

Lehrstuhl für Genomorientierte Bioinformatik
der
Technischen Universität München

A Bioinformatic Approach to the Metabolic and Functional Analysis of Biological High-Throughput Data

Matthias Fellenberg

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum
Weihenstephan für Ernährung, Landnutzung und Umwelt der
Technischen Universität München zur Erlangung des akademischen
Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Arne Skerra

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Hans-Werner Mewes
2. Univ.-Prof. Dr. Alfons Gierl
3. Univ.-Prof. Dr. Ralf Zimmer,
Ludwig-Maximilians-Universität München

Die Dissertation wurde am 14.03.2002 bei der Technischen
Universität München eingereicht und durch die Fakultät
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung
und Umwelt am 29.10.2002 angenommen.

Contents

Abstract	6
1 Introduction	8
1.1 Trends in Modern Molecular Biology	9
1.2 Trends in Bioinformatics	11
1.3 Approach to an Integrative Data Analysis	13
1.4 Acknowledgements	13
2 High-Throughput and other Biological Data	15
2.1 Gene Expression Data	17
2.1.1 Introduction	17
2.1.2 Aims of Gene Expression Analysis	18
2.1.3 Biotechnological Analysis Methods	19
2.1.4 Variations of Expression Rates	21
2.1.5 From Arrays and Chips to Data Matrices	22
2.1.6 Normalization	23
2.1.7 A Statistical Normalization Procedure	25
2.1.8 Gene Clustering	25
2.1.9 A SOM Approach to Gene Clustering	30
2.1.10 A Case Study: Human Heart Tissue	33

2.2	Protein-Protein Interactions	41
2.2.1	MIPS Yeast Interaction Tables	42
2.2.2	Visualization of Protein-Protein Interactions	42
2.3	Functional Annotations: The MIPS Functional Catalog	45
2.4	Metabolic Pathways	47
2.4.1	Introduction	47
2.4.2	Metabolic Databases and Resources	47
2.5	Dynamic Modeling of Metabolic Pathways	53
2.5.1	A Dataset for Dynamic Pathway Modeling	54
2.5.2	Metabolic Graphs	55
2.5.3	Computing Metabolic Networks	56
2.5.4	Linear Path Search	58
2.5.5	Constraint-based Pathway Construction	62
3	Integrative Data Analysis	69
3.1	Integrating Gene Expression Data with Functional Annotations	70
3.1.1	Methods	70
3.1.2	Results	74
3.2	Integrating Gene Expression Data with Metabolic Pathways	81
3.2.1	Pathway Projection	81
3.2.2	Metabolic Mapping of Gene Clusters	86
3.2.3	Discussion	94
3.3	Discussion of Functional Projection, Pathway Projection and Metabolic Mapping	97
3.4	Integrating Protein-Protein Interactions with Functional An- notations	99
3.4.1	Methods	99
3.4.2	Verification	100

3.4.3	Revealing the Functional Context of Uncharacterized Proteins	103
3.4.4	Discussion	107
4	The Software System Architecture	109
4.1	The <i>AMPhora</i> Pathway Modeling Application	111
4.2	<i>BioTalk</i> : Representing Pathways in XML	113
4.3	The <i>MapClusterer</i> application	114
4.4	The Graphical User Interface	114
4.4.1	GUI for gene expression analysis	114
4.4.2	GUI for <i>AMPhora</i> pathway modeling	118
5	Discussion	119
5.1	The approach to an integrative analysis	119
5.2	Related work	122
5.3	Relevance of the approach	123
A	Acronyms	125
	Bibliography	130

Abstract

This thesis describes the approach to a functional and metabolic analysis of biological whole-genome data. It focusses on three fields of biotechnology and bioinformatics: gene expression, protein-protein interactions and metabolic pathways. For each of them, an introduction is given that describes potential benefits of the technologies and that highlights the computational challenges that arise from their analysis. In this thesis, computational analysis methods for the respective data sets have been developed. These bioinformatic methods, the SOM clustering for gene expression data, the graph modeling of protein-protein interactions and the three methods for a dynamic modeling of metabolic pathways prepare the ground for the developed integrative methods.

For the further analysis and interpretation of the high-throughput data sets, a knowledge-based *integrative analysis approach* has been elaborated. The developed combinatorial and integrative analysis methods make use of existent knowledge in order to achieve qualitative and reliable results. Data sets are analyzed in the context of systematic, previously assembled facts, leading to a more holistic view of the subjects of analysis. Protein-protein interaction data is combined with systematic functional annotations, focussing the analysis of the interaction data on a specific biological context. This allows to scale the complexity of the large protein-protein interaction data sets and makes the results comprehensible. Moreover, it allows to hypothesize on the functional context of previously uncharacterized genes and proteins.

Biochemical reactions and textbook metabolic pathways are employed for the analysis of clustered gene expression data, which allows to analyze the metabolic properties and the changes in metabolism that have been captured by the respective gene expression experiment. Interesting features like co-regulated or conversely regulated pathways are highlighted by the integrative methods. Besides working with the established schemes and categories of textbook metabolic pathways, the elaborated methods allow to construct hypothetical pathways dynamically based on the gene expression profiles. From the structure of a hypothetical pathway, relations between parts of an

organism's metabolic network can be inferred that are conceptually distinct in the textbook pathways.

A method for the integration of gene expression data with functional annotations has been developed. An expression data set can be analyzed in the context of every of the various categories of a functional classification scheme. This functional projection is capable of identifying functionally related sets of genes that exhibit similar, correlated or anti-correlated expression profiles. Cellular processes that are co-ordinately switched on or off during a biological experiment are revealed. The relation between the experiment, e.g. a systematic variation of environmental conditions, and the genetic response of the analyzed organism becomes obvious. The analysis of overlapping groups of functionally related genes reveals how the genes of different functional categories relate, highlighting a larger biological context. The combination of the functional projection with the metabolic analysis methods allows to further investigate the identified co-regulated gene groups with a specific focus on aspects of intermediary metabolism.

The developed integrative methods are generically applicable to other types of high-throughput data, e.g. protein complexes, and for other systematically annotated facts about genes and proteins, e.g. subcellular localization, mutant phenotypes, protein classes, PROSITE motifs, signal transduction pathways and regulatory pathways.

The developed bioinformatic methods are compared with other approaches to a combinatorial and integrative analysis of whole-genome data sets. The benefit and the great potential of integrative methods is pointed out: the combination of different types of data can lead significantly towards a true systems biology that may allow the understanding and simulation of reasonably complex cellular processes or even whole cells.

Chapter 1

Introduction

The Bio2000 meeting [...] helped confirm in my mind that biotech is back with a vengeance and that bioinformatics will be a key part of the growth of this, the third, industrial revolution. [...] It's a privilege to be able to provide the picks and shovels to the miners in this goldrush.

– *Tim Littlejohn*¹

In the previous months and years, the biotechnological industry has been growing rapidly. Modern technologies in molecular biology have triggered the hope that one will find new drug targets and drugs and that it will be possible to find cures for severe diseases like cancer and diabetes. Many of the new technologies are *high-throughput technologies* that generate ever faster growing amounts of biological data. Effective and efficient intelligent methods are desperately needed in order to be able to manage, store, retrieve and analyze this heterogeneous data. *Bioinformatics* is the branch of computer science that deals with the development of informatic methods such as algorithms and data structures for biological problems especially of molecular biology.

This thesis describes a bioinformatic framework that realizes an *integrative analysis approach*. The approach integrates data from different sources. Data sets are combined and the individual sets are restricted or interpreted by means of the other. This allows to learn more from the data than the isolated analysis of any one kind of data could reveal. The focus is on the extraction of qualitative, not quantitative results from large-scale data sets.

This introductory chapter is not thought to be an introduction to molecular biology or bioinformatics. Brilliant textbooks can be found for further readings in both areas [ABL⁺89, APS99, Pev00]. This chapter introduces key

¹EMBnet News, 2000, 7(1):6-7.

terms of molecular biology and shall convey an idea of why the integrative analysis methods described in this thesis are useful. The subsequent chapters deal with biological concepts, the corresponding data types and data structures and the algorithms that work on these data structures. Chapter 2 describes the data types we use for the integrative analysis, their biological meaning and their computational aspects in detail. Some of the methods are part of the developments I have achieved at the Munich Information Center for Protein Sequences (MIPS) during the work that is the basis of this thesis. Sections that describe my contributions are marked. Chapter 3 is the core chapter of the thesis. The instances of my approach to an integrative analysis are described here. The system architecture and the user interface I have developed for this purpose are presented in Chapter 4. Finally, Chapter 5 discusses the results.

Some results of the analyses are publicly accessible at MIPS via the described suite of dynamic web pages [Fel01b].

1.1 Trends in Modern Molecular Biology

Cells occupy a halfway point in the scale of biological complexity. We study them to learn, on the one hand, how they are made from molecules and, on the other, how they cooperate to make an organism as complex as a human being.

– [ABL⁺89]

In order to clarify some terms of molecular biology, we partly follow the second chapter of [ABL⁺89] in this section. The book gives an easy to read and comprehensive introduction to the field.

The atomic unit of all living organisms is the *cell*. An individual organism may consist of only one cell, such as bacteria or fungi, or it may be composed of billions of cells of different types (e.g. there are 10^{14} cells in human). The cells can be viewed as chemical machines that interconvert thousands of molecules of eventually different types in a highly ordered, systematic way. The *metabolism* of the cell is an intricate network of biochemical reaction steps. Each of these reactions requires a catalytic molecule. Typically, the catalytic molecules are proteins (*enzymes*) – but in biology, there is no rule without exceptions. In rare cases, the catalyst is not a protein, but an RNA molecule. Enzymes are themselves products of the very same metabolic machinery. A small molecule, e.g. an amino acid, can typically be modified by six to ten different enzymes. It may be adenylated, degraded, acetylated, or transferred to a fatty acid. The corresponding *pathways* compete for the amino acid molecule. The whole network of bioreactions can be

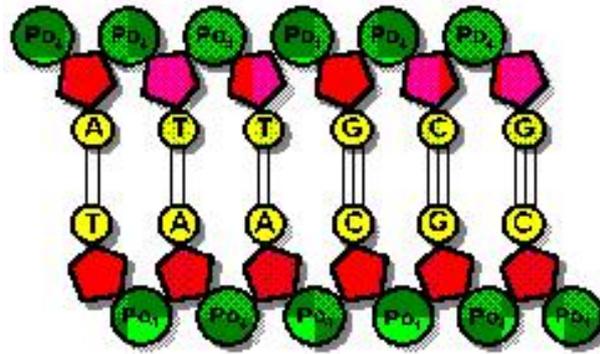


Figure 1.1: Two complementary DNA strands. The backbones of each strand consist of a sequence of alternating sugar (deoxyribose) and phosphate molecules. A base is bound to each sugar molecule. Hydrogen bonds between the bases link the two DNA strands (source: <http://www.fdl.cc.mn.us/>).

described in terms of pathways competing for small molecules. Despite of its complexity, the metabolism of the cells is very stable. It is regulated and coordinated by a network of control mechanisms on different levels.

The elements that carry the hereditary information of the cell are the *genes*. The totality of all genes is called the *genome*. The genes are made of deoxyribonucleic acid (DNA). DNA is a polymeric molecule that consists of two linear strands. The backbone of each strand is a sequence of alternating sugar (deoxyribose) and phosphate molecules. One of four kinds of base molecules (adenosine (A), thymine (T), guanine (G), and cytosine (C)) is bound to each sugar molecule of the strand. The monomeric building blocks of the DNA strands consist of one deoxyribose molecule, one phosphate molecule and one base molecule. They are called *nucleotides*. The linear sequence of the nucleotides encodes the genetic information. Two DNA strands are loosely linked via hydrogen bonds between complementary base pairs (A=T, C≡G) (Figure 1.1). The double stranded DNA molecules form the characteristic DNA double helix.

Whenever information from the DNA is needed, e.g. in order to construct an enzyme, a specific stretch of the coding DNA strand, the gene, is copied (*transcribed*) into a complementary ribonucleic acid (RNA) molecule. The entirety of the transcribed genes of a cell is called the *transcriptome*.

A number of different types of RNAs are distinguished. The ribosomal RNA (rRNA) builds cell organelles, the ribosomes. The messenger RNA (mRNA) is carried to the ribosomes, where the sequence of bases of the mRNA strand serves as a blueprint. It is specifically *translated* into a sequence of amino

acids that is called a *protein*. A third type of RNA is involved here, the transfer RNA (tRNA).

The totality of the proteins of a cell is called *proteome*. It accounts for more than a third of a cell's dry weight. Proteins determine the shape and structure of the cell. They are involved in molecular recognition and catalysis. The amino acid sequence of a protein is called *primary structure*. It determines the shape of the protein. Proteins fold into a specific *secondary structure* conformation of α -helices and β -sheets. These in turn pack together to form compactly folded *domains*. Functional proteins are either formed of a single domain or they are *protein complexes* that consist of a number of protein domains bound together by non-covalent interactions.

All different kinds of molecules mentioned above are subject to extensive research in molecular biology, biochemistry, biotechnology, and pharmaceuticals. While only a few years ago scientists in the laboratories dealt with one gene or one protein at a time, modern technologies now allow to deal with a whole genome, transcriptome, or proteome at the same time, within a single experiment. *High-throughput* technologies give rise to large amounts of data, what in turn triggers the development of bioinformatic methods, i.e. methods that allow the (semi-)automated analysis of the biological data by means of computers.

The most established field of high-throughput analysis is the *sequencing*, i.e. the determination of the nucleotide sequence of one strand of a DNA molecule and the subsequent *sequence analysis* of genomic DNA [DEKM98]). In order to learn which pathways of the large metabolic network are switched on or off under various conditions in various organisms, one either measures the expression rates of the individual genes of an organism or of a subset of genes of interest (cf. *gene expression analysis*, Chapter 2.1), or the protein abundance in the cell (*proteomics*). Since most proteins do not function isolated in a cell, but interact with one another and build protein complexes and protein machines, it is important to study protein-protein interactions on a large scale. Several high-throughput technologies are available that are capable of revealing nearly complete protein interaction maps of an organism (Chapter 2.2).

1.2 Trends in Bioinformatics

Major achievements are being made in biotechnology. The sequencing of the genomes of whole organisms, beginning with the brewer's yeast *Saccharomyces cerevisiae* in 1989 and the bacterium *Haemophilus influenzae* in 1994, has become routine work. Even the human DNA sequence is now

at hand [Int01, VAM⁺01]. The analysis of the huge amounts of sequence data as well as of the data produced by high-throughput technologies in the biotechnological laboratories has just begun. The data has to be stored efficiently. Scientists must be able to retrieve and handle the data they need. Since the number of elements in the biological databases has grown to an extent that makes it impossible to analyze them manually, automatic computational analysis methods have to be developed in order to destil information from the data. The needs of biotechnologists and molecular biologists push established computational analysis methods beyond their limits.

Since the human sequencing projects started, scientists speculated about the number of human genes. For quite some time estimations of about 100,000 human genes were agreed upon. In autumn 1999, when the draft versions of the sequences of the first human chromosomes were available, researchers changed their estimations from 100,000 to 130,000. Some groups even suggested three times as many genes. It was not before end of 2000 that the numbers were finally corrected to much smaller values. The number of genes predicted to exist in the human genome currently ranges from 30,000 to 40,000 [Cla01]. This uncertainty shows how insufficient our gene prediction methods still are. The definition of what a *gene* actually is, remains unclear [Att00]. It might turn out that the dramatic downcome of estimates to relatively small numbers of human genes does not mean that also the number of proteins and functions is as low. There are examples like immune globulines and signal transduction, which show that a single gene can have more than one function, that its transcript can be spliced in more than one way and that the resulting mRNAs can be edited before protein translation. On the protein level, complicated regulatory processes and biochemical modifications of proteins may turn out to be the rule, not the exception.

From recent bioinformatic publications one can tell that integrative analysis approaches will dominate the future of bioinformatics, e.g. [PMT⁺99, Pel01]. Looking at the sequence of a gene or the structure of a protein in isolation often does not allow to hypothesize on the function of the respective element. Integrating analysis results is like collecting evidence in an obscure criminal case. We can analytically link the genomic level with the protein level and the metabolite level. To promote the knowledge and understanding of what is going on in a living cell, the various fields of bioinformatics must go hand in hand.

1.3 Approach to an Integrative Data Analysis

Nature functions by integration, and the adoption of a more holistic view of complex biological systems is an essential step for bioinformatics.

– *Teresa K. Attwood* [Att00]

This thesis puts forward the approach of an integrative data analysis. The large-scale data sets resulting from high-throughput technologies of functional genomics and proteomics need to be analyzed automatically. These techniques produce data associated with certain genes (e.g. gene expression data) or proteins (e.g. protein-protein interaction data, protein expression data). The first step is to analyze the data sets by looking at the internal structure, distribution of values, and (numerical) peculiarities. A prominent example of such an analysis is the normalization and clustering of gene expression data. One will then analyze the data by comparison with other data sets of the same kind. The sequence similarity search using algorithms like BLAST and FASTA follows this approach: a DNA or amino acid sequence is compared with an entire database of sequences.

As I show in this thesis, a combinatorial and integrative analysis that combines different types of data sets can greatly promote the understanding of the data. Following this approach, the data and the structures derived from the data by internal analyses are further structured and analyzed by integrating them with other kinds of data. These may be other high-throughput data, but the most interesting qualitative results we obtain using data that systematically classifies genes and proteins. We use functional annotations, i.e. the MIPS functional catalog (Chapters 3.1, 3.4) and the grouping of genes and proteins to metabolic pathways (Chapter 3.2), both reference pathways and dynamically modeled hypothetical pathways and metabolic networks.

1.4 Acknowledgements

This doctoral thesis describes the main results of my work over the last three years at MIPS and Biomax. During this period many people, colleagues and external collaborators, helped to promote and realize my ideas and often introduced new ideas.

First of all, my advisor, Prof. H. Werner Mewes, has had great positive influence. He put me onto the *pathway project* at MIPS and later pushed me into the direction of gene expression analysis. The idea to combine the two threads and to follow the approach of an integrative analysis was born.

Dmitrij Frishman, senior scientist at MIPS and co-founder of Biomax, has had a more subtle but nevertheless very important influence on my work. He introduced me to many interesting people and arranged the contact that made me one of the teachers at the Bioinformatics Summer School (BiSS 2001) at Göttingen in September 2001.

Kaj Albermann, Jean Hani, and Alfred Zollner, colleagues of mine at Biomax, brought the protein-protein interactions into play. It was Kaj who initially set up the interaction tables at MIPS and Jean who had the idea of analyzing the interaction data set in the context of functional categories. Our common work was and still is efficient and pleasant. Only four months after Jean first came with his idea, we submitted the ISMB paper [FAZ⁺00].

Volkmar Liebscher of the biomathematical institute of the GSF Research Center in Neuherberg helped with questions concerning the statistics of gene expression analysis. He finally convinced me that *significant signals* in large data sets can best be found by combining mathematical methods with biological annotations. And that's exactly what this thesis is about!

An important scientific application of the methods I have developed was on the data that Hendrik Milting obtained at the Clinic of Thoracic and Cardiovascular Surgery of the Ruhr-University Bochum at the Heartcenter (HDZ) in Bad Oeynhausen. During the collaboration, I believe, we both learned a lot. A paper has been submitted for publication that resulted from the work of his group that I partly supported, again in cooperation with Volkmar Liebscher.

With my room mate at MIPS, Ole Bents, I had interesting and often effective discussions on technical issues of his and my work but also on science, bioinformatics and the many different possibilities to work as a computer scientist. With his many corrections and suggestions he also did a great job when reading this thesis as well as numerous articles over the years.

In March 1999, at a very early stage of my work, I partly joined Biomax Informatics, the MIPS spin-off company at Martinsried. This allowed and obliged me to present the work to the Biomax customers who gave very important feedback. Biomax with its CEO Klaus Heumann gave me the freedom to continue with the scientific work I had begun. Doing the work in parallel at MIPS and Biomax was sometimes strenuous but after all, the work was greatly promoted by the different and sometimes contradictory influences and necessities.

This is only a small subset of the people I would have to mention. Thank you all, your help is greatly appreciated!

Chapter 2

High-Throughput and other Biological Data

This chapter describes the data sets that the integrative analysis approach deals with. We distinguish between *high-throughput data* and other biological data. The latter may be termed *qualitative data* and includes information on the function of genes (annotations) and the structure of metabolic pathways. These data sets are often collected from various sources and mostly describe an organism independent from a special experiment, while high-throughput data sets can be produced in arbitrary number for different experimental setups. The most prominent type of high-throughput data is the *gene expression data* obtained by DNA microarrays and similar techniques. Beginning in early 1997 with a few ground breaking publications [DIB97, ESBB98], microarrays have become a routine technique utilized by biological and medical research institutes as well as pharmaceutical and biotechnological companies all over the world. Another forthcoming technique are two-hybrid assays, a high-throughput technology for analyzing *protein-protein interactions*. After a moderate development of the analysis of protein-protein interactions in the 1990s, a stunning number of publications describing two-hybrid experiments and computational approaches of protein-protein interaction analysis appeared in 2000, e.g. [UGC⁺00, FAZ⁺00, SUF00]. Both, gene expression data and protein-protein interaction data are described in the following sections 2.1 and 2.2.

A third important high-throughput approach, *proteomics*, is not discussed in this thesis. Although a very promising field, public protein expression data sets are still sparse. This is due to the very complex and time consuming process of identifying and quantifying proteins from the probes. In proteomics, 2D gels are employed to separate the different proteins. In contrast to the microarray technology, the spots on the gels are not predetermined but their

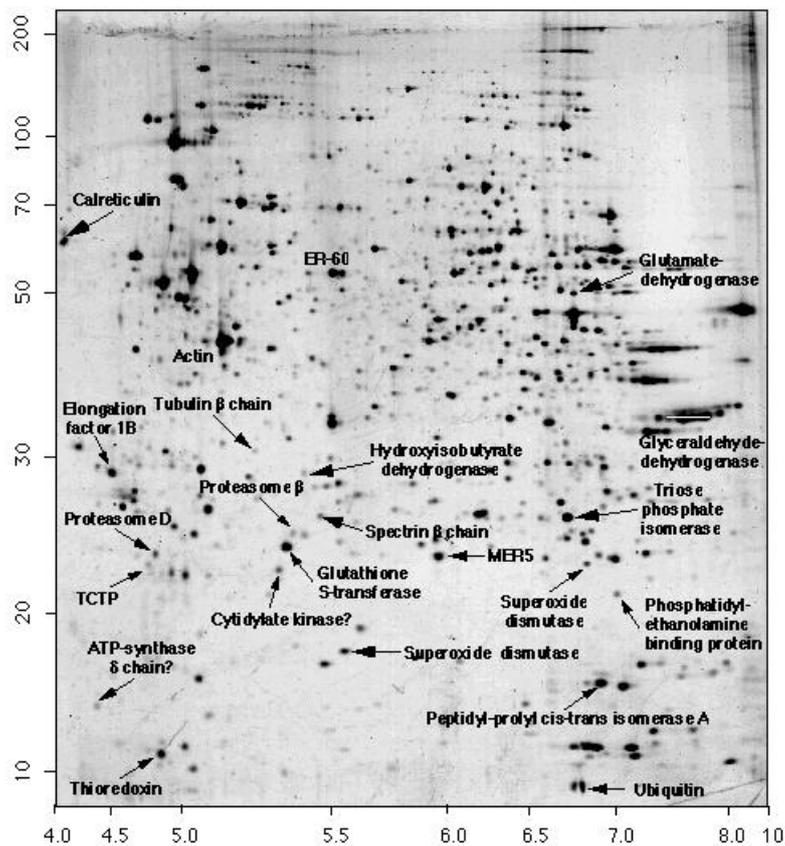


Figure 2.1: A 2D gel used for proteomic analysis. Some protein spots are labeled. Not only does such a *noisy* gel cause difficulties in image analysis. Assigning proteins to spots is costly and time consuming. The picture is taken from the ExPASy web pages and was originally published in [SAG⁺95].

position depends on the chemical and physical properties of the proteins. These properties of the native proteins are altered by post-translational modifications such as methylation and phosphorylation. Due to these modifications, a single protein may appear as two, four or even more spots on the gel. In order to determine the protein that accounts for a specific spot, *mass spectrometric analysis* has to be carried out.

Sections 2.3 and 2.4 describe the qualitative types of data discussed in this thesis: systematic *functional annotations* of genes and gene products and *metabolic pathways*. Systematic schemes for functional annotations are well developed. Therefore they are an ideal object of computational analysis approaches. Metabolic pathways are of particular interest in many fields, e.g. in metabolic engineering.

2.1 Gene Expression Data

Comparison of transcriptomes yields interesting information about the dynamics of total genome expression attributable to a change in environmental conditions or state of differentiation. In addition, it provides necessary clues to determine the function of those genes whose contribution to cellular life is still unknown.

– [KvZB⁺99]

With the genomic sequences of a substantial number of organisms available, the field of *functional genomics* is rapidly developing. Although the pure sequence information is theoretically sufficient to promote a full understanding of gene expression, gene function and regulation of gene activity, we are currently not able to interpret the sequence information to such an extent [LDB⁺96]. Cellular processes are genetically controlled. Signal transduction and other regulatory systems control the expression of the genes in each cell of an organism [PSE⁺00]. Changes in the physiology of an organism or a cell will be accompanied by changes in gene expression [vHVvH⁺00]. Therefore, knowledge of which genes are expressed in a certain organism and about the levels and timing of expression is important for a further understanding. Knowing when and where a gene is expressed allows to infer on its biological role while the pattern of the genes expressed in a cell allows to infer on its state [DIB97].

At MIPS I have developed a toolbox for the analysis of large-scale gene expression data. The tools and algorithms cover the analysis steps from normalization and clustering up to the integrative analysis of gene expression data in the context of other high-throughput data and systematically stored annotational data (Chapter 3).

2.1.1 Introduction

Large-scale gene expression studies are carried out with modern technologies like oligonucleotide arrays, DNA microarrays, and Serial Analysis of Gene Expression (SAGE). A brief introduction to these techniques and their specific differences is given below. With these technologies it is for the first time possible to measure gene expression on a genomic scale. For every gene of an organism the expression rates are determined in parallel under well defined experimental conditions. We call the resulting expression rates of all measured genes the *expression pattern*. A series of measurements is

typically carried out, either under systematically varied environmental conditions or at several successive time points during an experiment. The result is a vector of expression rates for each gene that we call *expression profile*. The expression profiles are time courses in the latter case. To allow statistical analyses and error estimation, the measurements for each condition/time point are typically repeated several times [LKWS00].

Studies of whole-genome messenger ribonucleic acid (mRNA) expression levels are also called *transcriptome analyses*. Consider a gene in a cell at a certain time point or under certain environmental conditions. The gene expression, i.e. the occurrence of the transcribed mRNA molecule that corresponds to the gene, serves as an indicator for the production of the corresponding protein in the cell.

2.1.2 Aims of Gene Expression Analysis

Having gene expression data available, a large variety of questions can be asked. One can identify at least the following groups of questions:

Single Gene Analysis. Focussing on a certain gene A , an obvious question arises: is gene A expressed in a certain context, e.g. mitosis? We can also compare the expression of gene A to other genes: which genes are expressed similarly to A (cf. Chapter 2.1.8)? Do they belong to a certain functional class (cf. Chapter 3.1)?

Single Genome Analysis. Which genes are (differentially) expressed? Which genes regulate and which are regulated in a time dependent process? Can we find meaningful clusters of similarly expressed genes? Do we find expected patterns of expression? How does a toxic compound effect gene expression? Gene annotations can be taken into account: Do we find clusters of genes that have a high rate of genes of a certain functional category or a certain biochemical pathway (cf. Chapters 3.1, 3.2.1)? Can we construct metabolic pathways from scratch, just analyzing the expression data (cf. Chapter 3.2)?

Cross-Genome Analysis. Do homologous genes share certain expression profiles?

Cancer Research. Can we distinguish between healthy and affected cells? How does the gene expression differ? Can we clearly identify different kinds of tumors via their expression patterns? This is regarded to be especially helpful for very similar and phenologically difficult to distinguish kinds of tumors. It is proposed that the diversity of tumors corresponds to a diversity in gene expression patterns. Capturing these differences by

measuring gene expression patterns on a genomic scale might lead to an improved understanding of cancer, different types of tumors, and their taxonomy [ABN⁺99, GST⁺99, PSE⁺00].

Gene expression analysis can shed light onto questions from various fields. It can aid the understanding of gene regulation, it can promote the understanding of changes that occur in a disease and it can help identifying genes not yet known to be involved in a disease (medicine, cf. Chapter 2.1.10). Gene expression analysis can help finding genes that might be of pharmaceutical relevance, e.g. in drug target discovery.

2.1.3 Biotechnological Analysis Methods

The following paragraphs describe very briefly the analysis methods used in biotechnology. The specific differences have to be considered when analyzing the resulting data sets. Differences include the type of resulting data, i.e. relative vs. absolute values, radioactive vs. fluorescent labeling, and of course potential errors and shortcomings. Oligonucleotide arrays, nylon filters and complementary DNA (cDNA) microarrays are based on hybridization of mRNA probes or cDNA probes, respectively, to a high density array of immobilized spots of target sequences that correspond to one gene each. SAGE uses a completely different technology based on sequencing.

Oligonucleotide Arrays

Lockhart et al. have developed oligonucleotide arrays, an approach that is based on hybridization to high-density arrays containing several thousand synthetic oligonucleotides [LDB⁺96]. Affymetrix, who produce the most popular oligonucleotide arrays, use 15 different 25-mers per gene. The arrays are designed based on sequence information and are synthesized *in situ* using a combination of photolithography and oligonucleotide chemistry. Oligonucleotide arrays have a large dynamic range. The detection of RNAs is quantitative over more than three orders of magnitude. RNAs present at a frequency of 1:300,000 are unambiguously detected. The method is readily scalable and allows the simultaneous monitoring of tens of thousands of genes. The equipment needed for manufacturing and reading the chips is not inexpensive.

Radioactively Labeled Nylon Filters

A well established method is the measurement of gene expression rates with arrays on nylon membrane using radioactive hybridization [LL91]. The

ORFs to be spotted onto the array are first polymerase chain reaction (PCR)-amplified from genomic DNA. The PCR products are arrayed at high density using robotic devices. The probes are made from total RNA preparations and are hybridized for the analysis of the transcriptional activity of the organism under investigation [HVS⁺98]. Nylon filters can be regenerated and can be reused for more than one hybridization.

cDNA Microarrays

cDNA Microarrays are produced on glass slides. Full-length DNA sequences are printed onto the slide using a robotic device. The cDNA sequences are prepared from the ORFs of the organism by PCR. We follow [DIB97] for the description of an experimental analysis: The organism that is to be analyzed is kept under the conditions defined by the experimental design. At predefined time points, samples are harvested and RNA is isolated from the samples. Fluorescently labeled cDNA is prepared from the RNA by reverse transcription of labeled deoxyuridine triphosphate (dUTP). Two different fluorescent labels are used: a green and a red one. The cDNA prepared from the samples at the individual time points are labeled red. A reference cDNA from samples harvested at a control time point is labeled green. Both cDNAs are mixed and hybridized to the microarray.

The described experimental design allows to measure current expression levels and reference levels in parallel on the same array. The hybridization of sample and reference sequences is competitive. The fluorescence intensities of the red and green markers are separately measured. For each pair, the relative intensity is calculated as a measure of the relative abundance of the corresponding mRNA in the two samples. The characteristic red/green arrays are the result of this technique (Figure 2.2). As described, the colors encode the relative abundance of the same sequences in two different samples. The arrays have to be evaluated electronically by means of image processing.

The introduction of fluorescent probes makes a miniaturization of arrays possible. Smaller amounts of starting material can be used and larger numbers of genes can be screened in parallel [vHVvH⁺00]. cDNA microarrays are cheaper to produce and easier to read than oligonucleotide arrays but they require handling full-length cDNAs instead of oligonucleotides.

SAGE

The SAGE technique, described in [KvZB⁺99], samples short sequences of 10-14 nucleotides (tags) of individual mRNAs. The sequence of these

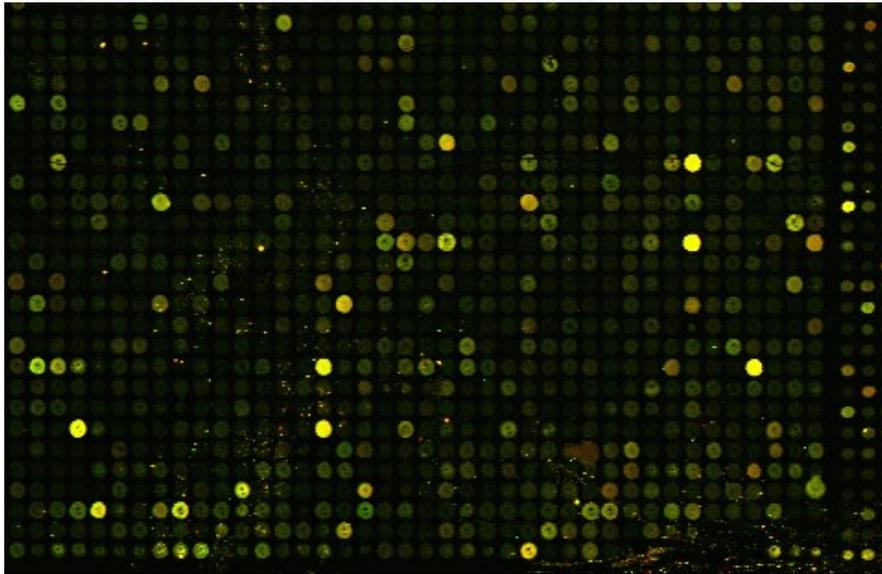


Figure 2.2: Detail of a microarray scan, taken from the diauxic shift experiment of DeRisi et al. The experiment data is accessible at <http://cmgm.stanford.edu/pbrown/explore/>.

tags has to be determined in order to allow the identification of the corresponding genes. Information about the level of gene expression is derived from the frequency of a tag. A SAGE protocol is described by Velculescu et al. [VZVK95]. SAGE involves complex sample preparation, requires extensive DNA sequencing and is not very sensitive [vHVvH⁺00]. An advantage of SAGE is that the genomic sequence of the genes does not have to be known *a priori*. SAGE is thus capable of determining the expression rates of previously unknown genes. Data obtained with the SAGE method is not considered in this thesis.

2.1.4 Variations of Expression Rates

With the described techniques it is still not possible to obtain a count of mRNA copies per cell. Gene expression measurements are relative by nature. Thus it is only possible to compare the expression levels of the same gene in different measurements or of different genes in the same measurement [BV00]. Computational processing, i.e. normalization and statistical evaluation of the data, play a crucial role. Biologically meaningful signals may be obscured by experimental noise and systematic errors. A statistical preprocessing has to be performed in order to get rid of the specific influences of the experimental methodology and in order to estimate the error

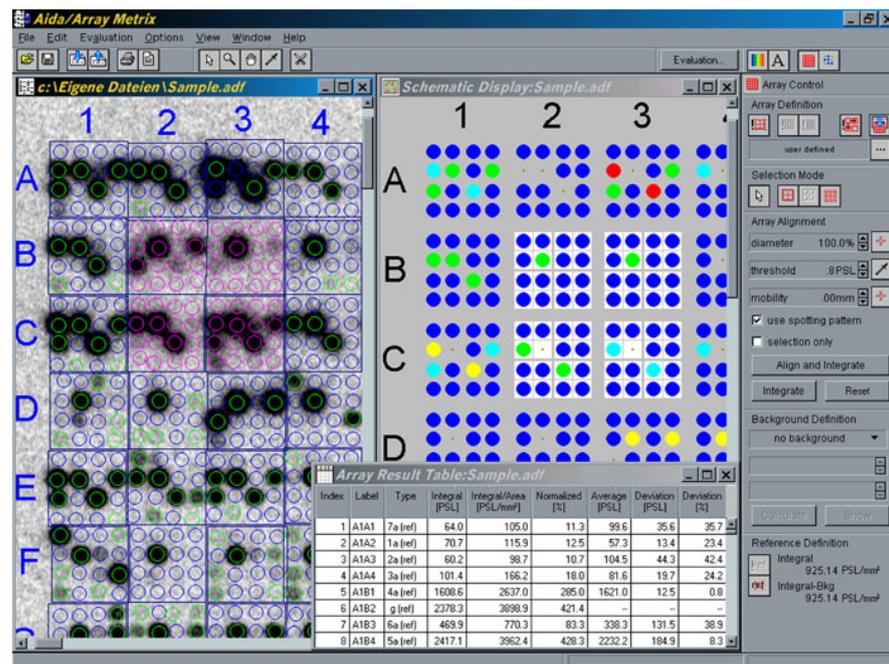


Figure 2.3: Example of an image analysis tool for gene arrays (AIDA Array Evaluation, Raytest Isotopenmeßgeräte GmbH, Germany). The array scan can be seen on the left with circles fitted to the spots. The color coded values for each spot are depicted in the abstract array view on the right.

rates for individual genes, measurements and experiments. The expression data and the results of their analysis always have to be interpreted with the shortcomings and potential errors in mind. Compared to directly interpreting quantitative results, heading for qualitative results has some advantages. The qualitative results are more robust against inherent measurement errors. Qualitative results can e.g. be obtained by the integrative analysis methods described in Chapter 3.

