplemental material is available in the Plant Subtilase Database (http://csbdb.mpimp-golm.mpg.de/psdb.html) as well as from the CSB.DB Homepage (http://csbdb.mpimp-golm.mpg.de/).

Introduction

Subtilisin-like proteases (subtilases), first charcterized in *Bacillus licheniformis*, are serine proteases with a catalytic triad of the three amino acids aspartate, histidine, and serine (Dodson and Wlodawer, 1998). Eukaryotic subtilases belong to the S8 serine protease family (http://merops.sanger.ac.uk) and can be grouped into the pyrolysins and the kexins. Nine subtilases, the pro-protein convertases, have been characterized in mammals. Of these seven belong to the kexin subfamily (Barr, 1991; Seidah and Chretien, 1999). More recently, two mammalian subtilases were identified within the pyrolysin subfamily. They carry out specific cleavage and processing reactions on sterol regulatory elements, binding proteins and pro-brain-derived neurotrophic factors, respectively (Sakai et al., 1998; Seidah et al., 1999). Currently, all identified plant subtilases are grouped into the pyrolysin subfamily within the S8 serine protease family (Siezen and Leunissen, 1997).

Despite the recent advances, our current understanding of subtilase functions in plants is still very limited. Currently, there are evidences for involvement of subtilases in both, general protein turnover (Bogacheva, 1999) as well as highly specific regulation of plant development (Berger and Altmann, 2000). A few proteases have been purified from plant tissues and classified as subtilases based on their catalytic properties and primary structure. For instance, macluralisin (Rudenskaya et al., 1995), taraxalisin (Rudenskaya et al., 1998), plantagolisin (Bogacheva et al., 2001), subtilases from green malt (Terp et al., 2000; Fontanini and Jones, 2002), developing tung fruits (Dyer et al., 1999), bean (Popovic et al., 2002), soybean (Beilinson et al., 2002; Boyd et al., 2002), and *Arabidopsis* (Hamilton et al., 2003). Most of these enzymes are highly abundant and exhibit broad substrate specificity. Thus, a functional involvement in general protein turnover was forecasted for these abundant proteins (Bogacheva, 1999). Cucumisin, which constitutes up to 10% of the soluble proteins in melon fruit, represents the primary example for a degrading functional role of subtilases. Cucumisin was characterised extensively enzymatically and was the first subtilase to be cloned from any plant species (Kaneda and Tominaga, 1975; Yamagata et al., 1994). The tomato subtilase P69, a pathogenesis-related protein, is one of several subtilases which are specifically induced following pathogen infection (Vera and Conejero, 1988; Tornero et al., 1996a; Jordá et al., 1999). P69 processes a leucin-rich repeat cell wall protein in virus-infected tomato plants and thus is one of the very few plant subtilases for which an endogenous substrate has been identified (Tornero et al., 1996b). The direct consequences of this processing event for pathogenesis are still unknown. The P69 enzymes form a distinct subgroup among the 15 subtilases that have hitherto been cloned from tomato (Meichtry et al., 1999).

Forward genetics has identified subtilases as highly-specific regulators of plant development. In the *Arabidopsis sdd1* mutant (*stomatal density and distribution 1*) the pattern of stomata formation is disrupted, resulting in clustering of guard cells as well as in a dramatic increase of stomatal density (Berger and Altmann, 2000). The SDD1 protease is expressed in meristemoids and guard mother cells, the precursor cells of stomates. SDD1 is probably secreted into the apoplast of the cells, where it probably acts as a processing protease in the generation of signals responsible for stomata density and distribution (von Groll et al., 2002). Likewise, the gene disrupted in the *ale1* mutant (*abnormal leaf shape 1*) was cloned and found to encode a subtilase. ALE1 is required for cuticle formation and epidermal differentiation during embryo development in *Arabidopsis*. A role for ALE1 was suggested in the generation of peptide signals required for proper differentiation of the epidermis (Tanaka et al., 2001). The mutant phenotypes of *sdd1* and *ale1* demonstrate that at least some subtilases carry out highly specific functions in plant development. Their modes of action in the regulation of the respective developmental processes are still unknown but SDD1 and ALE1 may be required for the generation of (poly-)peptide signals, which act non-cell autonomously to control plant development. Computational prediction based on sequence homology led to the identification of other 53 (in all 55) *Arabidopsis thaliana* subtilases (Beers et al., 2003).

Despite this recent progress, there is still uncertainty about the functions of the majority of plant subtilases including those of the model organism *Arabidopsis*. To investigate the physiological and developmental processes involving subtilase functions in *Arabidopsis*, a multinational Arabidopsis Subtilase Consortium (TASC) was initiated between laboratories in Germany, Great Britain, Spain, and the US. It is the goal of TASC to characterize the function of all the *Arabidopsis* subtilases (AtSBTs) in a functional genomics program by using experimental and computational approaches.

Here we describe first results obtained with gene knock-out mutants and expression analysis. The main focus of our report is directed towards the initial computational analysis of the so far uncharacterized *A.thaliana* subtilase gene family. We have extended common classification of gene families by sequence similarity towards investigation into co-responding synchronous changes of transcript levels (co-response analyses). These analyses enabled us to infer hypotheses about the respective functional involvement of subtilases, which were described here and will be the attracting point for further experimental-driven characterization.

Result and Discussion

The goal of this work was to investigate in the initial functional characterization of the *Arabidopsis* subtilase family. Assigning a basic function to novel discovered genes can be approached from a multitude of different scientific perspectives and therefore can be implemented by a variety of technologies developed (Vukmirovic and Tilghman, 2000). A traditional approach based on pairwise or multiple sequence comparisons and alignments by various algorithms (Hodgman, 2000) which

allowed functional prediction for genes or gene products by annotation transfer from homologous sequences (McGeoch and Davidson, 1986; Bork and Gibson, 1996). We applied this approach to identify and classify *Arabidopsis* subtilase genes according sequence homology. However, initial attempts to annotation transfer gave us no strong clues for experimental approaches due to the lack of characterized candidate genes. Moreover, verified homozygote knock-out lines revealed no visible phenotype under standard growth condition and therefore, do not support basic functional assignment.

The *Arabidopsis thaliana* Subtilase Family comprises 56 Genes (AtSBTs)

Our initial effort to identify subtilases was based on sequence comparisons with well known and characterized *Arabidopsis* subtilase genes. Subtilases contain a catalytic triad (S8 domain) of the amino acid residues aspartate (Asp, D), histidine (His, H) and serine (Ser, S), as well as an asparagine (Asn, N) suggested as substrate binding site. Sequence comparisons to identify sequences homologous were performed against AGI proteins [TAIR, (Rhee et al., 2003)] using the Blast algorithm (Altschul et al., 1990) with the S8 domain of the SDD1 amino acid sequence. The identified sequences were evaluated for the presence of the conserved D-, H-, S- and N-regions and resulting in 56 genes that encodes for subtilases (http:// csbdb.mpimp-golm.mpg.de/psdb.html). From the entire members 55 genes contain all conserved motifs, while At5g45640 (*AtSBT5.5*) lacks the central Asp residue of the D-region. Hence, the subtilase family is among the largest multigene protease family known in *Arabidopsis*. Recently 55 subtilase genes were identified for *A.thaliana* (Beers et al., 2003). We identified one further gene containing the S8 domain, namely At4g20850 (*AtSBT6.2*).

Beyond the sequence homology, the subcellular targeting of a gene product allows hints for a functional involvement. Primary structure analysis using TargetP (Emanuelsson et al., 2000) indicated that 46 *Arabidopsis* subtilases possess a signal sequence for targeting to the secretory pathway (see http://csbdb.mpimp-golm.mpg.de/psdb.html). Six subtilases do not contain any known protein targeting motifs. Three genes are predicted to be targeted to mitochondria and one to chloroplasts. Experimental data for the subcellular localization of *Arabidopsis* subtilases are presently available only for SDD1 and for ARA12, which were both shown to be exported to the apoplast (von Groll et al., 2002, Hamilton et al., 2003) for generation of (poly-)peptide mediated signals.

The Arabidopsis thaliana Subtilase Family consist of 6 Subfamilies

Analyzing all 56 *Arabidopsis* subtilase sequences we wanted to investigate whether groups of genes are more strongly associated based on sequence homology and therefore could have overlapping or similar functions. We performed a multiple alignment with the deduced complete amino acid sequences by using ClustalX (Thompson et al., 1997). This analysis revealed six distinct subtilase subfamilies in *A.thaliana* (Fig.1). The assignment of a gene to a specific subfamily was based primarily on the position within the phylogenetic tree, as defined by the degree of homology between the deduced full length amino acid sequences. When a gene could not be assigned to a particular clade

with a significant bootstrap value, the assignment to a certain subfamily was made by ranking BLAST search results of queries for family members against the gene. Repeating the analysis by comparing only the conserved peptidase S8 domain we could confirm the assignments for all *A.thaliana* subtilase genes into these six subfamilies. The assignments were further supported by distance matrices obtained by pairwise global alignments of the nucleic acid and amino acid sequences (http://csbdb.mpimp-golm.mpg.de/psdb.html).

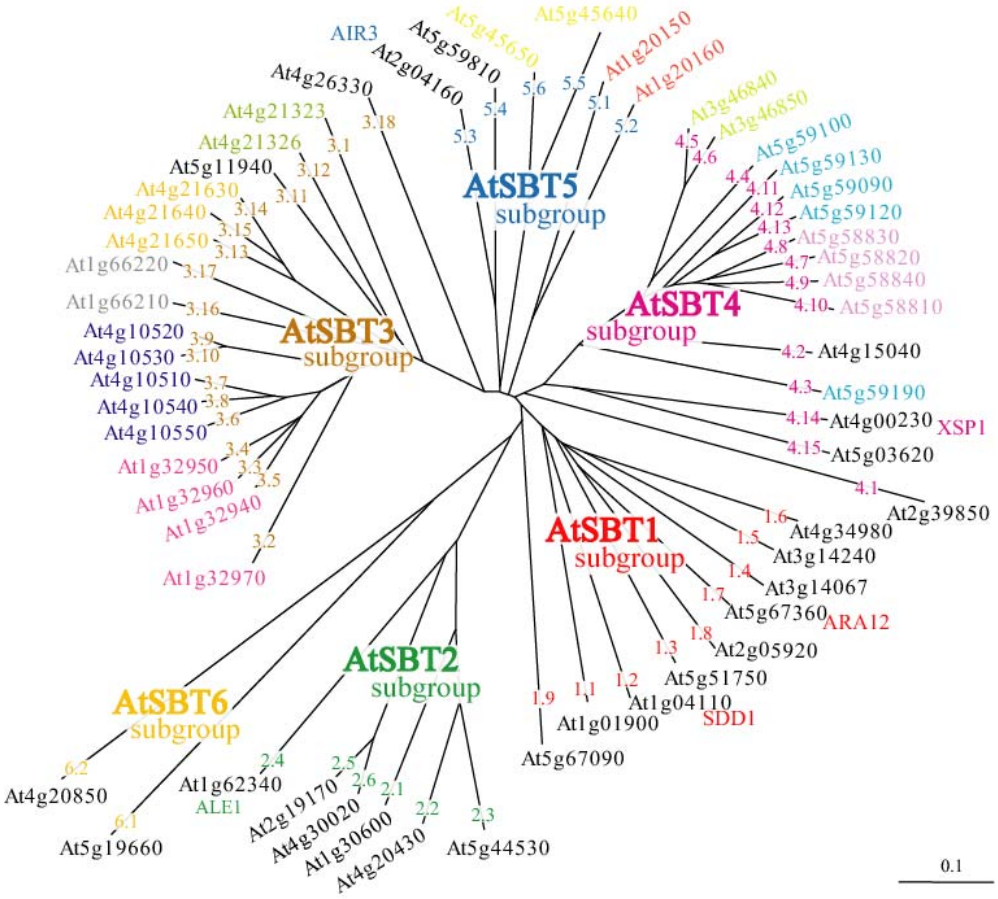

Fig. 1. Bootstraped neighbor-joining tree with 1,000 bootstrap replicates generated from an alignment of the predicted 56 AtSBT full-length protein sequences using ClustalX 1.81. The tree was displayed by TreeView and edited manually. Neighboring genes are distinguished by specific colors.

The protease associated domain (PA) is supposed to determine the substrate specificity of subtilases or to form protein-protein interactions (Mahon and Bateman, 2000; Luo and Hofmann, 2001). Most proteins of the subtilase family contain a sequence region of about 120 amino acids inserted into their catalytic domain. Therefore, to uncover similar substrate specificities within the *Arabidopsis* subtilase family, the PA domain was used for the assignment into subfamilies (Fig. 2). Apart from *AtSBT4.1*, *AtSBT 6.1*, and *AtSBT 6.2* all *Arabidopsis* subtilases contain an insertion, consistent with a PA domain. Apart from members of the heterogeneous subfamily 5, all subtilases were again assigned to the same subfamilies as before with only minor changes (compare Fig.1 and Fig.2).

The general consistency of phylogenetic trees derived from the full-length and the PA domain sequences suggests that the PA domain insertion was already present in the ancestral subtilases. These results are consistent with those reported by Beers et al. (2003), who performed sequence analysis of the *Arabidopsis* peptidase S8 serine, C1A cysteine and A1 aspartic protease families. The AtSBT1 and AtSBT2 subfamilies are identical with the S8-2 and S8-3 groups (Beers et al., 2003). The large heterogeneous S8-1 group, however, was subdivided further into AtSBT families 3, 4, and 5. The AtSBT6 subfamily includes just two members, i.e. AtSBT6.1, which had not been assigned to any group by Beers et al. (2003), and AtSBT6.2, a previously unrecognized *Arabidopsis* subtilase. Both genes are characterized by a stronger homology to (mammalian) kexins (http://csbdb.mpimp-golm.mpg.de/psdb.html). In yeast Kex2p, the first eukaryotic identified kexin, is required for the processing of the precursors of α-mating factor and of killer toxin (Fuller et al., 1988). In analogy to Kex2p kexin-like subtilases have been postulated in plants involved in killer toxin processing (Kinal et al., 1995). Mammalian kexin homologs have been identified required in formation of functional proteins from precursor polypeptides (Barr, 1991, Seidah and Chretien, 1999).

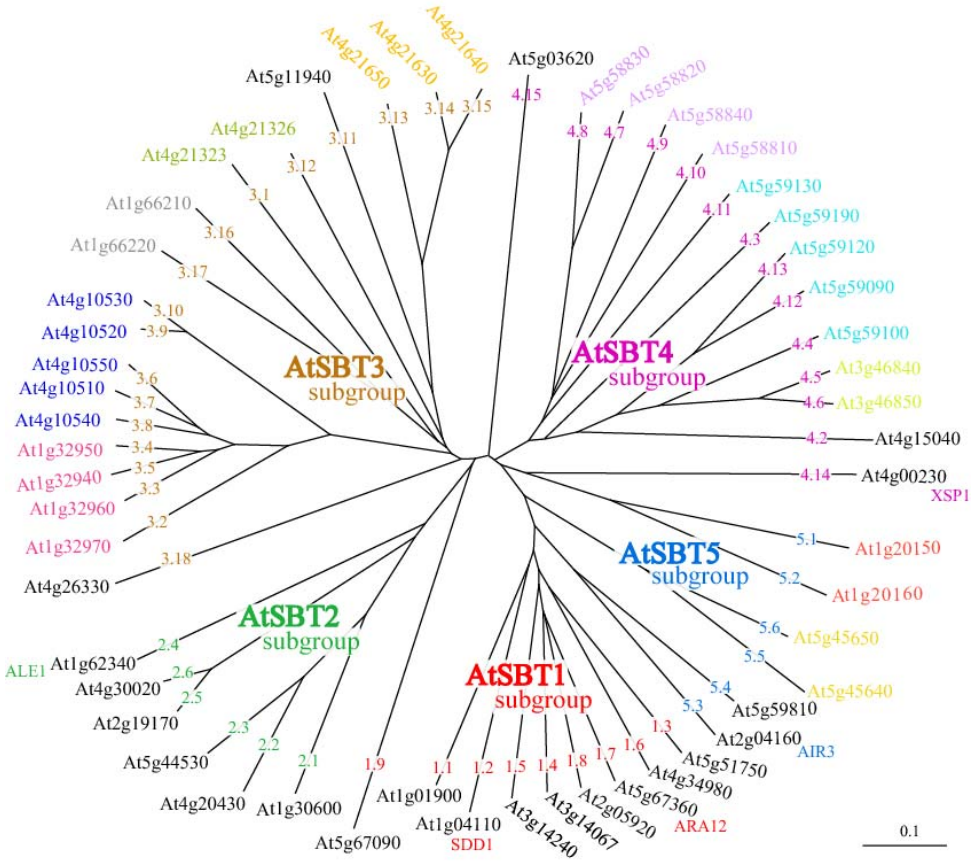

Fig. 2. Bootstraped neighbor-joining tree generated with 1,000 bootstrap replicates from an alignment of the predicted Protease-Associated (PA) domain using ClustalX 1.81. The tree was displayed by TreeView and edited manually. Notice that AtSBT6.1, 6.2 and 4.1 do not contain a PA domain. Neighboring genes are marked by specific colors.

