Facilitating functional interpretation of microarray data by integration of gene annotations in Correspondence Analysis

Dissertation
zur Erlangung des naturwissenschaftlichen Doktorgrades
der Fakultät für Biologie
an der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Christian Busold
Hameln

Würzburg, 2006

Eingereicht am ....

Mitglieder der Prüfungskommission:

Vorsitzender: Prof. Müller
1.Gutachter: Prof. Dandekar
2.Gutachter: Prof. Wiemann

Tag des Promotionskolloqiums:..........................

Doktorurkunde ausgehändigt am:........................

Für meine Eltern: Friedlinde und Klaus Busold

Contents

1 Introduction **1**
- 1.1 Microarray technology . . . 1
- 1.2 Ontology . . . 3
- 1.3 Current methods to analyze microarray data in context of annotation data . . 5

2 Integration of gene annotation data **8**
- 2.1 Gene Ontology . . . 10
  - 2.1.1 Overview . . . 10
  - 2.1.2 The structure of GO . . . 11
  - 2.1.3 Annotation of gene products . . . 12
  - 2.1.4 Exploiting the ’true path rule’ - i.e. how to associate genes to GO terms 13
- 2.2 Integration of gene annotation data in Correspondence Analysis . . . 14
  - 2.2.1 Correspondence Analysis . . . 14
  - 2.2.2 Interpretation of Correspondence Analysis plots . . . 15
  - 2.2.3 Visualizing the quality of display in CA . . . 17
  - 2.2.4 Gene annotation data in CA . . . 20
    - 2.2.4.1 Boolean implementation . . . 20
    - 2.2.4.2 Intensity based implementation . . . 22
    - 2.2.4.3 How to assign genes to a single, best-fitting GO term . . . 24
    - 2.2.4.4 Data as supplementary points vs. points with mass . . . 29
    - 2.2.4.5 Representation of a single gene by multiple features . . . 29
- 2.3 Filtering of GO annotation terms . . . 30
  - 2.3.1 Based on ontological characteristics . . . 33
  - 2.3.2 Based on gene characteristics . . . 35
  - 2.3.3 Receiver Operating Characteristic curves to evaluate filter performance 35
    - 2.3.3.1 Definition of ’standard annotations’ . . . 38
    - 2.3.3.2 Identification of optimal measure of homogeneity . . . 38
- 2.4 Biological validation of algorithms . . . 41
  - 2.4.1 *Saccharomyces cerevisiae* - glucose data set . . . 43
    - 2.4.1.1 Experimental setup . . . 43
    - 2.4.1.2 Intensity based coding . . . 43
    - 2.4.1.3 Application of Spearman-filter . . . 46

2.4.2 *Homo sapiens* 49
2.5 Analysis of transcription factors in CA 52
2.5.1 Transfac 52
2.5.2 Integration of transcription factors in CA 52
2.5.3 Incorporation of ChIP-chip data 53
2.5.4 Biological validation 54
2.5.4.1 *Saccharomyces cerevisiae* 54
2.5.4.2 *Homo sapiens* 61

3 Discussion **65**
3.1 Gene annotation data 65
3.2 Comparison of implementations 66
3.3 Applicability of annotation filters 68
3.4 Integration of transcription factors 71
3.5 Future prospects 72

4 Summary **75**

5 Zusammenfassung **77**

6 Publications **79**

Bibliography **81**

7 Appendix **97**
7.1 SQL query to extract father-child relations in GO 97
7.2 Experimental procedures for human cancer study 97
7.3 UML schema of the GO database 99
7.4 Snapshots of GUI 100
7.5 Comprehensive listing of annotations displayed in human cancer data set . . . 102
7.6 Unique assignments of gene products 103
7.7 Prediction of TF binding sites 107
7.8 Software used 107

Abbreviations **109**

1 Introduction

1.1 Microarray technology

Almost all cells in the human organism contain identical sets of chromosomes and thus also the same set of genes. Nevertheless this identical set gives rise to a huge variety of cells types fulfilling the most diverse functions. The majority of functionality in a cell is based on the activity of proteins, whereas the 'Central Dogma of Molecular Biology' identifies the DNA as the carrier of genetic information [1]. To translate this information into functionality (i.e. proteins) a transfer of information from DNA to an intermediate molecule, namely the mRNA, is carried out within the cell. After transport into the cell's cytoplasm the mRNA is translated into the corresponding protein. In general the presence and abundance of a particular mRNA regulates the presence and abundance of the encoded protein.