2.1.5 From Arrays and Chips to Data Matrices

After the laboratory part of a gene expression measurement is finished, the array, chip, or filter has to be evaluated. Therefore a non-compressed digital image (TIFF bitmap) of the device is used. Advanced techniques are necessary for a reliable transformation of the image into a data set, i.e. generating numerical values for each spot of the device [BV00]. The spots that correspond to genes have to be identified and their borders have to be determined. The intensity of the spot has to be measured. It has to be compared with the background intensity. The image analysis software tools, e.g. by Raytest (Figure 2.3), are often bundled with the scanner hardware. They identify the

spots by fitting an irregular grid onto the image. The background intensity is measured, either globally, i.e. for the whole array, or locally, i.e. for each spot or group of spots. The obtained values are corrected according to the background. From intensity, background intensity, position and shape of the spots, a trust score is assigned to each spot. Result of the image analysis process is a data vector. Repeated measurements and several conditions or time points lead to a data matrix. Within the matrix, each row represents a gene, each column represents a measurement. The focus of the MIPS group is on the data analysis. Our analyses start with the data matrix produced by the image analysis software.

The data matrix can be interpreted in two ways: by rows or by columns, either looking at gene expression profiles (*one gene* and *one genome analysis*) or looking at measurements (*cross genome comparison* and *cancer research*). Time course experiments are the typical case for a gene-wise analysis, e.g. a gene clustering (Chapter 2.1.8). Comparing cells of different phenotypes or tissues, one would look at the measurements, e.g. in a differential display [BFB⁺00]. In the context of this thesis we focus on gene expression profiles, mainly by performing gene clusterings as the first step.

2.1.6 Normalization

The normalization of the gene expression data is a crucial process in order to make measurements comparable. Even the measurements of a series of repetitions that have been performed with the same array type at the same laboratory by the same person vary significantly. The variations can easily be observed in a scatter plot (Figure 2.9, left) that reveals additive as well as multiplicative offsets. [SBM⁺00] name several sources of fluctuations in probe, target and array preparation, in the hybridization process, and in image processing (quantification). Without normalization, measurements could therefore not be compared in an analysis.

Several methods have been suggested for normalization of measurements, e.g. [SBM⁺00, BFB⁺00]. The methods have different characteristics, especially as far as robustness is concerned. The decision for an individual method depends partly on the type of array. One has to distinguish between whole genome measurements and partial arrays that contain only a specific subset of the genes of an organism.

Normalization to a Set of Constitutively Expressed Genes

A number of genes of an organism is considered to be constitutively expressed, i.e. these genes are unregulated so that the expression rate should

largely be constant. Up to 30% of the genes expressed in a cell are thought to be constitutively expressed. They are probably involved in respiration, cell growth, replication, and gene expression [DH97]. Whenever the expression rate of genes that are known to be constitutively expressed is measured on an array, one can assume that the expression rate of these genes does not change during the time line of the experiment. One can normalize the measurements onto the constitutively expressed genes. The procedure may lead to good results, still it is not very robust since it depends on the measured values of a small number of genes. Besides for normalization, constitutively expressed genes are also applied for an estimation of the reliability of gene expression data.

Normalization to a Set of Reference Sequences

Another strategy is to normalize onto a set of reference sequences (external spikes). The prerequisite for this method is that the array contains a number of sequences that do not appear in the gene set of the analyzed organism. Together with the probes, one then applies the external spikes in a predetermined concentration to the array. The sequences will hybridize to the corresponding spots. Since the concentration of the reference sequences is known and since it is kept constant throughout a whole series of measurements, the measurements can be normalized onto the values obtained for these external spikes. Although this method proved useful and is applied for example by Affymetrix, one of the leading chip providers, many other chip manufacturer do not use reference sequences. In these cases, it is not feasible for an experimenter to use external spikes on the standard array. Therefore this method has a limited field of application.

Purely Statistical Normalization

In order to normalize gene expression data by means of statistics on the data, one has to make certain assumptions on the characteristics of the data sets. An important and at least for whole-genome arrays largely accepted assumption is that on average the transcription rate and thus the mRNA abundance of the measured genes does not change. When employing arrays that contain specialized subsets of genes or particularly small gene sets one has to consider carefully, whether this assumption holds.

2.1.7 A Statistical Normalization Procedure

The statistical normalization procedure described here, I have developed for the analysis of gene expression data obtained from human heart tissue ([MBK⁺01], Chapter 2.1.10). In collaboration with the molecular biologist that set up the study, we agreed on the assumption that on average the gene expression rates would not change during the experiment. Thus we can use a statistical normalization method.

Normalizing gene expression data sets means that the individual statistical characteristics of a set of hybridizations are made similar without affecting the possibility to derive the desired signals. For a pair of measurements this can be achieved via a line fit. When normalizing a whole set of measurements, we adapt every single hybridization to a pseudo-measurement. For each gene the median of the individual values is computed. The resulting pseudo-measurement, i.e. the pattern that consists of the medians, is used to adapt the hybridization data sets via linear regression. The computation of a median measurement avoids the adaption to a single, arbitrarily chosen hybridization. We prefer the median over the average because it is more robust, i.e. outliers influence the median significantly less. The performance of the method can be assessed by computing the mean and the standard deviation of the data points of each measurement. We observe that the means of the normalized measurements are nearly equal. The standard deviations also become very similar. The skewness values vary between the hybridizations. Skewness is not affected by the normalization. The fact that they all have the same sign supports the conclusion that the data sets have a similar statistical characteristic. The case study, described in Chapter 2.1.10, provides an example for the usefulness of this method. Figure 2.9 shows scatter plots of the same measurements as raw values (left), and with both measurements adapted via a line fit to the pseudo-measurement of the medians of the 46 hybridizations of the same series.

2.1.8 Gene Clustering

A step towards the rapid and comprehensive interpretation of gene expression data is the clustering of the genes with respect to the expression profiles [ESBB98]. The individual genes are sorted into groups (clusters) by a clustering algorithm. One wants to find clusters of genes that are co-expressed. This is biologically relevant since co-expression may be caused by co-regulation according to similar protein binding sites or transcription factors.

The clustering can generally be achieved with every established clustering

method of data analysis. Eisen et al. use nearest neighbor joining, a hierarchical clustering method that results in a binary tree with the genes at its leaves [ESBB98]. A neural network approach, Kohonen's self-organizing map (SOM), is well suited for the analysis of multi-dimensional data. Tamayo et al. first used the SOM for clustering gene expression profiles. They name distinct advantages of the SOM clustering over the hierarchical clustering [TSM⁺99]. More methods are described in the literature, some developed especially for the application on gene expression data [MCA⁺98].

In essence, different clustering methods provide very similar results. Differences pertain to the possibilities of representing and assessing results and in the computational complexity of the methods. The distance measure is as relevant as the actual method used for the clustering. In this chapter we first describe several distance measures and their specific advantages and disadvantages. Then we present three clustering methods largely employed for clustering gene expression data. The methods are only briefly described here. The SOM clustering approach I have developed at MIPS is described in detail in the next section (Chapter 2.1.9).

Distance Measures for Gene Expression Profiles

In order to cluster genes according to their expression profiles, the distance between two expression profiles has to be computed. The expression profile of a gene can be interpreted as a n -dimensional vector or as a point in n -dimensional Euclidean space. Thus the Euclidean distance measure is applicable. Other distance measures have been suggested, among them the linear correlation coefficient and the L_1 distance measure. Rank correlation and mutual information have also been proposed. Obviously, there is no *best measure* and no *right choice*. At MIPS, I have implemented and employed the following four measures. In the equations below, X and Y denominate the expression profiles of two genes. Correspondingly, x_i and y_i are the i th components of the respective profile vectors.

The Euclidean distance $D_E(X, Y)$ of two n -dimensional expression vectors X and Y is defined as

$$D_E(X, Y) = \sqrt{\sum_{i=1, \dots, n} (x_i - y_i)^2} \quad (2.1)$$

The linear correlation coefficient $D_{LC}(X, Y)$ of two expression vectors is related to Euclidean distance if before applying the Euclidean distance the

expression vectors are standardized to mean zero and unit variance. The linear correlation is computed as

$$D_{LC}(X, Y) = \frac{\sum_{i=1, \dots, n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1, \dots, n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1, \dots, n} (y_i - \bar{y})^2}} \quad (2.2)$$

with \bar{x} and \bar{y} representing the mean of the x_i and y_i , respectively. The Manhattan distance (L1) $D_{L1}(X, Y)$ is defined as

$$D_{L1}(X, Y) = \sum_{i=1, \dots, n} |x_i - y_i| \quad (2.3)$$

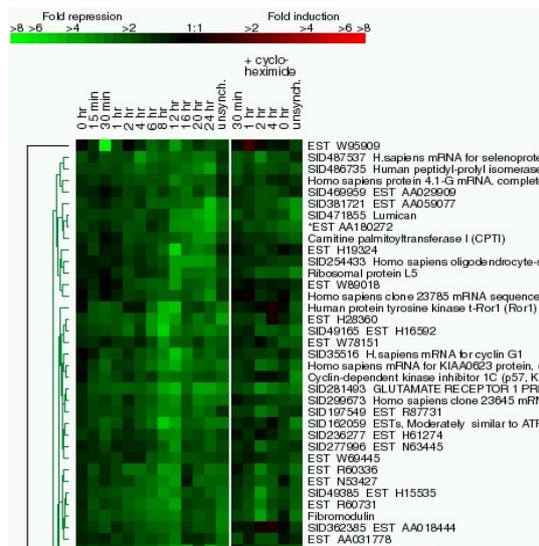
This is a more robust measure with respect to outliers compared to Euclidean distance and the linear correlation coefficient since the differences between the components of the expression vectors are linearly summed, not squared. To make the measures even more robust, we also tried to cap the components of the Manhattan distance. The capped Manhattan distance $D_{L1C}(X, Y)$ is defined as

$$D_{L1C}(X, Y) = \sum_{i=1, \dots, n} \max(T, |x_i - y_i|) \quad (2.4)$$

with T being the maximal value (*threshold*) of any addend of the sum. Thus, the maximum distance of two expression profiles is nT . The robustness stems from the fact that outliers resulting from measurement errors are weighted with a maximum of T .

Pairwise Average-Linkage Clustering

The clustering using *pairwise average-linkage cluster analysis* was introduced into gene expression analysis by Eisen et al. [ESBB98]. We follow their description of the algorithm here. Eisen et al. combine the clustering that results in a reordering of the list of genes, with a red/green representation of the expression profiles (Figure 2.4). The pairwise average-linkage cluster analysis is a form of *hierarchical clustering*. The relationship between the clustered entities is represented by a binary tree, called *dendrogram*. The branch lengths of the tree reflect the similarity of the entities that



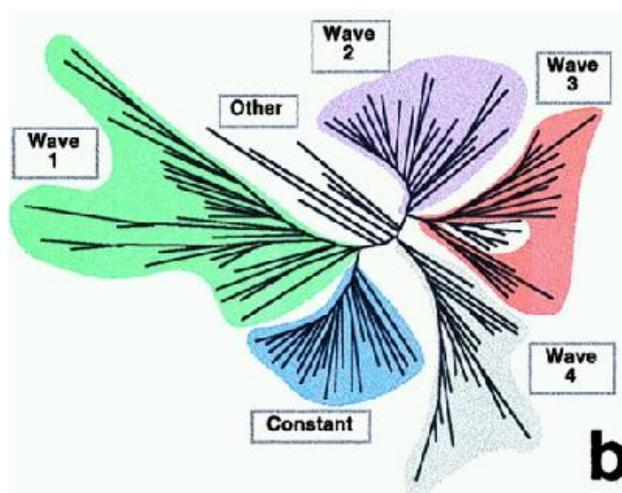


Figure 2.5: A clustering tree as it is produced by the algorithm described in [WFM⁺98]. The authors highlighted areas that correspond to an *expression wave* (shaded areas) to demonstrate the utility of the clustering method. The areas are determined manually by visual inspection. The figure is taken from [WFM⁺98].

SOM Clustering

The SOM is a neural network that provides a mapping from a multi-dimensional data space into a discrete two dimensional space [Koh95]. The SOM proved valuable in fields like engineering and medical image analysis [EFK⁺99]. The usage of the SOM algorithm for clustering expression profiles was first described by [TSM⁺99]. For the MIPS group, we also employ this very robust method for expression data produced in the EURO-FAN project [FM99] and for the human data of the case study described in Chapter 2.1.10.

The details of the SOM algorithm and the approach to gene clustering using the SOM I developed at MIPS are described in Section 2.1.9.

Other Clustering Methods

Other clustering methods have been described in literature, e.g. in [MCA⁺98]. Here, the Euclidean distance is employed as the distance metric between the multi-dimensional expression data vectors of the individual genes. The vectors are made up of the n expression values and the $n - 1$ slopes between them. Both types of vector components are reduced to a range [0;1]. By adding the slope values to the data vectors the

method accounts for parallel but offset expression profiles. The Fitch algorithm [FM67] that originates from phylogeny is used for the calculation of a *Euclidean distance tree* from the pre-determined Euclidean distance matrix. The method produces characteristic trees (Figure 2.5). From the tree, the cluster boundaries are determined manually by visual inspection. The authors point out that main branches separate *waves* of gene expression, i.e. groups of genes that have their maximum expression value at the same time point. The computational complexity is high compared with the SOM clustering and the pairwise average-linkage clustering.

2.1.9 A SOM Approach to Gene Clustering

This section describes the SOM algorithm in more detail. At MIPS, I developed a method for clustering genes according to their expression profiles using the SOM. This approach and the features that make the SOM better suited for gene clustering than other clustering methods are discussed.

The SOM is a neural network that provides a mapping from a multi-dimensional data space into a discrete two dimensional space. The method is robust, scalable, flexible, and reasonably fast. Additionally, the clusters are sorted according to the two dimensional regular discrete topology of the map. Thus, neighboring clusters are quite similar, while more distant clusters become increasingly diverse [Koh95]. The mapping is reached via an iterative *learning process*. One generally distinguishes between *supervised learning* and *unsupervised learning*. The learning process is called unsupervised, if no external teacher is involved that would give a correct output during learning. The neural network adapts its weights unsupervised (self-organizing) in response to an external stimulus ξ [Bar89]. The learning rule is often defined as the partial derivative of an *energy function* (also called *cost function*), such that the weight adaption minimizes the energy function. During the iterative weight adaption process, an internal representation of the input data is established that captures interesting features of the high-dimensional input data. The inherent reduction of dimensionality makes these features seizable. Applications for unsupervised learning include classification, clustering, estimation of probability distributions in the input data and the generation of topographic maps [HN90].

Neural networks define a mapping from an input space E into an output space A . The adaption of the network's weights changes this mapping. Let the output space A consist of a regular grid of N neurons. The mapping of the network assigns a reference vector $w_i \in E$ to each neuron $i \in A$. The components of the reference vectors are called the *weights* of the neural network (Figure 2.6). Defining distance measures in input and output

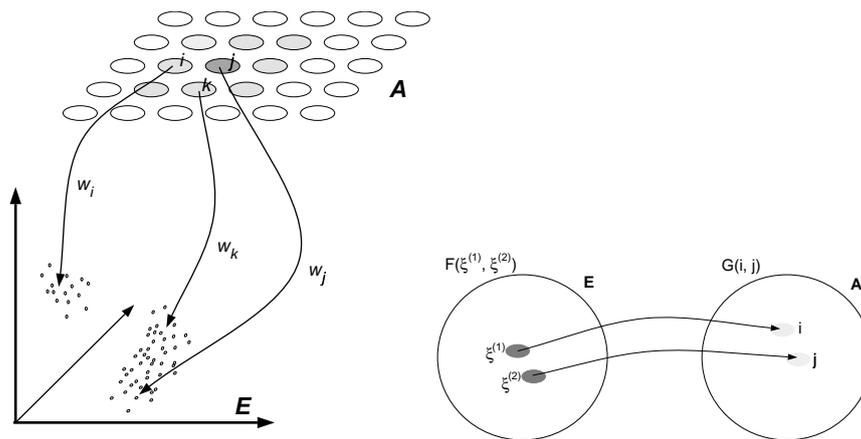


Figure 2.6: Left: Scheme of a neural map. The neurons are arranged on a grid A . Each neuron i has got a reference vector w_i that points into the input space (data space). Right: A topographic mapping of neighboring input data points (external stimuli) $\xi^{(1)}$, $\xi^{(2)}$ onto neighboring points i , j of the output space A . F and G are distance functions that define a measure for the neighborhood of two inputs and outputs respectively [GFS95]. In terms of neural maps the output space corresponds to the regular grid of neurons.

space allows to use the term *neighborhood* of input data and output neurons respectively.

Mappings defined by a neural network are called *topology conserving* by Villmann et al. [VDHM97], if neighboring reference vectors $w_i, w_j \in E$ belong to neighboring neurons $i, j \in A$ and neighboring neurons $i, j \in A$ are assigned to neighboring reference vectors $w_i, w_j \in E$. The mappings are also called *topographic mappings* [GFS95, GS96] (cf. Figure 2.6). Neural networks that establish a topographic mapping are called neural (feature) maps [BGPW96], a term that goes back to Kohonen's Self-Organizing Feature Map [Koh82], now called self-organizing map (SOM). The iterative procedure of learning can formally be described as a Markov process [RMS90]. Initially, the reference vectors of the neurons are randomly distributed over the input space E . At each step, a data point that is randomly chosen from the input data is presented to the network. This data point ξ is called input, or *stimulus*. The neural map adapts the reference vectors of the neurons according to the learning rule (Equation 2.6). A topographic mapping establishes iteratively. A predetermined number of learning steps is applied. The steps are denoted by t (*time*). The learning rule follows the principle of *competition* between neurons. The neuron i whose reference vector w_i is closest to the stimulus ξ is assigned the winner i^* of the competition. The inequality

$$|w_{i^*} - \xi| \leq |w_i - \xi| \quad (2.5)$$

holds for all $i = 1, \dots, N$. The components w_{ij} of the reference vectors w_i are changed according to the learning rule

$$\Delta w_{ij} = \eta \Lambda_t(i, i^*) (\xi_j - w_{ij}) \quad (2.6)$$

for all $i = 1, \dots, N$ and $j = 1, \dots, D$. Note that i^* always depends on the current stimulus ξ .

The neurons are ordered on a grid. For any two neurons, one can calculate their distance on the grid. The neighborhood function $\Lambda_t(i, i^*)$ determines how strong the neighborhood between neurons i and i^* is rated relative to their distance. Λ is a function of the time t and of the distance between the neurons i and i^* within the grid. At any time point t , Λ_t decreases monotonically with increasing distance of the neurons. During the learning process, the neighborhood is normally decreased globally. The learning rule moves the reference vector of the winning neuron i^* towards the stimulus (*Hebbian learning*). The reference vectors of the other neurons are also moved towards the stimulus, but not as much. The adaption rate decreases according to the neighborhood function. The factor η is called learning rate and decreases during the learning process, too. The following set of functions is a reasonable choice [Fel98]: Λ is a Gaussian function with logistically decreasing width σ . The learning rate η is a decreasing logistic function (Figure 2.7):

$$\Lambda_t(i, i^*) = \exp(-|i - i^*|^2 / 2\sigma(t)^2) \quad (2.7)$$

$$\sigma(t) = A_1 / (1 + \exp(C_1(t - B_1))) + D_1 \quad (2.8)$$

$$\eta(t) = A_2 / (1 + \exp(C_2(t - B_2))) + D_2 \quad (2.9)$$

For the calculation of $\sigma(t)$ and $\eta(t)$, t is mapped onto the interval $[-10; 10]$. This arbitrary predefinition eases the parametrization of the functions since the characteristic of the functions does not change when the number of learning steps is adapted.

Figure 2.8 shows the graphical representation of a SOM clustering. The analyzed gene expression data set has been published in [DIB97]. They have measured the expression rates for a set of 6153 genes at seven successive time points. The conditions varied monotonously during the experiment. Yeast was initially grown on a glucose rich medium. During the experiment, the glucose was depleted by the yeast and a reprogramming of the

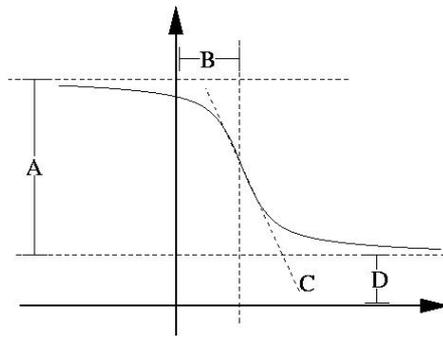


Figure 2.7: A logistic function. The role of the four parameters A, B, C, and D is indicated by the dashed lines. A and D determine the negative and the positive limes. B determines the horizontal transition and C relates to the steepness of the curve.

metabolism was necessary for the organisms to survive. This *diauxic shift* is well understood and therefore this experiment served as a proof of the usefulness of the microarray technique for measuring expression profiles on a genomic scale. Each diagram in the figure corresponds to a cluster. The clusters are shown in the explicit order imposed by the SOM clusterer, conserving the two-dimensional topological ordering featured by the neural map. Within the diagrams, time points or conditions respectively are on the abscissa, the values (logarithmized ratios, here) are on the ordinate. The topological order is clearly visible: neighboring clusters are similar in both dimensions. This similarity of neighboring clusters makes the representation comprehensive. The topological ordering is the basis of the integrative approaches described in Chapter 3. It allows to combine neighboring clusters according to heuristics induced by other data, e.g. functional annotations and metabolic pathways. We can compute a large number of clusters according to the numerical expression data that can be flexibly joined to overlapping, meaningful groups by further analysis steps. Even for large numbers of clusters the result remains comprehensible.

2.1.10 A Case Study: Human Heart Tissue

This section presents results of a cDNA array study that was carried out at the Clinic of Thoracic and Cardiovascular Surgery of the Ruhr-University Bochum at the Heartcenter (HDZ) in Bad Oeynhausen, Germany [MBK⁺01]. At MIPS I have employed the techniques described above to verify or prove wrong a specific working hypothesis formulated within the project.

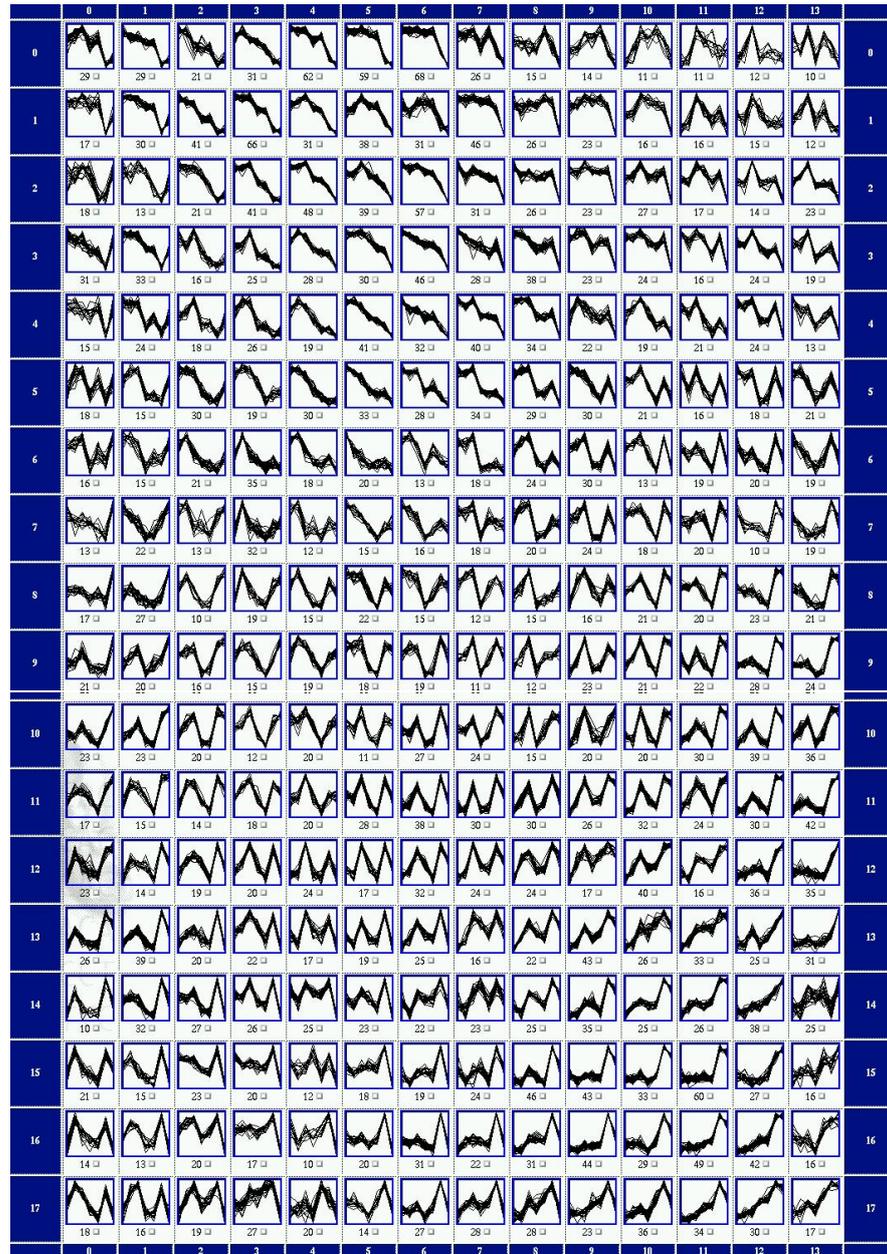


Figure 2.8: The graphical representation of a SOM clustering. Distance measure: Euclidean distance. The data set has been published by [DIB97]. Each diagram represents one cluster and shows a plot of the seven time points vs. the log relative expression levels. The topological order is clearly visible: neighboring clusters are similar.

Description of the Study

Heart failure is the most frequent cause of death in the industrial countries. The ultimate treatment for patients suffering from endstage heart failure is the orthotopic heart transplantation (HTx) despite efforts to develop alternative therapeutical approaches [LVdVC⁺93, KBIU⁺98]. However, because of the shortage of donor hearts a considerable number of patients dies on the waiting list. During the last decade the gap between demand and availability of donor hearts increased considerably. Since the mid of the nineteen-eighties Ventricular Assist Device (VAD) are implanted as a bridge to transplantation in specialized heart centers [DLEB⁺01]. The implantation of VADs leads to a short term recompensation of the patient's circulatory demands and to the recovery of peripheral organs as liver and kidney [FMM⁺94].

The main group of patients bridged to HTx by VAD so far suffers from dilated cardiomyopathy (DCM), which is also the main indication for HTx. The unloading of the failing heart by implantation of VADs has been shown to cause considerable changes of myocardial gene expression in different patients [BMS⁺99]. Apart from these observations the Heartcenter Bad Oeynhausen and other groups have experienced the successful weaning of patients from VADs leading to the idea that VAD might be used as a *bridge to recovery* for some patients. On the other hand, clinical experience has taught us, that in some cases myocardial recovery was not sustained on a time scale of weeks or months after weaning from the VAD [HMM⁺00]. Nevertheless transient VAD implantation is discussed as a putative alternative treatment for heart transplantation in a subgroup of HTx candidates suffering from DCM. But clinicians still lack criteria for identifying patients who might develop a sustained recovery of the myocardium under VAD support [HMM⁺00].

DCM is a disease with a heterogeneous etiology. Whereas 20-30% of the DCM patients are regarded to have a congenital disorder [AMP⁺00] another estimated 30% suffered from DCM as a consequence of a virus infection. DCM is associated with a complex remodeling process in the myocardial wall and reduced contractility of the myocytes caused by an increased wall stress in the enlarged ventricle. The clinical phenotype of DCM is the result of an adaptive process associated with a remodeling process in the myocardium leading to diastolic and systolic dysfunction. The implantation of VAD reduces the wall stress and might therefore induce a reversal of the remodeling process leading to partial recovery of the myocardium.

Using cDNA arrays we have analyzed the gene expression profiles of 588 genes in myocardial samples of ten transplantation candidates suffering from DCM [MBK⁺01]. In parallel we have analyzed in the same samples the

Ca^{2+} dependent ATPase activity as a marker of sarcoplasmic Ca^{2+} uptake. We show that unloading of the left ventricle by VAD induces a reverse remodeling process, which is independent of myocardial Ca^{2+} dependent ATPase activity. The cluster analysis of array data reveals distinct phenotypes in patients supported by VAD.

Ten patients of the heart transplantation program at the HDZ suffering from dilated cardiomyopathy and supported by VAD were analyzed. Seven patients were supported by Novacor LVAD and three by Thoratec VAD (2 LVAD, 1 BiVAD). Mean supporting time was 264 days (range 38-741 days), mean age 51 years (range 12-70 years).

RNA expression patterns of paired myocardial samples from VAD patients were analyzed by commercially available cDNA arrays (Clontech Laboratories; Palo Alto, CA, USA). The myocardial samples of each patient were obtained at the time of VAD *implantation* and at the time of *transplantation* (HTx). Hybridisation was performed in triplicate for each sample according to the manufacturers recommendations with modifications. Arrays, double spotted with 588 different cDNAs, were incubated in parallel with radioactive labeled first strand cDNA. The cDNA arrays were exposed to a phosphor imager screen for 3-5 days. Resulting data files were imported into the Atlas Image software (Clontech). Hybridization spots were aligned and the expression data sets were normalized. The data sets of 46 hybridizations from eight patients were already available in an early stage of the project. We generated a multi-dimensional expression data set from these comprising eight conditions, one per patient. A gene clustering was achieved using the SOM clustering procedure (Chapter 2.1.9).

Methods

Hybridization of cDNA arrays was done in triplicate from the same RNA sample [LKWS00]. Resulting data were evaluated under functional aspects and were analyzed in order to identify differences between expression patterns of patients and genes.

For the normalization we have employed the pure statistical procedure that is described in Chapter 2.1.7. For each gene the medians of all 46 values have been computed. The resulting *pseudo-measurement* of medians was used to adapt the 46 data sets via linear regression. Table 2.1 shows mean, median, standard deviation and skewness of the values of each of the measurements before and after the normalization, respectively. The effect is clearly visible: the mean values are nearly equal afterwards, the standard deviations vary significantly less than before. The skewness values remain practically unchanged by the normalization. As expected, they are not affected by the

Measurement	Mean	Median	Std Dev	Skew	Mean	Median	Std Dev
P1_220500_IMP	13734.5	11278	8092.16	1.52	15344.0	13119.91	7358.24
P1_220500_TRA	14879.4	12426	7221.84	1.66	15324.0	12797.14	7438.14
P1_231299_IMP	19114.8	18192	4312.66	3.75	15344.1	13279.95	9244.74
P1_231299_TRA	25258.3	22872	7376.90	1.87	15344.0	12876.22	7520.76
P1_270600_IMP	20344.3	18436	7575.03	1.51	15344.0	13431.14	7528.45
P1_270600_TRA	18312.7	15568	7852.74	1.70	15344.1	12690.94	7568.71
P2_070700_IMP	15127.4	13298	7845.24	1.57	15266.3	13420.01	7917.80
P2_070700_TRA	15973.6	13406	9425.23	1.22	15344.0	13294.04	7556.14
P2_130600_IMP	4439.9	1690	6270.48	2.40	15344.0	11962.73	7785.54
P2_130600_TRA	9071.9	5246	9345.30	1.40	15344.0	12342.53	7398.49
P2_180200_IMP	19216.7	16906	7696.34	1.66	15344.0	13014.20	7716.30
P2_180200_TRA	17968.8	15670	6498.51	2.10	15344.0	12613.65	7662.84
P3_110900_IMP	18797.6	16070	9359.38	1.23	15212.1	13004.65	7574.58
P3_110900_TRA	20191.9	18024	8746.24	1.34	15111.0	13159.06	7834.96
P3_2808002_IMP	17012.4	14500	9715.66	1.16	15111.0	13114.54	7740.48
P3_2808002_TRA	16088.6	13734	8818.41	1.37	14932.5	12746.97	8185.20
P3_280800_IMP	14497.4	11828	9022.38	1.44	15111.0	12804.79	7823.69
P3_280800_TRA	16303.8	13564	8938.03	1.43	14930.2	12421.08	8185.46
P4_030400_IMP	8736.6	5884	7916.26	1.76	15344.1	12630.58	7594.97
P4_030400_TRA	5015.3	2396	6093.82	2.52	15344.0	12077.50	7669.99
P4_130600_IMP	9995.2	7140	8425.86	1.48	15344.1	12841.59	7449.45
P4_130600_TRA	7643.8	4898	7423.17	1.68	15344.0	12590.74	7512.39
P5_191199_IMP	14506.0	13690	3146.54	4.12	15344.0	12994.18	8738.10
P5_191199_TRA	14429.5	13704	4054.33	3.52	15344.0	13755.96	8589.26
P5_220500_IMP	14186.7	11930	6971.28	1.68	15301.2	12866.99	7519.43
P5_220500_TRA	12753.6	9936	7808.99	1.98	15344.1	12631.75	7545.34
P5_300600_IMP	22154.9	19812	8207.37	1.19	15344.0	13198.75	7458.41
P5_300600_TRA	21256.3	19188	8432.73	1.36	15344.1	13508.90	7435.53
P6_100800_IMP	12630.4	10404	8009.02	1.53	15111.0	12881.53	8049.15
P6_100800_TRA	12248.7	9632	8495.90	1.56	15111.0	12720.23	7804.43
P6_200600_IMP	14648.0	11956	9360.95	1.15	15344.0	13208.69	7467.92
P6_200600_TRA	15796.4	13526	9858.74	1.16	15344.0	13627.84	7500.40
P6_290500_IMP	19501.5	17100	9566.59	1.06	15343.5	13453.92	7527.23
P6_290500_TRA	18106.4	15574	10141.2	1.11	14998.3	12900.56	8400.85
P7_100800_IMP	11177.3	8150	8247.81	1.63	15111.0	12283.14	7746.55
P7_100800_TRA	8957.0	5858	7525.21	1.93	15111.0	11927.16	7778.93
P7_2407002_IMP	19387.3	16696	9255.55	1.30	15374.9	13240.52	7340.43
P7_2407002_TRA	21109.9	18864	7781.32	1.65	15344.1	13180.40	7439.92
P7_240700_IMP	18816.6	16308	8748.24	1.45	15432.3	13374.79	7175.22
P7_240700_TRA	19582.4	17386	7600.98	1.76	15344.0	13148.77	7549.16
P8_200600_IMP	8391.8	5674	7373.85	2.02	15344.0	12612.75	7472.51
P8_200600_TRA	10578.3	7896	8211.96	1.61	15344.0	12944.06	7407.21
P8_240100_IMP	19041.2	17034	6906.06	2.09	15344.0	13078.25	7726.87
P8_240100_TRA	14988.0	13150	5681.87	2.57	15344.0	12779.63	7871.20
P8_290500_IMP	9866.6	7334	7322.60	1.95	15344.0	12753.15	7541.43
P8_290500_TRA	11882.1	9738	7276.82	1.87	15344.0	13149.63	7483.11

Table 2.1: Mean, median, standard deviation, and skewness of the 46 measurements from the HDZ study. Left: values as measured. Right: values after statistical normalization of the measurements.

normalization, except due to numerical inaccuracies. An alternative normalization procedure suggested by the array distributor is the normalization to the constitutively expressed gene GAPDH. Normalization to GAPDH was found to be in rough accordance with the statistical normalization. The statistical normalization over all genes resulted in no significant differences of GAPDH hybridization signals.

The normalized triplicate measurements have been combined by averaging the values of the individual genes, i.e. we have computed the mean expression value of each gene. A log ratio has been computed as the logarithm of the ratio *transplantation* divided by *implantation*. The ratios thus give the relative fold change of gene expression for the respective gene in each patient that occurred in the period between implantation and transplantation.

Results

Since a recovery of the heart muscle had been observed on unloading of the heart by the implanted VAD previously to this study, we hypothesized that the change in gene expression that can be observed between *implantation* and *transplantation* values would depend on the supporting time, i.e. the time that passed from implantation to transplantation. However, clustering using the SOM algorithm has revealed that the differences between patients are more important compared to the length of the period of VAD support.

The order of the patients' values in the data set has been rearranged according to increasing individual VAD support time. Thus the data set can be seen in analogy to a time course of expression values. The assumption is that if there is a significant support time dependence of the expression rate changes, a substantial number of the differentially expressed genes will exhibit almost monotonously increasing or decreasing expression profiles. Figure 2.10 shows the partition of the genes into 12×9 clusters. The clusters exhibit distinct expression profiles. Obviously, there are groups of co-expressed genes on the chip. However, no cluster can be found whose genes exhibit a significant monotonous increase or decrease of expression rates. The differences in gene expression between the patients seem to conceal common changes that occur due to the VAD implantation. This finding supports the characterization of dilated cardiomyopathy (DCM) as a heterogeneous disease also on the gene expression level. A typical common response in gene expression could not be found, neither globally nor for a subset of the measured genes.

Results obtained with methods other than SOM cluster analysis are not mentioned here. Please refer to the original paper for detailed descriptions of methods, results and a medical discussion [MBK⁺01].