Subtilase Families in Plants exceed Complexity in Animals

Phylogenetic relationships may help to unravel the basic functions of genes based on annotation transfer from orthologous sequences. Blast searches via the NCBI *O.sativa* BLAST page (Wheeler et al., 2004) by using the peptidase S8 domain and several full-length amino acid sequences of *Arabidopsis* revealed 34 non-redundant rice subtilase genes. A multiple alignment of the 14 known tomato (Meichtry, et al., 1999), the 34 identified rice, and the 56 *Arabidopsis* subtilases was performed to elucidate the phylogenetic relationships within the plant subtilase family. Four **m**ajor **c**lusters of **o**rthologous **g**roups (MCOG) were identified that includes all members of the AtSBT subfamilies 1-3 and 5, whereas *AtSBT4* seems to be a subfamily specific for *A.thaliana*. The obtained neighbour joining tree enabled us to identify putative orthologous gene pairs and groups of genes (see http:// csbdb.mpimp-golm.mpg.de/psdb.html). However, the lack of functionally characterized orthologs in the subtilase family among the three plants species gave us no strong hints for functional annotation. Interestingly, all three plant species are characterized by significant increase in the number of subtilases compared to animal organisms, e.g. human (9), *Caenorhabditis* (4) or *Drosophila* (3) obtained by blast search. This fact could suggest that the number of *Arabidopsis* subtilase genes is the results of duplication events with overlapping ('redundant') functions or strong functional diversification may occur in evolution.

Chromosomal Distribution and Gene Duplications of the AtSBTs

To unravel a possible redundancy we investigated in chromosomal distribution and gene duplication events. *Arabidopsis* subtilase genes are distributed over all 5 chromosomes (http://csbdb.mpimp-golm.mpg.de/psdb.html). The genes occur isolated or in tandem repeats, indicating that segmental and tandem duplication events may have contributed to the evolution of the *Arabidopsis* subtilase gene family. In contrast to the observed average of 17% on genome scale (AGI, 2000) 54% of AtSBT genes occur in clusters of 2 up to 5 genes. These arrangements suggest that also local duplications events contributed to the AtSBT family expansion. Furthermore, several highly similar sequences are found on different chromosomes. Similar situations indicative of a complex evolutionary history have been observed in other Arabidopsis gene families, too (Shiu and Bleecker, 2001; Mladek et al., 2003).

Macro-scale duplication and rearrangement of chromosomes as well as micro-scale translocation and duplication are thought to be the major modes of plant genome evolution (Bancroft, 2000). Analyses of the chromosomal distribution revealed that at least 18 AtSBT genes are located in suggested segmental duplicated regions within the Arabidopsis chromosomes.

The results confirm local and segmental duplication events as the cause for expansion of the subtilase gene family in the course of the *Arabidopsis* genome evolution. As the two copies of a duplicated gene initially are identical and functional redundant, the structure of the subtilase gene family poses the question to what extent the divergence of duplicated genes led to the acquisition of novel and specific functions of subtilases in *Arabidopsis*.

Mutant Identification and Evaluation

To elucidate the functions of all *Arabidopsis* subtilases, T-DNA insertion mutants have been collected and analyzed for morphological traits expressed under standard cultivation conditions. A total of 179 T-DNA insertion lines of 55 AtSBTs have been retrieved from the SIGnAL (Alonso et al, 2003), the GABI-KAT (Rosso et al, 2003), the SAIL (Syngenta Biotechnology Inc., NC), the INRA FLAGdb collections and the University of Wisconsin Knockout Facility. All lines were tested by PCR with gene-specific primers for the presence of the proposed insertion, which was confirmed in 144 lines. For 44 genes, more than one verified T-DNA line is available and for 55 AtSBT genes homozygous T-DNA insertion lines have been collected. Aerial organs of all homozygous lines grown in standard soil cultivation conditions were visually and microscopically examined at several developmental stages. Except for *AtSBT1.2* (*sdd1*), no visible phenotypic alterations linked to the insertion were detectable under these conditions. These observations suggest that either most AtSBT genes mediate specific, conditional responses or else that a large degree of functional redundancy exists among or within subsets of the subtilase family. Indications for the latter possibility of redundancy were obtained by the sequence analyses that identified groups or pairs of closely related genes (see above). To test for potential homology-based functional redundancies we created and confirmed double knockouts and knockout / RNAi lines (http://csbdb.mpimp-golm.mpg.de/psdb.html). However, none of the obtained transgenic lines exhibited any morphological phenotypic alterations upon growth under standard cultivation conditions. While further in-depth analysis will be necessary including monitoring of the responses to various environmental challenges and investigation of metabolic perturbations to complete the phenotypic characterization, these observations may indicate that (partial) functional redundancy may exist even among more family members showing higher sequence divergence. In order to obtain further indications as to which pairs or groups of genes may perform similar or overlapping functions despite low degrees of sequence similarity, and what their physiological roles may be, gene expression co-response analyses were performed.

Ubiquitous and Conditional Expression of AtSBTs

The increasing number of publicly available expression profiles analyzed in the frame of specific experiments enables scientists to use and to re-analyze the data for certain different questions. We investigated in such a cross-experimental approach by computational analyzes of the co-expression and co-response behaviour of subtilases using 123 gene expression profiles publicly available from NASCArrays (Craigon et al., 2004). The expression profile data were generated using the Ath1 gene chip technology platform (Affymetrix, La Jolla, CA), which contains oligonucleotides for 54 of the 56 annotated AtSBT genes. Not represented are *AtSBT3.10* and *AtSBT4.6*. We focused initially only on the AtSBT genes to compare the expression within the subtilase family. This analysis was first performed using the qualitative attributes “present”, “marginal”, and “absent” and was then extended to quantitative values of expression levels. In a third step, the co-response analysis was widened to

include all other genes allowing us to assign subtilases to certain functional classes based on their co-response behaviour.

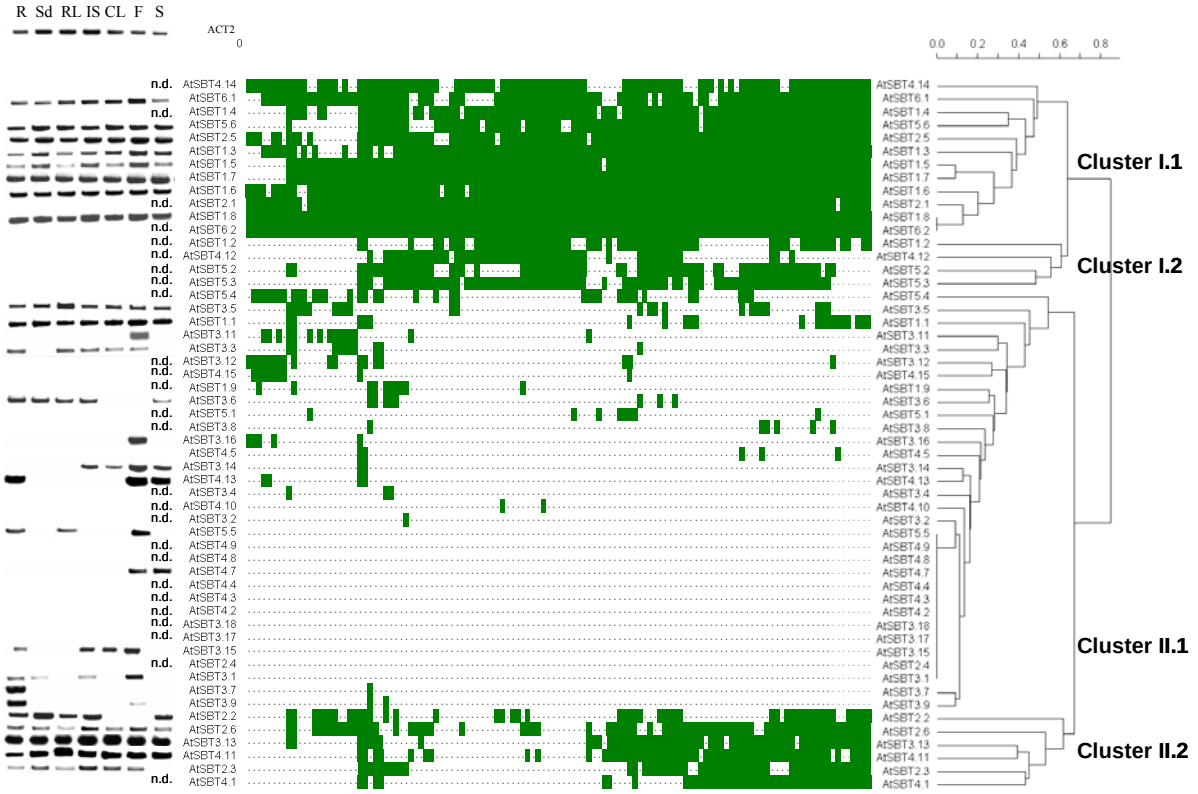

Fig. 3. Cluster tree from converted detection call matrix of 123 Affymetrix (Ath1) microarray experiments into Boolean values. The detection calls absent and medium were assigned to be 0 and present as one. Distances among genes were obtained by applying the S9 index of Gover and Legendre (1986). Sub-cluster SC1 (top main cluster) covers ubiquitously expressed genes whereas SC2 (bottom main cluster) represents low or specific expressed genes. These results were validated independently by semi-quantitative RT-PCR (sqRT-PCR) analysis: (root(R), seedling(Sd), rosette leaves(RL), inflorescence stem(IS), cauline leaves(CL), flower(F), siliques(S), and not determined (n.d.).

For the co-expression analysis we converted the detection calls into qualitative Boolean values: (i) absent and marginal detection calls were set to null and (ii) present calls to one. Pairwise distances among all genes were computed by use of the S9 index of Gower and Legendre (1986). The corresponding distance matrix was subjected to hierarchical clustering (HCA) of the genes. As a result of this analysis we identified two most distantly related AtSBT gene clusters (Fig. 3). The gene cluster I contains 16 (30%) of the 54 represented subtilase genes and showed the following sub-family representation: AtSBT1: 7 (78%), AtSBT2: 2 (33%), AtSBT3: 0 (0%), AtSBT4: 2 (14%), AtSBT5: 3 (50%), AtSBT6: 2 (100%). In contrast, gene cluster II contained 32 (70%) *AtSBT* genes, all belonging to the sub-families AtSBT2, AtSBT3 and AtSBT4. Whereas cluster I.1, a part of the main cluster I, mainly represents ubiquitously expressed genes and some of them expressed at high levels, the cluster

II.1, a part of the main cluster II, primarily contained genes with specific expression pattern and / or low expression levels. Moreover, both main clusters contained subsets of genes, namely cluster I.2 and II.2, which are well measured in an equal number of expression profile experiments. In contrast, genes of these two clusters show a co-expression in only 50% of the same experiments and therefore revealed slightly different expression patterns. To confirm the obtained co-expression behaviour of AtSBTs we performed semi-quantitative RT-PCR (sqRT-PCR) analyses on RNA extracted from different *Arabidopsis* organs. The obtained organ-specific expression patterns of the analyzed genes revealed ubiquitous expression for cluster I.1 genes (Fig. 3). The genes assigned to cluster II.1, on the other hand, exhibited expression primarily in one organ or in a subset of the analyzed organs. For some genes assigned to cluster II, namely *AtSBT2.6*, *3.13*, *4.11*, *2.6*, *3.5*, and *1.1*, we confirmed expression pattern for most of the analyzed organs. According to the results obtained by both analyses we concluded that the genes of cluster I.1 are constitutively expressed, both, in terms of organ specificity as well as according to various conditions. In contrast, the genes of cluster II mainly show specific expression patterns. Moreover, genes assigned to cluster I.2 and II.2 are ubiquitously expressed throughout all or most organs, but vary in expression in response to different conditions (Fig. 3).

Transcriptional Interrelation among AtSBTs revealed by Co-Response Analyses

While through the (qualitative) co-expression the global activity profiles of the AtSBT genes were revealed, the (quantitative) co-response analysis was performed to identify pairs or groups of AtSBT genes that show similar transcript changes among a multi-conditional set of expression. For our subsequent analyses we implicitly make the assumption that common transcriptional control of genes is reflected in co-responding, simultaneous changes in transcript levels (Steinhauser et al., 2004a). To investigate in co-response analysis the expression levels of genes have to vary across the data sets used and valid measures of expression, i.e. above the detection limit, have to be available for the genes in question in most, ideally all profiles. The three generate multi-conditional gene expression data matrices (replicates) consist of approximately 50 out of 123 expression profiles from an approximate equal contribution of each examined experimental condition. These matrices generated were thus maximised for the diversity of the represented experimental conditions. Each of them comprises approximately 10,000 genes, including 12 AtSBT genes, with valid measured transcript levels (see Materials and Methods).

Our numerical approach to detect transcript co-responses is based on the non-parametric Spearman's rank order correlation (r_s), which is a robust estimation of correlation. For bias estimation as well as for a more exact approximation of the statistical probability, we performed iterative computation of r_s based on bootstrap analysis. Test of homogeneity applied to compare the co-responses derived from the three data matrices revealed no significant differences among the pairwise transcript co-responses.

As the test of homogeneity can only detect larger differences we applied in addition the mantel test, performed as non-parametric Spearman correlation of matrices. This analysis showed highly significant correlations ($P << 0.001$) in the range of $0.87 \leq r_s \leq 0.90$ with an average of 0.89±0.02 among the data matrices. Both statistical tests revealed that similar information on transcript co-responses can be deduced from the data matrices. Therefore, the average of transcript co-response can be computed and used for hierarchical cluster analysis.