By measuring the abundance of mRNA molecules inferences on the activity of the encoding gene(s) can be made. DNA microarray technology [2, 3] allows to asses expression levels in a particular state of the cell for several ten-thousands of genes in a single experiment. To this end, the mRNA is extracted from cells and reversely transcribed to cDNA. During this process the cDNA is labeled by incorporation of labeled nucleotides. In the advent of the microarray technology, it was common to use radioactively labeled NTPs, whereas nowadays it is standard practice to use fluorescently labeled nucleotides [4]. In a consequent step, the cDNA is hybridized to a microarray.

The microarray itself consists of a solid support (glass-slide, nylon-membrane, silicon-chips or membrane-slides), on which single-stranded DNA fragments of different sequence have been immobilized at distinct, fixed locations. In case of expression profiling the length of the spotted DNA fragments can vary from as few as 10 bases (oligonucleotides) up to several thousand (cDNA) and are therefore referred to as oligo-microarrays and cDNA-microarrays respectively. The latter are commonly created by a robot depositing the DNA-fragments at specified locations. Oligo-arrays can be either spotted or the oligonucleotides can be synthesized directly on the chip by photolithographic means [5]. One prominent example of these are chips from Affymetrix [6].

Figure 1.1 provides an overview of the workflow of a typical microarray experiment: In short, mRNA is extracted from the samples of the experimental conditions and reversely transcribed into cDNA. The labeling can occur simultaneously with the reverse transcription (direct labelling) or in a subsequent step (indirect labeling). The labeled targets are combined and

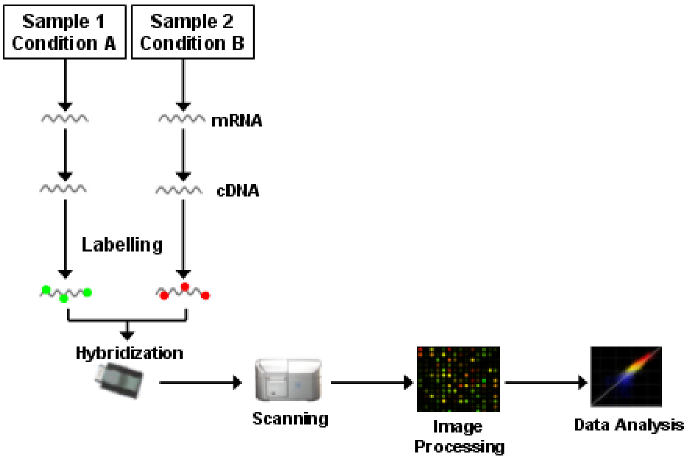


Figure 1.1: Workflow of a typical microarray experiment. mRNA is extracted from the sample(s) the experimental conditions that are to be compared and reversely transcribed into cDNA. The labelling can occur in the same step as the reverse transcription (direct labelling) or in a subsequent step as shown here (indirect labeling). The labeled targets or combined and hybridized to a microarray. After some postprocessing steps the arrays are scanned and the resulting images (commonly in tiff or bmp format) are processed to extract quantitative information on the genes' expression level for subsequent data analysis. This figure is reproduced from [7].

hybridized to a microarray. After some postprocessing steps the arrays are scanned and the resulting images are processed to extract quantitative information on the genes' expression levels for subsequent data analysis.