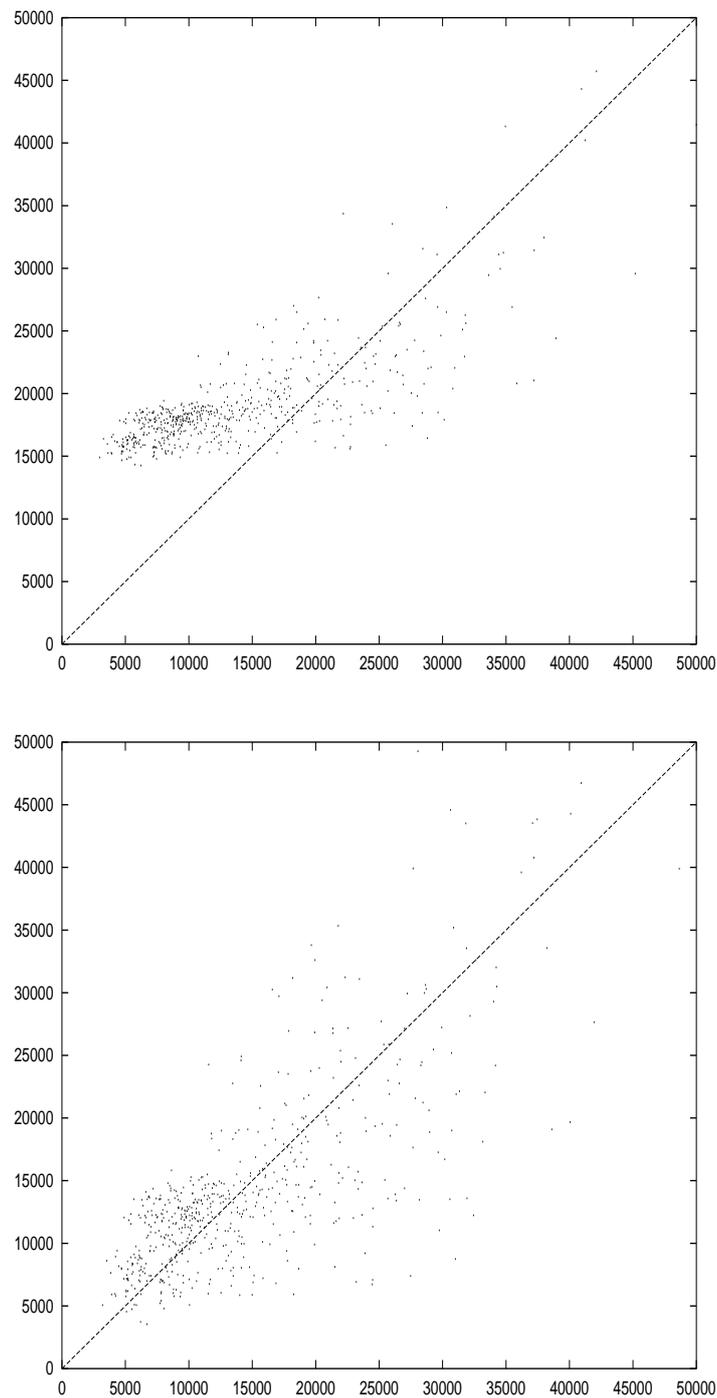


Figure 2.9: Scatterplots of two individual measurements. For each gene, the value of the first measurement is plotted against the value of the second measurement. Above: the raw values. Below: the same measurements after statistical normalization to the pseudo-measurement that consists of the medians of 46 measurements from the same series for each gene (cf. Chapter 2.1.7).

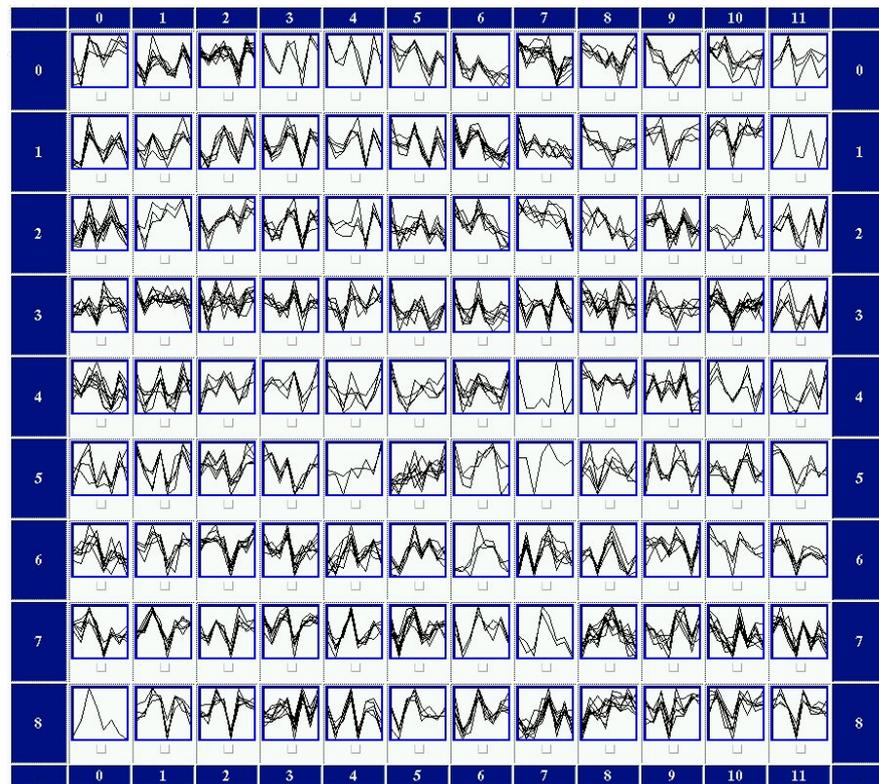


Figure 2.10: SOM clustering (12×9 clusters) of the human heart tissue data set discussed in this section. Each diagram represents a gene cluster showing a plot of the eight pseudo time points vs. the log relative expression levels. Here, no mean/variance standardization was applied to the data.

2.2 Protein-Protein Interactions

The knowledge of the whole genomic DNA sequence of organisms has started a new era in biological research. It is now for the first time possible to identify and analyze all genetic elements of a single organism. Beside the whole genome also the *proteome*, i.e. all proteins expressed by the genome of an organism, is becoming accessible with the further development of current analysis methods and the advent of new technologies like protein arrays [LHE⁺99]. The analysis of the proteome mainly depends on experimental biological data. A substantial amount of data is published, but in general not stored in sufficiently structured form.

Biological processes are mainly determined by *molecular interactions*, e.g. between DNA and proteins, proteins and proteins, or proteins and small molecules. Among these, *protein-protein interactions* play an especially important role since they are essential for virtually every biological process. Protein-protein interactions are the fundamental prerequisites for such complex phenomena as control of the cell cycle, DNA replication, transcription, metabolism and signal transduction. The knowledge of the biological context of a single protein, especially of its interactions with other molecules, is mandatory for a precise understanding of its function in the cell. Studying the functions of individual proteins in various organisms has shown that proteins do not function isolated in a cell but act either in *multi-protein complexes* or in *protein networks*. Often these multi-protein complexes act as highly efficient protein machines [AML92]. These protein machines are assemblies of different protein subunits in which the allosteric movement of individual components are coordinated to carry out complicated tasks which need temporal and spatial coordination.

Besides their importance for the formation of multi-protein complexes, protein-protein interactions are involved in a number of other essential features. Proteins are directed to the correct compartments of cells by binding to other proteins; protein messengers bind to protein receptors on the outer surface of cell membranes to exchange signals between cells; proteins form structural connections between cells; some inhibitors of enzymes are proteins; proteins are modified and degraded by interacting proteins, the enzymes; protein-protein interactions are involved in large-scale movements in organisms, such as muscle contraction. A vast amount of protein-protein interaction data has been generated during the last decades. Recently developed *high-throughput approaches* for a systematic analysis of *genome-wide protein-protein interactions* are widely used, producing large-scale data sets [FRRL97, ITM⁺00, UGC⁺00]. The final goal of studying protein-protein interactions in a given organism is to produce complete *protein interaction maps*.

2.2.1 MIPS Yeast Interaction Tables

One of the main challenges for the analysis and annotation of the genome of the baker's yeast *Saccharomyces cerevisiae* after completion of the sequencing project [GAAC⁺97] was to integrate all available gene-related information of the public domain into a comprehensive yeast genome database [AGH⁺01, MFG⁺00]. The information contained in MYGD is gathered from various sources, mainly the systematic functional analysis projects of yeast [OWKB98] and the yeast literature. Efficient integration of information from the literature requires the application of a standardized terminology as much as possible.

For the annotation of protein-protein interactions MIPS has developed the following format. Each interaction consists of 6 different annotation fields: first interactor, second interactor, type of interaction, method the interaction was detected with, references, and free text for additional information. Two types of interactions are distinguished: physical and genetic interactions. The type of interaction is annotated according to the experimental method applied. Physical interactions are detected by e.g. coimmunoprecipitation, two hybrid assay, and affinity purification; genetic interactions are revealed by methods like extragenic suppression, multicopy suppression, synthetic lethality, and transdominant inhibition. Genetic methods are often just a starting point for further biochemical or cell biological experiments since they only give indirect clues for the interaction of two proteins. This standardized annotation format allows the compilation of the gathered data into tables giving easy electronic access to the data [AHZ00]. So far data about interacting domains of individual proteins have not been systematically introduced into the data set. The MIPS Yeast Interaction Tables have been used for other types of presentation such as in the INTERACT database [EBPH99].

The MIPS yeast interaction tables consist of more than 1000 genetic interactions and more than 2500 physical interactions as of November 2000. For those cases where the interaction type could not be identified from the literature we generated a supplementary table that now contains 197 unclassified interactions.

2.2.2 Visualization of Protein-Protein Interactions

The binary protein-protein interaction compiled in the interaction tables need to be visualized in order to make them comprehensible and in order to be able to judge on the interactions of specific proteins more easily. The following paragraphs describe a procedure I have developed to transfer the

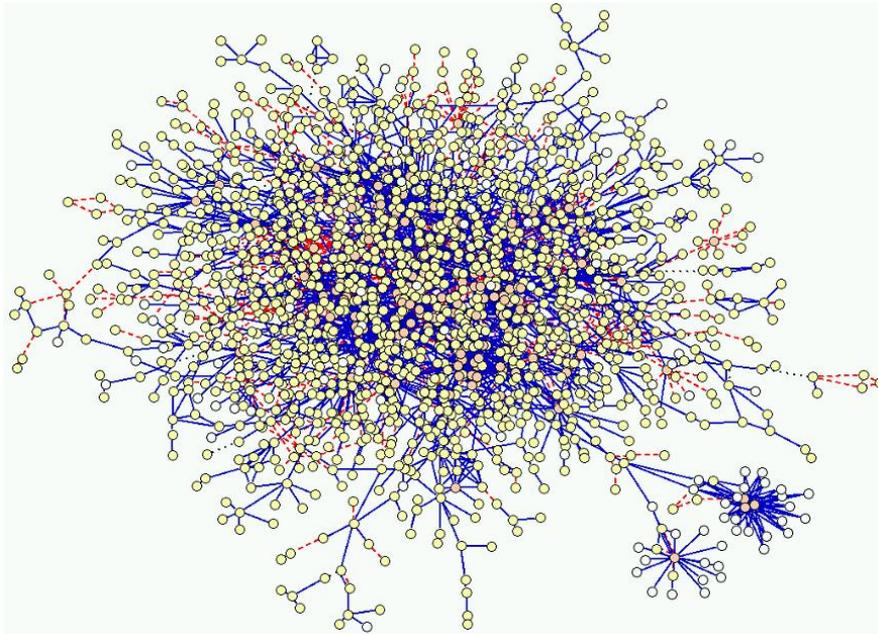


Figure 2.11: The 3500 protein-protein interactions of the MIPS Yeast Interaction Table [AHZ00] represented as a graph. Proteins are shown as nodes, interactions are represented by edges between them. Physical interactions are represented by solid blue edges, genetic interactions by dashed red edges. As the graph is too complex to be of direct use, we employ the integrative analysis approach to restrict the complexity and to focus on a specific biological context (Chapter 3.4).

information from the tables into graphs that can be visualized [FAZ⁺00]. Figure 2.11 shows an example.

Building Protein Clusters

In an iterative procedure we build *clusters of proteins* on the basis of the annotated interactions. Every single protein initially represents its own cluster. For every annotated interaction, where the interacting proteins are not in the same cluster, we join the two clusters involved. After the clustering, every cluster contains all the proteins that interact either directly or indirectly. In a graph theoretic sense, modeling proteins as nodes and interactions as edges, we build clusters of proteins that belong to the same connected component of the whole interaction graph.

Graph Representation

The genetic interactions as well as the physical interactions are binary relations. They are ideally suited for visualization as graphs. Genes and proteins are modeled as nodes, the interactions are represented as edges between the respective nodes. A graph editor tool-kit can be employed for displaying the interaction graphs. We have customized the LEDA graph editor [MN99] for the graphical visualization of interaction graphs. The nodes are labeled with the systematic names or the gene names if available. Edges correspond to interactions and are drawn according to the interaction mode. Physical interactions are represented by solid edges, genetic interactions by dashed edges. A color code can be applied for a deeper characterization of the different methods by which the interactions have been detected. The algorithms of the editor create a suitable layout for the complex graphs, resulting in a clear, easy to grasp picture of the displayed interactions. The user can alter the graph by moving nodes and by deleting nodes and edges.

2.3 Functional Annotations: The MIPS Functional Catalog

During the process of annotation of a sequenced genome, the identified genes and the corresponding proteins are characterized. Complex biological functions are assigned to the proteins. These include general aspects, e.g. naming a protein family that a certain protein belongs to. Other descriptions refer to the exact role of the protein in a special cellular process. The usage of free text for the systematic functional description of proteins is not adequate for computational tasks [Ril93]. In analogy to the established EC catalog [NI92] a hierarchical ordering of the gene products of a cell in terms of their function is an adequate solution for a systematic approach. Different levels of categories group together biochemical functionality according to their role in the organism in rough analogy to biochemical textbooks grouping the biochemical information in paragraphs, chapters and sections.

As the sequence of the yeast *Saccharomyces cerevisiae* was available in 1996 MIPS has generated a special *functional catalog* for yeast [ZHAM01, MAB⁺97]. Significant homologies of proteins to functionally characterized proteins as well as data from the literature derived from biochemical, genetic or phenotypic experiments have been used to assign functions. The yeast functional catalog is hierarchically organized. It contains 15 main categories, each containing 3 to 4 levels of subcategories. In total the catalog consists of more than 200 functional categories. Proteins can be assigned to more than one functional category. This allows a multi-dimensional annotation that takes into account different aspects of function. A protein that belongs to the *glycolytic pathway* category that is a subcategory of the *energy* main category also belongs to *carbohydrate utilization*. Additionally, it may be assigned to the category *cellular organization*. For 3793 out of 6359 yeast genes at least one of the functional categories is assigned, the remaining proteins are assigned to the category *unclassified proteins*.

An updated version of the functional catalog is currently being prepared. It will take into account the experiences made with the original version and will allow a more specific annotation. Recently, a review of functional annotation schemes has been published [RHT00]. According to this review the MIPS Functional Catalog, of all investigated schemes, provides the broadest coverage of cellular functions.

Systematic number	Description of category
01	METABOLISM
01.03.16	polynucleotide degradation
01.05.01	C-compound and carbohydrate utilization
01.05.04	regulation of C-compound and carbohydrate utilization
02	ENERGY
02.01	glycolysis and gluconeogenesis
02.13	respiration
02.19	metabolism of energy reserves (glycogen, trehalose)
03	CELL GROWTH, CELL DIVISION AND DNA SYNTHESIS
03.04	budding, cell polarity and filament formation
03.07	pheromone response, mating-type determination, sex-specific proteins
03.10	sporulation and germination
03.13	meiosis
03.16	DNA synthesis and replication
03.19	recombination and DNA repair
03.22	cell cycle control and mitosis
04	TRANSCRIPTION
04.01.04	rRNA processing
04.03.03	tRNA processing
04.05.01.04	transcriptional control
04.05.03	mRNA processing (splicing)
04.05.05	mRNA processing (5'-, 3'-end processing, mRNA degradation)
04.05.99	other mRNA-transcription activities
04.07	RNA transport
04.99	other transcription activities
05	PROTEIN SYNTHESIS
05.01	ribosomal proteins
05.04	translation (initiation, elongation and termination)
05.07	translational control
06	PROTEIN DESTINATION
06.04	protein targeting, sorting and translocation
06.07	protein modification (glycosylation, acylation, ...)
06.10	assembly of protein complexes
06.13.01	cytoplasmic degradation
08	INTRACELLULAR TRANSPORT
08.01	nuclear transport
08.07	vesicular transport (Golgi network, etc.)
08.13	vacuolar transport
08.19	cellular import
09	CELLULAR BIOGENESIS
09.25	vacuolar and lysosomal biogenesis
10	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION
10.01.05.11	key kinases
11	CELL RESCUE, DEFENSE, CELL DEATH AND AGEING
11.01	stress response
11.07	detoxification
11.11	ageing
11.13	degradation of exogenous polynucleotides
11.99	other cell rescue activities
99	UNCLASSIFIED PROTEINS

Table 2.2: An excerpt of the MIPS functional catalog. The systematic numbers appearing in Tables 3.4, 3.5, and 3.6 are described. The complete functional catalog is available at MIPS [ZHAM01].

2.4 Metabolic Pathways

In textbooks of molecular biology the metabolism of biological organisms is commonly divided into parts according to certain functional tasks performed. These parts are called *pathways*. Because the partition into distinct pathways is purely conceptual, they are referred to as *textbook pathways*. Along these pathways of the intermediary metabolism, organic substances, also called *compounds*, *reactants*, or *metabolites*, are interconverted, i.e. generated (*anabolism*) or degraded (*catabolism*).

2.4.1 Introduction

A major work in the systematic representation of metabolic pathways has been done by German researcher G. Michal, who created the *Boehringer Mannheim Biochemical Pathways Wall Chart* (Figure 2.12). It is available in electronic form and as a paper version [M⁺93]. A textbook version of the famous chart is also available [Mic99].

For metabolic databases, diverse pathway schemes exist. Most of them establish a hierarchy with three to six levels. Entities on the lowest level are then called pathways, higher levels group functionally related and neighboring metabolic pathways and the top level represents the whole metabolic network. Pathways can be specific for a certain organism or represent metabolic knowledge in a generalized form, i.e. containing all metabolic reactions known to exist in some organism.

Besides static information storage in metabolic databases (Chapter 2.4.2), dynamic approaches to metabolism have recently been undertaken. At MIPS I have developed a set of algorithms suited for a dynamic assessment of the metabolic capabilities of organisms for which a substantial number of bioreactions is known to be present. These algorithms are described in the remaining sections of this chapter.

2.4.2 Metabolic Databases and Resources

ENZYME

The Enzyme nomenclature database (ENZYME) is a repository of information relative to the nomenclature of enzymes [Bai00]. It contains an entry for every type of enzyme for that an Enzyme Commission (EC) number has been assigned. As of February 24, 2002 the Enzyme nomenclature database (ENZYME) database contained 3916 entries.

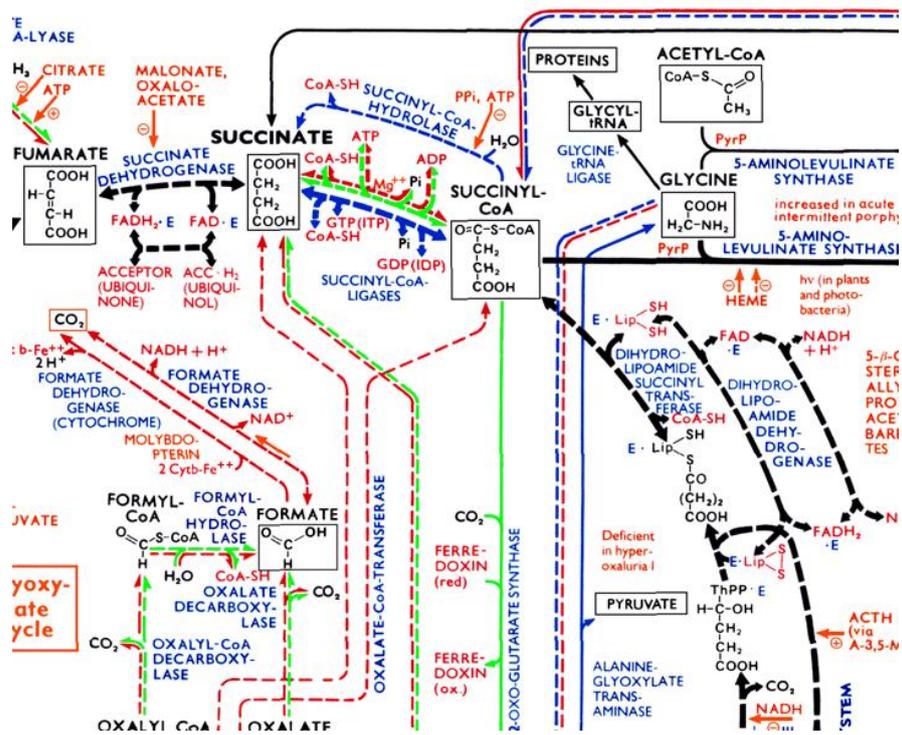
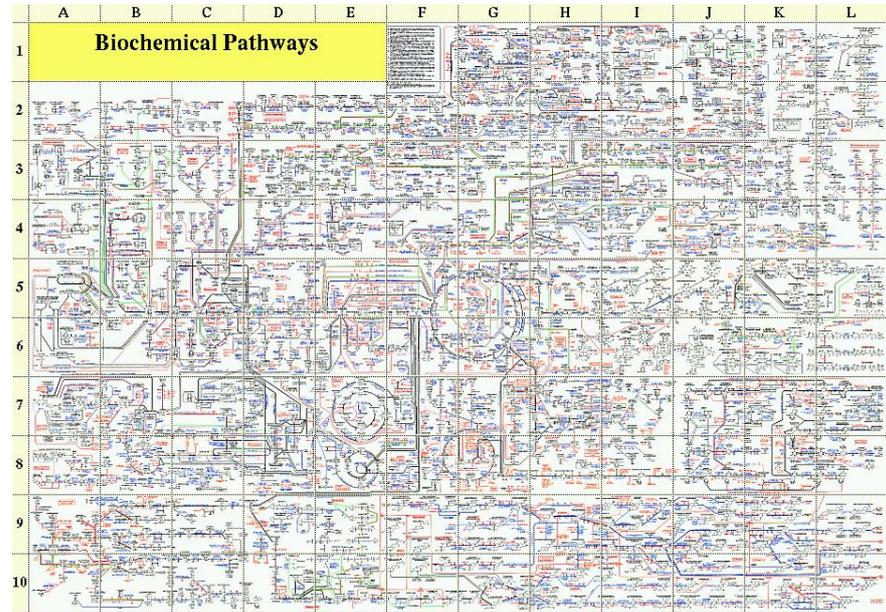


Figure 2.12: The famous *Boehringer Mannheim Biochemical Pathways Wall Chart* [M⁺93]. The detail below (section G5) shows a part of the *tri-carbon acid cycle* metabolic pathway.

```
ID 1.1.1.1
DE Alcohol dehydrogenase.
AN Aldehyde reductase.
CA An alcohol + NAD(+) = an aldehyde or ketone + NADH.
CF Zinc or Iron.
CC -!- Acts on primary or secondary alcohols or hemiacetals.
CC -!- The animal, but not the yeast, enzyme acts also on cyclic
    secondary alcohols.
PR PROSITE; PDOC00058;
PR PROSITE; PDOC00059;
PR PROSITE; PDOC00060;
DR BRENDA; 1.1.1.1.
DR EMP/PUMA; 1.1.1.1.
DR WIT; 1.1.1.1.
DR KYOTO UNIVERSITY LIGAND CHEMICAL DATABASE; 1.1.1.1.
DR P80222, ADH1_ALLMI; P49645, ADH1_APTAU; P06525, ADH1_ARATH;
DR P41747, ADH1_ASPFL; P12311, ADH1_BACST; Q17334, ADH1_CAEEL;
DR P43067, ADH1_CANAL; P48814, ADH1_CERCA; P23991, ADH1_CHICK;
...
```

Figure 2.13: Entry of the alcohol dehydrogenase of the ENZYME database. Field acronyms are the EC number (ID), the official name (DE, description), alternative names (AN), the reaction catalyzed (CA, catalytic activity), cofactors (CF), comments (CC), and cross references (PR, DR).

In principle, the known biochemical reactions are covered by the ENZYME entries. From a computational point of view however, the entries are too general as we can see from the example entry (alcohol dehydrogenase, Figure 2.13). Alcohol dehydrogenases can catalyze a whole set of bioreactions involving a number of different alcohols and a number of different ketones and aldehydes. The ENZYME entry does not enumerate all these specific alcohol dehydrogenase reactions. Thus for the computation of metabolic pathways the given information is not suitable. Even an elaborate hierarchical substrate scheme that defines, e.g., which substrates actually are alcohols or ketones would not help. Additionally, a mechanism is needed that allows to infer on the very ketone that a given alcohol would be converted to.

WIT

The *What Is There?* (WIT) system has been developed by Evgeni Selkov et al. at the Argonne National Laboratory, Illinois, USA. The WIT team produces metabolic reconstructions for completely or partially sequenced organisms. As of February 2002, the reconstructions of 39 organisms are accessible online [OS⁺01].

The WIT pathway diagrams are dynamically drawn. The diagrams link from the compounds of a pathway into the underlying database that provides detailed information on the compound, including a two dimensional diagram

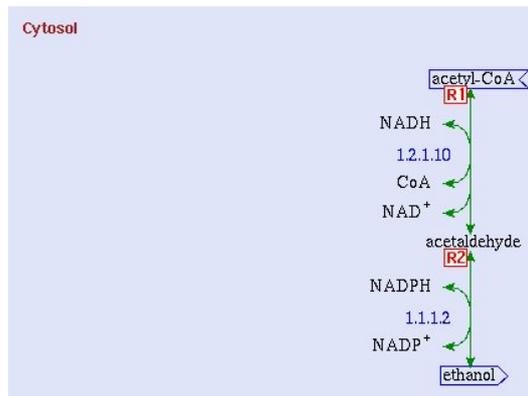


Figure 2.14: A pathway from WIT. The WIT classification scheme is very detailed with up to 6 levels of hierarchy. Some pathways consist of only one or two reactions. WIT pathways are generally linear sequences of biochemical reactions. The organisms for which a specific pathway is asserted are listed. *R1* and *R2* mark links to other pathways.

of its molecular structure. Links from the enzymes of the pathway diagrams lead to the KEGG/LIGAND databases.

KEGG/LIGAND

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive information resource of knowledge of molecular and cellular biology [K⁺01, GK00]. KEGG is largely known as a metabolic database that provides diagrams of manually created metabolic pathways. Besides the metabolic pathway part, KEGG contains some information on signal transduction and regulatory pathways. The KEGG metabolic pathways are internally called *reference pathways*. The term indicates the independence of the pathways from individual organisms. They encode general knowledge of metabolism. The pathways are hierarchically organized in three levels. The lowest level contains specific pathways like *glycolysis* or *cysteine metabolism*. These pathways are grouped on the next level according to a common functional context, e.g. *amino acid metabolism*. The top level integrates all groups of pathways in a node called *metabolism*. Figure 2.15 shows an example of the KEGG pathways on the three levels. Compared to WIT, the KEGG pathways consist of larger networks of biochemical reactions. In most cases, a single pathway contains more than a linear path from one metabolic compound to another.

The Kyoto Chemical Database of Enzyme Reactions (LIGAND) [GNK00] contains the information on enzymes, substrates, and biochemical reactions

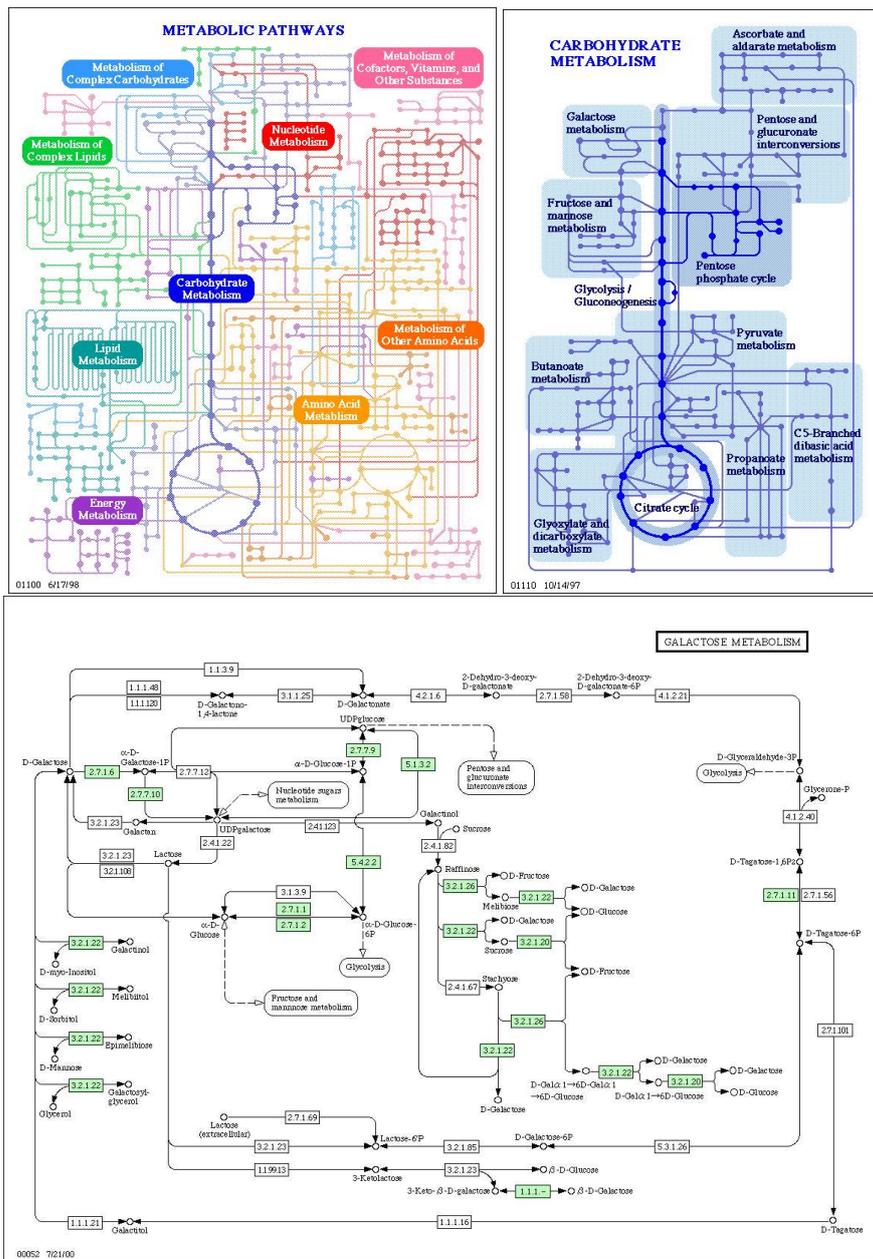


Figure 2.15: The KEGG pathways: an organism independent metabolic network is divided into pathways on three levels. (A) The whole metabolic network, shown top left. (B) The groups of functionally related pathways, here the *carbohydrate metabolism*, shown top right. (C) The individual pathways, here the *galactose pathway*, bottom. In this example, the enzymes that are known to appear in *S. cerevisiae* are marked, they appear shaded in the diagram [K⁺01].

```

ENTRY      EC 1.1.1.1
NAME       Alcohol dehydrogenase
           Aldehyde reductase
CLASS      Oxidoreductases
           Acting on the CH-OH group of donors
           With NAD+ or NADP+ as acceptor
SYSNAME    Alcohol:NAD+ oxidoreductase
REACTION   Alcohol + NAD+ = Aldehyde or Ketone + NADH
SUBSTRATE  NAD+
           Primary alcohol
           Secondary alcohol
           Cyclic secondary alcohol
           Hemiacetal
PRODUCT    Aldehyde
           Ketone
           NADH
COFACTOR   Zinc
COMMENT    A zinc protein. Acts on primary or secondary
           alcohols or hemiacetals; the animal, but not
           the yeast, enzyme acts also on cyclic secondary
           alcohols
           The insect enzyme is a member of the
           nonmetallo-short-chain alcohol dehydrogenase
           (ADH) family (Proc.Natl.Acad.Sci.USA(1991)
           88, 10064-10068).
PATHWAY    PATH: MAP00010 Glycolysis / Gluconeogenesis
           PATH: MAP00071 Fatty acid metabolism
           PATH: MAP00120 Bile acid biosynthesis
           PATH: MAP00350 Tyrosine metabolism
           PATH: MAP00561 Glycerolipid metabolism
GENES      ECO: b0356(adhC) b1241(adhE) b1478(adhP) b3589(yiaY)
           HIN: HI0185(adhC)
           XFA: XF1746 XF2389
....      ....

```

Figure 2.16: Entry of the alcohol dehydrogenase of the LIGAND database. The PATHWAY fields contain references to the KEGG pathway diagrams that contain an enzyme with the respective EC number. The GENES section lists the corresponding genes of various organism, here *Escherichia coli* (ECO), *Haemophilus influenzae* (HIN) and *Xylella fastidiosa* (XFA).

```

R00754:1.1.1.1: Ethanol <=> Acetaldehyde
R00633:1.1.1.1: 1-Alcohol <=> Aldehyde
R04805:1.1.1.1: 3alpha,7alpha,26-Trihydroxy-5beta-cholestane
=> 3alpha,7alpha-Dihydroxy-5beta-cholestan-26-al
R04880:1.1.1.1: 3,4-Dihydroxyphenylethyleneglycol
<=> 3,4-Dihydroxymandelaldehyde
R01035:1.1.1.1: D-Glyceraldehyde <=> Glycerol

```

Figure 2.17: The specific biochemical reactions LIGAND provides for an alcohol dehydrogenase (EC number 1.1.1.1). Only main compounds are listed while additional compounds like H_2O and $NADH$ are omitted.

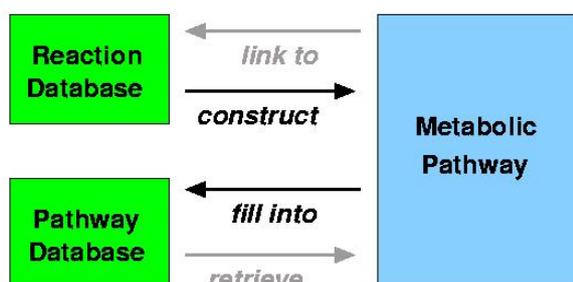


Figure 2.18: Information flow in static systems (metabolic encyclopedias, gray arrows) and in the *AMPhora* system for dynamic modeling of metabolic pathways (black arrows).

that serve as a the basis of the KEGG pathways. The KEGG pathway diagrams link to the respective LIGAND entries. In contrast to ENZYME, LIGAND enumerates the different reactions that enzymes with a single EC number may catalyze in a special file (Figures 2.16, 2.17). The enzyme entries also provide a list of the respective genes of a substantial number of organisms. A feature is implemented that allows to draw organism specific pathway diagrams by highlighting all enzymatic reactions in the pathway diagrams that are annotated to appear in the respective organism. LIGAND can be downloaded as a collection of flat files from the KEGG WWW pages [K⁺01]. We have parsed these files and transformed the LIGAND data into a format suitable for the dynamic pathway modeling approaches described below (Chapter 2.5 ff.).

2.5 Dynamic Modeling of Metabolic Pathways

As described above, metabolic databases today provide a static view of metabolic pathways. The metabolic network of a growing number of organisms is divided into pathways according to a historically developed scheme. These parts mainly correspond to the organization of metabolic pathways in textbooks of molecular biology. The metabolic data resources are also called *metabolic encyclopedias* since they provide the latest knowledge of the metabolism of the sequenced organisms. In contrast to these well established knowledge bases, dynamic approaches to metabolic pathways have not yet been followed much [GNK98].

The algorithmic framework *AMPhora* I have developed at MIPS allows to carry out computations on the set of known reactions of an organism and to construct metabolic pathways without the bias towards historically de-

veloped schemes. It constitutes a framework of methods that allow a very flexible and dynamic view of metabolism. Metabolic systems are described in terms of the stoichiometry of the involved bioreactions. Algorithms that handle sets of these reactions can be used to reveal the pathway structure of the metabolic system under investigation. The *AMPhora* tools are based on the data sets that provide the metabolic reactions and substrates needed for the modeling as described in the previous section. Figure 2.18 visualizes the information flow in both static and dynamic metabolic pathway systems.

We employ two different data structures for the technical representation of metabolic reactions and pathways, a *graph representation* and an *algebraic representation*. The graph representation is used for displaying and traversing the metabolic networks. Metabolic networks lend themselves naturally to being represented as graphs. We can take advantage of standard graph algorithms like depth first search (DFS), breadth first search (BFS), shortest path algorithms and the computation of the connected components of the graph. The algebraic representation maps metabolic reactions and pathways one-to-one onto vectors of some high-dimensional data space. This allows very efficient algebraic manipulations mainly used for the constraint-based pathway construction (Chapter 2.5.5). The representation of metabolic networks as *metabolic graphs* is described below. It is employed for the *analysis of metabolic networks* and the *linear path search*. The algebraic notation is described later in connection with the *constraint-based pathway construction*.

2.5.1 A Dataset for Dynamic Pathway Modeling

For the pathway modeling we need two non-redundant data sets, one set listing the metabolites involved and the other data set listing the biochemical reactions. The latter is organism specific and therefore has to be set up separately for every organism under investigation. We collect the data needed from various LIGAND source files, associating the data fields via the LIGAND keys that identify compounds, reactions, EC numbers, and the genes of the organisms.