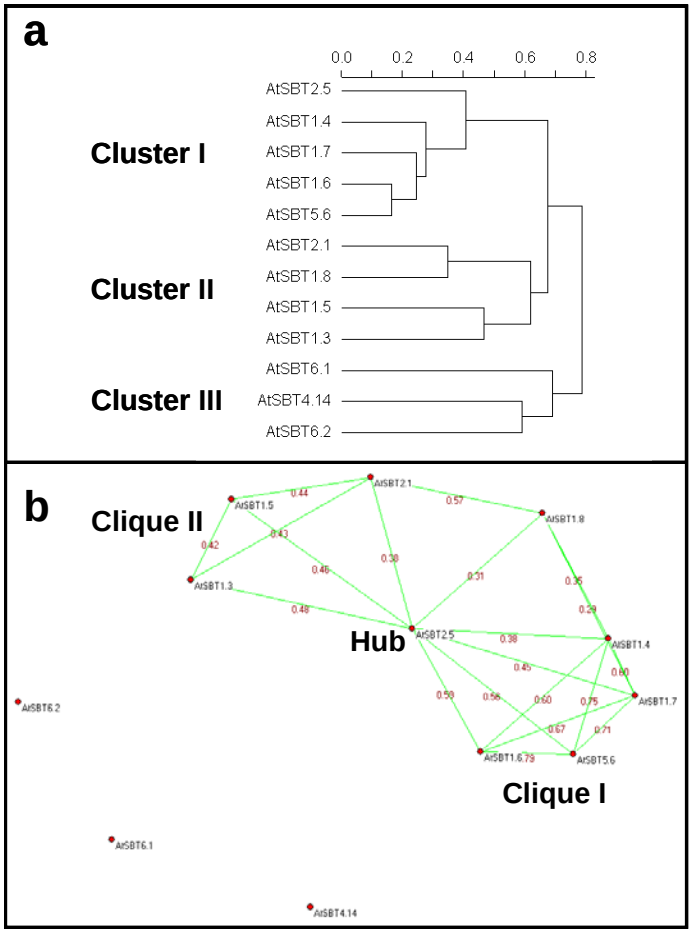

Fig. 4a-b. Fig. 4a shows a cluster tree of the correlated behaviour among transcript amount for ubiquitously expressed genes in multi-conditional 22k Affymetrix expression profiles. The correlations were obtained by computation of the Spearman correlation. Negative correlations were assigned to be most distant. The cluster tree based on the average Spearman correlation among the data matrices nasc0271, nasc0272 and nasc0273. Fig. 4b. shows a network of all significantly associated SBTs of the average Spearman correlation obtained from data matrices nasc0271, nasc0272, and nasc0273. Green coloured lines mark significant positive associations; red coloured lines significant negative correlations. Joint probabilities were obtained by combining probabilities from independent tests of significance. Significance threshold was 0.5 / number of combined probabilities, i.d. $p < 0.0167$.

Fig. 4a shows the cluster tree drawn on the basis of the average co-responses for the three data matrices. Negative co-responses were taken as measure of maximal distant pairs of genes, which showed opposing changes of transcript levels. The 12 represented AtSBT genes are grouped into three well separated clusters: (I) with *AtSBT2.5*, *AtSBT1.4*, *AtSBT1.7*, *AtSBT1.6*, *AtSBT5.6*, (II) with *AtSBT2.1, AtSBT1.8*, *AtSBT1.5*, *AtSBT1.3*, and (III) with *AtSBT6.1*, *AtSBT4.14* and *AtSBT6.2*. The joints are at relative large heights and reflect that the corresponding changes in transcript levels were not identical but similar among pairs and groups of AtSBT genes. For in-depth analysis we visualized only the Bonferroni corrected (Bonferroni, 1935) significant correlations among the AtSBT genes with the Pajek software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/, Fig. 4b). In conjunction with the cluster tree drawn on the basis of the average co-response (Fig. 4a) the obtained network revealed two cliques, where each gene member shows significant correlation to the other members. Clique I covers

AtSBT1.4, *AtSBT1.6*, *AtSBT1.7*, *AtSBT5.6* and *AtSBT2.5*, whereas the clique II enclose the genes *AtSBT1.3*, *AtSBT1.5*, *AtSBT2.1*, and *AtSBT2.5*. The subtilase *AtSBT2.5* is shared between both cliques and represents a hub, which shows significant connections to all genes of the two main cliques and interconnects both (Fig. 4a, b). The average co-response of *AtSBT2.5* to both cliques is 0.47±0.08. Exclusion of *AtSBT2.5* revealed an average co-response for the clique I of 0.69±0.08 and for the clique II of 0.43±0.01. Moreover, for *AtSBT1.8* similar correlations are detectable but this gene shows less connectivity to both cliques.

The statistical analyses revealed significant co-responses among AtSBT genes, but the causality of the interrelations remains to be shown. Non-parametric Kendall's tau (τ) correlation of *Escherichia coli* operon genes controlled by common *cis*-elements revealed a co-response distribution over a broad range (Steinhauser et al., 2004a). Considering the relationship of Spearman's r_s and Kendall's τ ($r_s \sim 3/2\ \tau$), the co-responses among AtSBT genes of the clique I are in the upper range of these distributions and therefore, allows assuming a biological relevance within the network. In conjunction with the results of sqRT-PCR and the co-expression analysis (see above, Fig. 3) we conclude that the genes of clique I are ubiquitously but not constitutively expressed and that they respond to similar cues. The revealed associations and the central positions of *AtSBT2.5* and *AtSBT1.8* in the network suggest that both genes might be involved in the same functional context and may complement each other. However, the amino acid (32.9%) and the nucleic acid sequences (50.0%) do not show any higher homology between these genes than to other AtSBT genes (avg. 34.7% / 50.5%; see http:// csbdb.mpimp-golm.mpg.de/psdb.html). In contrast, *AtSBT2.5* is highly related to *AtSBT2.6* (aa: 88.1%; nt: 83.7%), both are ubiquitously expressed and probably evolved from a sequential duplication. Consequently, they might be of redundant complementary function but a verified double homozygous T-DNA insertion line did not show any visible phenotype under standard cultivation conditions. Therefore, both genes may not share the same function which is supported by expression pattern analysis (Fig. 3). Similarly, *AtSBT5.6* is highly related to *AtSBT5.5* at the sequence level (aa: 62.3%; nt: 68.0%) but only *AtSBT5.6* is a member of clique I. Sequence similarities between *AtSBT5.6* and other member of clique I (avg. aa: 40.5%; avg. nt: 52.9%) are not notably higher than to other AtSBT genes (avg. aa: 36.9%; avg. nt: 51.3%). In contrast, *AtSBT1.4*, *AtSBT1.6*, and *AtSBT1.7* represent an example of evolutionary related genes with higher than average homology on amino acid (46.8 to 54.3%) and nucleic acid (57.7 - 59.7%) level that are members of the clique I and show significant co-regulation. Nevertheless, AtSBTs with even higher sequence homology but lower co-response are present in the subfamily 1. The co-response analysis of the AtSBT gene family thus revealed potential functional relationships, which in some cases clearly contradicted the predictions made on the basis of sequence analysis. In conclusion, we suggest that even minor differences in sequence similarity may confer functional divergence and that functional redundancy within the *Arabidopsis* subtilase family may be better revealed by transcriptional co-response analysis than by high sequence similarity. It is well conceivable that few amino acid changes could alter the substrate

specificity of a protease. A striking example of the consequences of a single amino acid change on the enzymatic properties is provided by the stilbene synthases (Suh et al., 2000).

Co-Response based Transcriptional Neighbourhood Search of AtSBTs

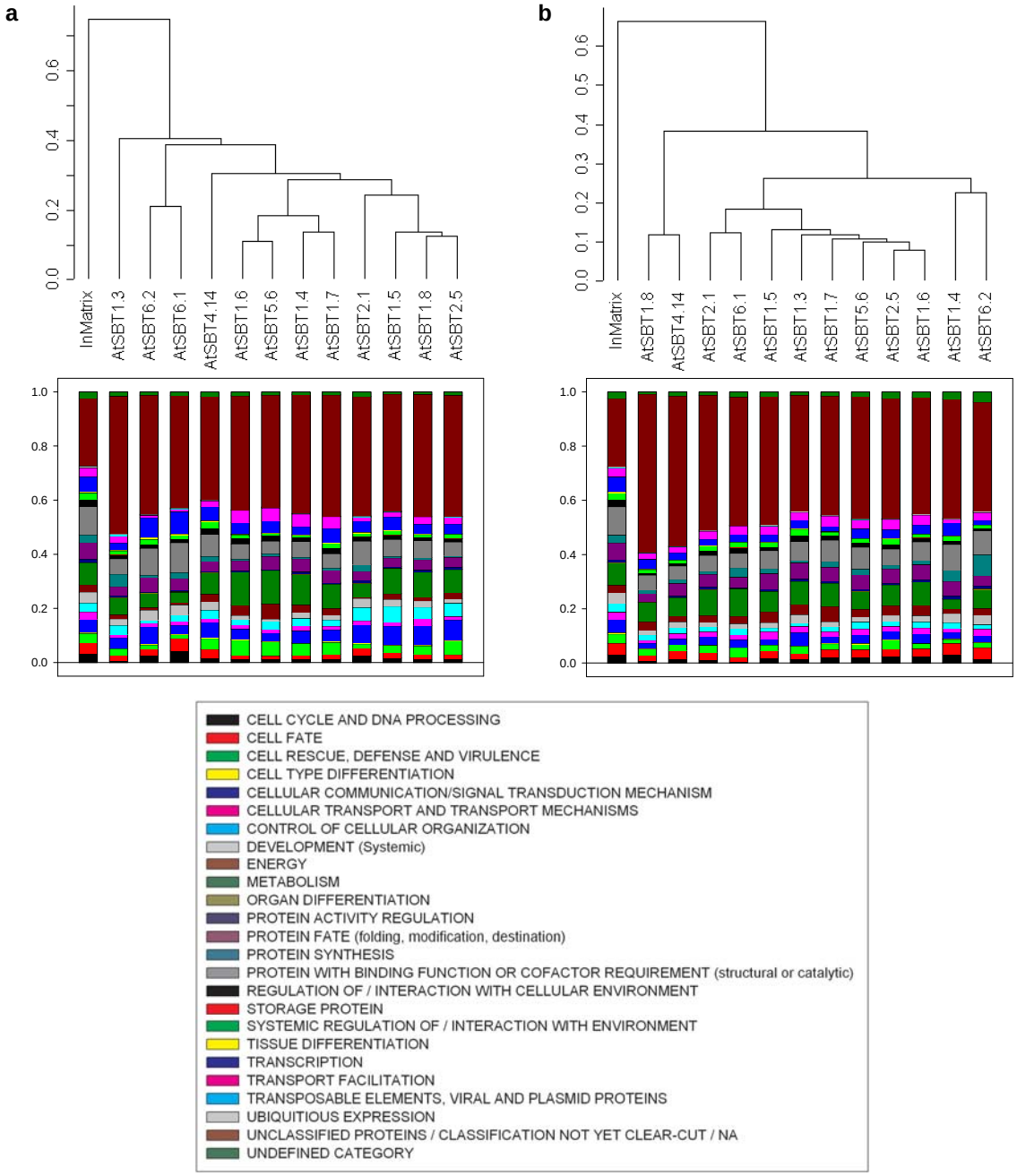

Fig. 5a - b. Result of the co-response based Transcriptional neighbourhood search for the best two percent of positive associated (Fig. 5a) and the best two percent of negative associated (Fig. 5b) correlated genes. The upper part represents the cluster tree resulting from our HCA analysis based on conversion of the enrichment of genes of particular functional categories into the Euclidean distance range. On the bottom a vertical stacked bar plot illustrate the distribution of associated genes to each represented AtSBT for the individual functional categories. For comparison the average distribution of genes belonging to functional categories among the underlying data matrices is shown on the left of each display.

As a third step we extended our co-response analyses to the characterization of the co-responses of AtSBT genes with all other genes represented in the underlying data matrices. This was done to identify sets of co-regulated genes, that are assigned to certain functional categories and that may provide information on the functional context of individual or groups of AtSBT genes.

Whereas the degrees of transcript co-responses may be influenced by the selection of the experiments used for generating the (multi-)conditional data matrices and predictions based on nearest neighbours may be of equivocal nature, we assumed that the enrichment of transcriptionally correlated genes of a certain functional category should be a more robust marker of the functional context of a gene of interest. To obtain such indications for the AtSBT genes the top two percent of the most strongly positive as well as negative correlated genes to each AtSBT gene were selected. The numerical computation of enrichment was done by summation of the relative impacts (RIs) of the genes assigned to particular functional categories, where the gene-specific RI was defined as the reciprocal of the number of assignments of a gene to different categories. As reference we calculated the enrichment as mentioned above over all genes represented in the underlying data matrices.

Application of the G-test of independence for the positive best two percent correlated genes (Fig. 5a) revealed that genes belonging to the category 'unclassified' are significantly enriched ($P << 0.001$) for each of the 12 AtSBT genes with an average of 1.78 fold. A significant ($P < 0.05$) enrichment of genes assigned to 'metabolism' and 'energy' was observed for *AtSBT5.6*. For *AtSBT1.6*, a member of the clique I (Fig. 4a, b). We detected a tendency ($P < 0.1$) of enrichment for 'metabolism' and for *AtSBT1.5*, a member of the clique II (Fig. 4a, b), a significant enrichment for 'control of cellular organisation'. To categorise AtSBT genes according to their neighbourhood we normalised each category-specific sum of RIs and expressed it as the fraction of the sum of all RIs over all categories. The co-responding matrix was subsequently used for hierarchical cluster analyses on the basis of the functional context in the neighbourhood by computing the Euclidean distances. According to the obtained cluster tree for positive associated neighbourhood (Fig. 5a), we suggest a similar functional context for *AtSBT2.5*, the major hub connecting the cliques I and II (Fig. 4b), and *AtSBT1.8*. Interestingly, analysis based on the two percent of strongly negative associated genes (Fig. 5b) revealed different neighbourhoods for the two genes. According to theses results and in conjunction with co-expression (Fig. 3) and co-response (Fig. 4a,b) analyses we suggest that *AtSBT2.5* and *AtSBT1.8* have overlapping but not identical functions. The hub *AtSBT2.5* and *AtSBT1.8* are characterized by an enrichment of positively correlated genes assigned to 'cellular communication / signal transduction mechanism' as well as 'cellular organization', which are ranked at position 2 and 3. Moreover, for the genes *AtSBT1.4* and *AtSBT1.7* as well as for *AtSBT1.6* and *AtSBT5.6*, the members of the cliques I (Fig. 4b), we observed early joining, according to the representation of functional classes by both, the strongly positive and, with exception of *AtSBT1.4*, the strongly negative associated genes. According to a significant enrichment of genes assigned to the functional category 'metabolism' (Fig. 4a, see http:// csbdb.mpimp-golm.mpg.de/psdb.html), we suggest that these genes

are embedded in the functional context of metabolism. These four genes were also correlated in expression with genes enriched for functions in 'cell rescue, defence and virulence' and in 'transport facilitation', which are ranked at positions 3 and 4 (see http:// csbdb.mpimp-golm.mpg.de/psdb.html). The correlated behaviour and similar functional neighbourhoods of this set of *AtSBTs* hints to an involvement within the physiological context of pathogen response or / and general stress related responses. The indications obtained for the functional contexts of these two sets of *AtSBT* genes lead us to suggest the hypothesis that *AtSBT2.5* and *AtSBT1.8* may be involved in sensing mechanisms or be early responsive elements and *AtSBT1.4*, *AtSBT1.6*, *AtSBT1.7* and *AtSBT5.6* may be related to more specific downstream components. The experimental verification of this hypothesis will be one of the goals of our continuing functional genomics project on the characterization of plant subtilases.

The Plant Subtilase Database (PSDB)

The multiple levels of comprehensive data accumulated in this project by members of TASC need a specialized web interface to store and distribute data related to plant subtilases. Accomodating this data we established the Plant Subtilase Database (PSDB), an associated database of CSB.DB, a comprehensive systems-biology database (http://csbdb.mpimp-golm.mpg.de; Steinhauser et al., 2004b). PSDB contains confirmed results of replicated experiments related to plant (*Arabidopsis*) subtilase genes and allows open access to the science community. PSDB will be regularly updated with the results of co-response analyses, performed on the increasing number of publicly available gene expression profiles. Furthermore, validated information of tissue specific expression patterns of AtSBT genes, cellular localisation of encoded proteins, as well as phenotype information of the mutants and transgenic plants will be displayed and regularly updated. Further information and supplemental material will be available at PSDB (http:// csbdb.mpimp-golm.mpg.de/psdb.html).