The quantified expression data can be represented as a matrix in which the rows depict the genes and the columns the individual hybridizations, the cells contain the corresponding expression intensities (a scheme of such a data-matrix from a simplified microarray experiment in provided in Table 3.1 on page 68). These intensities (in their non-preprocessed state) are commonly referred to as 'raw data'. This raw data as such, is not suitable for immediate analysis, since the amount of variation having accumulated in the data at the various experimental steps can be so predominant that the biological signals of interest are obscured. Technical variation can be introduced at almost every step in the production of a microarray, examples of which include the amount of DNA in each spot, spot shape, dye bias (i.e. decay rates and different labeling efficiencies), inhomogeneous slide surfaces, edge-effects, cross-hybridization and scanning parameters. This demonstrates the need for a normalization step prior to identification of regulated genes [8, 9, 10].

Besides the variety in normalization methods a wealth of procedures for subsequent analysis of the data is available . Clustering algorithms were among the first to be applied to microarray data. In general, clustering methods group genes or experiments such that the expression profiles within the groups are more alike than expression profiles across the groups. Among the frequently used clustering techniques are hierarchical [11, 12], k-Means [13], and self-organizing maps (SOMs) [14, 15]): where the former ones provide a series of nested clusters as results, whereas the latter generally find partitions with no nesting. However these methods require parameters (i.e. number of expected clusters for k-Means and map topology for SOM) and alterations in these can have significant impact on the results.

Another important aspect in microarray data analysis is classification. Here the samples are to be assigned to groups based on their expression profiles. To this end classification methods aim at identifying a small set of genes that can reliably be used as a predictor and furthermore can be generalized beyond the sample analyzed. Examples of methods applied for classification of microarray data include: discriminant analysis [16], tree-based algorithms (classification and regression trees [17]), nearest-neighbour [18], neural networks [19, 20] and support vector machines (SVM, [21]).

Apart from clustering and classification techniques, microarray data has been successfully analyzed with projection methods. In general these methods aim to find ’summary variables’ which can be used to display high-dimensional data in a small number of dimensions. Prominent examples of which are principal component analysis (PCA, [22]), multidimensional scaling (MDS, [23]) and Correspondence Analysis (CA, [24, 25]). PCA and MDS allow to project either the rows or the columns of the data matrix in a lower dimensional subspace, whereas CA allows for visualization of rows and columns in the same plot. For more details on CA please refer to 2.2.1.

1.2 Ontology

Ontology originates from the field of philosophy, as being the study of what is, of the kinds of objects, their structures, properties and relations in every area of reality. In other words ontology focuses on the nature of organisation of reality. This is often contrasted with Epistemology as a branch of philosophy which analyzes the nature and source of knowledge [26]. Ontology as a field of research can be traced back to Aristotle who defined ontology as the science of being as such. Aristotle developed 10 (top-level) ’Categories’ (namely Substance, Quantity, Quality, Relation, Place, Time, Situation, Condition, Action and Affection) which, from his point of view, are sufficient to describe anything that can be known about something.

Nowadays, especially in the context of computer science, an ontology is perceived in a narrower sense: commonly it refers to a working model of entities and their interactions in particular domain of knowledge or practise, such as molecular biology or bioinformatics.

While ontologies are already extensively used in areas like artificial intelligence research, lately their usefulness is being recognized by other fields as well. One of the first computational ontology to be constructed was Cyc, which was developed to describe ’everyday common-sense knowledge’ [27]. Another prominent example are the efforts to create a ’semantic web’. That is, to make the information that is captured in the internet accessible for computers and thereby significantly increasing the efficiency of software agents that are aimed at the retrieval of information. This example already demonstrates one of the strengths of an ontology: it renders the therein captured knowledge accessible to humans as well as to computers. Moreover it allows for consistent terminology in a specific field, to ensure that e.g. different laboratories use the same concepts with the same meaning to capture knowledge. The success of an ontology heavily depends on the involvement of the associated community, since it has to be agreed upon and ultimately be used by the community.

Ontologies are being used in biology to an ever increasing extent. Nowadays, the most prominent example of which is the gene product centered Gene Ontology (GO), which is discussed in detail in 2.1.2. Besides this, numerous other ontologies are created in the biomedical field. Examples of which include, the Sequence Ontology (capturing information on biological sequences [28]), STAR/mmCIF (structure of macromolecules [29]), MGED Ontology (to capture information about a microarray-experiment [30]) and Galen (clinical information, including human anatomy [31]). Finally the ’open medical ontologies’ (OBO, ’http://obo.sourceforge.net’), is an umbrella address for bio-medical ontologies that all commit to some shared requirements for the ontology.