The metabolites that appear in the reactions are classified as either *main metabolites* or *side metabolites*. Main metabolites are taken into account for the pathway modeling, while the side metabolites like H₂O, CO₂, ATP, and NADH are only displayed but never considered in the modeling procedures. This is according to the assumption that the side metabolites are available in abundance within the cell or the respective cellular compartment. In a semi-automated processing step we initially have to classify the metabolites as main and side metabolites.

The respective classification can be extracted from the LIGAND data files that describe the metabolic maps. Only those metabolites that do not appear in any of the maps have to be classified by other means. We first classify the metabolites whose name is at least seven characters long as main metabolites, all others as side metabolites. The resulting very rough classification is then corrected manually. Of course it is possible to reclassify any of the metabolites for a specific task or question formulation.

The metabolite data set comprises for every metabolite its name, an internal identifier and the main/side classification. The reaction data set lists for every reaction an identifier, the enzyme's EC number (if assigned), the enzyme name, and the substrates of the reactions. These are divided into reactants and products. For each, the internal metabolite number and the coefficient of the respective substrate in the reaction is given. In principle, biochemical reactions are reversible, i.e. the metabolic conversion can take place in both directions, depending on the conditions, especially the abundance of the respective metabolites. Under physiological conditions, some biochemical reactions are thought to run in one direction only. Our data set, based on the LIGAND data, does not provide information about these cases. We therefore model every reaction as being reversible. The direction of a specific reaction in the data set is mostly arbitrary, i.e. reactants and products might as well be swapped.

2.5.2 Metabolic Graphs

Biochemical reactions and pathways can be represented as directed graphs that we call *metabolic graphs*. Metabolic graphs are formally defined as

Definition (Metabolic Graph). Let \mathcal{M} be a set of (main) metabolites and let \mathcal{E} be a set of enzymes. A *metabolic graph* is a directed graph

$G = (V, E)$ with

nodes $V = \mathcal{M} \cup \mathcal{E}$ and

edges $E \subseteq (\mathcal{M} \times \mathcal{E}) \cup (\mathcal{E} \times \mathcal{M})$. □

Generating a metabolic graph from a set of reactions we transform the individual reactions into their graph representations. For every reaction, an enzyme node is created. The metabolites are also modeled as nodes. Edges leading from the reactants to the enzyme and from the enzyme to the products are added. Modeling all biochemical reactions of a given set within the same graph, only one node is created for each species of main metabolites. A reaction graph consists of an enzyme node along with its adjacent edges and

the metabolite nodes adjacent to these edges. The reaction graphs are automatically linked via common main metabolites to make up the metabolic graph.

Definition (Internal Metabolite, Internal Pathway). A metabolite that is neither accumulated nor consumed by a pathway, i.e. a metabolite that is stoichiometrically balanced within the pathway, is called *internal metabolite* of the pathway. A pathway that consists of internal metabolites only is called *internal pathway*. □

Definition (External Metabolite, External Pathway). A metabolite that is either accumulated or consumed by a pathway, i.e. a metabolite that is not stoichiometrically balanced within the pathway, is called *external metabolite* of the pathway. A pathway that contains at least one external metabolite is called *external pathway*. □

These definitions are first introduced in [SFD00].

The *AMPhora* framework consists of three analysis methods that are described in the remainder of this chapter. They have been developed with a focus on the metabolic interpretation of an eventually incomplete set of reactions, e.g. obtained via the automatic annotation of the genomic sequence of a recently sequenced organism.

2.5.3 Computing Metabolic Networks

A particular instance G of a metabolic graph for a set \mathcal{R} of bioreactions decomposes into a number of *connected components*¹. In a biological context we use the term *metabolic network* for a connected metabolic construct as it is more adequate than a graph theoretical term. In contrast to the term *pathway* we use *network* for a metabolic construct to indicate that it is not the result of a modeling step but results just from linking the reactions of a given set via their metabolites. As indicated above, virtually every metabolic reaction may occur in either of two directions. In a metabolic network, a reaction appears in the one direction that has arbitrarily been assigned to it in the database. Due to the fact that no real modeling step is involved in computing the metabolic networks, there is no biological context according to which one could assign a meaningful direction. This is in contrast to the pathways, where there is a certain context, e.g. a conversion of a metabolite A into another metabolite B , in which the directions of the individual reactions are biologically meaningful.

¹A graph is called *connected* if for all pairs of nodes there is a path from one to the other. A maximal connected subgraph of a graph is called *connected component* of the graph.

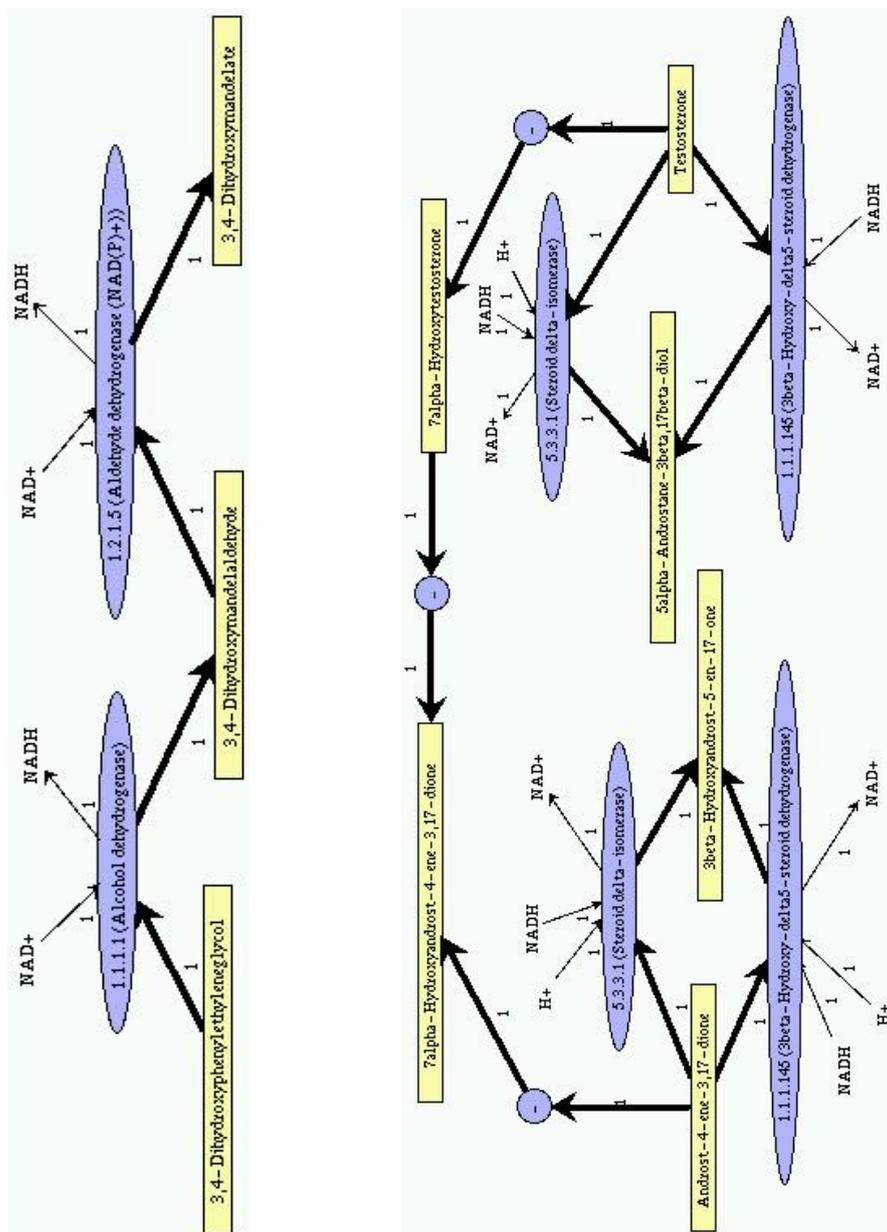


Figure 2.19: Two of the smaller metabolic networks found in the *S. cerevisiae* metabolic graph. These small networks are not linked to the remaining reactions because the main metabolites do not appear in any other reaction. This is due to a still incomplete knowledge and a consequently incomplete data set. The lower network gives an example of enzymes with the same EC numbers transforming different substrates. In the representation, enzyme nodes that are labeled with a dash ('-') indicate biochemical reaction steps that are known to exist though so far there are no enzymes identified that catalyze them. The numbers at the reaction arrows give the multiplicity of the respective metabolite in a reaction.

The generation of the connected components of a metabolic graph, i.e. the computation and the graphical display of metabolic networks is the basic method of the *AMPhora* pathway modeling techniques. Due to incomplete knowledge, the metabolic networks generated from the available reaction data sets are usually not completely linked. The metabolic network analysis is of particular use to investigate the smaller parts of the metabolic graph of an organism. It may also be applied whenever the reaction set to be analyzed consists of only a few biochemical reactions and the resulting metabolic networks keep reasonably small. This may be the case if not many enzymatic proteins of an organism are known or if the reaction set was restricted by methods combined with the pathway modeling in an integrative analysis (Chapter 3.2).

Let the reaction set that serves as an input to the metabolic network analysis consist of n reactions. Depending on the connectivity of the reaction set the metabolic graph may consist of one to n metabolic networks, representing the extreme cases that all reactions are transitively linked or none of the reactions are linked.

The yeast reaction set compiled from LIGAND consists of 1079 reactions interconverting 1141 metabolites². The resulting metabolic graph decomposes into 132 connected components. The largest component contains 847 reactions, i.e. more than 75%. The second largest consists of only 12 reactions. All but one network are thus small enough to be displayed and inspected directly whereas the large network is too big and too complex. Figure 2.19 shows an example of two small metabolic networks found in the metabolic graph of *Saccharomyces cerevisiae* (yeast) set up according to the reaction set distilled from LIGAND. Since in a biological metabolic network reactions would hardly be isolated the connecting reactions between these components and the large main metabolic network are obviously missing from the reaction set. The main metabolic network can be further analyzed by the more elaborate methods presented below.

2.5.4 Linear Path Search

With a growing number of biochemical reactions in the analyzed reaction set, the metabolic graph grows. Its connected components, the metabolic networks, also tend to grow. They become too large and too complex to be displayed and to be visually inspected as a whole. The mere display of the metabolic networks is often not sufficient. The graph-based path search algorithm that forms part of the *AMPhora* framework allows to investigate

²Numbers as of February 2001.

whether linear sequences of reactions exist in the network that transform one specific metabolite into another.

From the set of metabolites appearing in a reaction set, two metabolites S and T are selected as source and sink nodes of a path search. Using a BFS algorithm the length of the shortest path between S and T is computed if such a path at all exists. If it exists, the user is asked to specify the maximal length of the linear paths that should be computed. All paths between the two metabolites that do not exceed the specified maximum path length are determined using a recursive DFS algorithm. The result of the path search is a number of linear hypothetical pathways that can be displayed graphically, either separately (Figure 2.20) or all together in a single non-linear pathway linking the source and the target metabolite (Figure 2.21).

Both standard graph algorithms, the BFS as well as the DFS are adapted to the special requirements of the metabolic graphs. We distinguish between enzyme nodes and substrate nodes. The directed edges of the graph link enzymes with substrates and vice versa. The graph is traversed from substrate node to substrate node, at each step moving over an enzyme node that links the two substrates. All bioreactions of the reaction data set are thought to be reversible, i.e. the metabolic transformation may occur in either of both directions. We therefore have to make the metabolic graph bidirected before searching for paths by constructing the reverse edge $e' = (B, A)$ for every edge $e = (A, B)$. We label the edges according to their direction in order to distinguish between the original edges and the reverse edges. Traversing the graph, we have to take care of the edge directions. Traversing from a metabolite node to an enzyme node and from there to the next metabolite node, we have to use either original edges only, or reverse edges only. Otherwise, we would follow a path that is biologically impossible (Figure 2.22).

Both BFS and DFS, applied on a graph $G = (V, E)$ have a running time of $O(|V| + |E|)$ [MN99]. For the determination of the non-linear cumulative pathway (Figure 2.21) this holds true since we visit each node and each edge at most one time. For computing all linear paths, the situation is different. Potentially there is an exponential number of linear paths between two nodes of a graph. This is because any two linear paths may contain a common sub-path. Consider the graph shown in Figure 2.23. To reach metabolite T from metabolite S a pathway can follow either the upper or the lower reaction at metabolite S and at every metabolite node M_i . With n reactions in the network, this gives rise to $2^{n/2}$ different pathways of length $\frac{n}{2}$. Thus the computational complexity of the linear path search is exponential in the number of reactions. Theoretical and real computational complexity though deviate significantly. Time and space requirements for the computation of paths of a realistic maximum length are reasonable because biological metabolic systems contain long sequences of reactions but few parallel branches [Mav93].

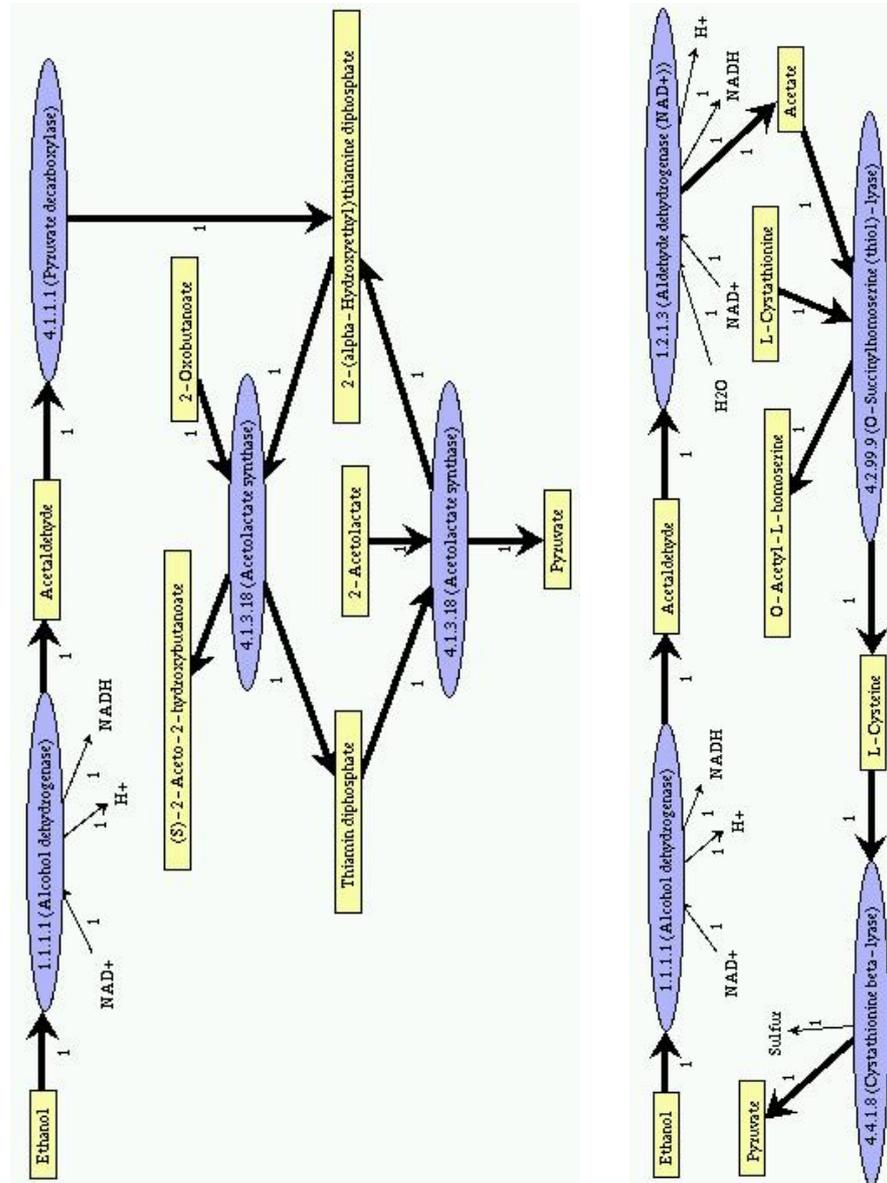


Figure 2.20: Two of 16 linear paths from Ethanol to Pyruvate of maximum length 4 found in the *S. cerevisiae* data set. The length of the shortest path is 3.

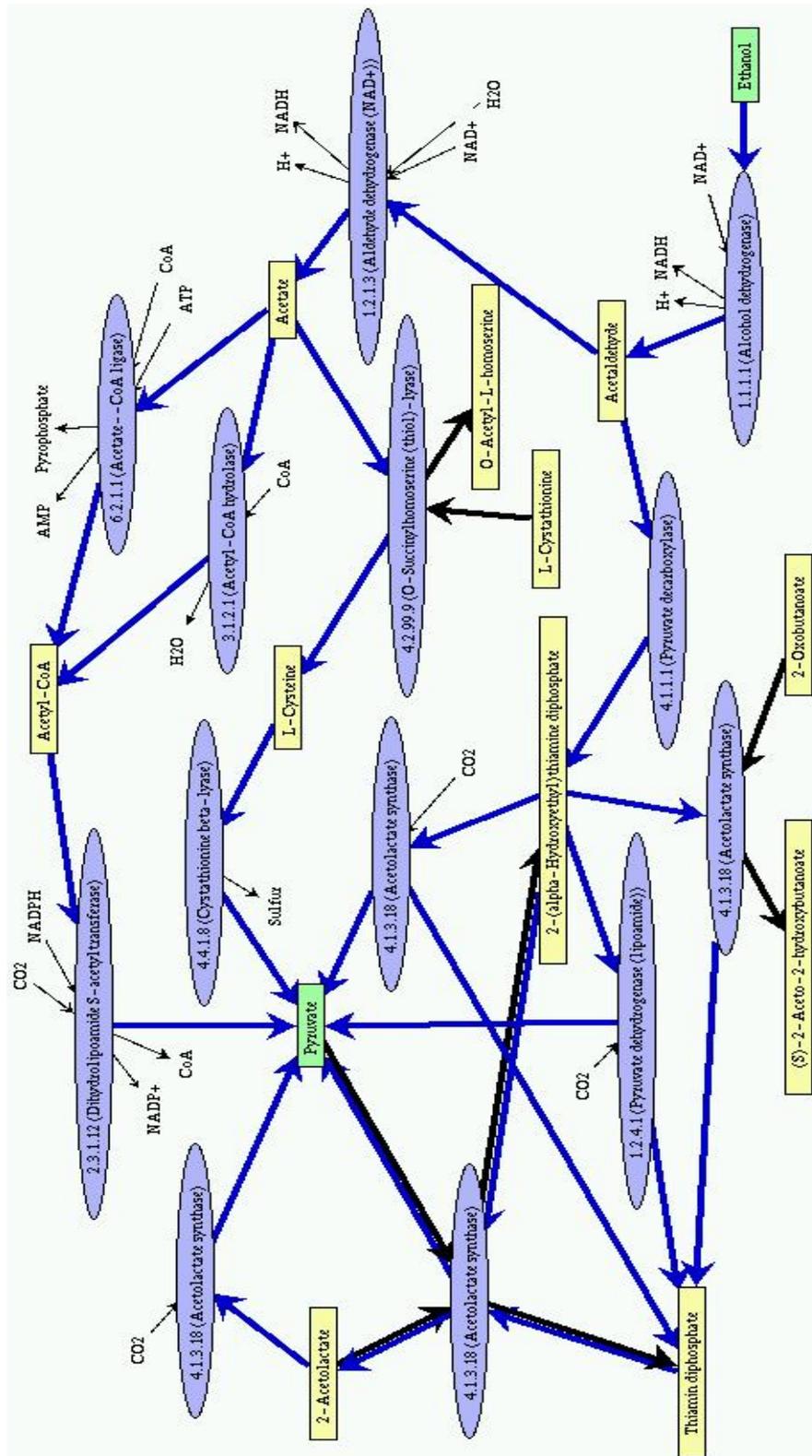


Figure 2.21: The 16 linear paths of maximum length 4 found in the *S. cerevisiae* data set from Ethanol to Pyruvate shown as a single graph.

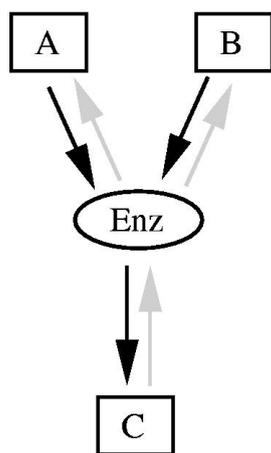


Figure 2.22: The schema visualizes a part of a metabolic graph that represents a single biochemical reaction. The edges that correspond to the original direction of the reaction (from the database) are shown in black. Since the reaction is reversible, we made the graph bidirected by adding the reverse edges shown in gray. Biologically, the enzyme either joins to metabolites, A and B , to form a third, C , or splits C into A and B . It does however not convert A and C into B or B and C into A . Traversing the graph, we have to consider this by either using original edges only or reverse edges only.

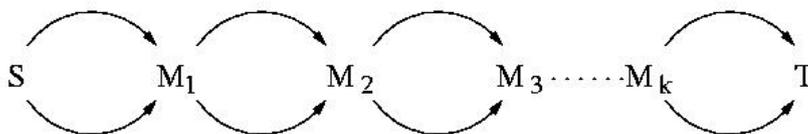


Figure 2.23: A network of reactions that gives rise to an exponential number of different linear metabolic paths.

2.5.5 Constraint-based Pathway Construction

Inspired by a theoretical paper [Red88] I have developed an efficient constraint-based algorithm for the construction of stoichiometrically balanced metabolic pathways. The problem solved by the algorithm is defined as

Definition (Stoichiometric Pathway Construction). Find all metabolic pathways within the whole metabolic network that convert a certain set of metabolites, the reactants, into another set of metabolites, the products, such that no other than the specified metabolites appear as external metabolites of the constructed pathways. \square

An algorithm that accomplishes a similar task is described by [Mav93]. The author shows that the number of pathways constructed by the algorithm can be exponential in the number of bioreactions. The same reasoning applies here as for the linear path search in the previous chapter. Though this observation holds for our problem definition, the employed model of the problem definition and the chosen representation of results allows an efficient computation of the pathways. A concurrent approach has been described in [SLP00] that employs virtually the same method as described here, despite that their approach is based on convex analysis and takes into account the *in vivo* irreversibility of some biochemical reactions. A similar

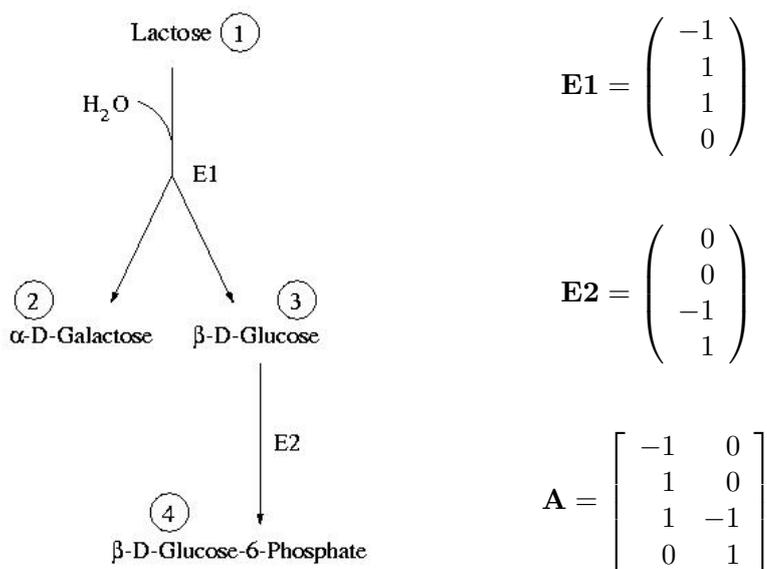


Figure 2.24: An example of the vector notation for reactions. The two reactions on the left interconvert four metabolites, numbered one to four. The vectors for the reactions **E1** and **E2** are given as well as the resulting matrix **A**.

approach to a mathematical description of bioreactions and pathways can be found in [Alb96] with a focus on thermodynamics. [Red88], [SFS99] and [HS98] also use a vector notation for structural and steady state analyses of metabolic networks.

I have developed an algorithm that provides a systematic algebraic description of the identified pathways rather than enumerating them. The length of this description is linear in the number of reactions. The reduction of complexity is achieved by constructing the minimal pathways that satisfy the constraints. These can be linearly combined to form the potentially exponential number of qualitatively different pathways satisfying the constraints.

For the computational handling of the reactions and pathways we employ a vector representation similar to [Red88]. We will show how the described construction problem can be formulated in terms of linear equation systems and how the basis of the solution space of an equation system has to be transformed in order to obtain a result that is intuitive and easy to understand.

The vector notation for metabolic reactions

Given m metabolites in an arbitrary but fixed order, reactions are represented by m -dimensional vectors. The value of a vector component defines the sto-

ichiometry of the corresponding metabolite in the reaction. Opposing signs indicate that the respective metabolites are on opposing sides of the reaction arrow. The definition of the components' signs is arbitrary and does not define the direction of the reaction with respect to its equilibrium. Relative to this technical direction given by the signs of the metabolite components, the coefficient of a reaction in a pathway vector determines which metabolites are reactants and which are products of the reaction within the corresponding metabolic pathway. Figure 2.24 shows an example of the vector notation for reactions.

A set of n bioreactions interconverting m metabolites corresponds to a set $\mathcal{R} = \{r_1, \dots, r_n\}$ of vectors of dimension m that form a $m \times n$ matrix $\mathbf{A} = [r_1 r_2 \dots r_n]$ we call *stoichiometry matrix*.

Metabolic pathways are completely described by their constituting reactions. Thus given n bioreactions in any order, a pathway is represented by an n -dimensional vector \mathbf{x} . Each component of the vector contains the coefficient of the corresponding reaction, where the absolute value defines the multiplicity and the sign defines the direction of the reaction within the pathway.

Multiplying the $m \times n$ stoichiometry matrix \mathbf{A} with an n -dimensional pathway vector \mathbf{x} we obtain an m -dimensional vector \mathbf{b} that corresponds to the overall stoichiometry of the pathway \mathbf{x} :

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \quad (2.10)$$

The vector \mathbf{b} has to be interpreted just like a reaction vector. It represents the metabolic conversion accomplished by the whole pathway.

Formalization of the construction problem

In the construction problem stated above, the set of reactions is given, so we can determine the stoichiometry matrix \mathbf{A} . Let k be the number of allowed reactants and products. All $m - k$ metabolites not specified as allowed reactants or products must be balanced in the constructed pathways, i.e. they must be *internal metabolites*. Considering only the internal metabolites, \mathbf{A} reduces to the $(m - k) \times n$ matrix \mathbf{A}' , the *internal metabolite stoichiometry matrix* [SFS99]. Then, equation (2.10) changes to the homogeneous linear equation system

$$\mathbf{A}' \cdot \mathbf{x} = \mathbf{b}' = \mathbf{0} \quad (2.11)$$

The constraints concerning the balances of the internal metabolites are modeled in this system. In general, the reaction vectors are linear dependent, i.e. $\text{rank } \mathbf{A}' < n$. This effects the dimensionality d of the solution space L , which is

$$d = \dim L = n - \text{rank} \mathbf{A}' \quad (2.12)$$

Normalizing the upper triangular matrix obtained from the Gaussian elimination algorithm leads to a matrix of the form

$$\mathbf{A}'' = \left[\begin{array}{c|c} \mathbf{I} & * \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \quad (2.13)$$

$$\mathbf{I} = \left[\begin{array}{cccccccc} 1 & 0 & * & 0 & \dots & 0 & * & 0 \\ 0 & 1 & * & 0 & \dots & 0 & * & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & * & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & 1 & * & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{array} \right] \quad (2.14)$$

in the notation above, the asterisk marks matrix positions that can carry any value. Leaving out the columns that contain asterisks, we get a square upper triangular matrix with all values on the main diagonal being 1. Between the columns of this reduced matrix may be additional columns as shown. To clarify the matrix structure, let us assume that the interchange of columns is allowed (in practice we would have to keep track of the changes and interchange the elements of the pathway vector in parallel). Then, we can rearrange the columns of the left part of \mathbf{A}'' such that a unit matrix is formed on the very left. The result is a matrix of the form

$$\mathbf{A}''^R = \left[\begin{array}{c|c|c} \mathbf{E} & * & * \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array} \right] \quad (2.15)$$

with \mathbf{E} being the unit matrix of size $\text{rank} \mathbf{A}' = \text{rank} \mathbf{A}'' = \text{rank} \mathbf{A}''^R$. Comparing \mathbf{A}'' and \mathbf{A}''^R , the right parts in the notation are the same while the left part of \mathbf{A}'' corresponds to the left and middle parts of \mathbf{A}''^R . The solution space of equation system 2.11 is

$$L = \left\{ \mathbf{b}; \mathbf{b} = \sum_{i=1, \dots, d} x_i \mathbf{p}_i \right\} \quad (2.16)$$

where the \mathbf{p}_i are basis vectors of L and the x_i are their coefficients in the linear combination vectors \mathbf{b} . The basis vectors are determined from the normalized upper triangular matrix \mathbf{A}'' . The solution space is a systematic description of the qualitatively feasible metabolic pathways satisfying the defined constraints. Each \mathbf{p}_i corresponds to a basic pathway. The basic pathways can be denominated smallest possible pathways satisfying the constraints. A basic pathway is either an *internal pathway* with all involved metabolites balanced (sometimes called a *futile cycle*), or it accumulates or consumes a subset of the allowed reactants and products. These metabolites are external metabolites in the context of this pathway and we call the pathway *external pathway*. By specifying the coefficients x_i , we determine the multiplicity of a basic pathway in the linear combinations of the \mathbf{p}_i . Considering all biochemical reactions as reversible, every linear combination of the basic pathways is a valid pathway with respect to the modeled biochemical properties of the reactions.

Figure 2.25 shows how an external basic pathway can be combined with internal basic pathways to form other external pathways that are qualitatively different from all other basic pathways. The external pathway in the example can be combined with 10 internal pathways. Thus, the systematic description consists of a total of 11 basic pathways, while from combining the external pathway with the possible subsets of the 10 internal pathways, up to 2^{10} qualitatively different external pathways result. Often there are less, namely when two or more internal pathways replace the same reactions in the original external pathway. In these cases, combining the original pathway with two internal pathways does not yield a qualitatively different pathway. Instead, the resulting pathway is the sum of the two pathways that are built by combining the original pathway with the individual internal pathways separately.

Correctness and completeness of the algorithm are directly connected to the correctness and completeness of the algorithm for solving the linear equation system, here Gaussian elimination. The constructed pathways satisfy the constraints since they correspond to solutions of the linear equation system that models the constraints. On the other hand, the vector of any pathway that satisfies the constraints is a solution of the system by definition.

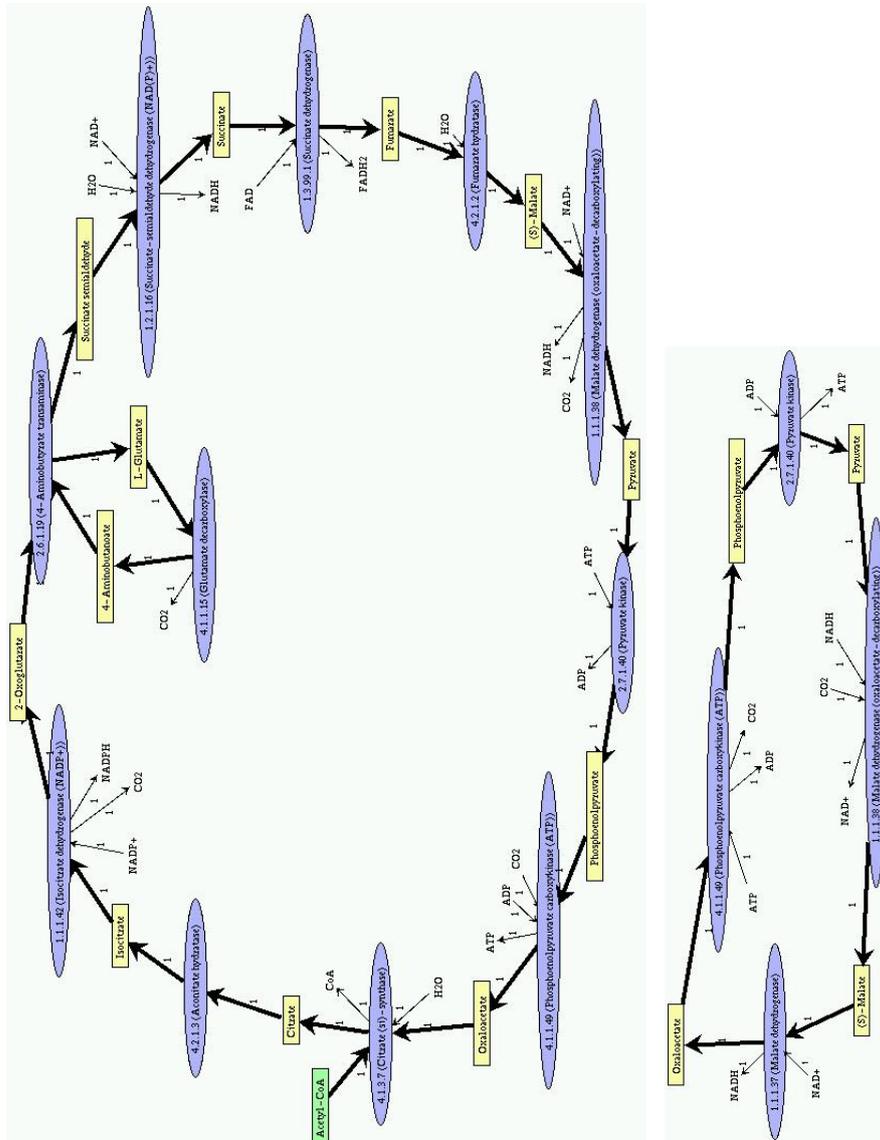


Figure 2.25: Using the constraint-based pathway construction: *Acetyl-CoA* has been defined as the only external metabolite. The upper diagram shows a pathway that consumes *Acetyl-CoA*, the lower pathway is an internal pathway. Linear combination of both pathways results in the reaction 4.1.1.49 in the upper diagram being replaced by the remaining three reactions of the internal pathway, forming a new qualitatively different external pathway. In total 10 internal pathways have been found that can be combined with the displayed external pathway. The linear combination gives rise to up to 2^{10} different external pathways.

Complexity of the algorithm

Avoiding an enumeration of solutions leads to a sub-exponential complexity of the algorithm. The potential number of individual solution pathways is of course still exponential, but the systematic description achieved makes their enumeration unnecessary. Large systems of more than a thousand bioreactions become manageable. The user gains an overview of the feasible pathways instead of being deluged with an enumeration of all pathways.

The core of the method is the algorithm for solving the linear equation system. I have implemented a Gaussian elimination algorithm with pivot search. It has a cubic complexity in the maximum of n and m , the numbers of metabolites and reactions, respectively. Since the number of metabolites and reactions will seldomly be greater than 10^3 by order of magnitude, the cubic complexity of the Gaussian elimination is fully acceptable for typical applications. Gaussian elimination applied on large matrices can lead to numerical inaccuracy due to rounding errors. This is not the case here since the metabolic matrices are sparse. About 99.6% of the entries are zero. Therefore the entries of the matrix do not get too big throughout the elimination and rounding errors have not been observed.

Chapter 3

Integrative Data Analysis

This chapter describes three approaches I have developed for an integrative analysis of biological high-throughput data. The integrative analysis methods are based on the analyses described for the different types of high-throughput data in the previous chapter.

The results show how a comparative and integrative analysis that combines different types of data sets can promote the interpretation of gene expression data and protein-protein interaction data. Following the approach, the data and the structures derived from an internal analysis are further structured and organized by integrating them with other kinds of data. The functional catalog (FunCat, Chapter 2.3) of computationally accessible functional annotations is used for both types of high-throughput data sets. For the interpretation of gene expression data, the systematic functional annotations are used in order to obtain reliable and meaningful groups of genes (*functional projection*). The resulting gene groups are further analyzed in terms of metabolic pathways using the *metabolic mapping* approach. In analogy to the functional projection, the *pathway projection* makes use of predefined textbook metabolic pathways. For the concise analysis of protein-protein interactions, the functional annotations are applied in order to focus on a specific biological context enabling the functional classification of previously uncharacterized genes and proteins.

The chapter is divided into four sections. The first two sections describe the integrative analysis of gene expression data. I have developed the presented methods in cooperation with Kaj Albermann and Jean Hani of Biomax Informatics AG. Their evaluation of the results from a biologist's point of view put the development in the right direction. Section 3.3 discusses the results of both approaches. The fourth section describes the functional analysis of protein-protein interactions that I presented at the *8th International Conference on Intelligent Systems for Molecular Biology (ISMB2000)* held in San

Diego, California, in August 2000. The basis of this section is the respective ISMB paper [FAZ⁺00].

3.1 Integrating Gene Expression Data with Functional Annotations

This section introduces a set of integrative methods for the functional analysis of gene expression data, subsumed under the term *functional projection*. A first step towards the rapid and comprehensive interpretation of gene expression data is the clustering of the genes with respect to the expression patterns [ESBB98]. The individual genes are partitioned into distinct clusters by a clustering algorithm. A neural network approach, the self-organizing map (SOM) [Koh95], is well suited for the analysis of multi-dimensional data. For the gene clustering that is a prerequisite of the functional analysis methods presented here, we use the SOM gene clusterer described in Chapter 2.1. The resulting gene clusterings are further analyzed by the integrative methods of the *functional projection* that make extensive use of the systematic annotations of the genes according to the functional catalog (Chapter 2.3). The *functional projection* can be used interactively as well as in conjunction with an automated group identification algorithm.