Material and Methods

Sequence analysis

Nucleic acid and amino acid sequences were retrieved by searching public databases with the BLAST algorithm (Altschul and Lipman, 1990; Altschul et al., 1997) at TAIR (http://www.arabidopsis.org/), TIGR (http://www.tigr.org/), NCBI (http://www.ncbi.nlm.nih.gov) and MIPS (http://mips.gsf.de/). Subcellular localization was predicted using TargetP (http://www.cbs.dtu.dk/services/TargetP/). The deduced amino acid sequences were aligned using the CLUSTALX program (Thompson et al., 1997) with the default parameter settings and manually improved in respect to all known conserved subtilase motifs. The phylogenetic tree was obtained with the neighbour-joining method (bootstrap values have

been generated from 1000 replicates). The tree representation was generated by using the TreeView application (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Plant material and growth conditions

Seeds of *Arabidopis thaliana* ecotype Columbia (Col-0), Wassilevskiya (Ws) and the appropriate T-DNA mutant lines were surface-sterilized and germinated on half-concentrated Murashige and Skoog (MS) medium (M02 555, pH 5.6; Duchefa, Haarlem, The Netherlands) supplemented with 1% Sucrose and solidified with 0.7% agar under a 16-h day (140 $\mu mol \cdot m^{-2} \cdot s^{-1}$, 22°C)/8-h night (22°C) regime. The plates were incubated at 21/15 °C day/night, under a 16/8 h light/dark. After two weeks plants were transferred to standard soil (Einheitserde GS90; Gebrüder Patzer, Sinntal-Jossa, Germany) and further grown in a growth chamber under a long-day light regime (16 h of fluorescent light [120 $\mu mol \cdot m^{-2} \cdot s^{-1}$] at 20°C and 60% RH/8 h of dark at 16°C and 75% RH).

Mutant collection, confirmation and phenotypic analysis

T-DNA insertion mutants were retrieved from the SIGnAL (Alonso et al., 2003), the GABI-Kat (Rosso et al., 2003), the Genoplante FST/FLAGdb (Balzergue et al., 2001), the SAIL collection (Syngenta Biotechnology Inc, http://www.tmri.org/en/partnership/sail_collection.aspx), and the University of Wisconsin Knockout facility (http://www.hort.wisc.edu/krysan/DS-Lox/). Genomic DNA was isolated using the Dneasy 96 Plant Kit (Qiagen, Hilden, Germany) and subsequently used for PCR analysis. The T-DNA insertion lines were screened for the appropriate insert using the required T-DNA and a gene specific primer. Gene specific flanking primers were used to confirm homozygousity. Primer sequences are available at PSDB (http://csbdb.mpimp-golm.mpg.de/psdb.html). In respect to phenotypic alterations, homozygous insertion lines were evaluated in the following developmental stages (Boyes et al., 2001): 1.03 for seedlings grown on synthetic media, 3.9 for rosette leaves and at stage 6.9 for inflorescence stem, cauline leaves, flower and siliques. Plants were investigated regarding to leaf number, shape and size, epidermal constitution in respect to trichome and guard cell number and distribution, flowering time as well as flower and silique constitution.

Data source and pre-processing

Transcript co-responses were retrieved from the CSB.DB - a comprehensive systems-biology database (http://csbdb.mpimp-golm.mpg.de; Steinhauser et al., 2004b) for data matrix nasc0271. Co-responses for the additional matrices nasc0272 and nasc0273 were computed within this work. 123 publicly available expression profiles from 22 experiments were obtained from NASCarrays (http://affymetrix.arabidopsis.info/; Craigon et al., 2004; October 2003) and used for the generation of the data matrices. The profiles were originated through the Affymetrix Ath1 chip technology (Affymetrix, La Jolla, CA). The number of Present and Marginal calls (according to the MAS 5.0

algorithm) was calculated for each profile. In the majority of cases 2 or 3 profiles per experiment with the highest numbers of Present and Marginal calls were selected for nasc0271. In analogy, nasc0272 and nasc0273 were generated by experiments with 2nd and 3rd highest numbers of Present and Marginal calls. Thus the data matrices comprised approximately 50 out of 123 experiments with approximately 10,000 out of > 22.000 genes: nasc0271: 51 experiments with 9694 genes, nasc0272: 51 experiments with 8927 genes, and nasc0273: 49 experiments with 8691 genes. Each well measured in at least 85% of the underlying expression profiles. Transcript co-responses were computed on data matrices with log base two transformed and range-normalised transcript intensities, i.e. log base two transcript intensities for each gene were in range of 0 to 1.

Co-Expression analysis

For co-expression analysis the detection calls were converted into Boolean values. The numerical value null was assigned to absent and marginal calls, whereas present call were set to be one. Pairwise distances among entities, i.e. genes, of the Boolean matrix were computed by use of the S9 index of Gower and Legendre (1986) and subsequently used for hierarchical cluster analysis (HCA, Mirkin, 1996). Computation was executed with the statistical software environment R (http://www.r-project.org) version 1.8.1. Distances were computed with the function 'dist.binary' of the 'ade4' package. HCA was performed as unweighted average linkage clustering algorithm (UPGMA) by use of the 'hclust' function implemented in the 'mva' package.

Semi quantitative RT-PCR expression analysis

Sample material of the appropriate organs from *Arabidopsis* plants (Col-0) were harvested at the same stage used for mutant screening (see above). Total RNA was isolated with TRIzol® reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol. 1µg of total RNA was pre-treated with DNaseI (Ambion, Austin, Texas) and reverse transcribed with SuperScriptII® reverse trascriptase (Invitrogen, Carlsbad, CA) and $d(T)_{15}$. The cDNA reaction was diluted 1:5 with water and 5 µl of the diluted cDNA was used as template for PCR analysis applying the Advantage 2 PCR Enzyme System (BD Biosciences, Palo Alto, CA) according to the manufacturer's protocol. In general, due to the low abundance of subtilase transcripts, up to 40 cycles were performed with a PTC-200 thermal cycler (MJ Research, Waltham, MA). Primers, AGI gene name, and the size of cDNA and genomic amplicons are available via PSDB (http:// csbdb.mpimp-golm.mpg.de/psdb.html). ACT2 was used as external standard. 20µl each PCR reaction was analysed by agarose gel electrophoreses.

Co-Response analyses

According the observation that a general bivariate normality can not be assumed for each pair of genes, respectively analysed with the Cramer-test (Baringhaus and Franz, 2004), transcript co-response analyses were performed by determination of non-parametric Spearman's rank order correlation (r_s)

(Sokal and Rohlf, 1995). Co-Response analysis among AtSBT genes were computed by non-parametric bootstrap analyses with 2,000 numbers of bootstrap samples (Efron and Tibshirani, 1993). Transcript co-responses of AtSBT genes with all other genes represented in the respective matrix were computed with cCoRv1.0 (Steinhauser et. al., unpublished). Mantel test and test on homogeneity (Sokal and Rohlf, 1995) were used to compare and compute the average of correlations among different co-response matrices. The test of homogeneity was performed with Microsoft Excel. The mantel test, computed as non-parametric Spearman correlation of (dis-)similarity matrices, was executed in R by use of the 'mantel.test' function of the 'vegan' package. The average Spearman correlations and the joint probabilities among the data matrices were calculated as recommended (Sokal and Rohlf, 1995). In order to generate normalised distance matrices for HCA correlations were converted into distance range according to Sokal and Rohlf (1995). Negative Spearman correlations were assigned to be most distant and were converted into the largest distances: distance = $1 - r_s$. Normalisation of the obtained distance matrix was done by dividing all distances with the obtained maximum distance. HCA was performed as mentioned above. Visualisation of significant associations among AtSBT genes was done with the software Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). The multiple comparison performed required the adjustment of α to accept significant associations, which was done by application of the Bonferroni correction $\alpha' = \alpha / k$. The corrected α' was 0.00416 (12 comparisons).

Transcriptional Neighbourhood Search

The assignment of gene products to functional categories was retrieved from MAtDB (Schoof et al., 2004; December 2003). The functional categorisation is tree-like structured and each category is subdivided into sub-categories. We used only the highest branch (level) for each category which yielded to 99 categories, 29 described with a category name. Categories without category name were merged into the 'undefined' category and the categories 40, 43, 45, 47 were merged into the class 'localisation'. Genes without assignment or with unclear classification were treated as 'unclassified'. Genes assigned into more than seven categories, which represents 5% of the whole annotation, were also treated as 'unclassified'. The relative impact (RI) of a gene with multiple assignments (n_{assign}) onto each category was defined as: $ri = 1 / n_{assign}$.

The transcriptional neighbourhood search was performed as follows: The best two percent of positively and negatively correlated genes to each represented AtSBT gene were extracted and grouped according their assigned functional category separately for each AtSBT gene. For calculation of the enrichment into functional categories the sum of all RIs for each category was computed. The sum of all RIs for each category over all represented genes was used as reference. Comparisons of the observed enrichment for each AtSBT genes to the reference were done by G-test of independence (Sokal and Rohlf, 1996) and separately for each functional category.

Acknowledgements

We thank the Salk Institute Genomic Analysis Laboratory, INRA (Versaille) and Syngenta Biotechnology Inc. for providing the sequence-indexed Arabidopsis T-DNA insertion mutants and for the T-DNA mutants that were generated in the context of the GABI-Kat program and provided by Bernd Weisshaar (MPI for Plant Breeding Research; Cologne, Germany).

We thank the staff of CSB.DB for the support with mathematical algorithms and special software. Technical assistance by Jutta Babo and Ursula Glück-Behrens is gratefully acknowledged. Furthermore, we acknowledge the NASCArrays for the establishment of public accessible repository for microarray data as well as all scientists who submitted transcript profile data to these databases and thereby enabled us to perform comparative investigations.

References

Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. et al. (2003) Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653-657.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990). Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403-410.

Bancroft,I. (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast*, **17**, 1-5.

Barr, P.J. (1991) Mammalian subtilisins: The long-sought dibasic processing endoproteases. *Cell*, **66**, 1-3.

Balzergue,S., Dubreucq.B., Chauvin,S., Le-Clainche,I., Le Boulaire,F., de Rose,R., Samson,F., Biaudet,V., Lecharny,A., Cruaud,C. et al. (2001) Improved PCR-Walking for Large-Scale Isolation of Plant T-DNA Borders. *BioTechniques*, **30**, 496-504.

Baringhaus,L. and Franz,C. (2004) On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190-206.

Beers,E.P., Jones,A.M. and Dickerman,A.W. (2003) The S8 serine, C1A cysteine and A1 aspartic protease families in *Arabidopsis*. *Phytochemistry*, **65**, 43-58.

Beilinson,V., Moskalenko,O.V., Livingstone,D.S., Reverdatto,S.V., Jung,R. and Nielsen,N.C. (2002) Two subtilisin-like proteases from soybean. *Physiol Plant.*, **115**, 585-597.

Berger, D. and Altmann,T. (2000) A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev.*, **14**, 1119-1131.

Bogacheva,A.M. (1999) Plant Subtilisins. *Biochemistry (Moscow)*, **3**, 287-293.

Bogacheva,A.M., Rudenskaya,G.N., Dunaevsky,Y.E., Chestuhina,G.G. and Golovkin,B.N. (2001) New subtilisin-like collagenase from leaves of common plantain. *Biochimie*, **83**, 481-486.

Bonferroni,C.E. (1935) Il calcolo delle assicurazioni su gruppi di teste. In Studi in Onore del Professore Salvatore Ortu Carboni, Rome, 13-60.

Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Methods Enzymol.*, **266**, 162-184.

Boyd,P.M., Barnaby,N., Tan-Wilson,A. and Wilson,K.A. (2002) Cleavage specificity of the subtilisin-like protease C1 from soybean. *Biochim. Biophys. Acta*, **29**, 269-282.

Boyes,D.C., Zayed,A.M., Ascenzi,R., McCaskill,A.J., Hoffman,N.E., Davis,K.R. and Gorlach,J. (2001) Growth Stage–Based Phenotypic Analysis of Arabidopsis. *Plant Cell*, **13**, 1499-1510.

Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC`s transcriptomics service. *Nucleic Acids Res.*, **32**, D575-D577.

Dodson,G. and Wlodawer,A. (1998) Catalytic triads and their relatives. *Trends Biochem. Sci.*, **23**, 347-352.

Dyer,J.M., Chapital,D.C., Lax,A.R. and Pepperman,A.B. (1999) Identification of a subtilisin-like protease in seeds of developing tung fruits. *J. Plant Physiol.*, **155**, 802-805.

Efron,B. and Tibshirani,R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York, London, 1-456.

Emanuelsson,O., Nielsen,H., Brunak,S., and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005-1016.

Fontanini,D. and Jones,B.L. (2002) SEP-1 - a subtilisin-like serine endopeptidase from germinated seeds of *Hordeum vulgare* L. cv. Morex. *Planta*, **215**, 885-893.

Fuller,R.S., Brake,A. and Thorner,J. (1989) Yeast prehormone processing enzyme (*Kex2 gene product*) is a Ca2+-dependent serine protease. *Proc. Natl. Acad. Sci. USA*, **86**, 1434-1438.

Gower,J.C. and Legendre,P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.

von Groll,U., Berger,D. and Altmann,T. (2002) The Subtilisin-Like Serine Protease SDD1 Mediates Cell-to-Cell Signaling during *Arabidopsis* Stomatal Development. *Plant Cell*, **14**, 1527–1539.

Hamilton,J.M., Simpson,D.J., Hyman,S.C., Ndimba,B.K. and Slabas,A.R. (2003) Ara12 subtilisin-like protease from *Arabidopsis thaliana*: purification, substrate specificity and tissue localization. *Biochem. J.*, **15**, 57-67.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145-154.

Hodgman,T.C. (2000) A historical perspective on gene/protein functional assignment. *Bioinformatics*, **16**, 10-15.

Jordá,L., Coego,A., Conejero,V. and Vera,P. (1999) A genomic cluster containing four differentially regulated subtilisin-like processing protease genes is in tomato plants. *J. Biol. Chem.*, **274**, 2360-2365.

Kaneda,M. and Tominaga,N. (1975) Isolation and characterization of a proteinase from the sarcocarp of melonfruit. *J. Biochem. (Tokyo)*, **78**, 1287-1296.

Kinal,H., Park,C.M., Berry,J.O., Koltin,Y. and Bruenn,J.A. (1995) Processing and secretion of a virally encoded antifungal toxin in transgenic tobacco plants: evidence for a Kex2p pathway in plants. *Plant Cell*, **7**, 677-688.

Luo,X. and Hofmann,K. (2001) The protease-associated domain: a homology domain associated with multiple classes of proteases. *Trends Biochem. Sci.*, **26**, 147-148.

Mahon,P. and Bateman,A. (2000) The PA domain: a protease-associated domain. *Protein Sci.*, **9**, 1930-1934.

McGeoch,D.J. and Davison,A.J. (1986) Alphaherpesviruses posses a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. *Nucleic Acids Res.*, **14**, 1765-1777.

Mladek,C., Guger,K. and Hauser,M.T. (2003) Identification and Characterization of the ARIADNE Gene Family in *Arabidopsis*. A Group of Putative E3 Ligases. *Plant Physiol*., **131**, 27-40.

Meichtry,J., Amrhein,N. and Schaller,A. (1999) Characterization of the subtilase gene family in tomato (*Lycopersicon esculentum* Mill.). *Plant Mol. Biol.*, **39**, 749-760.