The aim of an ontology is to describe what entities and interactions are relevant in a certain field of knowledge. In context of knowledge sharing, Gruber defines an ontology as ’the specification of conceptualisations, used to help programs and humans share knowledge’ [32].

The main components of an ontology are concepts, relations, instances and axioms. A concept represents a set of entities within a domain, for instance ’Enzyme’ is a concept in the domain of molecular biology. Concepts can be divided into two classes:

1. **primitive concepts:** these have only necessary conditions describing the properties of the concept. All entities belonging to this concept share these properties, but not all entities possessing these properties belong to this concept. E.g. proteins are composed of individual amino acids connected via peptide bonds, whereas not every molecule composed of amino acids will qualify as a protein (e.g. molecules of less than 100 amino acids are referred to as peptides).
2. **defined concepts:** are those where the conditions are not only necessary but also sufficient. As with primitive concepts all entities of this concept share the properties, moreover if an entity has the defined properties it belongs to this concept. E.g. eukaryotic cells, all eukaryotic cells have a nucleus and if in turn a cell has a nucleus, it is an eukaryotic cell.

Relations describing the potential interactions between the concepts can be categorized in two broad groups:

1. **Taxonomies** organize concepts into sub- / super-concept tree structures. The most frequently used forms are:

  - Specialisation relationships commonly known as ’is a kind of’ relationship. E.g. a *Ligase* is a kind of an *Enzyme* which is a kind of a *Protein*.
  - Partitive relationships describe concepts which are part of other concepts. E.g. the *Centromere* is part of the *Chromosome* which is part of the *Nucleus*.

2. **Associative** relationships which relate concepts across tree-structures. Some common examples are:
   - Nominative relationships describe the names of concepts
   - Locative relationships describe the location of the concept with respect to another
   - Associative relationships that represent, for example, the functions, processes a concept has or is involved in, as well as other properties of the concept

Finally, axioms can be used to put constrains on classes or instances, for instance chains of less than 100 amino acids connected via peptide bonds are categorized as peptides, rather than proteins.

1.3 Current methods to analyze microarray data in context of annotation data

In context of microarray data one can roughly divide the available annotation data into concepts describing genes and their functionality and those used for the description of the experimental setting. It is only in the recent years that experimental annotations are being recognized as essential information for the comprehensive description of a microarray study. This is documented by the fact that the major microarray data repositories (such as GEO and ArrayExpress) require the submission of experimental descriptions along with each microarray data set. Nevertheless the format, in which this information is submitted and stored, varies to a great extend between the repositories. In case of GEO, this information was originally stored in the SOFT (Simple Omnibus Format in Text) format, which allows the use of free text. Currently the experimental annotations are stored using a XML-based schema, which is named MINiML (MIAME Notation in Markup Language). Here certain tag-value pairs are defined, some of which are mandatory to be uploaded along with a data set in order to be in compliance with the ’Minimal Information about Microarray Experiments’ (MIAME) -standard. However only for some of these values constraints are provided, e.g. for the tag <Web-Link> a ’valid URL’ has to be provided. For the majority of tags however, no constraints are defined, rendering the annotations in a freetext-like format and thus resulting in large problems when computer-based extraction and analysis of the experimental annotations is desired.

This problem has been recognized by the Microarray Gene Expression Data (MGED) Society and lead to the development of the MGED Ontology (MO). Up to January 2006 experimental annotations submitted to ArrayExpress were stored, in compliance with MIAME standards, in the Microarray Gene Expression Model (MAGE-OM) which is based on the related XML format MAGE-ML. This format provides a common syntax for the exchange of data but lacks important ontological characteristics such as an explicit terminology (i.e. controlled vocabulary) with unambiguous definitions for each term as well as the semantic relationships between terms. Storing information about the experimental setup of a microarray experiment in a computer-accessible format is the first step to integrate this data into the statistical analysis. Up to now, there are no published methods available which fully exploit this source of information. Nevertheless, annotations describing the experimental settings can be stored by our ’in-house-software’ M-CHiPS. Even though, not being structured as an ontology, the information is captured by controlled vocabularies rendering it accessible for statistical analysis [33].