The SOM clustering defines a partition of the genes. It assigns each of the genes to exactly one cluster. The clusters are implicitly ordered on a 2-dimensional regular grid. Due to the topology conservation achieved by the SOM, the expression profiles of genes in neighboring clusters tend to be similar. The idea now is to compute a large number of clusters and to subsequently identify groups of neighboring clusters that contain a significant number of genes of the same functional category, either interactively by direct intervention of the user or in an automated fashion. The result of both applications of the *functional projection* is a non-partitional clustering of the genes that takes into account common biological properties. These properties are systematically annotated according to the functional catalog.

3.1.1 Methods

The basic method and the basis of all successive parts of the *functional projection* is the projection of gene functions onto the gene clusters. For a specific, selected functional category, the genes that are described as to belong to this category are determined and are identified in the gene clusters. The results of this computationally straight forward task are striking: for many functional categories it shows that the respective genes are accumulated in

certain regions of the grid of clusters. The projection can be graphically represented in a 3-dimensional plot (Figure 3.2) or in a 2-dimensional plot, where a color coding replaces the third dimension (Figure 3.3). For a given clustering, the functional category to be projected onto the grid of clusters can be determined by the user. The diagrams show the distribution of the respective genes over the clusters and the user can select groups of neighboring clusters according to this distribution. A group of selected clusters can be further analyzed as described below or by application of the metabolic analysis methods presented in the next section.

Finding Groups of Covariant Clusters According to Shared Functional Categories

Having a gene clustering and the projection technique described above at hand, the next step is the automated identification of functional categories whose genes accumulate in certain regions of the clustering map and the determination of tight groups of neighboring clusters that contain a significantly high number of genes of the respective category.

For this automated identification, I developed a greedy group identification algorithm that can be described as follows: Groups are identified separately for each functional category F . A gene g that belongs to F is randomly selected and its cluster c_g is determined. Cluster c_g is assigned to group G . Each of the neighboring clusters $c \in \text{neigh}(c_g)$ of c_g are processed: if c also contains at least one gene of category F and the linear correlation $lc(c, c_g)$ of the mean cluster profiles of c and c_g is above an empirically determined threshold (about 0.85 to 0.9), c is also assigned to G , enlarging the group. Its neighbors are recursively processed. If the group G cannot further be enlarged, a new group is built. The greedy algorithm stops when all the clusters that contain a gene $g \in F$ have been assigned to a group. The result is a set S_F of groups of clusters identified for category F . Figure 3.1 shows a scheme that describes the group finding process graphically.

P-value: Judging on the Significance of Identified Groups

The computed set S_F of cluster groups for a functional category F may contain groups that consist of just one cluster or groups that extend over more than half of all clusters. Only some of the groups represent what we are looking for: compact groups of clusters that contain a substantial number of genes of functional category F . In order to judge on the significance of the identified groups, we apply the following criteria:

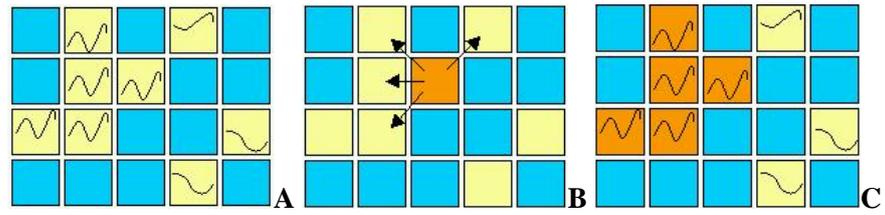


Figure 3.1: The process of identifying groups of clusters that contain genes of the same functional category F . The clusters of a group have to be neighboring. Each cluster has to contain at least one gene that belongs to F and the mean expression profiles of the clusters have to have a high linear correlation. In the fictive example, the yellow clusters contain genes of category F (A). From a seed cluster, the group of clusters is greedily extended to neighboring clusters that also contain genes of F (B). Three clusters are added to the group in the first iteration, another one in the second iteration. One of the considered clusters does not show a sufficient linear correlation with the seed cluster and is spared (C).

- groups must comprise less than 15% of the clusters (less than 30% for functional category 99 of uncharacterized genes).
- groups must have a p-value $P(G) < 2 \cdot 10^{-4}$.

The values for these criteria have been empirically determined. We compute the probabilistic score (*p-value*) according to the following *ad hoc* procedure: let group G comprise n_c clusters of a total of N_c clusters and n_g genes of a total of N_g genes of category F . The p-value $P(G)$ of a group G is the probability to find at least n_g genes of category F in n_c randomly selected clusters provided that the genes have been randomly distributed over the clusters:

$$P(G) = \sum_{i=n_g, \dots, N_g} \binom{N_g}{i} \left(\frac{n_c}{N_c} \right)^i \left(1 - \frac{n_c}{N_c} \right)^{N_g-i} \quad (3.1)$$

The p-value threshold has been determined by visual inspection: the groups G with a p-value $P(G) < 2 \cdot 10^{-4}$ showed to be interesting and would probably also be identified by a user who evaluates the projection diagrams manually. We visualize significant groups using the previously described 3-dimensional plot that shows the distribution of the genes of a certain functional category over the clusters (Figure 3.2). The identified groups of clusters are highlighted in this diagram. Also, we employ the color-coded 2-dimensional schema. Here, the clusters that belong to the respective group are pre-checked to allow the selection of these clusters in order to apply further analysis steps (Figure 3.3).

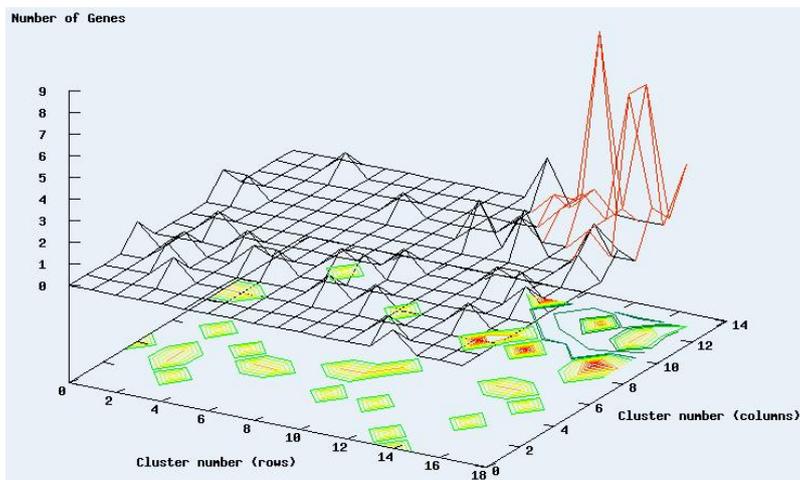


Figure 3.2: 3D-diagram showing the distribution of the genes of functional category *respiration* over the 14×18 clusters. The grid of clusters is drawn in the x/y-plane. The z-axis gives the number of respiratorial genes in each cluster. A cluster group that comprises 14 clusters and 38 respiratorial genes is highlighted in red.

0	0	1	0	0	0	0	0	0	0	10
0	0	0	0	0	2	0	0	0	2	11
0	0	0	0	0	0	1	0	1	0	12
0	0	0	0	0	2	0	1	1	1	13
0	0	1	0	0	0	0	2	9	0	14
0	0	1	0	0	0	1	1	1	0	15
0	1	0	0	1	2	0	7	7	0	16
0	0	0	0	0	1	0	2	1	3	17
4	5	6	7	8	9	10	11	12	13	

Figure 3.3: 2D-diagram showing the distribution of the genes of functional category *respiration* over the 14×18 clusters. The number of genes per cluster is color-coded from green (no genes) to red (maximum number of genes). The clusters of the same group as in Figure 3.2 are checked in this example. The numbers given for each cluster indicate the number of genes of the selected functional category.

Finding Overlapping Groups for Different Functional Categories

Having the significant groups identified for all functional categories one can ask whether groups identified for different functional categories overlap. We straight forwardly determine overlapping groups, i.e. groups containing at least one common cluster. Again, two modes of analysis are implemented: an interactive analysis that focusses on a specific selected group and shows all groups that overlap with the selected group and a batch analysis that identifies the most strongly overlapping pairs of groups of clusters over all functional categories.

We have developed a diagram for the visualization of overlapping groups of clusters based on the 2D-diagram shown in Figure 3.3. Up to three overlapping groups can be visualized according to this schema (Figures 3.4 to 3.6). The clusters of the three groups are colored red, yellow, and blue, respectively. Regions of overlap are shown in the combination color of the involved groups.

3.1.2 Results

To demonstrate the benefit of the *functional projection*, we analyzed public gene expression data sets available for *S. cerevisiae* (yeast). We used the publicly available data set of the diauxic shift in yeast [DIB97]. Similar results have been obtained for other data sets, e.g. [SSZ⁺98] and [CEM⁺98].

For the analysis of the genetic reprogramming associated with the diauxic shift in yeast, expression profiles of more than 6000 genes of *Saccharomyces cerevisiae* have been measured at seven successive time points by DeRisi et al. The shift from anaerobic metabolism (*fermentation*) to aerobic metabolism (*respiration*) is a potentially recurring cycle during the life of yeast. Inoculation of yeast into a sugar rich medium is followed by a rapid growth phase induced by fermentation and the production of ethanol. Exhaustion of the fermentable sugar makes the yeast cells turn to the previously produced ethanol as a carbon source for aerobic growth. The switch from anaerobic growth to aerobic respiration upon depletion of glucose is referred to as the *diauxic shift*. The shift is correlated with a dramatic change in the expression of genes involved in fundamental cellular processes such as protein synthesis, carbon metabolism and carbohydrate storage [JM92]. DeRisi et al. have performed a comprehensive investigation of the temporal program of gene expression accompanying the metabolic reprogramming using DNA microarrays [DIB97]. In the paper, several classes of genes are discussed, such as cytochrome C-related genes and genes involved in the TCA/glyoxylate cycle and carbohydrate storage that are coordinately induced on glucose exhaustion. Genes involved in processes related to protein

No.	Trend	Functional category	P-Value	#Genes	#Clusters
1		ribosomal proteins	7.868213e-104	120	14
3		respiration	1.193299e-28	38	14
4		translation	1.669822e-25	37	20
9		metabolism of energy reserves (glycogen, trehalose)	1.166582e-14	18	12
15		tricarboxylic-acid pathway	7.791692e-12	12	9
22		tRNA-synthetases	8.773009e-10	13	10
34		tricarboxylic-acid pathway	2.975183e-08	6	2
40		ribosomal proteins	4.598571e-07	31	16
52		glyoxylate cycle	9.820714e-06	3	2
53		glyoxylate cycle	9.820714e-06	3	2
57		glycolysis and gluconeogenesis	1.747843e-05	6	4

Table 3.1: The list shows 11 groups for the diauxic shift data set that have a p-value below the threshold of $2 \cdot 10^{-4}$. The properties of these groups and of the genes of their functional categories are discussed in the text.

biosynthesis, such as ribosomal proteins, tRNA synthetases and translation, elongation and initiation factors exhibit a coordinated decrease in expression. An interesting exception is that genes encoding for mitochondrial ribosomal proteins are generally induced rather than repressed after glucose exhaustion.

We have clustered and functionally analyzed the data of the DeRisi experiment with the previously described methods. Using the described parameter settings for the clustering of the diauxic shift data set and for the group finding in the clusterings, 62 significant groups of clusters can be identified. Those groups of clusters that contain genes of the functional classes that have been described by DeRisi et al. as being co-ordinately regulated are listed in Table 3.1. We have analyzed the properties of these groups of clusters in detail, applying the analysis of overlapping groups of clusters and the metabolic mapping approach (Chapter 3.2).

Identifying Groups of Clusters

Groups have first been computed without application of the constraints concerning the number of clusters and the number of genes of a group. This led to 3929 groups that contained an average of 4.7 genes and that have an average p-value of 0.2.

As described in detail below, groups of clusters could automatically be identified for the functional categories and metabolic pathways discussed by

DeRisi et al. The identified groups correspond very well to the manually achieved description in the original publication [DIB97].

Ribosomal proteins. Two groups of clusters (#1, #40) have been found for the *ribosomal proteins*, consisting of 14 and 16 individual clusters. The groups represent the repressed cytoplasmatic and the induced mitochondrial ribosomal proteins, respectively. In total, 151 of the 187 ribosomal proteins are contained in these two groups.

Translation (decreased). The functional catalog lists 63 ORFs of the functional category *translation*. 39 of these genes can be found in a single group of clusters (#4), containing genes which are down-regulated during the diauxic shift. Among these are 35 of the 42 known cytoplasmic translation factor complex proteins. The 24 missing genes are either not contained in the diauxic shift data set (1 gene), were wrongly assigned to the functional category translation (4 genes), are found in nearby clusters that exhibit similar expression profiles (6 genes), are involved in mitochondrial translation processes showing different expression profiles (6 genes), or are involved in cytoplasmic translation but show a different expression profile than the other genes (7 genes).

tRNA synthetases (decreased). Of the 36 *tRNA synthetase* genes of *S. cerevisiae* 15 are known to code for cytoplasmic tRNA synthetases, 12 genes are known to code for mitochondrial tRNA synthetases and 3 genes are known to code for both, cytoplasmic and mitochondrial tRNA synthetases. For the remaining 6 genes, the localization is not yet known. 13 of the 36 tRNA synthetases can be found in a single group of clusters (#22), of which 10 are cytoplasmic, 2 act in cytoplasm and in mitochondria and one has an unknown localization. Additional 5 genes can be found in nearby clusters (4 cytoplasmic, 1 unknown). All of the mitochondrial tRNA synthetases are found in clusters showing a different expression profile than the genes of group #22. Only one known cytoplasmic tRNA synthetase is not found in the clusters of group #22 or in a nearby cluster.

Respiration (induced). The functional category *respiration* contains 71 genes. These genes can be found in 43 different clusters. A group of 14 clusters has been identified by our analysis method that contains 38 respiratory genes (group #3). Among them are all but four of the nuclear-encoded genes known to belong to the respiratory chain complexes. The missing genes were either not in the set of genes analyzed by DeRisi et al. (2 genes) or are assigned to clusters that are near the respiration group of clusters (2 genes). Genes that are required for the assembly of the individual respiration complexes are also found in the respiration group of clusters.

TCA/glyoxylate cycles (induced). In *S. cerevisiae* at least 15 proteins are directly involved in the metabolic conversions of the *TCA cycle*. All but one

of them can be found in either of two directly neighboring groups identified for the category TCA cycle (#15, #34). Two groups of clusters (#52, #53) have been identified that contain the 5 genes known to be directly involved in the *glyoxylate pathway* of yeast. The sixth gene of these groups that is assigned to the category *glyoxylate cycle* has recently been shown to belong to another pathway [LKS⁺00] and is thus wrongly annotated.

Glycogen/trehalose metabolism (induced). The functional category *metabolism of energy reserves (glycogen, trehalose)* contains 17 genes in total, 7 are involved in glycogen metabolism, 7 in trehalose metabolism and 3 are involved in both pathways. Of these, 11 are found in the group of clusters identified for the functional category. Of the remaining 6 genes, 5 can be found in nearby clusters. These 16 genes exhibit expression profiles that indicate an induction during the diauxic shift. The last gene shows a qualitatively different expression profile and is contained in a cluster of repressed genes.

Overlaps of the discussed groups of clusters

The analysis of overlapping groups has been performed to identify possibly co-regulated groups of genes of different functional categories. The previously discussed groups of induced genes revealed the following significantly overlapping groups of clusters (Figures 3.4, 3.5): The groups of clusters identified for the glyoxylate pathway is a subset of the larger TCA group. This indicates that the glyoxylate pathway genes are co-expressed with a part of the TCA cycle genes in this experimental context. The group that contains respiratory genes shows overlaps with all other discussed groups that show an induced expression during the diauxic shift. Looking for significant overlaps in the repressed groups of clusters shows that the group identified for the mitochondrial ribosomal proteins overlaps with the translation and tRNA synthetases groups (Figure 3.6).

This kind of functional analysis elucidates co-ordinated regulation of the genes of different functional categories. Not only can functional categories be identified whose genes are largely co-regulated, but a greater context is shown by pointing out the relation of different functional categories that are eventually involved in the same biological cellular processes.

The interpretation of the overlapping gene groups identified here is continued using the *metabolic mapping* approach that is described in the next section.

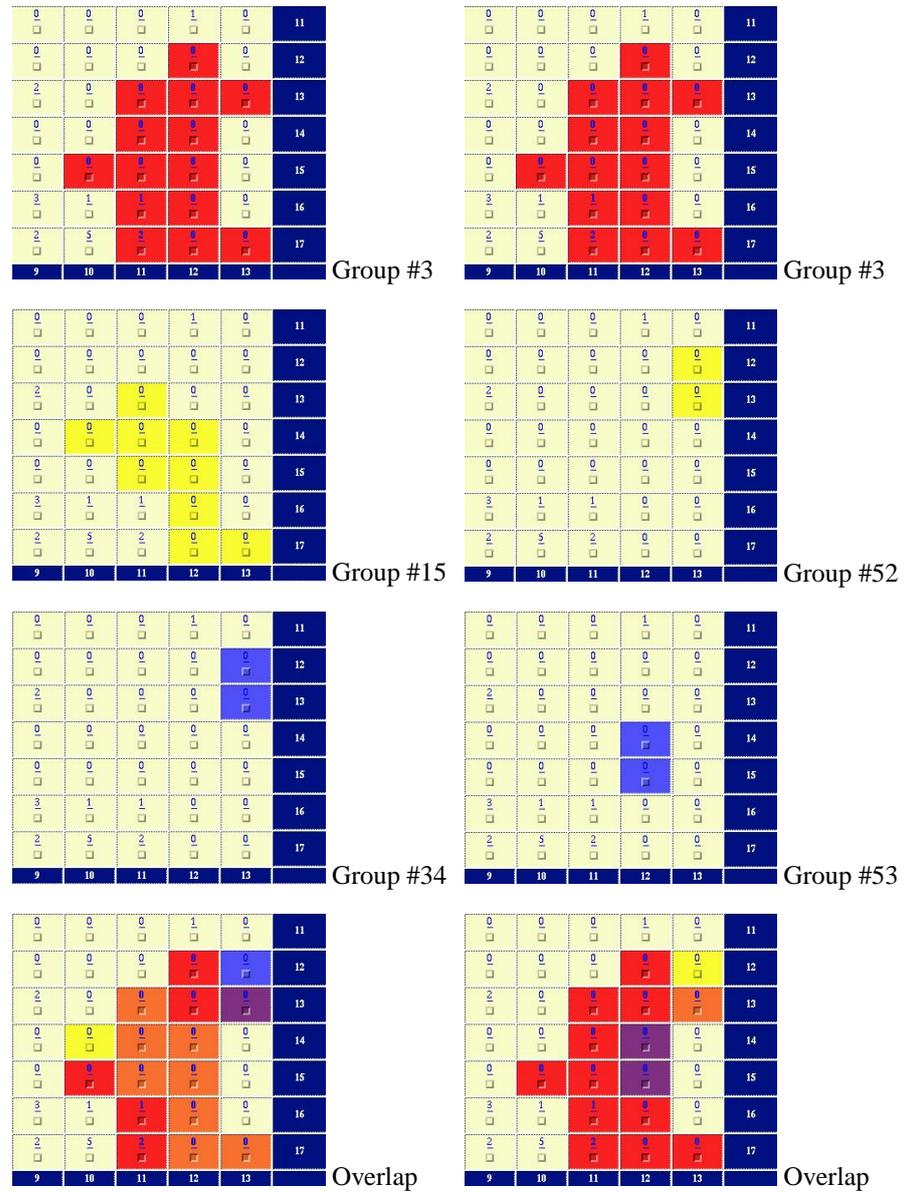


Figure 3.4: Groups of the functional categories with induced genes as discussed in the text. Shown in red is group #3 (*Respiration*). Left: The groups #15 and #34 (*TCA cycle*) are shown in yellow and blue. Right: The groups #52 and #53 (*Glyoxylate cycle*) are shown in yellow and blue. The bottom-most diagrams show the color-coded overlaps. Clusters that belong to more than one group are shown in the respective combination color (red and yellow = orange, red and blue = purple, blue and yellow = green). Those clusters that belong to all three groups are shown in white.

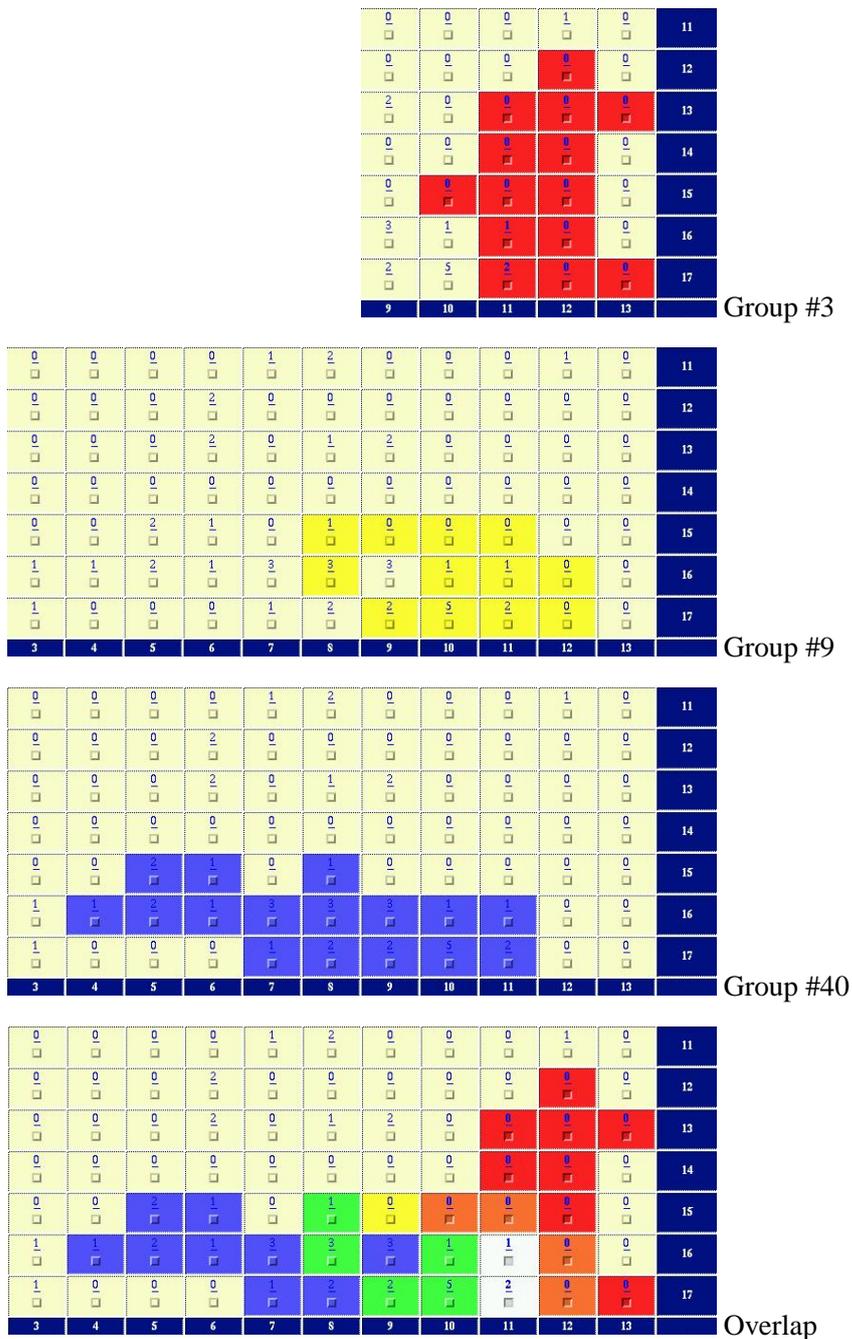


Figure 3.5: Groups of the functional categories with induced genes as discussed in the text (cont'd). Shown in red is group #3 (*Respiration*). The groups #9 (*Glycogen and trehalose metabolism*) and #40 (*Ribosomal proteins, mitochondrial*) are shown in yellow and blue. The bottom-most diagram shows the color-coded overlap.

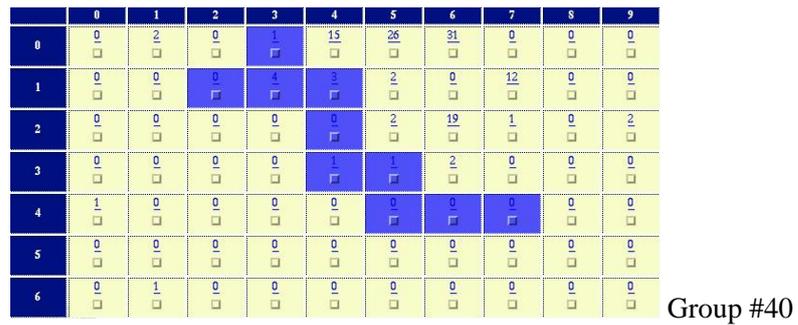
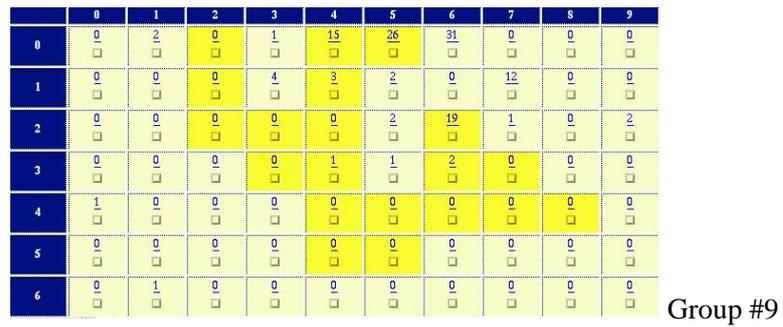
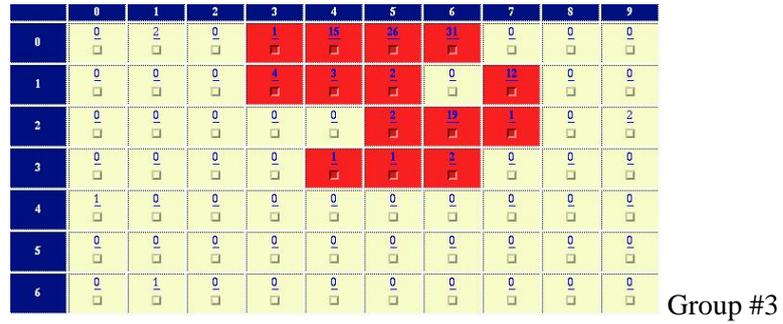


Figure 3.6: Groups of the functional categories with repressed genes as discussed in the text. Shown in red is group #1 (*Ribosomal proteins*, cytoplasmatic). The groups #4 (*Translation*) and #22 (*tRNA synthetases*) are shown in yellow and blue. The bottom-most diagram shows the color-coded overlaps.

3.2 Integrating Gene Expression Data with Metabolic Pathways

This chapter describes two approaches to a metabolic analysis of gene expression data, the *pathway projection* and the *metabolic mapping*. The pathway projection maps predefined textbook pathways onto the grid of clusters of co-expressed genes in analogy to the functional projection. The metabolic mapping is a method of inferring on hypothetical and asserted biochemical pathways that are affected by changes in gene expression. Time courses of expression data are intuitively the most suitable input for such analyses, but expression profiles assembled from non-time course measurements may as well be subject to the described analysis method. We have first described the *metabolic mapping* approach in [FM99] in an early stage. The *metabolic mapping* approach constructs hypothetical metabolic pathways from scratch. As an input it uses known biochemical reactions and the annotations of the measured genes that are related to metabolism. Analyzing the expression data we construct metabolic networks from a set of bioreactions whose corresponding genes are co-expressed.

3.2.1 Pathway Projection

A straight forward way of integrating gene expression data with metabolic knowledge is described by DeRisi et al. They annotate the enzymes in the maps of textbook pathways with the corresponding, color encoded gene expression levels [DIB97]. The gene expression data stems from a series of seven successive measurements of yeast. During the measured period, the glucose concentration in the growing medium is decreased. Thus the genes involved in *glycolysis* and the subsequent metabolic pathways *tri-carbon acid cycle* (TCA) and *pentose phosphate cycle* are subject to substantial changes in gene expression. DeRisi et al. provide a pathway diagram that indicates the change in gene expression at one significant time point compared with a reference hybridization (Figure 3.7). The colors of the enzyme boxes encode upregulation and downregulation. The multi-dimensional expression profiles are not considered in this representation.

The pathway projection follows the inverse direction: the genes that belong to a specific metabolic pathway are projected onto the topologically ordered grid of gene clusters. This approach is in direct analogy to the functional projection. Provided that the functional catalog used contains categories for textbook metabolic pathways, the pathway projection is in fact a special case of the functional projection.

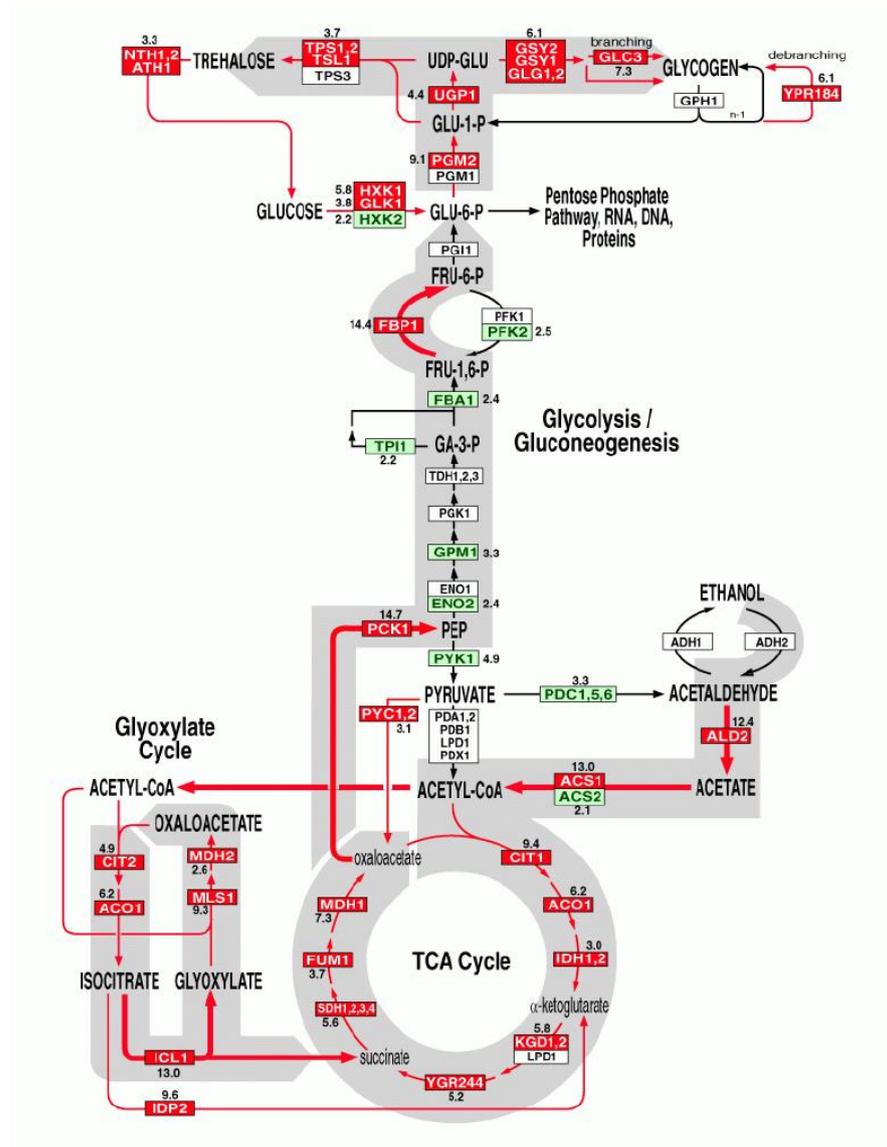


Figure 3.7: Diagram of the glycolysis, TCA cycle and pentose phosphate cycle metabolic pathways from [DIB97]. The colors of the enzyme boxes encode upregulation (red) and downregulation (green) at the most significant time point in a series of gene expression measurements. This is a simple and intuitive way of integrating gene expression data with metabolic knowledge. Nonetheless multi-dimensional data sets cannot be considered.

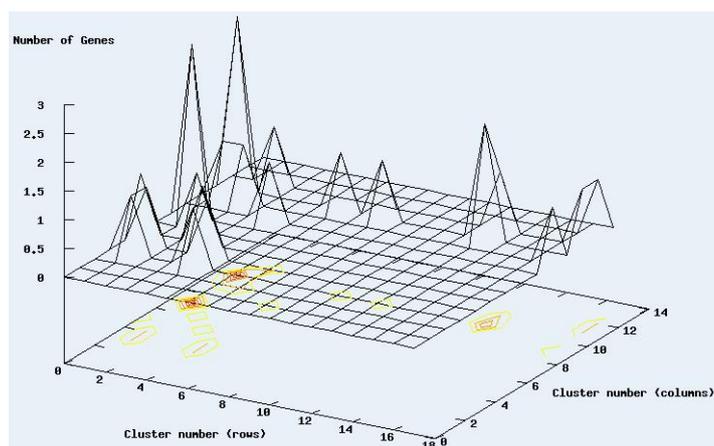


Figure 3.8: 3-dimensional plot showing the distribution of the genes of the *glycolysis* textbook pathway in yeast over the gene clusters resulting from clustering the *diauxic shift* data set [DIB97], Figure 2.8. The regular grid of 18×14 clusters is shown in the plain, while the z-axis represents the number of genes of the pathway in the respective clusters.

The distribution of the genes of the respective pathway over the clusters is shown in a 3-dimensional plot (Figure 3.8). A co-ordinated expression of the genes of a whole pathway or parts of it becomes directly visible in this diagram. According to the projection, a group of clusters can be selected that contain a substantial number of the pathway genes. This group of clusters can in turn be the starting point of a more detailed analysis. The clusters can be put into a functional context other than metabolism using the functional catalog 3.10. The metabolic context of the selected group of clusters can be explored with the *metabolic mapping* approach that is described in the next section.

The textbook pathways are derived from an extended functional catalog that has been developed at Biomax: as of August 2001, 628 pathways have been defined on two hierarchical levels. Pathways from the higher level (147 *super-pathways*) each subsume a number of pathways of the lower level (481 pathways in total). Table 3.2 shows a fraction of the Biomax textbook pathways.

Results

Figures 3.8 to 3.10 give an example for an application of the *pathway projection*. It is applied to the gene clustering of the *diauxic shift* data set to show the distribution of the genes of the *glycolysis* textbook pathway over

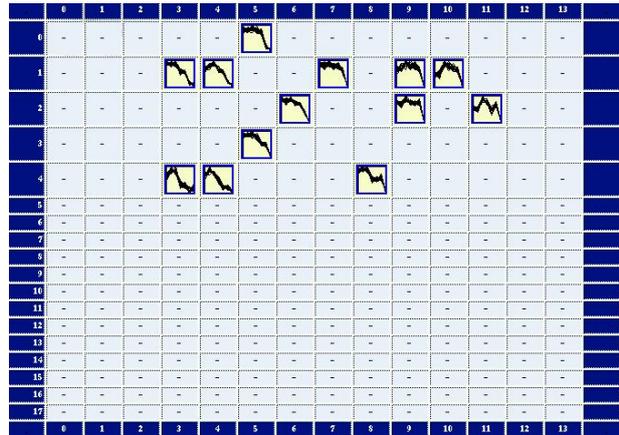


Figure 3.9: The diagram shows a group of selected clusters in the grid of all 18×14 clusters. The clusters of the group do not share highly correlated expression profiles but the expression of the genes of all selected clusters is repressed during the experiment and all clusters contain genes that belong to the *glycolysis* pathway. The clusters form the peaks visible in the upper left region of the 3-dimensional plot (Figure 3.8).

Category	# of Genes	Ratio within Clusters	Ratio within Category	Description of Category
Sort	Sort	Sort	Sort	Sort
30.03	121	27.3%	21.7%	organization of cytoplasm
99	118	26.7%	4.6%	UNCLASSIFIED PROTEINS
05.01	67	15.1%	32.7%	ribosomal proteins
30.10	48	10.8%	6.3%	nuclear organization
01.05.01	31	7.0%	12.0%	C-compound and carbohydrate utilization
03.22	21	4.7%	6.8%	cell cycle control and mitosis
02.01	17	3.8%	48.6%	glycolysis and gluconeogenesis
01.01.01	16	3.6%	13.4%	amino-acid biosynthesis
03.04	15	3.4%	8.9%	budding, cell polarity and filament formation
08.07	13	2.9%	12.1%	vesicular transport (Golgi network, etc.)
03.07	12	2.7%	7.6%	pheromone response, mating-type determination, sex-specific proteins
04.05.01.04	12	2.7%	3.7%	transcriptional control
05.04	11	2.5%	17.7%	translation
01.06.01	11	2.5%	10.0%	lipid, fatty-acid and isoprenoid biosynthesis
30.07	10	2.3%	6.5%	organization of endoplasmic reticulum

Figure 3.10: This table shows the functional categories that are best represented within the 447 genes of the selected clusters. Though a large fraction is uncharacterized (26.7%), still 3.8% of the genes in the clusters belong to the functional category *glycolysis*, nearly half of all genes annotated to belong to this pathway. 7% of the genes belong to the category *carbohydrate utilization* that also the *glycolysis* genes fall into.