Mirkin,B. (1996) Nonconvex Optimisation and Its Application, Mathematical Classification and Clustering, Vol 3. Kluwer Academic Publishers, 1-428.

Popovic,T., Puizdar,V. and Brzin,J. (2002) A novel subtilase from common bean leaves. *FEBS Lett.*, **23**, 163-168.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224-228.

Rosso,M.G., Li,Y., Strizhov,N., Reiss,B., Dekker,K. and Weisshaar,B. (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol. Biol.*, **53**, 247-259.

Rudenskaya,G.N., Bogacheva,A.M., Preusser,A., Kuznetsova,A.V., Dunaevsky,Y.E., Golovkin,B.N. and Stepanov,V.M. (1998) Taraxalisin - a serin proteinase from dandelion *Taraxacum officinale* Webb s.l. *FEBS Lett*., **437**, 237-240.

Rudenskaya,G.N., Bogdanova,E.A., Revina,L.P., Golovkin,B.N. and Stepanov,V.M. (1995) Macluralisin- a serine proteinase from fruits of *Maclura pomifera* (Raf.) Schneid. *Planta*, **196**, 174-179.

Sakai,J., Rawson,R.B., Espenhade,P.J., Cheng,D., Seegmiller,A.C., Goldstein,J.L. and Brown,M.S. (1998) Molecular identification of a sterol-regulated luminal protease that cleaves SREBSs and controls lipid composition in animal cells. *Mol. Cell*, **2**, 505-515.

Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,F.X. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acid Res.*, **32**, D373-376.

Seidah,N.G. and Chrétien,M. (1999) Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides. *Brain Res.*, **845**, 45-62.

Seidah,N.G., Mowla,S.J., Hamelin,J., Mamarbachi,A.M., Benjannet,S., Toure,B.B., Basak,A., Munzer,J.S., Marcinkiewicz,J., Zhong,M., Barale,J.C., Lazure,C., Murphy,R.A., Chretien,M, and Marcinkiewicz,M. (1999) Mammalian subtilisin/kexin isozyme SKI-1: A widely expressed proprotein convertase with a unique cleavage specificity and cellular localization. *Proc. Natl. Acad. Sci. USA*, **16**, 1321-1326.

Shiu,S.H. and Bleecker,A.B. (2001) Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proc. Natl. Acad. Sci. USA*, **98**, 10763–10768.

Siezen,R.J. and Leunissen,J.A.M. (1997) Subtilases: The superfamily of subtilisin-like serine proteases. *Protein Science*, **6**, 501-523.

Steinhauser,D., Junker,B.H., Luedemann,A., Selbig,J. and Kopka,J. (2004a) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, **20**, 1928-1939.

Steinhauser,D, Usadel,B, Luedemann,A, Thimm,O. and Kopka,J. (2004b) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* (in press).

Sokal,R.R. and Rohlf,F.J. (1995) Biometry: The principles and practice of statistics in biological research. Ed 3. W.H. Freeman and Company, New York, 1-887.

Suh,D.Y., Fukuma,K., Kagami,J., Yamazaki,Y., Shibuya,M., Ebizuka,Y. and Sankawa,U. (2000) Identification of amino acid residues important in the cyclization reactions of chalcone and stilbene synthases. *Biochemical Journal*, **350**, 229-235.

Tanaka,H., Onouchi,H., Kondo,M., Hara-Nishimura,I., Nishimura,M., Machida,C. and Machida,Y. (2001) A subtilisin-like serine protease is required for epidermal surface formation in *Arabidopsis* embryos and juvenile plants. *Development*, **128**, 4681-4689.

Terp,N., Thomsen,K.K., Svendsen,I., Davy,A. and Simpson,D.J. (2000) Purification and characterization of hordolisin, a subtilisin-like serine endoprotease from barley. *J. Plant Physiol.*, **156**, 468-476.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, *408*, 796-815.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876-4882.

Tornero,P, Conejero,V. and Vera,P. (1996a) Primary structure and expression of a pathogen-induced protease (PR-69) in tomato plants: Similarity of functional domains to subtilisin-like endoproteases. *Proc. Natl. Acad. Sci. USA*, **93**, 6332-6337.

Tornero,P., Mayda,M., Gomez,M.D., Caña,S.L., Conejero,V. and Vera,P. (1996) Characterization of LRP, a Leucine-Rich Repeat (LRR) protein from tomato plants that is processed during pathogenesis. *Plant J.*, **10**, 315-330.

Vera,P. and Conejero,V. (1988) Pathogenesis-related proteins of tomato, P-69 as an alcaline endoprotease. *Plant Physiol.*, **87**, 58-63.

Vukmirovic, O.G. and Tilghman, S.M. (2000). Exploring genome space. *Nature*, **405**, 820-822.

Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35-40.

Yamagata,H., Masuzawa,T., Nagaoka,Y., Ohnishi,T. and Iwasaki,T. (1994) Cucumisin, a serine protease from melon fruits, shares structural homology with Subtilisin and is generated from a large precursor. *J. Biol. Chem.*, **269**, 32725-32731.

Chapter VI - From Genome to Metabolome:
- GMD@CSB.DB: The Golm Metabolome Database -

Joachim Kopka[1], Nicolas Schauer[1], Stephan Krueger[1], Claudia Birkemeyer[1], Björn Usadel[1], Eveline Bergmüller[2], Peter Dörmann[1], Yves Gibon[1], Mark Stitt[1], Lothar Willmitzer[1], Alisdair R. Fernie[1] and Dirk Steinhauser*[1]

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

[2]Institute of Plant Sciences, Swiss Federal Institute of Technology, 8092, Zurich, Switzerland

Abstract

Summary: Metabolomics, in particular gas chromatography – mass spectrometry (GC-MS) based metabolite profiling of biological extracts, is rapidly becoming one of the cornerstones of functional genomics and systems biology. Metabolite profiling has profound applications in discovering the mode of action of drugs or herbicides and in unravelling the effect of altered gene expression on metabolism and organism performance in biotechnological applications. As such the technology needs to be available to many laboratories. For this, an open exchange of information is required, like that already achieved for transcript and protein data. One of the key-steps in metabolite profiling is the unambiguous identification of metabolites in highly complex metabolite preparations from biological samples. Collections of mass spectra, which comprise frequently observed metabolites of either known or unknown exact chemical structure, represent the most effective means to pool the identification efforts currently performed in many laboratories around the world. Here we present GMD - The Golm Metabolome Database, an open access metabolome database, which should enable these processes. GMD provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools, e.g. regarding to methods, spectral information or compounds. The main goal will be the representation of an exchange platform for experimental research activities and bioinformatics to develop and improve metabolomics by multidisciplinary cooperation.

Availability: The Golm Metabolome Database can be accessed through the following URL http://csbdb.mpimp-golm.mpg.de/gmd.html.

Contact: Steinhauser@mpimp-golm.mpg.de

Introduction

The sequencing and annotation of whole genomes of various organisms (Goffeau et al., 1996; Blattner et al., 1997; Arabidopsis Genome Initiative, 2000; Lander et al., 2001) facilitate the development of technology platforms to monitor the cellular inventory (Fiehn et al., 2000; Lockhard and Winzeler, 2000; Corbin et al., 2003). Since the dawn of genomic technology in the past decade in conjunction with enhancing genomic information a vast amount of diverse data has been generated and released to the public community. The improving knowledge of gene functions in concurrence with global expression analyses is allowing phenotypes to be linked to their co-responding genomic data. However, our knowledge of the molecular basis of biological functions and their respective contribution to observed phenotypes is, as yet, relatively rudimentary. Recently, the mining and exploitation of data by multi-parallel 'omics technologies open up the possibility to gain comprehensive insight into understanding biological systems (Kitano, 2002; Oltvai and Barabási, 2002; Fernie et al, 2004). The flood of information obtained worldwide by scientists for this purpose urgently requires user friendly public data access. In the past decades much progress has been made on the storage of information derived from the various levels of the cellular hierarchy. For instance, databases like BRENDA (Schomburg et al., 2004), KEGG (Kanehisa et al., 2004) or MetaCyc (Krieger et al., 2004) harbour information concerning metabolic pathways, chemical reactions including inventory of the genes and enzymes involved. Genomic databases, such as MIPS (Mewes et al., 2004), TAIR (Rhee et al., 2003) and TIGR (Quackenbush et al., 2000), provide public access to protein sequences based on whole genome analyses, maps of protein-protein interactions, protein localization and many further features. Recent developments in transcript profiling technologies have led to the adoption of largely similar experimental platforms that are used worldwide. The commonality of experimental approach facilitated the establishment of expression profile related databases, such as the Stanford Microarray Database [SMD, (Gollub et al., 2003)], TAIR or NCBI-GEO (Edgar et al., 2002). Similarly the availability of proteome data has driven the establishment of various databases [e.g. SWISS-PROT, (Boeckmann et al., 2003)] or initiatives [e.g. HAP, (Hermjakob et al., 2004)] focussing on the functional annotation of proteins.

In contrast to the multitude of well established databases which comprises information gathered on genome, transcriptome and proteome level, no attempt has been made to store the flood of data arising from metabolome analyses of biological samples. As already outlined, metabolites have an enormous range of structures. These are measured using a wide range of technology platforms (Kopka et al., 2004). There is an urgent need for publicly accessible metabolome databases that harbours underlying information on metabolites. Here we describe the Golm Metabolome Database (GMD), an open access metabolome database for exchange and presentation of metabolomic and related information. In the current build the main emphasis focuses on GC-MS (Roessner et al., 2000), the most advanced and widespread technology platform for metabolomics. The collected information (i) covers knowledge

concerning analytical technologies (Analytics), (ii) harbours underlying evidence and information that supports unequivocal metabolite identification (MSRI). In addition, GMD (iii) provides access to stored metabolite profiles (Profiles).

GMD - Structure and Data available

Affiliation and Implementation

The Golm Metabolome Database (GMD) platform is affiliated to CSB.DB - a comprehensive systems-biology database, which is hosted at the Max-Planck-Institute of Plant Molecular Physiology, Golm-Potsdam, Germany. GMD complements the currently available transcriptional co-response databases and uses a similar system for data storage and handling as described earlier (Steinhauser et al., 2004).

Analytic – information concerning analytical technologies

The highly complex nature and the enormous chemical diversity of compounds obtained when analyzing the metabolome of organisms constitutes one of the main challenges in metabolomics (Oksman-Caldentey et al., 2004; Fernie et al., 2004). Current estimations vary however it is thought that between 4000 - 25000 compounds may represent the metabolome of any given organism (Trethewey, 2004; Fernie et al., 2004). The plant kingdom as a whole is believed to have in excess of 200,000 metabolites (Fiehn, 2002; Trethewey, 2004). The range of the highly diverse chemical characteristics in conjunction with the vast amount of potential measurable compounds has large implications for metabolite extractability and stability. Any one protocol for measurement thus represents a balance between accuracy and coverage of metabolites. The GMD analytic pages allow access to expert knowledge for an overview of applied methods by the GMD contributors. They include information related to different technology platforms, publicly available methods, as well as contact information for individual, tailor-made knowledge exchange. Furthermore, an overview of the potential available resources is given for scientists who are unfamiliar in this field of experimental biology or interested in setting-up a metabolomics facility.

MSRI – mass spectra and retention time index libraries

Following analytical measurements, data processing algorithms are applied to detect peaks in spectral data. The identification and characterization of the hundreds to thousands of metabolites obtained from diverse biological samples represents a major challenge in metabolomics. These identification efforts required a large-scale analysis of standard substances to generate customized spectral libraries for further identification of unknown metabolites. To overcome this current limitation of individually customized mass spectral libraries the GMD mass spectra information pages are developed to exchange information on the underlying evidences that supports metabolite identification in complex

GC-MS profiles from diverse biological sources. The MSRI web platform provides access to customized mass spectral and retention time index (MSRI) libraries, which were generated using identical capillary GC columns using two different electron impact ionization GC-MS technologies, namely quadrupole GC-MS (Fiehn et al., 2000; Roessner et al., 2000) and GC-TOF (time of flight)-MS (Wagner et al., 2003). Currently, five downloadable libraries are available, which may be imported into the NIST02 mass spectral search program or AMDIS, the automated mass spectral deconvolution and identification system (National Institute of Standards and Technology, Gaithersburg, MD, USA). The (pre-computed) libraries are split according to the technology platform and the degree of manual mass spectral identification. The Q_MSRI and T_MSRI libraries contain MSTs, which were either generated on four identically configured quadrupole (Q_MSRI) GC-MS systems or on a single time of flight (T_MSRI) system which run with identical settings but slight modifications. Mass spectral libraries, which exclusively consist of manually evaluated, identified or classified MSTs are assigned to ID-libraries. In contrast, libraries which were generated by automated deconvolution using AMDIS software were assigned to NS libraries, indicative of the non-curated mode of construction. The currently available libraries covering data from mammals, yeast, corynebacterium, model plants, crop plants and related wild species, as well as from suitable non-sample controls. More than 2000 evaluated mass spectra data from the two technology platforms featuring 1089 non-redundant and 360 identified MSTs are included in the current available version of these libraries.

GMD Profiles – the metabolite profiling platform

The recent maturity of GC-MS (and other) technology platforms has facilitated the development of metabolomics into an important technology for functional genomic efforts. The vast amount of complex data obtained from metabolite profiling experiments in conjunction with the ongoing developments on analytical technologies requires the public availability of those data for cross-comparisons and cross-experimental approaches. According to these demands we started to implement metabolic fingerprinting and metabolite profiling experiments, which can be currently searched by compound names or browsed by a list of experiments (see below).

For the exchange of the highly complex experimental background information and data from metabolite profiling experiments we implemented the MIAMET description, the **M**inimum **I**nformation **A**bout a **MET**abolomics experiment, as suggested by Bino et al. (2004). Similar to the MGED effort to standardize microarray data by the MIAME standard (Brazma et al., 2001) MIAMET may evolve to a general accepted and recommended MIAME format for the metabolomics field.

GMD - Features and Queries

Content Browsing: The GMD content can be explored by browsing the HTML content through lists or a simple site map, a hierarchical tree representation, which is linked to the available second level of HTML pages. Information regarding downloadable MSRI libraries as well as related supplementary information, such as technologies, method descriptions and acknowledgements, is made accessible. Both, the MSRI libraries as well as the currently integrated metabolite profiling experiments are presented in a list format, which provides links to associated detailed information.

A more sophisticated way to explore the GMD content is offered through the available query pages. Currently, five different types of queries are implemented which can all be accessed by the GMD site map.

MSRI Compound Search: The compound search tool allows searching by compound name and provides access to the linked mass spectral information harboured at GMD. Various filter options can be applied to restrict the query results, for example to the available technology platforms, particular libraries or methods. The retrieved mass spectral entries are presented as a table which contains basic mass spectral information for a particular compound, such as compound role, i.e. metabolite or internal standard, observed retention time index (RI) and technology platform. This basic information can be sorted upon user invocation. All information is linked to the detailed physicochemical characteristics of the available mass spectra. This final level of information facilitates the identification of a particular compound in profile analyses. The in-depth mass spectral information encompasses in addition the recommended quantifier and qualifier masses, and access to available replicate mass spectra of the same compound.

MSRI Mass Spectrum Search: For analysis of user provided mass spectra we implemented a query tool which allows comparison to all available curated mass spectra of our libraries. Mass spectra may be submitted in either NIST02 or AMDIS format (Ausloos et al., 1999; Stein, 1999). The search is performed by computing the fragment-intensity agreement, measured as dynamically normalized Euclidean distance [Euclid], as S12 [s12] index (Gover and Legendre, 1986; Ochiai, 1957), Hamming (Hamming, 1950) and Jaccard distance (Jaccard, 1908). The result set is presented as a sortable HTML table containing information such as the rank, the identifier for each spectrum, the retention time index (RI), the method information, the compound name in case of identified metabolites, and all computed similarity measures. All types of information can be used for sorting. Moreover, additional criteria for comparison are given based (1) on absolute RI differences to the observed RI taken from an optional user input and (2) alternatively to the best hit found. If available, analysis of occurrence of qualifier as well as quantifier masses is performed. A head-to-tail plot of the query and selected hit spectra can be invoked. Depending on the chosen sorting a colour-coded graphical representation of the 10 best hits is generated below the result table. The graphical out-put is similar to a typical BLAST (Altschul et al., 1990) result. The ratio plot mirrors the occurrence of the masses and their co-responding ratios of

intensities in comparison to the query spectrum. The result table can be downloaded by an exporter function as a tab delimited and zipped file. The file contains the identical information as the HTML table but extends beyond the HTML display to all available mass spectra. Various filter options, especially restriction to a pre-defined RI window or set of major fragments, can be invoked by the user to limit the search to relevant results.

MSRI Customized Library Generation: In contrast to the (pre-computed) MSRI libraries GMD allows the user to generate customized mass spectral library from the list of curated mass spectral entries. The results can be downloaded as a zipped text file and handled and used like a pre-computed MSRI library (see above). The search input is restricted to the MPIMP-IDs which can be obtained by the above-mentioned queries or by using the compound name converter (see below). The result can be limited according to the technology platform or the methods used to obtain a curated spectrum.

Profile Compound Search: As mentioned above GMD has started to integrate metabolite profiling experiments generated on a quadrupole GC-MS technology platform. Currently, 69 profiles of nine replicate sets are included describing metabolic changes under different light conditions. The profiles can be queried by a particular compound name and allows searching for the co-responding compound levels within an experiment. Various filter options are available to restrict computation to high-quality mass traces by using the default or user modified values. Moreover, the user can select between parametric or non-parametric statistics for the dynamically computation of the treatment-control comparisons. The result set covers information regarding to the experimental backgrounds, comparisons done as well as information concerning the probability of the difference in the observed compound levels. Furthermore, treatment-control ratios are given and colour coded to mark decreased or increased levels. In analogy to the Affymetrix oligonucleotide technology platform we use different masses as representatives for a particular compound. Each of the used masses is represented in the results tables as well as is characterized by their co-responding behaviour in relation to each other mass.

Compound Name Converter: According to the usage of different (identifier for) compound names for a metabolite we implement a converter which allows converting user compound names to MapMan names for stand-alone visualization of the results with the MapMan software (Thimm et al. 2004) or to MPIMP-IDs for customized library generation.

Outlook

GMD will frequently be updated with new mass spectra, metabolite identifications, mass spectral libraries of biological samples and metabolite profiling experiments. GMD is intended as a repository for experiments performed at the Max-Planck-Institute of Molecular Plant Physiology and for data made available through collaborating scientists. We offer our already well characterised GC-MS technology platforms specifically for co-operations on metabolite identification in complex biological

samples. As suggested by Bino et al. (2004) we envision to share biological samples and metabolite identifications between laboratories engaged in GC-MS metabolite profiling. Thus we provide a public platform for future advances and developments in metabolomic science. In-depth analysis and understanding of metabolome data at systems level will require a multidisciplinary effort, especially integration of proteome and transcriptome data. Such interdisciplinary cooperation and data mining is in preparation and in the case of steady state transcript analysis already in place (Steinhauser et al., 2004). We are convinced that GMD will represent a crucial building block for CSB.DB (http://csbdb.mpimp-golm.mpg.de). CSB.DB, a comprehensive systems-biology database project, will harbour and allow joined access to metabolome, proteome and transcriptome data.

Thus CSB.DB will develop into a highly useful and informative public resource for researchers focusing on experimental biology as well as for computational biology and bioinformatics.

Acknowledgements

We appreciate the work of all scientists, who contributed samples and submitted mass spectral information or metabolite profiling experiments to GMD. Detailed acknowledgements and affiliations are made accessible through GMD (http://csbdb.mpimp-golm.mpg.de/gmd.html). We are grateful to Prof. Mark Stitt and Prof. Lothar Willmitzer, the Max-Planck-Institute of Molecular Plant Physiology and the Max-Planck-Society for long standing and continuous support of the Golm Metabolome Database.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403-410.

Ausloos,P., Clifton,C.L., Lias,S.G., Mikaya,A.I., Stein,S.E., Tchekhovskoi,D.V., Sparkman,O.D., ZaikinV., Zhu,D. (1999). The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287-299.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1462.

Bino,R.J, Hall,R.D, Fiehn,O., Kopka,J., Saito,K., Draper,J., Nikolau,B.J., Mendes,P., Roessner-Tunali,U., Beale,M.H. et al. (2004) Potential of metabolomics as a functional genomics tool. *Trend Plant Sci.* (in press).

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365-370.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. et al. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.*, **29**, 365-371.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.,Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S. et al. (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. *Proc. Natl. Acad. Sci. USA*, **100**, 9232-9237.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207-210.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.*, **18**, 1157-1161.

Fiehn,O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155-171.

Fernie,A.R., Trethewey,R.N., Krotzky,A.J., Willmitzer,L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol*, **5**, 763-769.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. et al. (1996) Life with 6000 Genes. *Science*, **274**, 546-567.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. et al. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94-96.

Gower,J.C. and Legendre,P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.

Hamming,R.W. (1950) Error Detecting and Error Correcting Codes. *Bell System Tech. Journal*, **9**, 147-160.

Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. et al. (2004) The HUPO PSI's molecular interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnol.*, **22**, 177-183.

Jaccard, P. (1908) Nouvelles recherches sur la distribution florale. *Bull Soc. Vaud Sci. Nat.*, **44**, 223-270.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acid Res.*, **32**, D277-280.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662-1664.

Kopka,J., Fernie,A., Weckwerth,W., Gibon,Y. and Stitt,M. (2004). Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acid Res.*, **32**, D438-442.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzttugh,W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.

Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acid Res.*, **32**, D41-44.

Oksman-Caldentey,K.-M., Inzé,D. and Orešič, M (2004). Connecting genes to metabolites by a systems biology approach. *Proc. Natl. Acad. Sci. USA*, **101**, 9949-9950.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763-764.

Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224-228.

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

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acid Res.*, **32**, D431-433.

Stein,S.E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/ mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770-781.

Steinhauser,D, Usadel,B, Luedemann,A, Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* (in press).

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914-939.

Trethewey,R.N. (2004) Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opi. Plant Biol.*, **7**, 196-201.

Wagner,C., Sefkow,M. and Kopka,J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887-900.

Chapter VII - General Discussion & Outlook:
- Systems Biology: From Inside to Outside -

Preface

The work of this PhD thesis was done within the context of the rapidly evolving field of systems biology as the gaining number of references mirror. The following chapter introduces systems biology and outlines the investigations made in conjunction with the trends in systems biology. For discussion of the chapter specific topics the critical reader is refereed to the discussion section of the respective chapter.

Introduction

Biological systems are generally regarded as complex systems, with an enormous potential to adjust intracellular processes in relation to internal and external signals of biotic or abiotic origins (Kitano, 2002a; Oltvai and Barabási, 2002). The organism specific responses generated will be in the range of the 'physiological competences' of the individual species and are influenced by multiple environmental and genetic factors (Trewavas, 2000). Furthermore, these responses are strongly dependent on the status of an organism, such as the developmental stage or the energy status (Trewavas, 2000). The complexity of organization and systems responses increase from prokaryotic to eukaryotic organisms and from a single cell to multicellular organisms. The structural complexity of multicellular organisms is a result of the variation and function of the cells (Esau, 1953 in Edwards and Coruzzi, 1990). The bases of the cellular complexity and their responses are included and executed by the basic sets of cellular elements, namely the genome, the transcriptome, the proteome and the metabolome (Oltvai and Barabási, 2002).

In the past decades with the arising of molecular biology the function of genes or their respective gene products have been mainly studied by approaching single or small sets of cellular elements. Nowadays molecular tools and their application in unravelling biological questions have enabled the discovery of regulatory processes and their basic biological principles (see Buchannan et al., 2000). Despite this highly successfully applied component-centric approach many important biological processes are still unknown or only partially understood. For instance, the observed differences among 'identical' cells from the same tissue (Fricke et al., 1994; Kehr, 1998) can not be explained by the function of a particular gene or more general an element. The surprisingly low estimate of up to 25.000 genes of the human genome (The International Human Genome Sequencing Consortium, 2004) compared to *Arabidopsis thaliana* with approximately 26.000 genes (The Arabidopsis Genome Initiative, 2000) in conjunction with the observed complexity pointed to further more complex processes and mechanisms

despite single gene function. Arising from the demand to unravel complex biological mechanisms the concept of 'systems biology' was re-discovered in the past years (Kitano, 2002b). The renewed interests to look at the entire biological system(s) is currently treated as equivalent to a (r)evolution in biological science (Butler, 2001; Westerhoff and Palsson, 2004). The focus of understanding biological systems at holistic level is not a new field in biology (Kitano, 2002b) and may be one of the important challenges in the *post-genomic* era (Forst, 2002). The movement into systems biology is mainly driven by the availability of whole genome sequences for various organisms (Pennisi, 2003). The annotation of entire genomes facilitates ongoing developments and the maturation of multiplex, high-throughput profiling technology platforms (Ideker et al., 2001; Hood and Perlmutter, 2004). Instead of analyzing single or small sets of cellular elements these technology platforms enable scientists to simultaneously monitor the whole or a vast amount of the cellular compounds, such as transcripts (Lockhart and Winzeler, 2000) or metabolites (Kopka et al., 2004a). First it was thought that high-throughput technology driven analyses will be sufficient for systems level understanding. Currently, the necessity emerges to access the cellular level by integrative approaches (Kitano, 2002b,c; Oltvai and Barabási, 2002). In combination with the ongoing efforts to unravel the molecular basis of organism responses at full systems level a shift in the central paradigm for molecular biology is supposed to be absolutely necessary (Henikoff, 2002; Oltvai and Barabási, 2002; Westerhoff and Palsson, 2004).

Understanding the biological complexity and modelling the cellular systems represents the driving force to move away from component-centric focus to a systems level quest (Nature, 2000; Hwa, 2004). Whereas 'systems biology' is not consentiently defined it represents an analytical approach to unravel the interrelations of the cellular elements of a biological system (Weston and Hood, 2004; Hood and Perlmutter, 2004). The main objective is to offer a comprehensive knowledge backbone for understanding the basic principles of biological systems without abstracting the essential aspects in biology (Kitano, 2002b,c). To understand how a limited number of cellular elements work together and form complex processes research at different areas need to be carried out (Kitano, 2002b; modified):

(1) Genomics and molecular biology research,
(2) Network analysis to unravel the structure, properties and dynamic of biological networks (e.g. transcript or metabolite networks) over time and under various conditions,
(3) Analyses related to robustness and control of biological systems,
(4) Advances in simulation, modelling, algorithm and easily-useable tool development,
(5) Ongoing development of high-precision, high-throughput and comprehensive measurements technologies, and
(6) Databases, allowing public access to data from all cellular levels as well as to statistical and computational results.

Research at these different areas demands cross-disciplinary scientists, expertises from different research activities and specialists in various fields of natural science, e.g. biology, physics, mathematics, chemistry, computer science. Furthermore, high-throughput facilities are required for implementing systems approaches. These requirements represent the most challenging aspects in the evolving field of system biology (Hood and Perlmutter, 2004).

The following part of this thesis will give a brief overview of previous and recent developments and investigations to integrate the different cellular levels for improved analyses and more holistic understanding of biological systems. Requirements will be defined to open up the possibility to understand and successfully predict complex biological processes in conjunction with the recent and future developments of CSB.DB - a comprehensive systems-biology database and in general with our investigations.

CSB.DB - a comprehensive systems-biology database

CSB.DB - a comprehensive systems-biology database - in conjunction with the underlying comprehensive systems-biology project (see Chapter II) focus on the generation of easily accessible knowledge and hypotheses about apparent interactions of elements of the cellular inventory (Steinhauser et al, 2004a, Kopka et al., 2004b; see Chapter II, VI).

Recently, the majority of high-throughput multi-parallel transcript profiling technologies (Lockhart and Winzeler, 2000; see Chapter I) developed to a common and worldwide used tool to uncover biological questions conveying the generation of thousands of transcript profiles. For instance, the AtGenExpress is a multinational coordinated effort to generate a comprehensive set of transcript profiles of *Arabidopsis thaliana* based on the Affymetrix ATH1 full genome chip. The full generated data set comprises samples from various biotic and abiotic treatments and multiple developmental stages, tissues and organs (http://www.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm). The public availability and the continuously increasing amount of expression profiles analyses (Edgar et al., 2002; Gollub et al., 2003; Rhee et al., 2003; AtGenExpress Consortium) initially determine the focus on apparent gene-to-gene interactions. Thus, CSB.DB - a comprehensive systems-biology database was developed to open this information to a broad audience.

In the current build CSB.DB integrates two built-in modules, namely the CoR (Steinhauser et al., 2004a) and the GMD module (Kopka et al., 2004b). Whereas the CoR module allows access to transcriptional co-responses of various key model organisms (see Chapter II) the GMD module harbours information regarding to metabolite measurements, metabolite identification as well as protocols (Chapter IV). For further developments of CSB.DB into a highly useful and informative public resource for researchers from various scientific fields investigations must be made regarding different areas. The basic levels of the cellular inventory and the interrelation among and within these

levels represent the points of origin for our future improvements. These improvements and novel investigations will focus on:

(1) Collection of results from component-centric approach,
(2) Transcriptome analyses to allow access to transcript level measurements over time, under various growth and treatment conditions as well as developmental stages,
(3) Proteome analyses to supplement the CoR and GMD modules,
(4) Metabolite profile analyses and compound identification effort to extend and improve the GMD module in parallel to the transcript profile module
(5) Co-Response analyses in conjunction with promoter and motif analyses, and
(6) Combined analyses among the different cellular elements and levels.

'Component Centric' driven Systems Biology – A Contradiction?

The evolution of systems biology is characterized by a trend from component-centric to a systems level quest approach (Hwa, 2004). Whereas this shift is important to decipher complex biological questions a component-centric research delivers basic information regarding pieces of the whole, e.g. promoter activities or enzymatic properties of proteins. Gathering of information from component-centric driven research is important for various reasons:

(1) describing the properties and characteristics of particular elements of the cellular inventory,
(2) assisting, i.e. validate or invalidate, results of high-throughput assays (e.g. metabolic profiling, microarray analyses) for particular elements and therefore allow reassigning a new quality (independently confirmed) to an element, and
(3) providing essential information or hypotheses for system quest driven modelling and validation.

Therefore, collecting and providing access to data from component-centric research does not represent a contradiction to systems level quest. Recently, various databases provide access to enzymatic properties, protein characteristics and sequence information derived from component-centric research, multinational annotation projects or literature (e.g. Schomburg et al., 2004; Karp et al., 2002a,b; Wheeler et al., 2004). Despite these various activities fewer attempts have been made to gather component-centric data regarding to e.g. protein localization or promoter activity. Such efforts were accomplished for yeast with the main emphasis to provide access to protein localization data (http://yeastgfp.ucsf.edu/; Huh et al., 2003; Ghaemmaghami et al., 2003) but seem to be lacking for plants. According to the aforementioned reasons the public access to component-centric experimental results is essential for integrative systems biology approaches and represents one of the future challenges of CSB.DB. In analogy to the (suggested) standardization of transcript [MIAME, (Brazma et al., 2001)] or metabolite [MIAMET, (Bino et al., 2004)] profile analyses integration of component-centric data require standardization of the experimental parameters. The standardization effort will be not trivial, requires strong interaction of computer and experimentally focused scientists and will be a

long-term goal. Furthermore, novel tools for an efficient search will be necessary. One of the possible solutions may be 'text-mining' in conjunction with manual curation efforts (i) to group data according to their contexts and (ii) to allow a context-dependent search of the harboured data.