Biosynthesis of the serine family

Biosynthesis of serine
Biosynthesis of cysteine
Biosynthesis of glycine

Biosynthesis of the aspartate family

Biosynthesis of aspartate
Biosynthesis of asparagine
Biosynthesis of threonine
Biosynthesis of methionine

Biosynthesis of lysine

Diaminopimelin acid pathway
Aminoadipic acid pathway

...

Glucose catabolism

Embden-Meyerhoff-Parnas pathway
Embden-Meyerhoff-Parnas pathway via EC 5.1.3.3 EC 2.7.1.2
Embden-Meyerhoff-Parnas pathway BPG-independent
Glycosomal glycolytic pathway
UDP glucose metabolism

Galactose catabolism

Galactose oxidation
De Ley-Doudoroff pathway via EC 2.7.7.10 EC 5.4.2.8
Galactose to glycerol
Galactose to sorbitol

...

Glycolipid biosynthesis

Glucosylceramide biosynthesis

Fatty acid biosynthesis

Fatty acid biosynthesis 1
Fatty acid biosynthesis 2
Palmitate anabolism
Palmitoyl-ACP anabolism
Palmitoyl-CoA anabolism
Holo-ACP synthase reaction
Propanoyl-CoA anabolism

...

Metabolism of energy reserves glycogen, trehalose

Glycogen metabolism
Trehalose metabolism

Glyoxylate cycle

Glyoxylate cycle

Table 3.2: A fraction of the 628 standard pathways and super-pathways as defined in the Biomax extension of the functional catalog. The definition of these textbook pathways serves as the basis for the *pathway projection*.

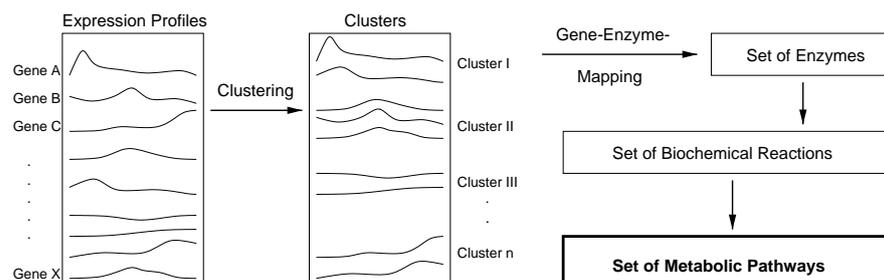


Figure 3.11: *Gene–enzyme mapping.* The genes for that expression levels have been determined in a microarray experiment are partitioned into clusters of similarly expressed genes. One or more clusters are selected. Those genes of the selected clusters that are known to code for an enzyme are mapped onto the reaction that is catalyzed by the respective enzyme. Hypothetical metabolic pathways are constructed from this set of biochemical reactions.

the 252 (18×14) clusters. A group of clusters that contain genes of this pathway is selected from the upper part of the cluster map. As the listing of the best represented functional categories shows, there are 17 genes in the group of selected clusters that belong to glycolysis. These also belong to the functional category *C-compound and carbohydrate utilization* of that 31 genes can be found in the clusters. The carbohydrate utilization defines the larger functional context of the glycolysis genes. Genes of other functional categories can also be found in the selected clusters.

3.2.2 Metabolic Mapping of Gene Clusters

The aim of the *metabolic mapping* is to interpret the gene expression data in terms of metabolism. A prerequisite is the clustering of the genes of an organism. We apply the SOM algorithm as described in Chapter 2.1.9. The topology conservation achieved by the SOM ideally supports the *metabolic mapping*. The clusters of co-expressed genes are used for the dynamic construction of metabolic pathways. We map each of the genes that are known to code for an enzyme to the metabolic reaction or family of reactions that the respective enzyme catalyzes (*gene–enzyme mapping*, Figure 3.11). The reaction database is the basis for this mapping. A discussion of reaction databases can be found in Chapter 2.4.2, the origin of the reaction data set used here is described in Chapter 2.5.1.

After a clustering of the genes of a gene expression data set has been achieved, each cluster corresponds to a set of biochemical reactions according to the *gene–enzyme mapping*. This is realized via the annotated EC numbers. The EC numbers that are carried by at least one of the genes of a cluster are mapped onto the biochemical reactions that correspond to the

EC numbers. One EC number may correspond to exactly one enzymatic reaction or to a set of reactions that vary in their co-factors (the side metabolites like H_2O , CO_2 or $NADH$) or their main substrates. Some enzymes potentially convert a group of different metabolites, e.g. alcohol dehydrogenases that may convert different alcohols into aldehydes or ketones. Using the annotation of EC numbers and the reaction database, the *gene-enzyme mapping* translates a set of genes into a corresponding set of biochemical reactions. Note that the inverse mapping is not possible since in most organisms there are numerous metabolic conversions that are coded for by more than one gene, i.e. different genes may have the same EC number.

The reaction sets are subject of the subsequent pathway construction. Unions of reaction sets can be analyzed by selecting more than one gene cluster. This allows to carry out various analyses: the selected clusters may origin from the same region of the topologically ordered grid of clusters. In this case, all genes of the union are more or less co-expressed, having similar expression profiles. Clusters may also be selected according to other heuristics. A cluster that contains a particular gene of interest can be selected along with those clusters in the surrounding that show similar expression patterns. Or one selects a group of clusters that contain a substantial number of genes of a particular functional category (*functional projection*, Chapter 3.1) or of a textbook metabolic pathway (*pathway projection*, Chapter 3.2.1).

The set of biochemical reactions resulting from the *gene-enzyme mapping* is transformed into a graph representation according to the procedure described in Chapter 2.5. The connected components of this graph, the *metabolic networks*, are computed. These metabolic networks allow a comprehensive analysis of gene expression data in the context of metabolism. They consist solely of enzymatic reactions whose encoding genes showed to be co-expressed in the experiment for that the expression data has been obtained. For such a metabolic analysis of gene expression data, knowledge of the biochemical reactions of the examined organism has to be available. The genes must be annotated, i.e. they must be functionally classified and an EC number has to be assigned to those genes that code for an enzyme. The annotation of the genes may be derived from public databases or can be assigned by an automatic bioinformatic method. Tools like the PEDANT system [FAH⁺01] achieve annotations of large sequence stretches. They extract potential coding regions (ORFs) and assign annotations by sequence comparison with already annotated genes. This is based on the assumption that the cellular function of a gene can be inferred from the genes of other organisms by sequence homologies on the DNA level. Automatic annotation of sequences is of particular interest, if the expression profiles have been obtained for a set of functionally uncharacterized ESTs or if the sequences of an organism are the property of a company, such that no information on

the sequence set is available in public databases.

The genes of a number of selected clusters are mapped onto the corresponding set of enzymes by the *gene–enzyme mapping*. The set R of bioreactions that these enzymes catalyze is used as the input for the pathway modeling that results in a number of metabolic networks. With R containing n reactions, the number of constructed networks ranges from one network containing all reactions up to n networks, each consisting of only one reaction. The metabolic mapping is computationally efficient. The *gene–enzyme mapping* can be done in linear time as far as the number of genes is concerned. The LEDA algorithm for computing the connected components of the metabolic graph has a complexity of $O(|V| + |E|)$ [MN99], i.e. linear complexity in the number of nodes and edges of the metabolic graph.

Results

Analyzing publicly available data sets with the metabolic analysis methods described here, we obtained biologically meaningful and surprisingly detailed results. The *metabolic mapping* approach generated hypothetical pathways that often showed parts of textbook pathways surrounded by single reactions of neighboring metabolic pathways. An example is shown here for another data set, the yeast cellcycle data set of Spellman et al. [SSZ⁺98]. The graphical overview of the 252 gene clusters that have been determined with the SOM clusterer is shown in Figure 3.12. We used the pathway projection to determine those clusters that contain genes belonging to the *fatty-acid biosynthesis* pathway. The 3D projection identifies two clusters that are close to each other and that contain 11 of the 15 genes involved in this pathway (Figure 3.13). One of them, the cluster that contains 6 pathway genes, is selected using the 2D projection diagram (Figure 3.14). Applying the *metabolic mapping* to the genes of this cluster, we generated two large metabolic networks along with a number of smaller networks that consist of only one or two biochemical reactions. Figure 3.15 shows the biggest network. It shows an impressively large linear fraction of the fatty acid biosynthesis pathway. Some conversion steps are catalyzed by two parallel biochemical reactions. A single enzyme, the *fatty-acyl–CoA-synthase* (EC 2.3.1.86), appears numerous times in this hypothetical pathway. The surplus value of such a dynamically created pathway diagram compared with a representation of a textbook pathway is that all the genes that code for one of these enzymes showed to be co-expressed in the cellcycle experiment. Here, the *metabolic mapping* approach revealed that a large linear fraction of the fatty-acid biosynthesis pathway is co-ordinately regulated during the yeast cellcycle.

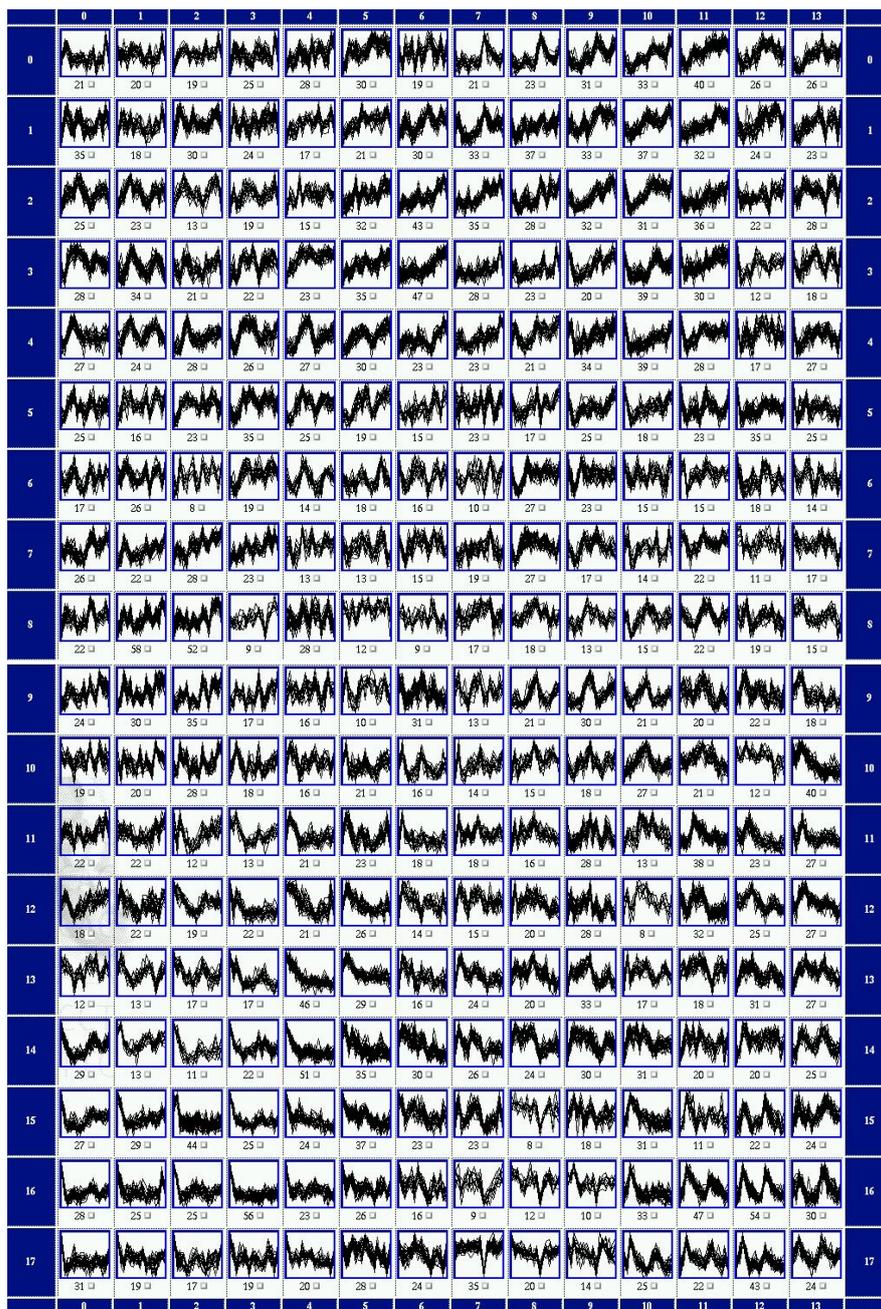


Figure 3.12: Overview representation of the gene clustering obtained from the cell-cycle experiment of Spellman et al. [SSZ+98]. 252 clusters have been computed. The data set consists of gene expression data from sixteen time points. During the experiment, the yeast culture under observation completed two cellcycles. The expression profiles of many clusters show a periodical pattern that reflects the cellcycles.

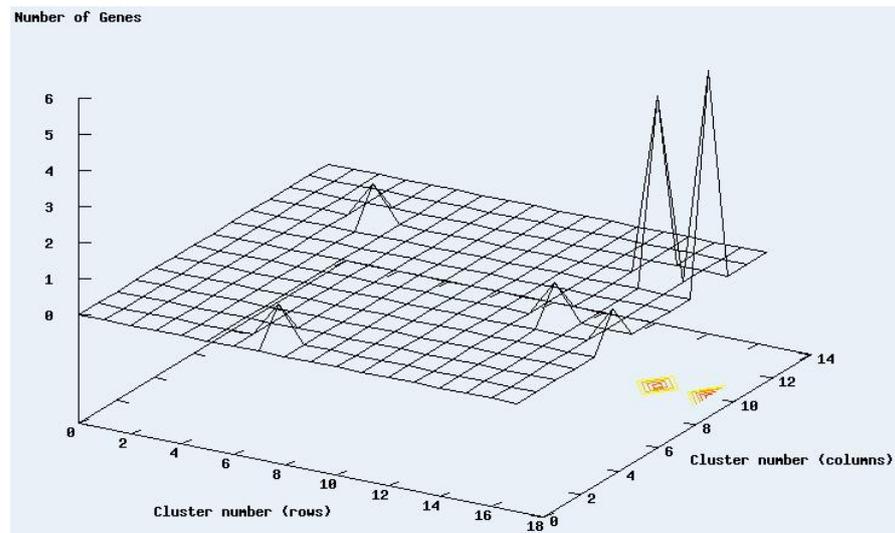


Figure 3.13: Pathway projection: 3D plot of the projection of the *fatty-acid biosynthesis* pathway. The grid of 18×14 clusters is drawn in the plain. The z-axis shows the numbers of pathway genes in the individual clusters. 11 of 15 genes can be found in two clusters that are additionally close to each other. This strongly supports the assumption that the *fatty-acid biosynthesis* pathway is co-ordinately regulated in yeast during the cellcycle.

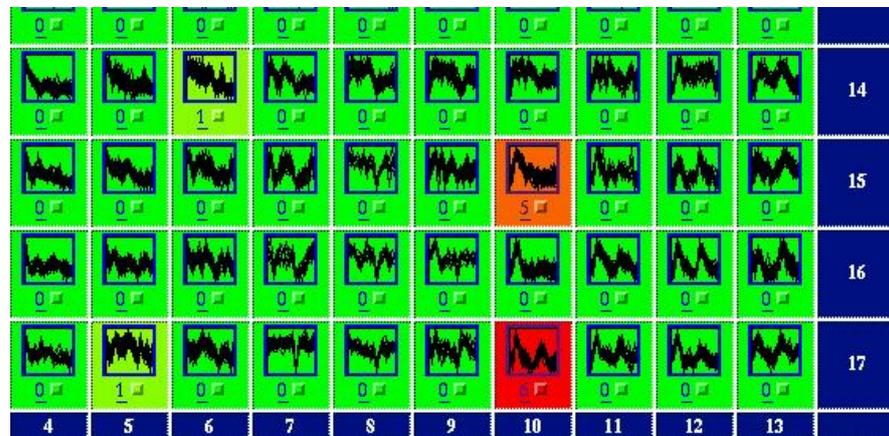


Figure 3.14: Pathway projection: The same projection as in Figure 3.13 is shown here in a 2D diagram. The third dimension has been replaced by a color coding: green encodes minimal values, red stands for maximal values. Intermediate values are shown in transitional colors from green via yellow to red. This representation allows to clearly identify and to select the clusters of interest. Here, the cluster that is shown to contain 6 genes of the projected pathway has been selected.

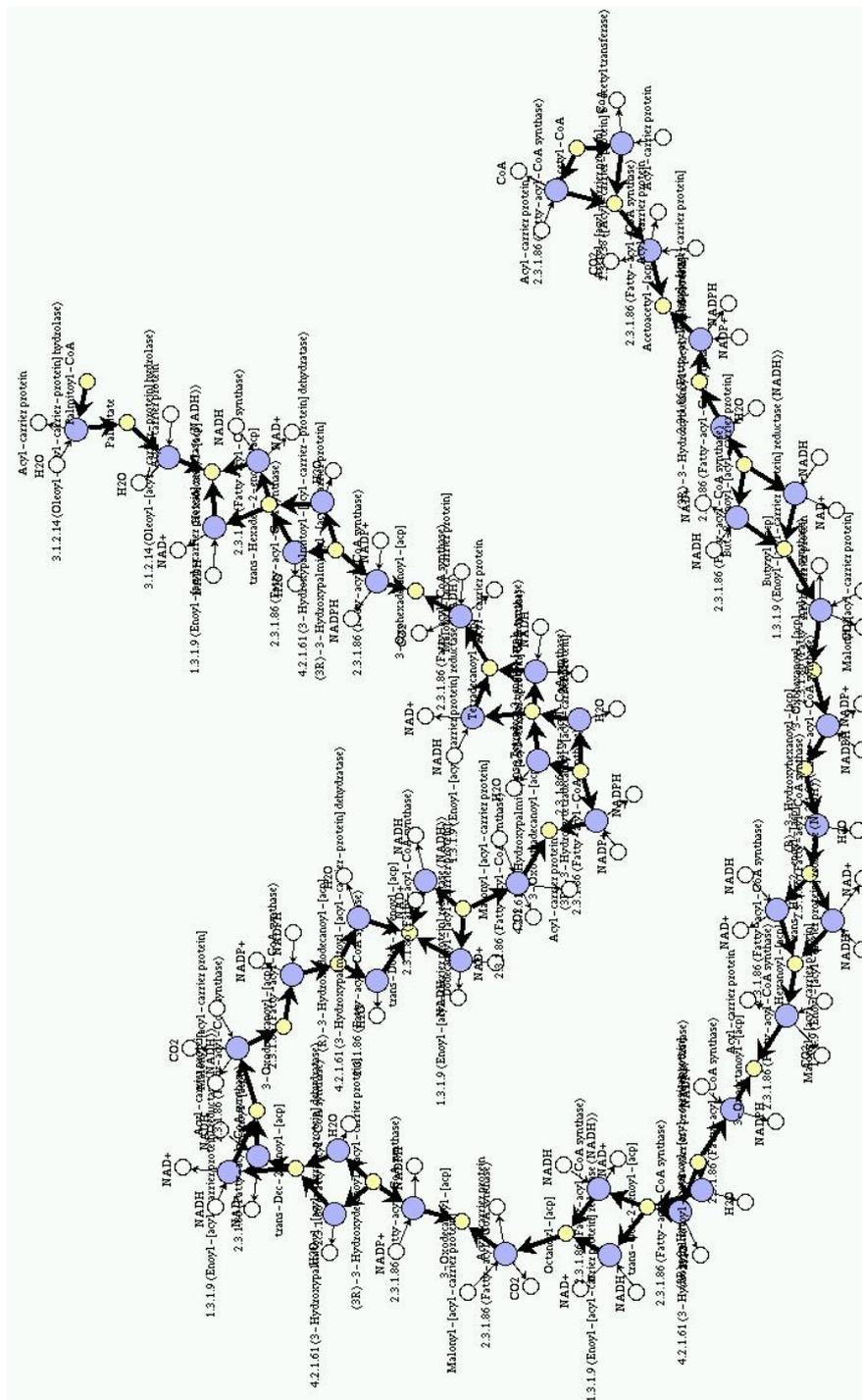


Figure 3.15: One of the pathways that emerge from the *metabolic mapping* of the previously selected cluster. It shows a large linear sequence of metabolic reactions. Two reactions, *fatty-acyl-CoA-synthase* (EC 2.3.1.86) and *enoyl-[acyl-carrier protein] reductase (NADH)* (EC 1.3.1.9), appear numerous times in this hypothetical pathway. It is up to the user to judge whether the proposed pathway is actually reasonable. Facts that are not taken into account may prove to contradict the proposition, e.g. different cellular localizations of the involved enzymes.

The *metabolic mapping* approach has also been applied in order to obtain an intuitive picture of the metabolic pathways that are co-ordinately regulated during the diauxic shift. Using the *functional projection* (Chapter 3.1), we found rather large hypothetical pathways that correspond to the respective textbook metabolic pathways for the induced groups of glycogen/trehalose metabolism, TCA cycle and glyoxylate metabolism as automatically identified. As expected, these pathways contained additional reactions that represent parts of connected pathways (Figures 3.16 and 3.17). The metabolic mapping for the group of glycogen and trehalose metabolism produces a pathway that besides the trehalose and glycogen network shows the connecting reactions to glycolysis/gluconeogenesis. This shows that a part of the glycolysis pathway is co-ordinately induced with the trehalose and glycogen metabolic pathway, though the larger part of glycolysis is repressed at the same time. Two results can directly be derived here: during the diauxic shift, the glycolysis pathway falls into two sections that are contrarily regulated. The one section is co-ordinately regulated with the trehalose and glycogen metabolic pathway.

3.2.3 Discussion

In this chapter, I present two approaches to a pathway analysis of gene expression data that can be used separately or in conjunction. Based on the expression data, metabolic networks are extracted that are co-ordinately regulated in a certain biological context. The metabolic analysis depends on the quality of the gene expression data as well as on the quality of the gene annotations, i.e. the correctness and the degree of completeness of the EC catalog. Furthermore, the quality of the reaction database that associates certain metabolic conversions with the EC numbers is crucial. Problems and inaccuracies caused by the reaction database are mainly due to wrong or missing assignments of EC numbers. As far as the gene expression data is concerned, the problems mentioned for gene expression data analysis in general apply (Chapter 2.1). An important weakness of any metabolic analysis of gene expression data is the implicit assumption that relative gene expression rates can be used to extrapolate the corresponding protein levels. It has been shown in yeast that mRNA levels may vary significantly for steady-state protein levels and vice versa [GRFA99]. Transcription is a reaction of the cell to some stimulus. The transcription of a gene is not necessarily correlated to the concentration of the respective proteins. The conversion steps from transcription to the end product, the protein, and the properties of the intermediates are not considered in this approach. Many regulative and regulated steps are involved in gene expression, the chain that leads from the read out of the DNA to the functional protein, e.g. mRNA processing

(splicing, editing), translation, and protein modification. Additionally, the half-life period of RNAs and proteins differs significantly [GRFA99].

This dynamic approach is designed to complement existing approaches that map genes onto diagrams of textbook pathways like it is realized within the KEGG system [K⁺01, GK00]. While these more static approaches provide the user with a familiar representation of the metabolic pathways and can support the search for single, unknown or erroneously annotated enzymes within the pathways, the dynamic approach presented here can lead to the development of new ideas and concepts of metabolic conversions, especially in organisms whose metabolism is not yet well analyzed. It supports the *metabolic reconstruction* process that often follows sequencing and annotation of whole genome sequences.

Gene expression data and metabolic pathways can be further integrated with information on regulatory units like promoters to support the analysis of gene expression data. First approaches in this direction have already been published [JCCS01, BLS01]. The authors are looking for common sequences in the upstream regions of genes that showed to be co-expressed in microarray experiments in order to identify potential new regulatory sequences. The use of a promoter database, e.g. TransFac [WCF⁺01], can further support such an analysis.

Zien et al. explore metabolic databases by calculating pathway scores based on an expression data set [ZKZL00]. They extract linear pathways between a source and a target metabolite from the reaction network that is inherent in the reaction database [KZL00]. The pathways are reduced to lists of EC numbers. Each of the EC numbers is mapped to the set of encoding genes. For each list of EC numbers, this mapping leads to a set of gene lists. The gene lists are scored, i.e. the correlation of the corresponding expression profiles is evaluated. A p-value that indicates the significance of a score is also proposed. The approach of Zien et al. is able to integrate metabolic pathways and gene expression data without the need for a clustering of genes according to their expression profiles. The starting point of the pathway scoring approach is a metabolic conversion. By some means, metabolic pathways have to be selected for scoring. This can be done by dynamic extraction of pathways from a metabolic network, e.g. by the methods described here or by the approach of [KZL00], but also textbook pathways are suited. The selected pathways that accomplish a specific conversion are evaluated by means of a gene expression data set. The pathway scoring complements the pathway projection described in this thesis. Both methods can well be combined. The pathway scoring approach would be used to select the most interesting, i.e. high scoring pathways from the given set of textbook pathways. These pathways can then be projected onto the clustering.

The pathway scoring can be seen as an inverse approach to the metabolic mapping. Both methods achieve a metabolic interpretation of a gene expression data set. While the pathway scoring starts from a metabolic conversion or a set of metabolic pathways, the metabolic mapping approach starts from the gene expression data, i.e. from the gene clusters that have been determined according to gene expression profiles. It allows to find metabolic networks of co-expressed genes without the need to predetermine a metabolic conversion. On the other hand, the pathway scoring approach leads to results for reasonably defined metabolic pathways while the metabolic networks that result from the metabolic mapping approach may or may not be functional in the sense of a metabolic pathway. They can be regarded as a hypothesis that has to be tested.

3.3 Discussion of Functional Projection, Pathway Projection and Metabolic Mapping

The described methods for a function-guided clustering and a metabolic analysis of gene expression data allow a fully automatic identification and elucidation of non-partitional groups of co-expressed genes that are annotated to have a role in the same biological context. Functional categories whose genes are tightly co-expressed are effectively identified by means of a numeric analysis of large-scale gene expression data sets and the successive automatic construction of corresponding groups of clusters. Additionally, overlapping groups found for different functional categories are visualized. This reveals cases where genes of related but distinct functional categories are co-expressed. Co-ordinately affected metabolic pathways are extracted from the network of intermediary metabolism and connections to co-expressed neighboring metabolic conversions are identified.

Such an integrative analysis is largely dependent on the correctness and completeness of the annotation of the genomes. Functional categories have to be consistently assigned to the genes and the enzymatic activities have to be mapped onto the genes, i.e. the EC numbers have to be assigned. Only with these prerequisites the integrative approach generates reliable and reasonable results. Since a significant number of organism-specific high quality databases have been established or are under development, the approach can be immediately applied to these organisms (FlyBase [Con99], MIPS [AGH⁺01], SubtiList [MGD95], EcoCyc [KRP⁺99]). For organisms for that only the genomic sequence and the genetic elements are available it is possible to achieve an annotation by transferring structured information from preferably manually annotated genes to previously uncharacterized genes on the basis of sequence homology (Genequiz [HLB⁺00], Pedant [FAH⁺01]) in a systematic and fully automatic way.

Functional properties and metabolic activities of genes with respect to specific environmental conditions can be derived. This is achieved by combining numerical analysis methods with additional analysis steps that take into account external information. The integration of the numerical data with systematic annotations allows to guide the clustering/grouping according to existing knowledge and significantly reduces the effort for an interactive mining of the clustering results. The identification of interesting groups of genes can now be automated. The result of the analysis is a list of co-expressed, i.e. potentially co-regulated, groups of genes that allow a direct biological interpretation of the underlying gene expression experiment.

Functionally uncharacterized genes are put into a detailed functional context. The additional integration with other methods capable of suggesting

functional roles for uncharacterized genes, e.g. based on protein-protein interactions (Chapter 3.4, [FAZ⁺00]) can promote the revelation of the functional roles of previously uncharacterized proteins by significantly reducing the search space and by providing further evidence.

The *pathway projection* has been established in analogy to the functional projection. Instead of the genes of a functional category, the genes of a textbook pathway are mapped to the gene clusters. Like the functional projection, also the pathway projection can be automated by a group finding algorithm. The method would identify groups of neighboring clusters that each contain genes of the same metabolic pathway. The analysis of overlapping groups would reveal pathways whose genes are co-expressed in the underlying biological experiment. Applying the group identification to the pathway projection, the focus of the analysis would specifically shift to the intermediary metabolism. With the new FunCat that comprises a finer classification on more hierarchy levels and that includes categories for metabolic pathways, the pathway projection can be seen as a specialized, refined version of the functional projection.

3.4 Integrating Protein-Protein Interactions with Functional Annotations

We have developed a method for the integrative analysis of protein interaction data [FAZ⁺00]. It comprises clustering, visualization and data integration components. The method is generally applicable for all sequenced and annotated organisms. In this section we describe in detail the combination of protein interaction data in the yeast *Saccharomyces cerevisiae* with the functional classification of all yeast proteins. We evaluate the utility of the method by comparison with experimental data and deduce hypotheses about the functional role of so far uncharacterized proteins. Further applications of the integrative analysis method are discussed. The method presented here is powerful and flexible. We show that it is capable of mining large-scale data sets.

3.4.1 Methods

The combination of different data sets leads to a more detailed level of information. Here we show how data of physical and genetic interactions can be combined with other types of biological data, especially annotations according to systematic functional categories (MIPS functional catalog). The combination of different data sets leads to a more detailed level of information.

The MIPS functional catalog (Chapter 2.3) is based on a common-sense terminology of biological function and is therefore well suited for selecting subsets of genes with related functions. We restrict the gene set to those genes sharing functional categories. The result is an interaction graph that consists of the genes belonging to the selected functional categories (nodes drawn with thick borders) and those directly interacting with them (nodes drawn with thin borders). Most of the categories of the MIPS functional catalog contain a large number of genes due to the fact that biological processes in general involve a substantial number of proteins. The catalog does not describe the exact role of single genes in the cell. The combination of diverse data sets generates the information necessary for an exact assignment of the genes to cellular processes.

Protein-protein interaction data in turn is not sufficient for a comprehensive description of the functional context. If no additional information, e.g. about the time, localization or function of the individual interactions is taken into account, i.e. using the pure clustering method described above, the proteome of an organism tends to be in a single cluster.

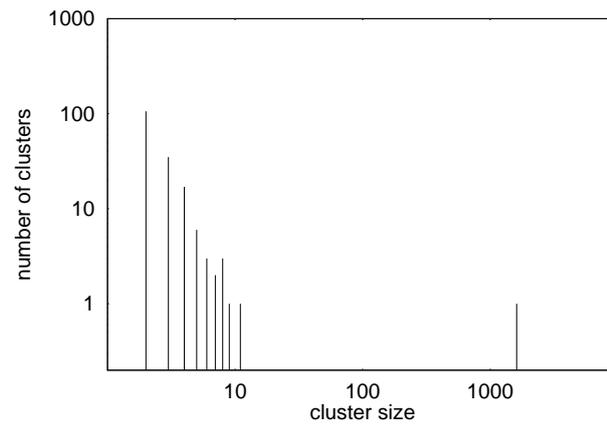


Figure 3.18: Histogram of cluster sizes. Cluster sizes are shown on the abscissa, the number of clusters of the respective cluster sizes are shown on the ordinate. Both scales are logarithmic. For the clustering, the ORFs of all functional categories have been considered. Though there are 174 small clusters with less than 12 ORFs, these contain a total of only 562 ORFs, while one cluster comprises 1721 ORFs.

functional category	genes in this category	genes with interaction	genes in biggest cluster	interaction in all clusters
all	6359	2283	1721	6158
04	749	460	732	2416
04.05	542	336	558	1862
04.05.05	40	32	62	259

Table 3.3: Number of genes found in different categories and properties of the resulting clusters. Categories: *Transcription* (04), *mRNA processing* (04.05) and *5'-, 3'-end processing, mRNA degradation* (04.05.05).

3.4.2 Verification

The MIPS yeast interaction tables contain 6158 individual interactions annotated for 2283 ORFs as of February 2000. Clustering these genes as described above leads to 175 clusters. While 106 clusters consist of just two genes and 35 clusters comprise three genes, there is one cluster containing 1721 genes (main cluster). The remaining 33 clusters consist of 4 to 11 genes (Figure 3.18). A reasonable visualization and interpretation can be computed for the small clusters while the interaction graph of the large one is too complex to produce a clear, easy to grasp representation. Thus the interaction graph of the vast majority of the genes cannot be shown. One possibility to reduce the complexity of the data is to combine the set of genes to be analyzed to functionally categorized genes.

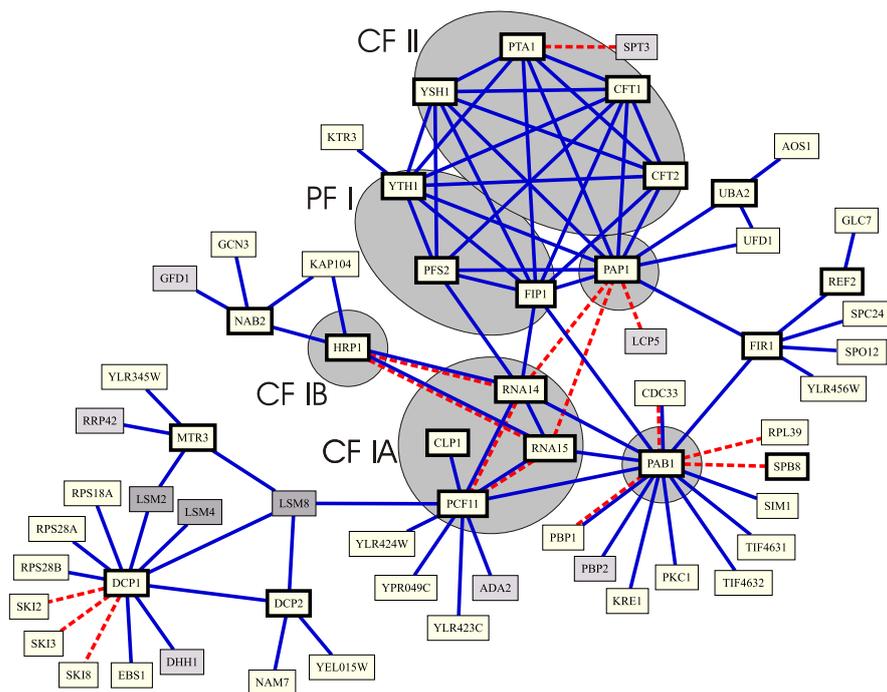


Figure 3.19: Display of the largest interaction cluster based on the functional category *mRNA processing (5', 3'-end processing, mRNA degradation)*. Proteins of the clustered category appear with thick borders, those of other functional categories have thin borders. Solid lines indicate physical interactions, dashed lines indicate genetic interactions. The genes with dark grey boxes belong to functional category *mRNA processing (splicing)*. Those with light grey boxes are genes of category *transcription*, the super-category of mRNA processing. Factors of the processing complex as described in [ZHM99] are enclosed within ellipses (cf. Table 3.4).

For the functional category *transcription* and two of its subcategories, we show the properties of the gene set restricted to the respective category and the resulting cluster sizes (Table 3.3).

Restricting the interaction to the lowest hierarchical level, i.e. category *mRNA processing (5', 3'-end processing, mRNA degradation)* leads to a gene set that consists of 40 genes, 32 of these with annotated interactions. We use this well described category to verify the utility of the method. For the clustering all interactions annotated for the genes of this category are used, including those interactions with a gene not belonging to the category. 8 clusters result from this process, the largest cluster containing 62 proteins, 23 from the specified functional category, and 99 interactions between the proteins.