CSB.DB: From Transcriptome to Metabolome

Transcriptome: In the current build of CSB.DB the main emphasis of transcriptional analyses focuses on apparent gene-to-gene interrelation. Despite the possibility to invoke various bi-plots (see Chapter II) we have made fewer attempts to usually applied transcript profile analyses, such as up-/down regulation under particular condition because various databases and public resources allow access to those results, like TAIR (Rhee et al., 2003) or the Stanford Microarray Database (Gollub et al., 2003). Despite this highly competitive field basic analyses of changed transcript level are necessary for improved validation of observed gene-to-gene interrelations. Our future investigation will be directed towards single genes and set of genes regarding to (i) visualization and (ii) statistical analyses of the measured transcript levels. First investigations, mainly initiated by B.Usadel, focus on false-colour code visualization of expression levels on diagrams. MapExpress, a MapMan-based web-application (Thimm et al., 2004), superimposes graphical maps with observed expression levels and allows easy overview across various transcript profiles for favourite genes.

Proteome: As proteins are generally regarded to determine cellular function the exploitation of the proteome is one of our future prime interests. The expression of a protein-encoding gene does not need to necessarily result in a functional protein and protein function. The path [gene – transcript – protein – protein function] is controlled by complex regulatory processes (see Taiz and Zeiger, 2000). For instance, Gibon et al., (2004) revealed for various enzymes that changes in transcript levels typically led to strongly damped changes of the enzyme activity. Hence, understanding gene functions requires deciphering the interrelation of transcripts and proteins as well as understanding post-transcriptional and post-translational mechanisms. Analyses regarding to the proteome or joint analyses of transcript and proteins have been applied to few organisms, for instance to *Escherichia coli* (Yoon et al., 2003; Corbin et al., 2003) and *Saccharomyces cerevisiae* (Gygi et al., 1999). Recently, the limited number of available proteome profile data restricts those analyses. Novel investigations and developments in proteomics, such as mass-spectrometry-based proteomics or protein microarrays (Pandey and Mann, 2000; Aebersold and Mann, 2003), will lead to a rapid increase of available proteome profiles. These profiles will be targets for investigations regarding protein relationships or protein-transcript interrelations.

Metabolome: Metabolites represent the end products of various cellular processes and can be signals for further responses. Their levels mirror responses of biological systems to environmental or genetic perturbations. Recently, the substantial progress made in metabolomics by gas chromatography combined with mass spectrometry (GC-MS) allow quantitative determination of hundreds of known or unknown metabolites (Fiehn et al., 2000; Roessner et al., 2001a,b, Kopka et al., 2004b). In the current

build of the GMD module we have started to integrate profiles obtained from quadrupole GC-MS technology platform (Kopka et al., 2004a). The profile technology platform will be frequently updated with new metabolite profiles from the quadruple GC-MS platform as well extended to GC-MS-TOF (time-of-flight) platform.

Co-Responses: Context and Causality

The first step to gain insight into understanding biological systems at entire level has been done with the implementation of the transcriptional co-response databases. By scanning for best co-responses among changing transcript levels the transcriptional co-response databases allow to infer hypotheses on functional interaction of genes. The basic assumption underlying this analysis is that common transcriptional control of genes is reflected in co-responding, synchronous changes in steady-state transcript levels. To investigate in apparent gene-to-gene interaction we are currently using publicly available transcript profiles generated on two-colour cDNA or oligonucleotide technology platforms (Lockhart and Winzeler, 2000; see Chapter I). Most of the underlying experiments are based on steady-state transcript measurements. The steady-state level of a transcript depends on the rate of synthesis (transcription) and degradation, which is influenced by the transcript stability. Synthesis and degradation can be transcript specific and therefore, can influence the co-response. For instance, stability of transcripts and expression in different time frames can mimic co-responding, synchronous changes in transcript levels which can lead to significant co-responses. Moreover, microarray measurements are noisy which can lead to imprecise transcript and resulting co-response measurements. Although these facts can influence the strength of a particular co-response, in general this analysis allows inferring precise hypotheses as demonstrated (Chapter III, IV, and V). Moreover, the ever growing amount of transcript profiles generated used to study the influence of factors by time-series analyses will help to address the aforementioned limitations.

In co-response analyses we want to determine whether two variables, such as transcripts, are interdependent or covary, i.e. vary together. A typical statistical assumption of co-response analyses is that both variables are effects of a common cause. The computed coefficient reflects the strength of co-response among two variables, but the common cause still remains unknown. To gain insight into unravelling causality variables can be grouped according their co-response behaviour (see Chapter I). Based on our initial assumption, that common transcriptional control is reflected in co-responding transcript changes, we can use these analyses to extract common motifs within promoter elements. Such analyses have been successfully accomplished for few key model organism, such as *Escherichia coli* (Pilpel et al., 2001; Shen-Orr et al., 2002) and *Saccharomyces cerevisiae* (Segal et al., 2003). Deciphering the transcriptional control in *Arabidopsis thaliana* will be one of our future focuses, which will include joint promoter and co-response analyses as it was shown earlier (e.g. Segal et al., 2003). In addition, joint analyses allow us to elucidate (overlapping) functional modules and therefore may lead to a better understanding of gene function. The analyses of transcription units in *E.coli*

revealed context dependent promoters usage, gene expression and transcript co-response (see Chapter III). Currently a project is running to compare co-responses across conditional data matrices to contextualize co-responses of *Arabidopsis thaliana*. Such analyses may help to better understand the stability of co-responses as well as the dynamics of networks.

The 'Cellular Inventory' Project

The substantial developments and maturation of recent multi-parallel high-throughput assays in conjunction with the dawn of genomic technology will encourage us to take the understanding of biological systems one step further (see Chapter I, and VI). To overcome the evidently contrary viewpoints of assigning a single given gene function and unravelling control points in complex biological processes the flood of information from the different technologies needs to be made publicly available to all researchers without demanding bioinformatics expertise. Combined and integrative analyses of data derived from all levels of the cellular inventory are required to understand biological systems from an entire point of view (Kitano, 2002a-c, Oltvai and Barabási, 2002). Data from the same biological sample needs to be collected and analysed to gain insight into deciphering the interrelation among and within these cellular levels.

The 'Cellular Inventory' project is a joint and multidisciplinary effort aimed at addressing these complex biological questions. It is a united initiative of CSB.DB (see chapter II) and GMD (see chapter VI) and involved scientist from the field of experimental and computational research as well as cross-disciplinary scientists. The main emphasis is directed towards paired analyses of the transcriptome and the metabolome of biological samples. Whereas various attempts have been made to combine transcript and protein data (Gygi et al., 1999; Yoon et al., 2003; Corbin et al., 2003), only few studies successfully integrated transcript and metabolite measurements to extract biological meaningful information. Askenazi et al. (2003) applied an integrated approach for improved fungal strain engineering and correlated transcript and metabolite measurements. Similar to this approach Urbanczyk-Wochniak et al. (2003) applied transcript-to-metabolite correlations for inferring complex hypotheses on multi-parallel high-throughput measurements of *Solanum tuberosum*.

Within the framework of the cellular inventory project we investigate in transcript-metabolite and metabolite-metabolite co-response of plant organisms as described earlier (Urbanczyk-Wochniak et al., 2003). Thus, a step further into better understanding of biological systems may be done by extension the transcriptional co-responses towards metabolite or transcript-metabolite co-responses. The bases for such investigations are represented by the particular databases harbouring profile measurements of the different cellular elements (see above).

In conjunction with this project and the possible extension to proteomics we may able to investigate in comprehensive network analyses within and among the cellular levels. Beyond it, further investigations regarding flux analyses may open up the possibility to modelling pathways at holistic level.

From Bacteria to Plants

CSB.DB - a comprehensive systems-biology database - is an open access tool which may serve as the basis for more sophisticated means of elucidating gene function. Once developed, it was indented to apply the implemented algorithms and developed tools (e.g. cCoRv1.0; Steinhauser et al., unpublished) to proof the concept, that common transcriptional control is reflected in synchronous changes of transcript levels. To investigate, we selected the prokaryotic operon structure of the well characterized key model organism *Escherichia coli* (Steinhauser et al., 2004b; Chapter III). Because, genes which are co-regulated in physical units of common polycistronic messenger RNA (mRNA) can be expected to reveal high correlations in transcriptome analyses. High correlations should be observed independent of the nature of underlying biological experiments (see Chapter III). In conjunction with the work of Bockhorst et al. (2003a,b) and Yamanishi et al. (2003) we found that prediction accuracy of transcriptional analyses can be significantly increased by using additional information, such as genomic information (e.g. sequence data, intergenic distances [Bockhorst et al., 2003a; Yamanishi et al., 2003; Steinhauser et al, 2004b]) or complex models (Bockhorst et al., 2003b).

Using biological facile organisms to infer hypotheses

According to the aforementioned observation we can conclude, that learning from the facile can help to better approach complex organisms (see Chapter IV, V). Beyond it, most of the basic functional assignments of gene discovered by genome sequencing projects have been realised by annotation transfer from homologous sequences (McGeoch and Davidson, 1986; Bork and Gibson, 1996; Bork et al., 1998). Many orthologous genes can be found across species and often informational pathways in biological facile organisms are similar to those in complex organisms. Furthermore, Ueda et al. (2004) have shown that gene expression dynamics follows the same principle from bacteria to human. The observed principle, namely that gene expression changes are proportional to their expression levels, regenerate the complexity and dynamic organization of the transcriptome of various phyla (Ueda et al., 2004). Analyses of gene co-expression networks of human, yeast, worm and fly deciphered the existence of more than 22,000 conserved genetic modules (Stuart et al., 2003). These modules are characterized by pairs of genes whose expression is significantly correlated in multiple organisms. Stuart and co-worker (2003) have shown that multiple-species networks enabled discarding spurious gene interrelations from functional relevant associations. Thus, it is feasible to use genetically and biologically facile organisms to infer hypotheses on gene function and functional interrelation of genes from higher, more complex organisms. In contrast, comparing of functions or interrelation among different phyla may enable fundamental insights into mechanisms of evolution or physiology, e.g. as it was shown for the TCA cycle (Forst, 2002).

From biological facile to complex organisms

The using of biological facile organisms and comparisons across various phyla (see above) open up the possibility to extract and characterize universal and general biological processes and phenomena. Despite these opportunities various processes or genes are organism specific and recently, can not be approached efficiently by global techniques.

A widespread phenomenon in eukaryotic organisms is the coordinated expression of genes with common function (DeRisi et al., 1997; Eisen et al., 1998; Niehrs and Pollet, 1999). These sets of genes, called synexpressed groups, showing parallels to operons and may enable to understand the evolutionary changes to eukaryotic diversity (Niehrs and Pollet, 1999). Despite this, the transcriptional co-responses among genes transcribed as polycistronic mRNA (operons) are strongly overlapped with non-operon genes. As we have shown, prediction accuracy for genes transcribed as polycistronic mRNA (operons) can be significantly increased by use of additional information, i.e. intergenic distances between the genes (Chapter III). The choice of selecting the appropriate additional information depends on the biological question which scientists have and the availability of such data or background knowledge. In a first collaboration regarding the identification of brassinosteroid-related genes (Lisso et al., 2004) we queried for transcriptional co-responses with the BR-signalling components BRI1 and BAK1 (chapter IV) instead of using additional information. This approach strongly reduced the list of candidate genes which share common co-response to both and enabled the identification and confirmation of 72 genes showing BR-dependent expression. Our second collaboration focused on basic functional assignment of *Arabidopsis* subtilase genes. The lack of additional information as well as of anchor points, i.e. genes, restricts the general application of the above-mentioned approaches. Classification based on the co-response based functional neighbourhood enabled us to assign a basic function to the ubiquitous expressed subtilase genes (Chapter V). In contrast to the successful applications there are some pitfalls. For instance, the assumption that transcription factors exhibit a strong co-response to their targets may lead to a wrong set of possible regulated genes. In general, we can assume that regulator and target genes may not be correlated in steady-state expression profiles. To overcome this limitation Segal et al. (2003) described and successfully applied an iterative approach to identify regulatory modules and their regulators to yeast expression data. They identified 50 functional modules and described novel regulators by simultaneous partitioning of genes into modules and identification of the co-responding regulation program. Furthermore, transcript stability and degradation can mimic or mask biological meaningful co-responses. Genes regulated by common *cis*-elements need not necessarily reflect strong co-responses based on steady-state transcript measurement if the transcript stability different.

Despite the entire factor influencing transcriptional co-responses our applied method in conjunction with the developed web platform CSB.DB enabled easy access to large-scale statistical analyses of apparent gene-to-gene interactions. Furthermore, it allows inferring hypotheses for characterized as

well as uncharacterized gene which can not be accessed by sequence homology. These hypotheses can be successfully tested as shown in the chapters III-V.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.

Askenazi,M., Driggers,E.M., Holtzman,D.A., Norman,T.C., Iverson,S., Zimmer,D.P., Boers,M.E., Blomquist,P.R., Martinez,E.J., Monreal,A.W. et al. (2003) Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.*, **21**, 150-156.

Bino,R.J, Hall,R.D, Fiehn,O., Kopka,J., Saito,K., Draper,J., Nikolau,B.J., Mendes,P., Roessner-Tunali,U., Beale,M.H. et al. (2004) Potential of metabolomics as a functional genomics tool. *Trend Plant Sci.*, (in press).

Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F.R. and Craven,M. (2003a) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34-i43.

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003b) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227-1235.

Bork,P and Gibson,T.J. (1996) Applying motif and profile searches. *Methods Enzymol.*, **266**, 162-184.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting Function: From Genes to Genomes and Back. *J. Mol. Biol.*, **282**, 707-725.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. et al. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.*, **29**, 365-371.

Buchannan,B.B., Gruissem,W. and Jones,R.L. (2000) Biochemistry & Molecular Biology of Plants. American Society of Plant Physiologists (ASPP), Maryland.

Butler,D. (2001) Genomics. Are you ready for the revolution? *Nature*, **409**, 758-760.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S., Soupene,E. and Hunt,D.F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl. Acad. Sci. USA*, **100**, 9232-9237.

DeRisi,J.L., Vishwanath,R.I. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on genomic scale. *Science*, **278**, 680-686.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207-210.

Edwards,J.W. and Coruzzi,G.M. (1990) Cell-specific gene expression in plants. *Annu. Rev. Genet.*, **24**, 275-303.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.*, **18**, 1157-1161.

Forst,C.V. (2002) Network genomics – A novel approach for the analysis of biological systems in the post-genomic era. *Mol. Biol. Rep.*, **29**, 265-280.

Fricke,W., Pritchard,J., Leigh,R. and Tomos,D. (1994) Cells of the upper and the lower epidermis of barley (*Hordeum vulgare* L.) leaves exhibit distinct patterns of vacuolar solutes. *Plant Physiol.*, **104**, 1201-1208.

Gibon,Y., Blaesing,O.E., Hannemann,J., Carillo,P., Höhne,M., Hendriks,J.H.M., Palcios,N., Cross,J., Selbig,J. and Stitt,M. (2004) A robot-based platform to measure multiple enzymes activities in *Arabidopsis* using a set of cycling assays: comparison of changes of enzymes activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell*, (in press).

Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737-741.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. et al. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94-96.

Gygi,S.P., Rochon,Y., Franza,B.R. and Aebersold,R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, **19**, 1720-1730.

Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686-691.

Hwa,T. (2004) New Insights from a Classical System. *Science*, **305**, 345.

Henikoff,S. (2002) Beyond the Central Dogma. *Bioinformatics*, **18**, 223-225.

Hood,L. and Perlmutter,R.M. (2004) The impact of systems approaches on biological problems in drug discovery. *Nature Biotechnol.*, **22**, 1215-1217.

Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decoding life: systems biology. *Ann. Rev. Genomics Hum. Genet.*, **2**, 343-372.

Karp,P.D., Riley,M., Paley,S.M. and Pellegrini-Toole,A. (2002a) The MetaCyc Database. *Nucleic Acids Res.*, **30**, 59-61.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002b) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56-58.

Kehr,J. (1998) Mikroanalyse einzelner Zellen und Kompartimente transgener Pflanzen mittels biophysikalischer Methoden. Dissertation, Universität Potsdam.

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

Kitano,H. (2002b) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet.*, **41**, 1-10.

Kitano,H. (2002c) Systems Biology: A Brief Overview. *Science*, **295**, 1662-1664.

Kopka,J., Fernie,A., Weckwerth,W., Gibon,Y. and Stitt,M. (2004a). Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

Kopka,J., Schauer,N., Krueger,S., Birkemeyer,C., Usadel,B., Bergmüller,E., Doermann,P., Gibon,Y., Stitt,M., Willmitzer,L., Fernie,A.R. and Steinhauser,D. (2004b) GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics*, (submitted).

Lisso,J., Steinhauser,D., Altmann,T., Kopka,J. and Müssig,C. (2004) Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Plant Physiol.*, (submitted).

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.

McGeoch,D.J. and Davison,A.J. (1986) Alphaherpesviruses posses a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. *Nucleic Acids Res.*, **14**, 1765-1777.

Nature (2002) Systems biology's multiple maths. *Nature*, **407**, 819.

Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483-487.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763-764.

Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837-846.

Pennisi, E. (2003) Systems biology. Tracing life's circuitry. *Science*, *302*, 1646-1649.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153-159.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224-228.

Roessner,U., Willmitzer,L., and Fernie,A.R. (2001a). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.*, **127**, 749-764.

Roessner,U., Luedemann,A., Brust,D., Fiehn,O., Linke,T., Willmitzer,L. and Fernie,A. (2001b) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11-29.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acid Res.*, **32**, D431-433.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166-176.

Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64-68.

Steinhauser,D, Usadel,B, Luedemann,A, Thimm,O. and Kopka,J. (2004a) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* (in press).

Steinhauser,D., Junker,B.H., Luedemann,A., Selbig,J. and Kopka,J. (2004b) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, **20**, 1928-1939.

Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249-255.

Taiz,L. and Zeiger,E. (2000) Physiologie der Pflanzen. Spektrum Akademischer Verlag GmbH Heidelberg – Berlin.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

The International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914-939.

Trewavas,A. (2000) Signal Perception and Transduction. In Biochemistry & Molecular Biology of Plants, Buchannan,B.B., Gruissem,W. and Jones,R.L, ed. American Society of Plant Physiologists (ASPP), Maryland, 930-987.

Ueda,H.R., Hayashi,S., Matsuyama,S., Yomo,T., Hashimoto,S., Kay,S.A., Hogenesch,J.B. and Iino,M. (2004) Universality and flexibility in gene expression from bacteria to human. *Proc. Natl. Acad. Sci. USA*, **101**, 3765-3769.

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

Westerhoff,H.V. and Palsson,B.O. (2004) The evolution of molecular biology into systems biology. *Nature Biotechnol.*, **22**, 1249-1252.

Weston,A.D. and Hood,L. (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J. Proteome Res.*, **3**, 179-196.

Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35-40.

Yamanishi,Y., Vert,J.-P., Nakaya,A. and Kanehisa,M. (2003) Extraction of correlated gene clusters from multiple genomic data by generalised kernel canonical correlation analysis. *Bioinformatics*, **19**, i323-i330.

Yoon,S.H., Han,M.J., Lee,S.Y, Jeong,K.J. and Yoo,J.S. (2003) Combined transcriptome and proteome analysis of *Escherichia coli* during high cell density culture. *Biotechnol. Bioeng.*, **81**, 753-767.

Deutsche Zusammenfassung

Die vergangenen Jahrzehnte waren gekennzeichnet durch umfangreiche Bemühungen, die Genomsequenz verschiedener Organismen vollständig zu entschlüsseln. Die Verfügbarkeit vollständiger genomischer Daten löste die Entwicklung von modernen Hochdurchsatzmethoden aus, welche die gleichzeitige Messung von verschiedenen Transkripten, Proteinen und Metaboliten erlauben. Mittels genomischer Informationen und Hochdurchsatztechnologien erlaubt eine hoch parallelisierte experimentelle Biologie die Erforschung von Gesetzmäßigkeiten, welchen biologischen Systemen zugrunde liegen. Das Verständnis biologischer Komplexität durch Modellierung zellulärer Systeme repräsentiert die treibende Kraft, welche heutzutage den Element-zentrierten Focus auf integrative und ganzheitliche Untersuchungen lenkt. Das sich entwickelnde Feld der Systembiologie integriert Entdeckungs- und Hypothesen-getriebene Wissenschaft um ein umfangreiches Wissen durch Computermodelle biologischer Systeme bereitzustellen.

Im Kontext der sich neu entwickelnden Systembiologie investierte ich in umfangreiche Computeranalysen zur Transkript Co-Response bezüglich ausgewählter prokaryotischer und pflanzlicher eukaryotischer Organismen. CSB.DB - *a comprehensive systems-biology database* - (http://csbdb.mpimp-golm.mpg.de/) wurde initiiert, um freien Zugang zu den biostatistischen Ergebnissen als auch zu weiterem biologischem Wissen zu bieten. Die Datenbank CSB.DB ermöglicht potentiellen Anwendern die Hypothesengenerierung bezüglich der funktionalen Wechselbeziehungen von Genen von Interesse und kann zukünftig die Grundlage für einen fortgeschrittenen Weg der Zuordnung von Genfunktionen darstellen. Unter Verwendung chromosomaler Distanzen und Transkript Co-Response konnte das Konzept und CSB.DB angewandt werden, um bakterielle Operons in *Escherichia coli* erfolgreich vorherzusagen. Darüber hinaus werden Beispiele gezeigt, die andeuten, dass die Transkript Co-Response Analyse eine Identifizierung differentieller Promoteraktivität in verschiedenen experimentellen Bedingungen ermöglicht. Das Co-Response Konzept wurde, mit dem Schwerpunkt auf die eukaryotische Modellpflanze *Arabidopsis thaliana*, erfolgreich auf komplexere Organismen angewandt. Die durchgeführten Untersuchungen ermöglichten die Identifizierung neuer Gene hinsichtlich physiologischer Prozesse und darüber hinaus die Zuweisung von Genfunktionen, welche nicht durch Sequenzhomologie ermöglicht werden kann. GMD - *The Golm Metabolome Database* - wurde initiiert und in CSB.DB implementiert, um Metaboliten Informationen als auch Metaboliten Profile zu integrieren. Dieses neue Modul ermöglicht die Ausrichtung auf komplexere biologische Fragen und erweitert die derzeitige systembiologische Fragestellung in Richtung Phänotypus-Zuordnung.

Acknowledgements

First and foremost I would like to thank Prof. Dr. Lothar Willmitzer and Prof. Dr. Mark Stitt for giving me the possibility to do my PhD project in the Max Planck Institute of Molecular Plant Physiology and for long standing and continuous support of CSB.DB - a comprehensive systems-biology database as well as my work.

I would particularly like to thank Dr. Joachim Kopka for all his helpful supervision in discussing experiments and in proof reading of all my written work.

Then I would like to thank Prof. Dr. Joachim Selbig (Potsdam), Prof. Dr. Dierk Scheel (Halle), and Prof. Dr. Uwe Sonnewald (Gatersleben) for agreeing to examine my thesis.

Further I wish to thank Björn Usadel for his continuous collaboration regarding CSB.DB, the criticism, our 'disagreements', and… He deserves special thanks for his permanent moral support and his friendship.

I would thank all my colleagues in the 'lab', i.e. office, for their support. I am especially grateful to Alexander Luedemann for his help in educating me PERL programming and his patience with my initial 'DAU' trait.

I am especially grateful to Dr. Dirk Büssis, Dr. Carsten Müssig, and Dr. Leonard Krall for their critical discussion and proof reading my written work.

I am indebted to the MPI Infrastructure groups for their invaluable support, especially to Eckhard Simmat, Wilfried Grauholz, and Carsten Bochan to take care of my electronic bacteria and plants.

Last but not least I thank all people from the MPI for always supporting me and my work and for making work enjoyable in the last three years, especially Amelie, Annette, Berit, Björn, Björn, Carsten, Carsten, Claudia, Diana, Dirk, Jelena, Len, Maren, Oliver, Oliver, Regina, Regina, Stephanie, Stephan, Yves…

Curriculum vitae

Name:	Dirk Steinhauser
Date of Birth:	October, 9th, 1974
Place of Birth:	Zossen
Nationality:	German
Marital Status:	divorced

1981 - 1990	Primary School: Polytechnische Oberschule (POS), Baruth / Mark
1990 - 1993	Secondary School: Erweiterte Oberschule Ludwigsfelde
06/1993	General certificate of education: A-level (Abitur)
10/1993 - 09/1995	Instructor and combat troop leader in the German Federal Armed Forces
10/1995 - 03/1996	Unemployment
03/1996 - 09/1996	Practical course and free-lancing job in project work
10/1996 - 08/2001	Studies in biology at the University of Potsdam
10/2000 - 07/2001	Max Planck Institute of Molecular Plant Physiology, Golm, Germany Diploma thesis under the supervision of Dr. J. Fisahn 'Molekularbiologische und elektrophysiologische Charakterisierung am Saccharosetransport beteiligter Carrier auf Einzelzellebene'
08/2001	Diploma: 'Diplom-Biologe' in Biology, specialization Physiology and Biochemistry
08/2001 - 09/2001	Free-lancing job in project work
10/2001 - 09/2004	Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany PhD thesis under supervision of Prof. Dr. L. Willmitzer and Dr. J. Kopka 'Inferring Hypotheses from Complex Profile Data - By Means of CSB.DB, A Comprehensive Systems-Biology Database -'
since 10/ 2004	Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany Scientist in the group of Prof. Dr. L. Willmitzer

List of Publications

Articles (press):

Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* (in press).

Steinhauser,D., Junker,B.H., Luedemann,A., Selbig,J. and Kopka,J. (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, **20**, 1928-1939.

Steinhauser,D. (2002) Untersuchungen zur Ökologie der Mopsfledermaus, *Barbastella barbastellus* (Schreber, 1774), und der Bechsteinfledermaus, *Myotis bechsteinii* (Kuhl, 1817), im Süden des Landes Brandenburg. *Schriftenr. Landschaftspflege und Naturschutz*, **71**, 81-98.

Steinhauser,D. (1999) Erstnachweis einer Wochenstube der Nordfledermaus (*Eptesicus nilssonii*) im Land Brandenburg mit Hinweisen zur Ökologie dieser Fledermausart. *Nyctalus (N.F.)*, **7**, 208-211.

Steinhauser,D. (1999) Erstnachweis einer Wochenstube der Bechsteinfledermaus (*Myotis bechsteinii*) im Land Brandenburg. *Nyctalus (N.F.)*, **7**, 229.

Steinhauser,D. (1997) Systematische Erfassungen zur Nordfledermaus, *Eptesicus nilssoni* (Keyserling u. Blasius, 1839), im Süden des Landes Brandenburg - Reproduktionsnachweis. *Nyctalus (N.F.)*, **6**, 375-389.

Haensel,J., Arnold,D., and **Steinhauser,D.** (1994) Vorkommen der Nordfledermaus (*Eptesicus nilssoni*) am Rande des Baruther Urstromtales - Bestätigung durch Lebendfund! *Nyctalus (N.F.)*, **5**, 213-217.

Articles (submission):

Kopka,J., Schauer,N., Krueger,S., Birkemeyer,C., Usadel,B., Bergmüller,E., Doermann,P., Gibon,Y., Stitt,M., Willmitzer,L., Fernie,A.R. and **Steinhauser,D.** (2004) GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics*, (submitted).

Lisso,J., **Steinhauser,D.**, Altmann,T., Kopka,J. and Müssig,C. (2004) Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Plant Physiol.*, (submitted).

Rautengarten,C., **Steinhauser,D.**, Stinitzi,A., Büssis,D., Schaller,A., Kopka,J. and Altmann,T. (2004) Inferring Hypotheses For Gene Functions: The *Arabidopsis thaliana* Subtilase Gene Family. *Plant Physiol.*, (submitted).

Junker, B.H., Wuttke, R., **Steinhauser, D.**, Büssis, D., Willmitzer, L. and Fernie, A.R. (2004) Expression of an invertase in the vacuole of potato tubers provides circumstantial evidence for the endocytotic trafficking of sucrose between the apoplast and vacuole. *Planta*, (submitted).

Book chapters:

Desbrosses,G., **Steinhauser,D.**, Kopka,J., and Udvardi,M.K. Lotus metabolome analysis using GC-MS. In Lotus japonicus handbook, A.J. Márquez, ed (Dordrecht: Kluwer; in press).

Proposals:

Steinhauser,D., Krüger,S., Birkemeyer,C. and Kopka,J. (2003) Integrative systems biology approach for analysis of metabolic limits of root growth and genetic engineering to enhance plant growth and plant fitness. GABI-II proposal.

Talks:

12.06.2004: 'CSB.DB - a Comprehensive Systems-Biology Database-'. In. Workshop 'AtGenExpress', Arabidopsis 2004 Meeting, 11. - 14.07.2004, Berlin, Germany.

21.04.2004: 'CSB.DB - a Comprehensive Systems-Biology Database-'. Munich Information Center for Proteins Sequences (MIPS-GSF), Munich, Germany.

24.03.2004: 'CSB.DB - a Comprehensive Systems-Biology Database-'. Institute for Biochemistry, Cologne, Germany.

05.09.1997: 'Telemetrische Untersuchungen zur Biologie und Ökologie der Mopsfledermaus, *Barbastella barbastellus* (Schreber, 1774), im Süden des Landes Brandenburg'. In. International Workshop ‚*Zur Situation der Mopsfledermaus in Europa*', 05. - 07.09.1997, Mansfeld, Germany.

Funding & Projects:

1997: Subproject ‚Telemetrische Untersuchungen zur Biologie und Ökologie der Mopsfledermaus, *Barbastella barbastellus* (Schreber, 1774), im Süden des Landes Brandenburg' as part of the ‚F+E-Vorhaben: Schutz und Erhaltung von Fledermäusen in Wäldern' funded by the Bundesamt für Naturschutz (BfN).

1998: Subproject ‚Telemetrische Untersuchungen zur Biologie und Ökologie der Bechsteinfledermaus, *Myotis bechsteinii* (Kuhl, 1817), im Süden des Landes Brandenburg' as part of the ‚F+E-Vorhaben: Schutz und Erhaltung von Fledermäusen in Wäldern' funded by the Bundesamt für Naturschutz (BfN).

2001: Untersuchungen zum Status der Mopsfledermaus (*Barbastella barbastellus*), Bechsteinfledermaus (*Myotis bechsteinii*) und Teichfledermaus (*Myotis dasycneme*) im Land Brandenburg by a grant of the Ministerium für Ländliche Entwicklung, Umwelt und Verbraucherschutz (MLUV) of the federal state Brandenburg.

Appendix

Further information and supplemented material is available from the CSB.DB Homepage:
http://csbdb.mpimp-golm.mpg.de.