As expected, the interaction graph of this cluster (Figure 3.19) corre-

ORF	Gene	Description	Functional categories	Factor
YDR448w	ADA2	general transcriptional adaptor	04.05.01.04	
YPR180w	AOS1	Smt3p activating enzyme subunit	06.07;06.13.01	
YOL139c	CDC33	translation initiation factor eIF4E	03.10;05.04	
YDR301w	CFT1	pre-mRNA 3'-end processing factor CF II subunit	04.05.05	CF II
YLR115w	CFT2	pre-mRNA 3'-end processing factor CF II subunit	04.05.05	CF II
YOR250c	CLP1	cleavage/polyadenylation factor IA subunit	04.05.05	CF IA
YOL149w	DCP1	mRNA decapping enzyme	04.05.05	
YNL118c	DCP2	suppressor protein of a yeast pet mutant	02.13; 04.05.05	
YDL160c	DHH1	strong similarity to RNA helicases of the DEAD box family	04.99	
YDR206w	EBS1	similarity to EST1 protein	03.16	
YJR093c	FIP1	component of pre-mRNA polyadenylation factor PF I	04.05.05	PF I
YER032w	FIR1	interacts with the poly(A) polymerase	04.05.05	
YKR026c	GCN3	translation initiation factor eIF2B, 34 KD, alpha subunit	05.04	
YMR255w	GFD1	nuclear pore complex protein	04.07;08.01	
YER133w	GLC7	ser/thr phosphoprotein phosphatase 1, catalytic chain	01.05.04;02.19;03.04; 03.13;03.22;05.07	
YOL123w	HRP1	cleavage/polyadenylation factor IB	04.05.05 ;04.05.99;04.07	CF IB
YBR017c	KAP104	beta-karyopherin	08.01	
YNL322c	KRE1	cell wall protein	01.05.01	
YBR205w	KTR3	alpha-1,2-mannosyltransferase	01.05.01;06.07	
YER127w	LCP5	Ngg1p interacting protein	04.01.04	
YBL026w	LSM2	snRNP-related protein	04.05.99	
YER112w	LSM4	U6 snRNA associated protein	04.05.03	
YJR022w	LSM8	splicing factor	04.05.03	
YGR158c	MTR3	involved in mRNA transport	04.01.04; 04.05.05 ;04.07	
YGL122c	NAB2	nuclear poly(A)-binding protein	04.05.05 ;04.07	
YMR080c	NAM7	nonsense-mediated mRNA decay protein	01.03.16;05.07	
YER165w	PAB1	mRNA polyadenylate-binding protein	04.05.05 ;05.04	Pab 1
YKR002w	PAP1	poly(A) polymerase	04.05.05	Pap 1
YGR178c	PBP1	Pab1p interacting protein	04.05.05	
YBR233w	PBP2	Pab1p interacting protein	04.99	
YDR228c	PCF11	component of factor CF I	04.05.05	CF IA
YNL317w	PFS2	component of pre-mRNA polyadenylation factor PF I	04.05.05	PF I
YBL105c	PKC1	ser/thr protein kinase	03.04;03.22; 10.01.05.11;11.01	
YAL043c	PTA1	pre-mRNA 3'-end processing factor CF II subunit	04.03.03; 04.05.05	CF II
YDR195w	REF2	RNA 3'-end formation protein	04.05.05	
YMR061w	RNA14	component of factor CF I	04.05.01.04; 04.05.05	CF IA
YGL044c	RNA15	component of factor CF I	04.05.05	CF IA
YJL189w	RPL39	ribosomal protein L39.e	05.01	
YDR450w	RPS18A	ribosomal protein S18.e.c4	05.01	
YOR167c	RPS28A	ribosomal protein S28.e.c15	05.01	
YLR264w	RPS28B	ribosomal protein S28.e.c12	05.01	
YDL111c	RRP42	rRNA processing protein	04.01.04	
YIL123w	SIM1	involved in cell cycle regulation and aging	03.22;11.11	
YLR398c	SKI2	antiviral protein and putative helicase	11.07	
YPR189w	SKI3	antiviral protein	11.99	
YGL213c	SKI8	antiviral protein	03.13;03.19;11.13	
YJL124c	SPB8	suppressor of PAB1	04.05.05	
YMR117c	SPC24	spindle pole body protein	03.22	
YHR152w	SPO12	sporulation protein	03.10;03.13	
YDR392w	SPT3	general transcriptional adaptor	03.07;04.05.01.04	
YGR162w	TIF4631	mRNA cap-binding protein (eIF4F), 150K subunit	05.04	
YGL049c	TIF4632	mRNA cap-binding protein (eIF4F), 130K subunit	05.04	
YDR390c	UBA2	E1-like (ubiquitin-activating) enzyme	04.05.05 ;06.07;06.13.01	
YGR048w	UFD1	ubiquitin fusion degradation protein	06.13.01	
YLR345w		similarity to 6-phosphofructo-2-kinases	01.05.04;02.01	
YLR277c	YSH1	pre-mRNA 3'-end processing factor CF II subunit	04.05.05	CF II
YPR107c	YTH1	component of pre-mRNA polyadenylation factor PF I	04.05.05	PF I

Table 3.4: The functionally classified proteins of the main cluster of category *mRNA processing (5'-, 3'-end processing, mRNA degradation)* (04.05.05), displayed in Figure 3.19. The description of the factors can be found in [ZHM99]. Functional categories are explained in Table 2.2.

sponds to the description of the *pre-mRNA 3'-end processing* in the literature [ZHM99]. According to this review, the *yeast cleavage/polyadenylation complex* consists of 15 identified proteins. For 14 of them the corresponding genes are known. The complex is subdivided into five functionally distinct activities. CF IA, IB and II are described as sufficient for the cleavage reaction, and specific poly(A)-addition are described as to require CF IA and IB, Pap1, Pab1, and PF I. These functional factors can be identified in the interaction graph. The graph contains additional interactions to proteins of other functional categories, linking the cleavage/polyadenylation complex with neighboring cellular processes like splicing and transcription (Table 3.4).

The fifteenth, missing protein has been identified by purification of the PF I complex. It is a protein of 58kd called Pfs1p, which has not been published yet. PFS1 is an essential gene containing a zinc knuckle [ZHM99]. We did not find interaction data pointing to an ORF fulfilling these requirements. Thus it is not possible to speculate about this protein on the basis of the protein interaction data available.

3.4.3 Revealing the Functional Context of Uncharacterized Proteins

So far, roughly a third of the yeast ORFs are not functionally described [MAB⁺97]. Several systematic approaches have been developed in order to functionally classify these ORFs [OWKB98]. Since 1997 about 7% of the ORFs then called *proteins of unknown function* have been functionally described (cf. MYGD [AGH⁺01]).

We show here how the usage of protein interaction data can provide substantial clues as to the biological context of unknown proteins.

Restricting the gene set to the 2563 non-characterized proteins gives rise to 177 clusters. 39 of these clusters contain more than two unclassified genes. The biggest cluster, comprising 93 non-characterized and 66 characterized proteins, contains the ORFs YNR053c and YGL161c. We show the capabilities of our integrative method for the prediction of functions for so far non-characterized proteins by means of these ORFs. Evaluating a protein interaction graph, functional relationships between proteins can be deduced depending on the distance of the respective proteins. Two proteins that directly interact are most likely to be involved in the same biological process or pathway [WSL⁺00]. We thus concentrate on the direct surrounding of the considered ORFs. Therefore the interaction graphs of YNR053c (Figure 3.20) and YGL161c (Figure 3.21) have been cut at the second and third interaction level, respectively.

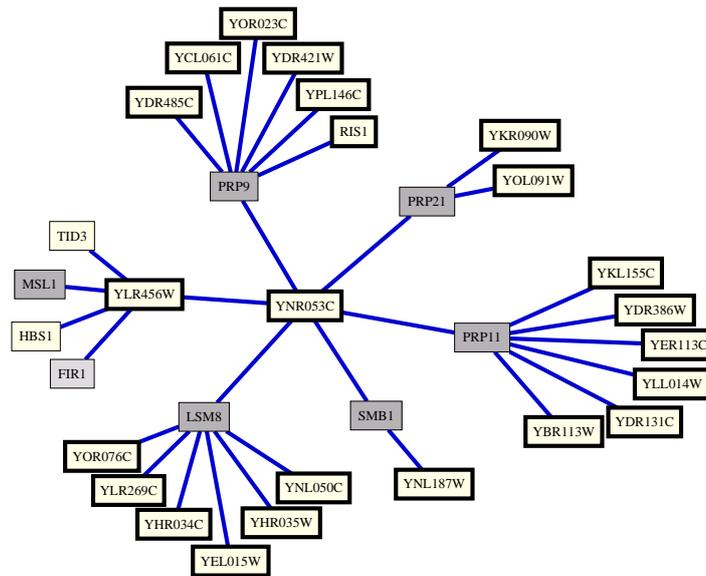


Figure 3.20: Part of an interaction cluster resulting from the gene set restricted to the non-characterized proteins. YNR053c directly interacts with five genes belonging to the functional category *mRNA processing (splicing)*, shown in dark grey. Another gene of this category is found in the second interaction level: *MSL1*. *FIR1* belongs to the same biological context, *mRNA processing (5'-, 3'-end processing, mRNA degradation)*. It is indicated in light grey. The graph is cut at the second interaction level with respect to YNR053c. Non-characterized proteins appear with thick borders, those with annotated functional categories have thin borders. The functional categories of the individual proteins are listed in Table 3.5. The solid lines indicate physical interactions.

The functional context of YNR053c

YNR053c is known to interact directly with six other proteins [FRRL97], five of which have a known functional classification (Figure 3.20). These five all belong to the category *mRNA processing (splicing)* (cf. Table 3.5). All but four of the indirectly connected proteins of the second level are uncharacterized. The four classified proteins comprise Msl1p that is also involved in the mRNA splicing and Fir1p that is involved in 3'-end mRNA processing. The remaining proteins, Tid3p and Hbs1p belong to other functional categories. The central position of YNR053c in the described protein interaction network is a strong clue as to its functional role in mRNA splicing. For YLR456w, a non-characterized protein that directly interacts with YNR053c, a functional prediction is more difficult to make. The five interactors of this ORF belong to more diverse categories. Two of the interacting proteins are known to be involved in mRNA transcription (Msl1p, Fir1p).

ORF	Gene	Description	Functional categories	Interaction level
YDL030w	PRP9	pre-mRNA splicing factor	04.05.03	1
YDL043c	PRP11	pre-mRNA splicing factor	04.05.03	1
YER029c	SMB1	associated with U1 snRNP	04.05.03	1
YER032w	FIR1	interacts with the poly(A) polymerase	04.05.05	2
YIL144w	TID3	Dmc1p interacting protein	03.22	2
YIR009w	MSL1	strong similarity to snRNPs	04.05.03 ;06.10	2
YJL203w	PRP21	pre-mRNA splicing factor	04.05.03 ;06.10	1
YJR022w	LSM8	splicing factor	04.05.03	1
YKR084c	HBS1	eEF-1 alpha chain homologue	05.04	2
YLR456w		strong similarity to YPR172w	99	1

Table 3.5: The interacting partners of the non-characterized protein YNR053c. All interacting proteins of level 1 and those of level 2 with a described function are listed. The functional categories are described in Table 2.2.

For the third interactor, YNR053c, there are strong indications for the participation in mRNA splicing as described above. The two remaining ORFs are involved in cell cycle control and translation, respectively. Considering the whole interaction context of YLR456w, allows to hypothesize on a potential cellular function in mRNA splicing.

The functional context of YGL161c

For YGL161c seven direct interactions with other proteins are known [ITM⁺00, UGC⁺00], three of them have a known function (Figure 3.21). These three proteins, Vam7p, Yip1p, and Pep12p, are involved in intracellular transport. Additionally, two of three functionally classified proteins of the second interaction level, Yip3p and Akr2p, are also intracellular transport proteins. The third, Ktr3p, is an alpha-1,2-mannosyltransferase involved in glycosylation of proteins to be secreted. Ktr3p is likely to be localized in the golgi apparatus [SPC95]. Thus the cellular role of Ktr3p is closely linked to intracellular transport processes. Even on the third interaction level of YGL161c one more protein, Ypt1p, is described to be involved in intracellular transport processes. Considering all these data one can assume that YGL161c is involved in intracellular transport. In addition, there is some evidence for the non-characterized proteins YIF1, YHR105w, and YPL246c to be involved in the context of intracellular transport. All three have a central position in the described interaction network of intracellular transport processes.

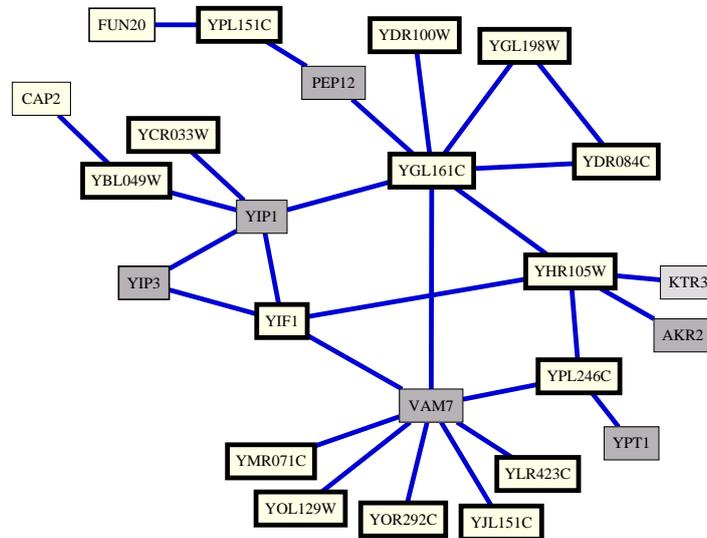


Figure 3.21: Part of an interaction cluster resulting from the gene set restricted to the non-characterized proteins. The graph is cut at the third interaction level with respect to YGL161c. Non-characterized proteins appear with thick borders, those with annotated functional categories have thin borders. The functional categories of the individual proteins are listed in Table 3.6. The solid lines indicate physical interactions.

ORF	Gene	Description	Functional categories	Interaction level
YAL032c	FUN20	required for RNA splicing	04.05.03	3
YBR205w	KTR3	alpha-1,2-mannosyltransferase	01.05.01;06.07	2
YFL038c	YPT1	GTP-binding protein of the rab family	08.07	3
YGL212w	VAM7	vacuolar morphogenesis protein	08.13 ;09.25	1
YGR172c	YIP1	golgi membrane protein	08.07	1
YIL034c	CAP2	F-actin capping protein, beta subunit	03.04;03.07	3
YNL044w	YIP3	involved in ER to golgi transport	06.04; 08.07	2
YOR034c	AKR2	involved in constitutive endocytosis of Ste3p	08.19	2
YOR036w	PEP12	syntaxin (T-SNARE), vacuolar	06.04; 08.13	1

Table 3.6: The interacting partners of the non-characterized protein YGL161c. All interacting proteins of level 1 and those of levels 2 and 3 with a described function are listed. The functional categories are described in Table 2.2.

3.4.4 Discussion

We have developed a method for the integrative analysis of protein interactions [FAZ⁺00] that combines protein interaction data on *S. cerevisiae* proteins, systematically collected from the literature, with the functional classification data of all yeast proteins. A clustering is performed to find maximal groups of proteins that are directly or indirectly connected to interaction networks. These networks are graphically represented using a graph editor toolkit. The visualization of protein interaction networks supports the comprehensive analysis of these large data sets.

The characteristic of the resulting cluster sizes (Figure 3.18) suggests that the more interactions are known for an organism the larger the clusters of genes become. We suspect that if all true interactions would be known, virtually every protein of that organism would be in a single cluster of proteins that results from the described clustering method. It is thus necessary to focus on a subset of the whole set of genes. A more efficient and exploratory interpretation of the protein interaction data is enabled by the concentration on a certain biological context that is achieved by the use of the systematic functional categorisation of all yeast ORFs.

This combination of different data sets results in the necessary reduction of complexity. Our results show that the integrative analysis with a hierarchically organized data set allows to scale the complexity of the interaction graphs (Table 3.3).

The utility of the presented method has been shown by applying it to the well described functional context of mRNA processing. The analysis of the biological context of the uncharacterized proteins YNR053c and YGL161c shows the relevance of the method for mining the protein interaction data and formulating hypothesis about a functional classification of so far uncharacterized proteins.

It has been shown that interactions among proteins are conserved between the homologous proteins of various organisms [UGC⁺00]. In cross-genome analysis the presented method can be used for the prediction of protein interactions in other species. This is of particular interest with respect to sequences resulting from whole genome sequencing projects, e.g. the human genome project. Recently an algorithm for an interaction-based protein interaction map prediction has been published [WS01].

The integrative analysis method presented here is easy to use and allows a comprehensive overview of the protein interaction data. It is very flexible, i.e. it can easily be applied to other types of annotation data sets to be combined. It is also possible and very promising to combine the protein

interaction data with other data produced by high-throughput methods, the most prominent being expression data produced by DNA microarray technology [DIB97] and cellular localization via GFP tagging [BLR⁺00].

Besides the careful assignment of functional categories to the genes two key features make the functional catalog so well suited for an efficient integration with high-throughput data like gene expression data and protein-protein interaction data. These are the hierarchical structure and a clear, systematic, and consistent definition of distinct categories. Similar catalogs can be set up for other annotational aspects of genes, proteins, and metabolites. The MIPS yeast project group has additionally defined catalogs for subcellular localization, protein complexes, mutant phenotypes, protein classes, EC numbers and PROSITE motifs. These are likewise suited for an integrative analysis. The only restrictions are imposed by semantical dependencies, e.g. it would probably make no sense to integrate the protein complexes catalog with protein-protein interaction data.

Chapter 4

The Software System Architecture

The tools that are described in the previous chapters have been developed and implemented in parallel on a Compaq Tru64 Unix system and on Intel PCs running Linux. A layered system architecture has been chosen. This modular architecture ensures that the structure of the code is relatively easy to understand and that maintenance is reasonably simple. Additionally, the layered, modular architecture allows to implement the individual layers successively. It provides the flexibility that is highly demanded by a research project where a complete specification of the software cannot be achieved in the beginning since the direction of development depends on intermediate results and is adapted accordingly throughout the time of the project. During development, the user interface of the tools has been a command line interface. At a point where the tools stabilized and could finally be integrated, a consistent user interface was set up using a web-based client-server architecture. The client side GUIs have been designed with Hyper Text Markup Language (HTML) and some Java for the highly interactive parts. Figure 4.1 shows a diagram of the layered system architecture.

The client-server architecture was chosen to achieve maximal flexibility and independence of the individual modules (modular programming). Using web browsers as clients enables the user of the system to freely choose the operating system of the client machine. Exporting the services over the internet makes them widely accessible. The involved standard protocol, Hyper Text Transfer Protocol (HTTP), is easy to use. Since character streams are only sent over the net using the standard port 80, HTTP transfer is not influenced by fire walls in contrast to the more elaborate Remote Method Invocation (RMI) protocols like Common Object Request Broker Architecture (CORBA). These protocols depend on other ports that are typically

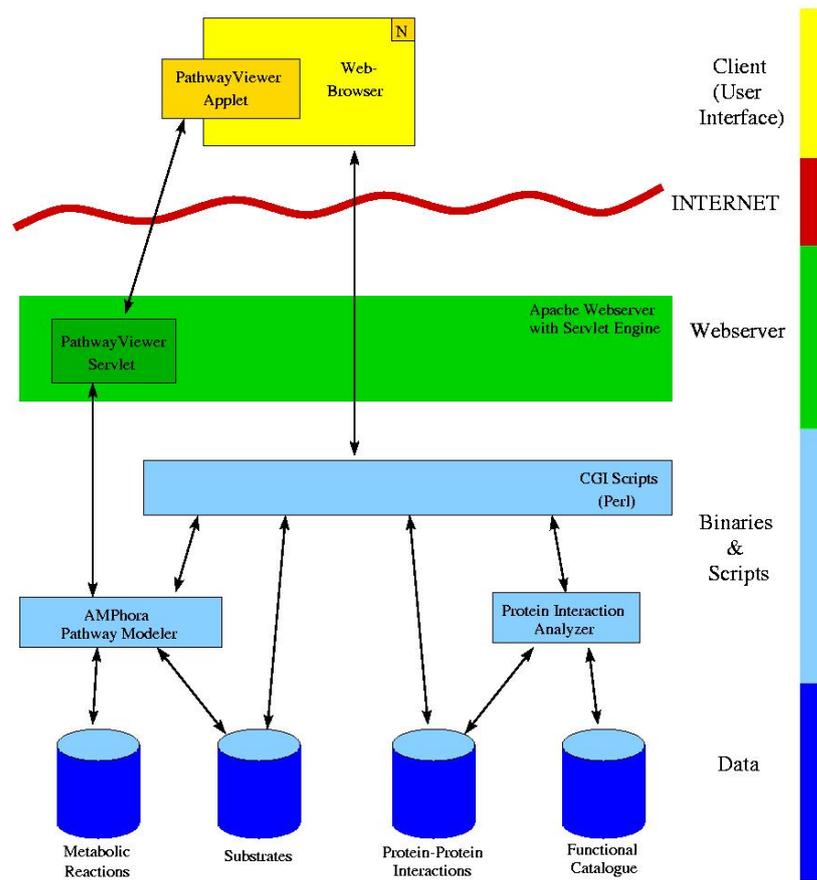


Figure 4.1: The layered architecture of the client-server system for a metabolic and functional analysis of biological high-throughput data.

blocked by a firewall. The Common Gateway Interface (CGI) technology allows to create the HTML documents dynamically. The possibilities and modes of user interaction provided by HTML are sufficient for most applications. Wherever a particular high degree of user interaction is needed, Java applet technology is at hand. In the described system the pathway viewer was implemented as an applet. Java servlets have been used as a server side counterpart to allow efficient communication with the pathway viewer applet. On the script/binary layer, Perl and C++ have been applied.

The server part of the system is not realized as a monolithic piece of software. Instead, it consists of a heterogeneous mixture of binary applications and scripts. This allows to extend the functionality of the system by adding new server side binaries and scripts and to extend the functionality of existing modules by replacing a binary or script by a new version. Such flexibility proved especially useful during the development phase. Scripts are implemented using the Perl programming language. Perl is excellent for rapid prototyping and for all tasks that involve a lot of input and output and that have a rather low computational complexity. This is the case with virtually every CGI script. The more complex tasks have been implemented in C++ in order to generate a fast running native binary application. In particular, the *AMPPhora* application for pathway modeling and the *MapClusterer* application for gene clustering (Chapter 2.1.8) are implemented in C++.

4.1 The AMPhora Pathway Modeling Application

The *AMPPhora* application realizes the computation of metabolic networks, the graph-based search for linear pathways and the constraint-based construction of metabolic pathways. It has a built-in pathway viewer that requires a connection to an X graphics server for command line use and can export pathways in *BioTalk* for propagation over the internet and representation in the applet pathway viewer. Biotalk is a language that is defined in the eXtensible Markup Language (XML). Figure 4.2 shows the information flow between the *AMPPhora* modules. The Library of Efficient Data Structures and Algorithms (LEDA) is employed for the pathway modeling application. The LEDA data structures used include strings, dictionaries, linear lists and labeled graphs. Especially the graph data structure and the algorithms working on graphs eased the development of *AMPPhora*.

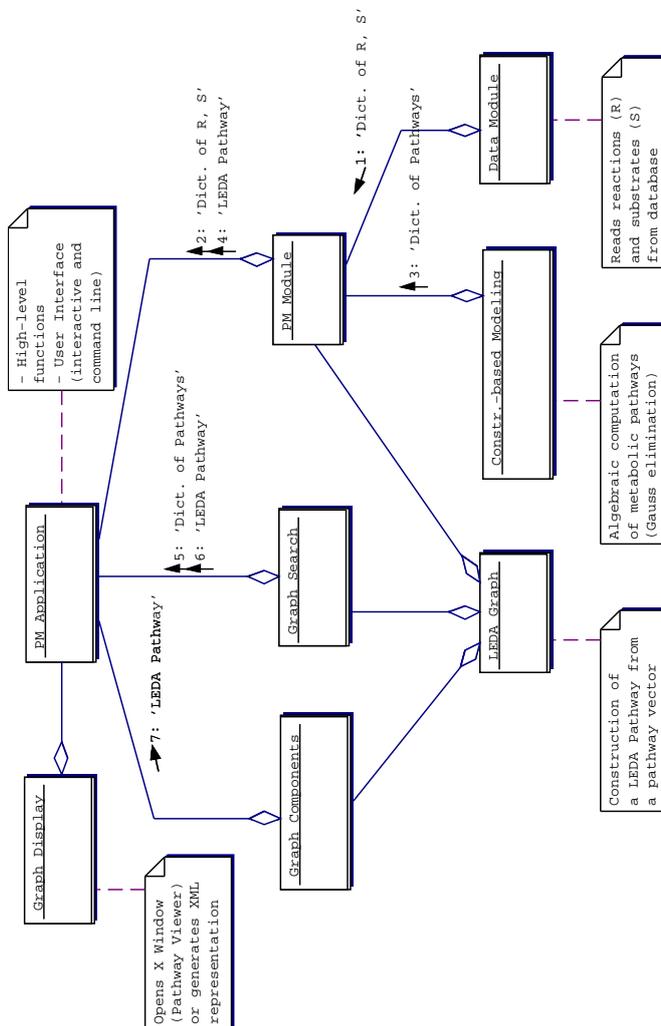


Figure 4.2: Modules and information flow in the *AMPhora* pathway modeling (PM) application. The main module (*PM_Application*) stages a command-line user interface and provides high-level functions for the three modes of pathway modeling. Pathways are represented in the modules and exchanged between modules as either pathway vectors or LEDA pathways (metabolic graphs). Pathway vectors can be compiled in dictionaries. The graph display module can generate XML output from the input pathway or open a pathway viewer window (uses X protocol, not available when used in client-server mode).

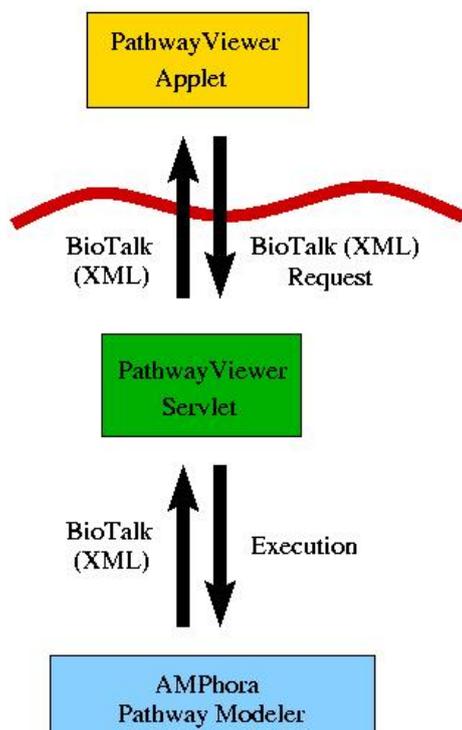


Figure 4.3: Communication between the pathway viewer Java applet and the *AMPhora* servlet is accomplished via the XML-defined language *BioTalk*. The servlet executes the *AMPhora* binary application that directly produces *BioTalk* output. The *BioTalk* output is passed on by the servlet to the applet, where an instance of the respective graph class is initialized from it.

4.2 *BioTalk*: Representing Pathways in XML

The transfer of metabolic graphs between server and client as well as the transfer of complex *AMPhora* requests are accomplished via XML. We employ the *BioTalk* language developed by Dieter Voges at *Biomax Informatics AG*. *BioTalk* is defined in XML. Instances of Java classes that are derived from the base class *BTOBJECT* can serialize themselves, i.e. they can produce a *BioTalk* output describing the values of their attributes. The other way round, a new instance of the same class can be initialized from the respective *BioTalk* code.

HTTP communication is generally organized in such a way that a client initiates a request. The request is processed and answered by the addressed server and the answer is sent back to the client. No mechanism is provided for a server to send data without answering a request. Therefore the pathway viewer has to request a pathway graph in *BioTalk* representation from the *AMPhora* servlet. The pathway viewer uses a request class that can export itself into *BioTalk* representation. The pathway viewer client instantiates the request class and sets the attributes according to the request. The *BioTalk* output of this request object is transferred to the servlet, where a

new instance of the request class is initialized from the *BioTalk* file. The *AMPhora* application is called by the servlet with the corresponding options set. It carries out the requested pathway modeling steps and finally produces a *BioTalk* document that matches the Java graph class of the viewer. The servlet passes the *BioTalk* output through to the client. The pathway viewer client initializes a new instance of a graph class from the *BioTalk* stream and the corresponding graph is displayed in the canvas of the viewer.

4.3 The *MapClusterer* application

The second binary application is the *MapClusterer* application that implements a SOM neural network and is used for clustering genes according to gene expression profiles. Input to the clusterer is a flat, tab-delimited textfile that contains one line for each gene. The first column of each line gives the gene name (or an internal identifier), the other columns contain the expression values for the respective gene. The SOM parameters are read from parameter files. These files provide reasonable parameter settings for different sizes of the neural grid, i.e. the number of clusters. A command line interface allows the calling user or script to select the data set to be clustered, the number of clusters to be computed, the normalization to be applied, and the distance measure to be used. The name of the resulting clustering has to be specified. The clusterer produces an output that complies to the format read by the CGI scripts of the graphical user interface.

4.4 The Graphical User Interface

The Graphical User Interface (GUI) of the system except the *AMPhora* pathway viewer is realized in HTML. The HTML code is generated dynamically via Perl CGI scripts. The GUI is organized in two major parts. The first part is for gene expression analysis, including the integrative analysis of gene expression data with functional projection, pathway projection, dynamic pathway construction (metabolic mapping) and construction of protein-protein interaction networks. The second part is for the *AMPhora* pathway modeling.

4.4.1 GUI for gene expression analysis

Figure 4.4 shows a site map of the GUI for gene expression analysis. There are two entry points to the system: the *List of Clustered Experiments* page

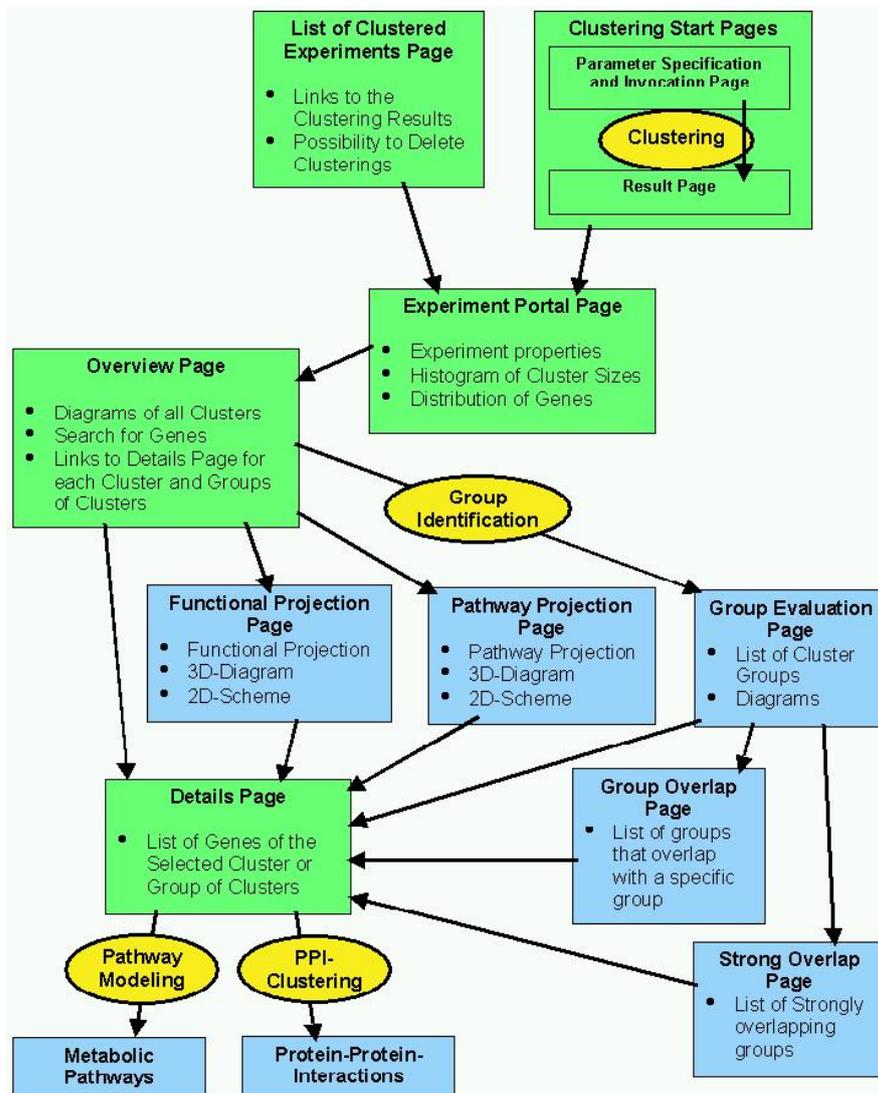


Figure 4.4: Site map of the GUI for gene expression analysis.

and the *Clustering Start Pages*. The *Clustering Start Pages* allow to invoke the clustering of a gene expression data set. The user has to select a gene expression dataset from a list of previously imported data sets and can specify the parameters for the clustering. The number of clusters to be computed and the distance measure have to be selected. The user has to decide whether or not to apply the standardization to mean zero and unit variance and whether or not to use logarithm transformed expression values. He invokes the clustering by submitting the settings. The clustering is performed on the server machine. On completion the result page is presented. A 3-dimensional projection of the number of genes per cluster roughly indicates the success of the clustering. A link to the detailed description of the clustering is provided.

The second entry point, the *List of Clustered Experiments* page, allows to access previously calculated clusterings. The available clusterings are shown, each with a link to the detailed results that are accessible from the *Experiment Portal* page. The experiment portal shows the parameter setting used for the clustering, the projection of gene numbers that also appeared on the *Result* page and a histogram of the cluster sizes. From here, the *Overview* page can be reached as well as the pages for the integrative analyses. The *Overview* page shows the regular grid of clusters with a 2-dimensional diagram for each cluster (Figure 2.8). For each cluster and any group of clusters, the *Details* page can be accessed, showing a larger diagram of the cluster or a schematic overview of the group of clusters (Figure 3.9) and a list of the genes that are in the cluster or in the group of clusters.

The *Details* page represents the hub of the GUI. From here, the dynamic construction of metabolic pathways and of protein-protein interaction networks are invoked. The *Details* page is reached from the *Overview* page as well as from the functional projection and the pathway projection. When reaching it from the functional projection, the *Details* page also shows a list of functional categories that are present in the selected cluster(s) (Figure 3.10).

The *Functional Projection* page, the *Group Evaluation* page and the *Pathway Projection* page are similarly constructed: they allow to select a functional category, a group of clusters that has been identified according to a certain functional category or a standard pathway. A 3-dimensional projection plot shows the distribution of the selected category or pathway (Figure 3.8). A 2-dimensional representation of that projection, where the third dimension is replaced by colors, allows to select clusters or groups of clusters of interest and to further analyze them on the *Details* page.

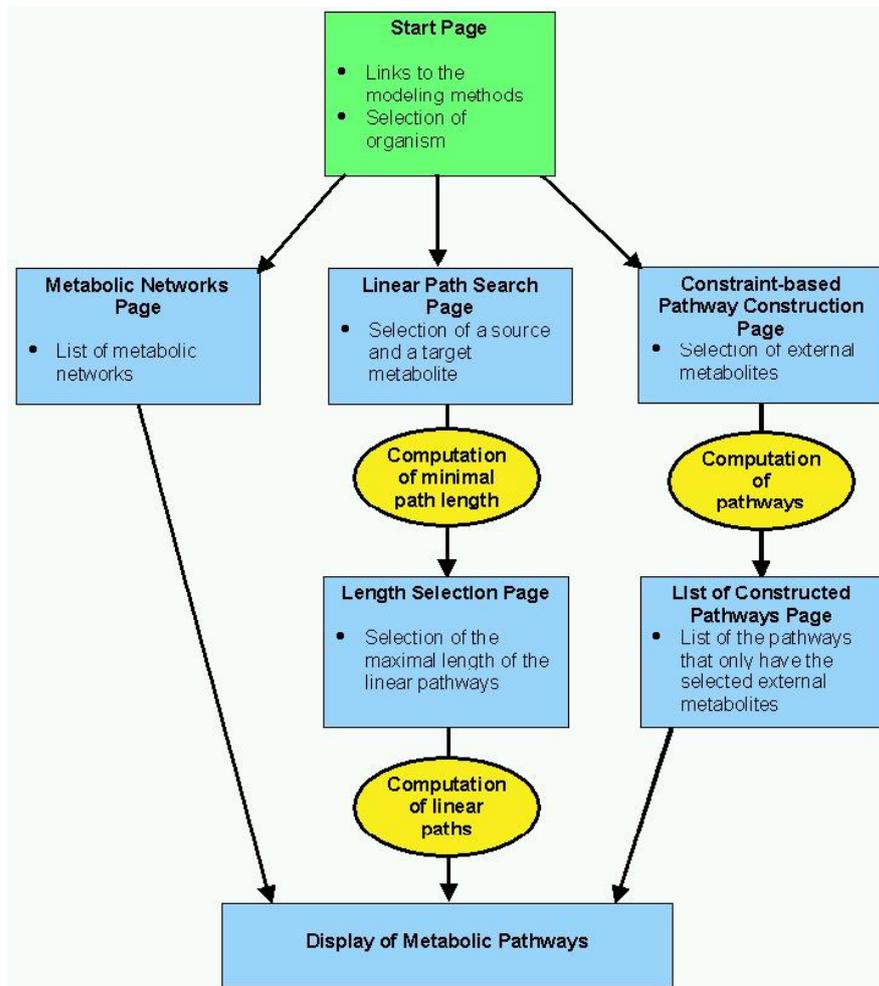


Figure 4.5: Site map of the GUI for the *AMPhora* pathway modeling tool.

4.4.2 GUI for *AMPhora* pathway modeling

The site map of the GUI of the *AMPhora* pathway modeling is shown in Figure 4.5. On entering via the start page, the user can select the organism for that the metabolic pathways shall be modeled. This selection restricts the set of EC numbers to those that are annotated to appear in the selected organism. The pathway modeling consists of three parts. There is a link on the start page for each of them that leads to the respective GUI section.

When selecting the *Computation of Metabolic Networks*, no parameters have to be set. The user directly gets to a page that lists the metabolic networks and that provides a link to the graphical representation for each network.

For the *Linear Path Search* the user has to select a source and a target metabolite. A search option allows to restrict the long list of metabolites to a subset matching the specified regular expression. On submission of the metabolite selection, feasibility and minimal length of the specified metabolic conversion are determined. The user has to specify the maximal length of the pathways to be shown. The metabolic pathways of at most the specified maximal length between the source and target metabolites are shown.

The *Constraint-based Pathway Construction* requires the user to specify a number of external metabolites. This is realized as a two step process in order to make the search for metabolites via regular expressions applicable here, too. The user selects a number of metabolites in the first list that he may restrict by specifying a regular expression. The selected metabolites are transferred into the second selection field. This can be repeated. From the second list, the user selects the external metabolites and invokes the pathway computation by submitting the selection. A list of the pathways that fulfil the metabolite constraints is presented, providing a link to the graphical representation for each of the pathways. Additionally, external pathways can be combined with matching internal pathways by checking the appropriate boxes on the left and clicking the *Show selected pathways* button.

Chapter 5

Discussion

This thesis describes the approach to a functional and metabolic analysis of biological whole-genome data. High-throughput techniques like the described methods for gene expression analysis and identification of protein-protein interactions produce data sets on a large scale. An efficient and comprehensive analysis of such huge data sets is only possible in an automatic way, making extensive use of computer systems. The isolated analysis of any kind of data has to be the first step in a chain of analytic steps. This includes basic processing methods like quality control, error correction, and normalization, but also high-level methods like the clustering of gene expression data.

5.1 The approach to an integrative analysis

Following an introduction to the problem domain, Chapter 2 of this thesis focusses on three fields of biotechnology and bioinformatics that are currently of great interest as one can tell from the large number of publications dealing with these topics: gene expression, protein-protein interactions and metabolic pathways. For each of them, I give an introduction that describes potential benefits of the technologies and that highlights the computational challenges that arise from the analysis. I explain suitable computational analysis methods for the respective data sets and describe in detail the analysis techniques I developed in this thesis. All of these methods, the SOM clustering for gene expression data, the graph modeling of protein-protein interactions and the three methods for a dynamic modeling of metabolic pathways, have been developed with the combinatorial analysis approaches in mind. They prepare the ground for the integrative methods that are described in Chapter 3.

The knowledge-based *integrative analysis methods* are used on a second, more abstract level of analysis. Large data sets of one type are analyzed in conjunction with another type of data. Combinatorial and integrative analysis methods make use of biological knowledge in order to achieve qualitative and reliable results. Data sets are analyzed in the context of systematic, previously assembled facts, leading to a more holistic view of the subjects of analysis.

Combining protein-protein interaction data with systematic functional annotations, we are able to focus on a specific biological context. This allows to scale and to reduce the complexity of the large protein-protein interaction data sets and makes the results comprehensible. More importantly, we show how the integration of protein-protein interaction data with functional annotations allows to hypothesize on the functional context of previously uncharacterized genes and proteins [FAZ⁺00]. This way, the intuitive graph representation and the integrational approach allow to assess large lists of protein-protein interactions efficiently.

Numerous scientific papers assess the management, statistical processing, normalization and clustering of gene expression data, e.g. [ARC00, BFB⁺00, BV00, YR01, BL01, NSH02]. This thesis describes the use of biochemical reactions (*dynamic modeling of metabolic pathways*) and textbook metabolic pathways (*pathway projection*) for the analysis of statistically evaluated and clustered gene expression data. It allows to analyze the metabolic properties and the changes in metabolism that have been captured by the respective gene expression experiment. Interesting features like co-regulated or conversely regulated pathways are highlighted by the integrative methods. Employing the set of integrative methods that are the subject of this thesis, it is possible to work with the established schemes and categories of textbook metabolic pathways or to construct hypothetical pathways dynamically based on the gene expression profiles. The hypothetical pathways do not necessarily correspond to textbook pathways. From the structure of a hypothetical pathway, relations between parts of an organism's metabolic network can be inferred that otherwise appear in distinct textbook pathways.

The *functional projection*, i.e. the integration of gene expression data with functional annotations has a broader, less specific range of application. An expression data set can be analyzed in the context of each of the various categories of a functional classification scheme like the MIPS functional catalog [MFG⁺00]. The functional catalog covers a very broad range of diverse categories that focus on different aspects of protein function. The specificity of a functional projection analysis can be varied due to the hierarchical organization of the catalog. The functional projection is capable of identifying functionally related sets of genes that exhibit similar, correlated or anti-correlated expression profiles. Cellular processes that are co-ordinately

switched on or off during a biological experiment are revealed. The relation between the experiment, e.g. a systematic variation of environmental conditions, and the response of the analyzed organism on a molecular level becomes obvious. The analysis of overlapping groups of functionally coupled genes reveals how the genes of different functional categories relate. This sheds light onto a larger biological context. The combination of the functional projection with the metabolic analysis methods allows to further investigate the identified co-regulated gene groups with a specific focus on aspects of intermediary metabolism. Again, the integrative analysis extracts and reveals interesting phenomena from data sets that could not be found without making use of additional, previously assembled knowledge.

All these methods are generically applicable also for other types of high-throughput data and for other systematically annotated facts about genes and proteins. Two recent publications report on a comprehensive analysis of yeast protein complexes [GBK⁺02, HGH⁺02]. Data on protein complexes can easily be used for an integrative analysis, e.g. in a *protein complex projection* (analogous to the pathway projection) or to restrict protein-protein interaction data to protein complexes instead of functional categories. In addition to the functional catalog, the MIPS yeast project group has defined catalogs for subcellular localization, mutant phenotypes, protein classes, and PROSITE motifs. All these systematic collections of biological knowledge can further promote integrative analysis methods.

- Before using one of the described metabolic analysis methods, a cellular localization catalog allows to restrict the gene set to those genes that are expressed in the same cellular compartment.
- Once the data is available, the use of signal transduction pathways and regulatory pathways will support the analysis of gene expression data. Signal transduction pathways as well as regulatory pathways can be modeled in much the same way as metabolic pathways. Hence the same integrative approaches can be applied, but the focus will be more general. Using metabolic pathways, an analysis is restricted to aspects of intermediary metabolism. Using pathways that describe signal transduction cascades and regulatory networks, an analysis may focus on arbitrary cellular processes.
- Mutant phenotype information, in conjunction with a specially designed gene expression experiment, potentially allows to reveal underlying metabolic mechanisms and properties of a respective phenotype.

The analysis of the proteomes of organisms (proteomics) will play an increasingly important role in molecular biology and biotechnology. As indicated above, new technologies have been described that will replace the

expensive and unhandy 2D gels now used in proteomics. Quantitative proteomics will produce protein expression data on a large scale. The methods for the integrative analysis of gene expression data that I describe in this thesis can readily be applied to protein expression data. The metabolic mapping and the pathway projection approaches are even more appropriate for protein expression data since an interpretation of the results does not rely on the partly incorrect assumption of a direct correspondance between mRNA abundance and protein abundance in the cells. Though the techniques of proteomic analysis have not even matured to a state where they would represent a broadly applied standard, the benefits of an integration of genomic analyses, proteomic analyses, and supplementary information has been reviewed [Fel01a].

5.2 Related work

Bioinformatic approaches to a combinatorial and integrative analysis of whole-genome data, including those developed during these doctoral studies, have a great potential. Recently a growing number of scientific papers have been published that introduce integrative analysis methods. This rather new category of methods will have significant impact on the further development of bioinformatic methods. It is now widely accepted that only the combination of different types of data can lead significantly towards a true systems biology that may allow the understanding and simulation of reasonably complex cellular processes or even whole cells.

- The group of David Eisenberg published a combined algorithm for genome-wide prediction of protein function [PMT⁺99]. They collect functional evidences from various sources and create *protein phylogenetic profiles* in order to predict function [MPT⁺99].
- Ralf Zimmer and Thomas Lengauer analyze dynamically derived metabolic pathways in the context of gene expression profiles [KZL00, ZKZL00]. Their approach is discussed in more detail in Chapter 3.2.
- Noordewier and Warren point out that the efficient use of gene expression measurements requires the integration of other sources of biological information [NW01].
- The correlation between gene expression profiles and biological function, assessed in Chapter 3.1 of this thesis, is also being studied by other groups. Instead of systematic functional annotations, computer

linguistics is employed to derive statistically significant terms from literature that describe biological function [OBH⁺00, MWF⁺01].

- Ideker et al. demonstrate an integrated approach to build, test, and refine a model of a cellular pathway using cDNA microarrays, quantitative proteomics and protein-protein interactions [ITR⁺01].
- Grigoriev reports on the relationship between gene expression and protein interactions on the proteome scale [Gri01].
- Mark Gerstein published a number of papers that discuss different data integration approaches. His group investigates how gene expression data relates to protein structure and function [GJ00]. For the study they also make use of the MIPS functional catalog. Another publication addresses the relation between gene expression data and protein-protein interactions [JGG02].

5.3 Relevance of the approach

Before having bioinformatics at hand, biologists already approached problems of the discussed kind. Instead of employing high-throughput techniques, experiments had to be carried out manually, investigating a single gene or protein a time. That way, experiments can be planned specifically in order to find answers to explicit questions. The results are to a very high degree reliable since the significance and reliability of each measurement and each partial result can be directly judged by the biologist. Whole-genome or whole-proteome data sets obtained by high-throughput methods cannot be verified to a comparable extent. They are faced with the trade-off between sensitivity (not missing too many signals) and selectivity (not getting too many false positive signals). The reliability of the results is assessed by automatically run protocols. These are often related with statistical methods that assign confidence scores to whole data sets as well as single data points. Bayesian statistics is widely employed. Careful experimental design and the development of intelligent and powerful analysis software can reduce the ratio of false results.

Integrative analysis methods support and enhance the analysis of noisy data. The additional context provided by a second kind of data, especially systematic annotations of the genes and proteins, helps to distinguish between real signals and insignificant artificial signals. When interpreting the results of an integrative analysis method, one has to keep the characteristics and shortcomings of the underlying experiments and data sets in mind. Implicit assumptions have to be taken into account, e.g. the discussed assumption of

a tight correlation of gene expression levels and protein levels that is made when integrating metabolic pathways and gene expression data. A potential source of misinterpretation are errors in the databases that are employed. Some database entries, especially of manually curated databases may be biased by the personal view of the curators. In many cases, database entries do not contain information on reliability. An important issue influencing the quality of a database entry is, whether attributes have been experimentally verified or whether they are just derived from other database entries that may themselves be derived from others instead of being experimentally verified. In other database entries, the underlying experimental data may be incorrect, e.g. the data that stems from two-hybrid assays that are known to generate a large number of false positive signals [LS00].

Despite of these weaknesses, the strong sides of high-throughput biology in combination with an advanced bioinformatics framework are significant. The whole-genome approaches allow to test, i.e. to prove or to disprove, a large number of hypotheses. By bioinformatic methods, massive amounts of large data sets can be mined automatically, often making extensive use of knowledge stored in biological databases. Data mining is capable of extracting facts and interesting features from data sets that the scientist did not explicitly ask for. This enables the scientist to formulate new hypotheses and to establish predictions that he can test by other high-throughput experiments or by classical laboratory experiments that focus on a single or a few genes or proteins.

Leroy Hood introduced the term *systems biology* for the approach to systematically disturb biological systems and to monitor the gene, protein, and information pathway responses [IGH01]. Bioinformatics is an inevitable part of this high-throughput biology. Intelligent combinatorial and integrative bioinformatic methods that interrelate various kinds of data sets either to verify a hypothesis or to suggest new hypotheses will prove crucial as post-genome biology moves further towards systems biology.

Appendix A

Acronyms

BFS breadth first search. Algorithm that visits all nodes or edges of a graph exactly once. Given a starting node, all adjacent edges are visited first.

DFS depth first search. Algorithm that visits all nodes or edges of a graph exactly once. Given a starting node, the adjacent edges of the node are visited. For every adjacent edge, the DFS algorithm is called recursively with the target node of the respective edge as the starting node.

cDNA complementary DNA. DNA that is synthesized, by reverse transcriptase, from an mRNA template, and therefore has no introns.⁽¹⁾

CGI Common Gateway Interface. A specification for transferring information between a World Wide Web server and a CGI program. A CGI program is any program designed to accept and return data that conforms to the CGI specification. The program could be written in any programming language, including C, Perl, or Java. CGI programs are the most common way for Web servers to interact dynamically with users. Many HTML pages that contain forms use a CGI program to process the form's data once it is submitted.⁽²⁾

CORBA Common Object Request Broker Architecture. An architecture that enables pieces of programs, called objects, to communicate with one another regardless of what programming language they were written in or what operating system they are running on. CORBA was developed by an industry consortium known as the Object Management Group (OMG).⁽²⁾

DCM dilated cardiomyopathy. An acquired disease characterized by the progressive loss of cardiac contractility of unknown cause. As cardiac contractile function is progressively lost, there is a decrease

in cardiac output. Increased blood volume and pressure within the chambers causes them to dilate, most dramatically evident in the left atrium and left ventricle. In response to the poor contractility and decreased cardiac output, the sympathetic nervous system and the renin-angiotensin-aldosterone axis are activated. As with degenerative valve disease, these compensatory mechanisms are initially beneficial, however their chronic activation becomes deleterious. With time, the enlarged heart gradually deteriorates, causing congestive heart failure.

DNA deoxyribonucleic acid. A macromolecule formed of repeating deoxyribonucleotide units linked by phosphodiester bonds between the 5'-phosphate group of one nucleotide and the 3'-hydroxy group of the next. DNA appears in Nature in both double-stranded (the *Watson-Crick* model) and single-stranded forms and functions as a repository of genetic information. The information is encoded in its base sequence.⁽¹⁾

dUTP deoxyuridine triphosphate. Substance that is used for the red/green labeling of cDNA in microarray experiments.

EC Enzyme Commission. The IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) is responsible for assigning each enzyme a recommended name and number, the EC number, to allow it to be identified. The list so obtained is published at intervals. Its most recent printed edition is *Enzyme Nomenclature*, published by Academic Press for IUBMB in 1992. Several supplements have also been published.

EST Expressed Sequence Tag. A partial coding sequence isolated at random from a cDNA library; like a sequence-tagged site for mapping total genomic DNA, used for identification and mapping of coding sequences, for discovery of new genes and (by reference to sequence data banks) for discovery of identities with other genes. (Venter, C. (1993) *Nature Genet.* 4:373-380.⁽¹⁾)

ENZYME Enzyme nomenclature database. ENZYME is part of the ExPASy database collection.

EUROFAN European Network for Functional Analysis of Yeast. A network of a large number European biochemical laboratories organized in 23 nodes. The project is focused on the systematic deletion of all the genes of yeast. The knockout strains are evaluated in order to reveal altered characteristics and possibly assign a function to unknown gene sequences. MIPS is a participant of EUROFAN.

ExPASy Expert Protein Analysis System. ExPASy contains a number of databases produced at the Swiss Institute of Bioinformatics (SIB) in

Geneva, such as SWISS-PROT, PROSITE, SWISS-2DPAGE, SWISS-3DIMAGE, ENZYME, CD40Lbase and SeqAnalRef, as well as other cross-referenced databases.

GUI Graphical User Interface. A program interface that takes advantage of the computer's graphics capabilities to make the program easier to use. Well-designed graphical user interfaces can free the user from learning complex command languages.⁽²⁾

HTML Hyper Text Markup Language. The authoring language used to create documents on the World Wide Web. HTML defines the structure and (to a certain extent) the layout of a Web document by using a variety of tags and attributes.⁽²⁾

HTTP Hyper Text Transfer Protocol. The underlying protocol used by the World Wide Web. HTTP defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands.⁽²⁾

HTx orthotopic heart transplantation. Surgical removal of the heart and its replacement with a new donor heart precisely in the place where the old one pumped blood.

KEGG Kyoto Encyclopedia of Genes and Genomes. KEGG is an initiative of the Institute for Chemical Research, Kyoto University, Japan. KEGG is available at www.genome.ad.jp/kegg/.

LEDA Library of Efficient Data Structures and Algorithms.

LIGAND Kyoto Chemical Database of Enzyme Reactions.

MIPS the Munich Information Center for Protein Sequences. MIPS is part of the GSF, Department IBI, Neuherberg, Germany. The group has been located at the Max-Planck-Institute f. Biochemistry, Martinsried, Germany, until March 2001.

mRNA messenger ribonucleic acid. The RNA that contains the coded information, as sequences of codons, for protein synthesis.⁽¹⁾

PEDANT Protein Extraction, Description and ANalysis Tool.

ORF open reading frame. One of three possible reading frames in which an mRNA is potentially translated into protein. In analysis of a DNA sequence, an ORF is characterized by the sequence of nucleotides that when transcribed into mRNA results in a series of triplet codons that is not interrupted by a translation termination codon.⁽¹⁾

- PCR** polymerase chain reaction. A technique to amplify a specific region of double-stranded DNA. An excess of two amplimers, oligonucleotide primers complementary to two sequences that flank the region to be amplified, are annealed to denatured DNA and subsequently elongated, usually by a heat-stable DNA polymerase from *Thermus aquaticus* (*Taq* polymerase). Each cycle involves heating to denature double-stranded DNA and cooling to allow annealing of excess primer to template and elongation of the primers by the *Taq* polymerase; the number of amplicons, i.e. the target sequence fragments between flanking primers, doubles with each cycle.⁽¹⁾
- RMI** Remote Method Invocation. A set of protocols being developed by Sun's JavaSoft division that enables Java objects to communicate remotely with other Java objects. RMI is a relatively simple protocol, but unlike more complex protocols such as CORBA, it works only with Java objects.⁽²⁾
- RNA** ribonucleic acid. A macromolecule formed of repeating ribonucleotide units linked by phosphodiester bonds between the 5'-phosphate group of one nucleotide and the 3'-hydroxy group of the next. RNA has several biological functions, most of which depend upon its ability to form sequence-specific interactions with DNA. RNA comprises the genome of some viruses.⁽¹⁾
- SAGE** Serial Analysis of Gene Expression.
- SOM** self-organizing map. The SOM was invented by Teuvo Kohonen. It is a self-organizing neural network providing a mapping from some high-dimensional data space into the two-dimensional discrete space of the map. The map is organized in a regular square grid with a neuron on every grid crossing.
- TCA** *tri-carbon acid cycle*. A central pathway of intermediary metabolism involved in the degradation of glucose.
- VAD** Ventricular Assist Device. A left ventricular assist device is a mechanical pump-type device that is surgically implanted. It helps to maintain the pumping ability of a heart that cannot effectively work on its own due to pathological procedures. The VAD is sometimes referred to as a *bridge to transplant*.
- WIT** *What Is There?* WIT is a world wide web-based system to support the curation of function assignments made to genes and the development of metabolic models.
- XML** eXtensible Markup Language. A recommendation of the World Wide Web Consortium (W3C) (www.w3.org), XML is a subset of SGML.

XML is a meta-language, which enables a general availability and interchange of information that is structured according to its content (www.xml.org).

Some of these definitions origin from (1) the David M. Glick *Glossary of Biochemistry and Molecular Biology* (<http://db.portlandpress.com/glick/search.htm>), (2) the *Webopedia* Online Dictionary for Computer and Internet Technology (<http://webopedia.internet.com/>) or the WWW pages of the described institutions and databases. The respective literal citations are marked.

Bibliography

- [ABL⁺89] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell*. Garland Publishing, 1989.
- [ABN⁺99] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of the Sciences of the USA*, 96:6745–6750, 1999.
- [AGH⁺01] K. Albermann, U. Güldener, J. Hani, M. Münsterkötter, S. Stocker, A. Zollner, and H. W. Mewes. The MIPS yeast genome database. <http://mips.gsf.de/proj/yeast/>, 1996–2001.
- [AHZ00] K. Albermann, J. Hani, and A. Zollner. The MIPS yeast interaction tables. <http://mips.gsf.de/proj/yeast/tables/interaction/>, 1996–2000.
- [Alb96] R. A. Alberty. Calculation of biochemical net reactions and pathways by using matrix operations. *Biophysical Journal*, 71(1):507–515, 1996.
- [AML92] B. Alberts and R. Miake-Lye. Unscrambling the puzzle of biological machines: the importance of the details. *Cell*, 68(3):415–20, 1992.
- [AMP⁺00] E. Arbustini, P. Morbini, A. Pilotto, A. Gavazzi, and L. Tavazzi. Familial dilated cardiomyopathy: from clinical presentation to molecular genetics. *Eur. Heart J.*, 21:1825–1832, 2000.
- [APS99] T. K. Attwood and D. J. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.
- [ARC00] J. Aach, W. Rindone, and G.M. Church. Systematic management and analysis of yeast gene expression data. *Genome Research*, 10:431–445, 2000.
- [Att00] T. Attwood. The babel of bioinformatics. *Science*, 290:471–473, 2000.
- [Bai00] A. Bairoch. The ENZYME database in 2000. *Nucl Acids Res*, 28(1):304–305, 2000.
- [Bar89] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

- [BFB⁺00] T. Beißbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J. M. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schtz, A. Poustka, and M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16(11):1014–1022, 2000.
- [BGPW96] H.-U. Bauer, T. Geisel, K. Pawelzik, and F. Wolf. Selbstorganisierende neuronale Karten. *Spektrum der Wissenschaft*, 4:38–47, 1996.
- [BL01] P. Baldi and A. D. Long. A bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.
- [BLR⁺00] A. Brachat, N. Liebundguth, C. Rebischung, et al. Analysis of deletion phenotypes and GFP fusions of 21 novel *Saccharomyces cerevisiae* open reading frames. *Yeast*, 16(3):241–253, 2000.
- [BLS01] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–171, 2001.
- [BMS⁺99] B. Bartling, H. Milting, H. Schumann, D. Darmer, L. Arusoglu, M. M. Koerner, A. El-Banayosy, R. Koerfer, J. Holtz, and H. R. Zerkowski. Myocardial gene expression of regulators of myocyte apoptosis and myocyte calcium homeostasis during hemodynamic unloading by ventricular assist devices in patients with end-stage heart failure. *Circulation*, 100:II216–223, 1999.
- [BV00] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.
- [CEM⁺98] S. Chu, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.
- [Cla01] J. M. Claverie. Gene number. what if there are only 30,000 human genes? *Science*, 291(5507):1255–1257, 2001.
- [Con99] The FlyBase Consortium. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Res*, 27(1):85–88, 1999.
- [DEKM98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
- [DH97] P. M. Dey and J. B. Harborne, editors. *Plant Biochemistry*. Academic Press, 1997.
- [DIB97] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.
- [DLEB⁺01] M. C. Deng, M. Loebe, A. El-Banayosy, E. Gronda, P. G. M. Jansen, M. Vigano, G. M. Wieselthaler, B. Reichart, E. Vitali, A. Pavie,

- T. Mesana, D. Y. Loisanca, D. R. Wheeldon, and P. M. Portner. Mechanical circulatory support for advanced heart failure – Effect of patient selection on outcome. *Circulation*, 103:231–237, 2001.
- [EBPH99] K. Eilbeck, A. Brass, N. Paton, and C. Hodgman. INTERACT: an object oriented protein-protein interaction database. In *Proc of the Seventh Int Conf on Intelligent Systems for Mol Biol*, pages 87–94. AAAI Press, 1999.
- [EFK⁺99] S. G. Erberich, M. Fellenberg, T. Krings, S. Kemeny, W. Reith, K. Willmes, and W. Oberschelp. Unsupervised time course analysis of functional magnetic resonance imaging (fMRI) using self-organizing maps (SOM). In Chin-Tu Chen and Anne V. Clough, editors, *Proc. of Physiology and Function from Multidimensional Images, Medical Imaging*. SPIE Press, 1999.
- [ESBB98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of the Sciences*, 95:14863–14868, 1998.
- [FAH⁺01] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H. W. Mewes. Functional and structural genomics using PEDANT. *Bioinformatics*, 17(1):44–57, 2001.
- [FAZ⁺00] M. Fellenberg, K. Albermann, A. Zollner, H. W. Mewes, and J. Hani. Integrative analysis of protein interaction data. In R. Altmann, T.L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck, and H. Weissig, editors, *Intelligent Systems for Molecular Biology*, volume 8, pages 152–161. AAAI Press, 2000.
- [Fel98] M. Fellenberg. Selbstorganisierende Neuronale Karten: durch Topologieerhaltung gesteuertes Lernen. Diploma thesis, Computer Science Department, University of Dortmund, Germany, 44221 Dortmund, Germany, 8 1998.
- [Fel01a] D. A. Fell. Beyond genomics. *Trends in Genetics*, 17(12):680–682, 2001.
- [Fel01b] M. Fellenberg. The MIPS gene expression analysis web site. <http://mips.gsf.de/proj/expression/>, 2000, 2001.
- [FM67] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.
- [FM99] M. Fellenberg and H. W. Mewes. Interpreting clusters of gene expression profiles in terms of metabolic pathways (Poster). In *Proceedings of the German Conference on Bioinformatics*, pages 185–187, 1999.
- [FMM⁺94] O. H. Frazier, M. P. Macris, T. J. Myers, J. M. Duncan, B. Radovancevic, S. M. Parnis, and D. A. Cooley. Improved survival after extended bridge to cardiac transplantation. *Ann Thorac Surg*, 57:1416–1422, 1994.

- [FRRL97] M. Fromont-Racine, J. C. Rain, and P. Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet*, 16(3):277–82, 1997.
- [GAAC⁺97] A. Goffeau, R. Aert, M. L. Agostini-Carbone, et al. The yeast genome directory. *Nature*, 387(6632 Suppl):1–105, 1997.
- [GBK⁺02] A. Gavin, M. Bösch, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. S., B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 2002.
- [GFS95] G. J. Goodhill, S. Finch, and T. J. Sejnowski. A unifying measure for neighbourhood preservation in topographic mappings. In *Proceedings of the 2nd Joint Symposium on Neural Computation, University of California, San Diego and California Institute of Technology, Pasadena, CA*, volume 5, pages 191–202, La Jolla, CA, 1995. Institute for Neural Computation.
- [GJ00] M. Gerstein and R. Jansen. The current excitement in bioinformatics – analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Op. in Structural Biol.*, 10:574–584, 2000.
- [GK00] S. Goto and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.
- [GNK98] S. Goto, T. Nishioka, and M. Kanehisa. LIGAND: chemical database for enzyme reactions. *Bioinformatics*, 14(7):591–599, 1998.
- [GNK00] S. Goto, T. Nishioka, and M. Kanehisa. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res*, 28(1):380–382, 2000.
- [GRFA99] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730, 1999.
- [Gri01] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucl Acids Res*, 29(17):3513–3519, 2001.
- [GS96] G. J. Goodhill and T. J. Sejnowski. Quantifying neighbourhood preservation in topographic mappings. In *Proceedings of the 3rd Joint Symposium on Neural Computation, University of California, San Diego and California Institute of Technology, Pasadena*, volume 6, pages 61–82, Pasadena, CA, 1996. California Institute of Technology.

- [GST⁺99] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- [HGH⁺02] Yuen Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, Sally-Lin Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Srensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–183, 2002.
- [HLB⁺00] S. Hoersch, C. Leroy, N. P. Brown, M. A. Andrade, and C. Sander. The GeneQuiz web server: protein functional analysis through the web. *Trends Biochem Sci*, 25(1):33–35, 2000.
- [HMM⁺00] D. N. Helman, S. W. Maybaum, D. L. Morales, M. R. Williams, A. Beniaminovitz, N. M. Edwards, D. M. Mancini, and M. C. Oz. Recurrent remodeling after ventricular assistance: is long-term myocardial recovery attainable? *Ann Thorac*, 70:1255–1258, 2000.
- [HN90] R. Hecht-Nielsen. *Neurocomputing*. Addison-Wesley, Reading, MA, 1990.
- [HS98] R. Heinrich and S. Schuster. The modelling of metabolic systems. Structure, control and optimality. *Biosystems*, 47(1-2):61–77, 1998.
- [HVS⁺98] N. Hauser, M. Vingron, M. Scheideler, B. Krems, K. Hellmuth, K.D. Entian, and J.D. Hoheisel. Transcriptional profiling on all Open Reading Frames of *Saccharomyces cerevisiae*. *Yeast*, 14:1209–21, 1998.
- [IGH01] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372, 2001.
- [Int01] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.
- [ITM⁺00] T. Ito, K. Tashiro, S. Muta, et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci USA*, 97(3):1143–1147, 2000.
- [ITR⁺01] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292:929–934, 2001.

- [JCCS01] L. M. Jakt, L. Cao, K. S. E. Cheah, and D. K. Smith. Assessing clusters and motifs from gene expression data. *Genome Research*, 11:112–123, 2001.
- [JGG02] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12:37–46, 2002.
- [JM92] M. Johnston and M. Carlson. Gene expression. In E. W. Jones, J. R. Pringle, and J. R. Broach, editors, *The Molecular Biology of the Yeast *Saccharomyces**, page 193ff., Cold Spring Harbor, NY, 1992. Cold Spring Harbor Laboratory Press.
- [K⁺01] M. Kanehisa et al. Kyoto Encyclopedia of Genes and Genomes. <http://www.genome.ad.jp/kegg/>, 1998–2001.
- [KBIU⁺98] A. T. Kawaguchi, J. Bergsland, H. Ishibashi-Ueda, T. Ujiie, S. Shimura, S. Koide, T. A. Salerno, and R. J. Batista. Partial left ventriculectomy in patients with dialated failing ventricle. *J Card Surg*, 13:335–342, 1998.
- [Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.
- [Koh95] T. Kohonen. *Self-Organizing Maps*. Number 30 in Springer Series in Information Sciences. Springer-Verlag, Berlin, 1995.
- [KRP⁺99] P.D. Karp, M. Riley, S.M. Paley, A. Pellegrini-Toole, and M. Krummenacker. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Research*, 27(1):55–58, 1999.
- [KvZB⁺99] A.J. Kal, A.J. van Zonneveld, V. Benes, M. v. d. Berg, M.G. Korkamp, K. Albermann, N. Strack, J.M. Ruijter, B. Dujon, W. Anson, and H.F. Tabak. Dynamics of gene expression revealed by comparison of SAGE Transcript Profiles from yeast grown on different carbon sources. *Molecular Biology of the Cell*, 10:1859–1872, 1999.
- [KZL00] R. Küffner, R. Zimmer, and T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (dmd). *Bioinformatics*, 16(9):825–836, 2000.
- [LDB⁺96] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(12):1675–1680, 1996.
- [LHE⁺99] A. Lueking, M. Horn, H. Eickhoff, K. Büssow, H. Lehrach, and G. Walter. Protein microarrays for gene expression and antibody screening. *Analytical Biochemistry*, 270:103–111, 1999.
- [LKS⁺00] M. A. Luttk, P. Kotter, F. A. Salomons, I. J. van der Klei, J. P. van Dijken, and J. T. Pronk. The *Saccharomyces cerevisiae* ICL2 gene encodes a mitochondrial 2-methylisocitrate lyase involved in propionyl-coenzyme a metabolism. *J. Bacteriol.*, 182(24):7007–7013, 2000.

- [LKWS00] M. T. Lee, F. C. Kuo, G. A. Whitmore, and Jeffrey Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA*, 97:9834–9839, 2000.
- [LL91] G. G. Lennon and H. Lehrach. Hybridization analyses of arrayed cDNA libraries. *Trends Genet*, 7(10):314–317, 1991.
- [LS00] P. Legrain and L. Selig. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Letters*, 480:32–36, 2000.
- [LVdVC⁺93] C. M. Lucas, F. H. Van der Veen, E. C. Cheriex, R. Lorusso, M. Havenith, O. C. Penn, and H. J. Wellens. Long-term follow-up (12 to 35 weeks) after dynamic cardiomyoplasty. *J Am Coll Cardiol*, 22:758–767, 1993.
- [M⁺93] G. Michal et al. Boehringer Mannheim metabolic pathways wallchart. <http://www.expasy.org/cgi-bin/search-biochem-index>, 1993.
- [MAB⁺97] H.W. Mewes, K. Albermann, M. Bähr, et al. Overview of the yeast genome. *Nature*, 387(6632 Suppl):7–65, 1997.
- [Mav93] M.L. Mavrovouniotis. Identification of qualitatively feasible metabolic pathways. In L. Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 325–364. AAAI Press / MIT Press, 1993.
- [MBK⁺01] H. Milting, A. El Banayosy, A. Kassner, L. Arusoglu, A. Sczyrba, M. Fellenberg, R. Thieleczek, V. Liebscher, and R. Koerfer. Myocardial gene expression analysis by cDNA-arrays in patients suffering from dilated cardiomyopathy and supported by ventricular assist devices. *Circulation*, 2001.
- [MCA⁺98] G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi. Cluster analysis and data visualization of large-scale gene expression data. In *Pacific Symposium on Biocomputing*, volume 3, pages 42–53, 1998.
- [MFG⁺00] H.W. Mewes, D. Frishman, C. Gruber, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 28(1):37–40, 2000.
- [MGD95] I. Moszer, P. Glaser, and A. Danchin. SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology*, 141(Pt 2):261–268, 1995.
- [Mic99] G. Michal, editor. *Biochemical Pathways*. Spektrum Akademischer Verlag, 1999.
- [MN99] K. Mehlhorn and St. Näher. *The LEDA Platform of Combinatorial and Geometric Computing*. Cambridge University Press, 1999.
- [MPT⁺99] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, 1999.

- [MWF⁺01] D. R. Masys, J. B. Welsh, J. L. Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–326, 2001.
- [NI92] NC-IUBMB. *Enzyme Nomenclature – Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB)*. Academic Press, New York, 1992.
- [NSH02] J. P. Novak, R. Sladek, and T. J. Hudson. Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, 79(1):104–113, 2002.
- [NW01] M. O. Noordewier and P. V. Warren. Gene expression microarrays and the integration of biological knowledge. *Trends in Biotechnology*, 19(10):412–415, 2001.
- [OBH⁺00] J. C. Oliveros, C. Blaschke, J. Herrero, J. Dopazo, and A. Valencia. Expression profiles and biological function. *Genome Informatics*, 11:106–117, 2000.
- [OS⁺01] R. Overbeek, E. Selkov, et al. What is there. <http://wit.mcs.anl.gov/WIT2>, 1997–2001.
- [OWKB98] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends Biotechnol*, 16(9):373–8, 1998.
- [Pel01] M. Pellegrini. Computational methods for protein function analysis. *Current Opinion in Chemical Biology*, 5(1):46–50, 2001.
- [Pev00] P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. Bradford Book, 2000.
- [PMT⁺99] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of the Sciences of the USA*, 96:4285–4288, 1999.
- [PSE⁺00] C.M. Perou, T. Sorlie, M.B. Eisen, et al. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.
- [Red88] C. Reder. Metabolic control theory: A structural approach. *Journal of Theoretical Biology*, 135:175–201, 1988.
- [RHT00] S.C.G. Rison, T.C. Hodgman, and J.M. Thornton. Comparison of functional annotation schemes for genomes. *Functional & Integrative Genomics*, 1(1):56–69, 2000.
- [Ril93] M. Riley. Functions of the gene products of escherichia coli. *Microbiol Rev*, 57(4):862–952, 1993.
- [RMS90] H. J. Ritter, T. M. Martinetz, and K. J. Schulten. *Neuronale Netze: Eine Einführung in die Neuroinformatik selbstorganisierender Abbildungen*. Addison-Wesley, München, 1990.
- [SAG⁺95] J.-C. Sanchez, R. D. Appel, O. Golaz, C. Pasquali, F. Ravier, A. Bairoch, and D. F. Hochstrasser. Inside SWISS-2DPAGE database. *Electrophoresis*, 16:1131–1151, 1995.

- [SBM⁺00] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, 28(10):47e, 2000.
- [SFD00] S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18:326–332, 2000.
- [SFS99] T.W. Simpson, B.D. Follstad, and G. Stephanopoulos. Analysis of the pathway structure of metabolic networks. *Journal of Biotechnology*, 71:207–233, 1999.
- [SLP00] C.H. Schilling, D. Letscher, and B.O. Palsson. Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J theor Biol*, 203:229–248, 2000.
- [SPC95] G. Sipos, A. Puoti, and A. Conzelmann. Biosynthesis of the side chain of yeast glycosylphosphatidylinositol anchors is operated by novel mannosyltransferases located in the endoplasmic reticulum and the golgi apparatus. *J Biol Chem*, 270(34):19709–15, 1995.
- [SSZ⁺98] P. T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.
- [SUF00] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.
- [TSM⁺99] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of the Sciences*, 96:2907–2912, 1999.
- [UGC⁺00] P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–27, 2000.
- [VAM⁺01] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.
- [VDHM97] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz. Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.
- [vHVvH⁺00] N.L.W. van Hal, O. Vorst, A.M.M.L. van Houwelingen, E.J. Kok, A. Peijnenburg, A. Aharoni, A.J. van Tunen, and J. Keijer. The application of DNA microarrays in gene expression analysis. *J of Biotechnology*, 78:271–280, 2000.

- [VZVK95] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.
- [WCF⁺01] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283, 2001.
- [WFM⁺98] X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, 95:334–339, 1998.
- [WS01] J. Wojcik and V. Schächter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17:296S–305S, 2001.
- [WSL⁺00] A. J. M. Walhout, R. Sordella, X. Lu, et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287:116–122, 2000.
- [YR01] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.
- [ZHAM01] A. Zollner, J. Hani, K. Albermann, and H. W. Mewes. The MIPS Functional Catalogue for *S. cerevisiae*. <http://mips.gsf.de/proj/yeast/catalogues/funecat/>, 1996–2001.
- [ZHM99] J. Zhao, L. Hyman, and C. Moore. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Mol Biol Rev*, 63(2):405–45, 1999.
- [ZKZL00] A. Zien, R. Küffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. In *Intelligent Systems for Molecular Biology*, volume 8, pages 407–417, 2000.