While experimental annotations are inevitable to, for instance, reliably identify technical artifacts in the data, the majority of available annotation data is focused on gene (protein)-properties. This data describe a huge variety of gene-centered features including, sequence properties, chromosomal location, homology, transcription factor binding sites, methylation status, relevance in disease, pathway-membership and functional properties - just to name a few. Tools that focus on analysis of individual properties, like sequence information (e.g. prediction of transcription factor binding sites), identification of homologies or prediction of 3-dimensional protein structures are readily available. Nevertheless in the recent years it has become more and more apparent, especially due to the success of high-throughput technologies such as microarrays, that the functional interpretation of this data is a major bottleneck. Inevitably, the ultimate goal of each experiment is to generate data in order to validate predefined hypothesis or to develop new ones. In context of microarrays, a wealth of methods is available to extract significantly regulated genes, which commonly are then represented in spreadsheet-like lists. Based on these, functional properties of the genes have to be gathered (if not already provided in the local database) and common functional properties have to be identified. While this might prove feasible for smaller number of genes in few experimental conditions, the list of regulated genes in a typical microarray experiment can easily be comprised of up to several hundreds of genes. These numbers render a non-computer based analysis not only tedious and time-consuming, but also prone to errors.

This bottleneck in data analysis has lead to the development of various tools to analyze microarray data in context of gene annotations. A large number of available methods make use of the Gene Ontology project (explained in detail in 2.1) - a listing of available tools can be found at [34]. The vast majority of these tools focuses on the identification of significantly over- or under-represented annotation terms from a set of regulated genes. Thus the analysis is a two step process, in which initially a set of regulated genes has to be calculated and subsequently submitted to a tool that calculates a list of significant annotations. The variety of methods for this task ranges from systems like DAVID [35], which is based on a database integrating annotations from Gene Ontology, KEGG and PFAM protein domains to, web-based tools like FatiGO [36]. DAVID however is restricted to a specific level of the ontology and the resulting annotations are presented in a list labeled with the percentages of regulated genes. Tools like FatiGO on the other hand are easy to use and do not require any software download and installation but have other limitations. In case of FatiGO the analysis is restricted to one of the main categories in the Gene Ontology and the resulting terms are only represented as lists. Moreover only two sets of genes can be compared per analysis.

In summary, none of the currently available tools allows for a highly integrated visualization of the data, such that associations among genes, experiments and gene annotations can be deduced from a single plot.

2 Integration of gene annotation data

The term ’gene annotation data’ can be perceived as any information that characterizes a property of a gene, some examples of which are provided in Table 2.1. Even though not representing an exhaustive overview, it demonstrates the wealth of available gene annotation data along with the heterogeneity of sources.

Despite the amount of available annotation data, not all of it is of immediate relevance for the functional interpretation of microarray data. Here the researcher commonly ends up with a list of regulated/interesting genes, which can be comprised of several hundreds of genes. The subsequent step in the analysis is time-consuming and labour-intensive: This list of regulated genes has to be associated to functional annotation data (e.g. a pathway or a biological objective like ’mitosis’), in order to deduce a biological conclusion.

To this end heterogeneous annotation data sources are combined by querying the corresponding databases, sometimes even on a gene-by-gene basis. Commonly this information is added to the extracted list of genes in a spreadsheet like format as well. In this approach the annotation data is entered as free text, not making use of any provided IDs or a controlled vocabulary. Moreover the use of free text increases the probability of errors like misspelling and thus could lead to incorrect analysis. The extraction of properties that are common to a set of genes is consequently done by eye, leaving this not only a cumbersome task but also prone to errors.

Having clarified the problem, two central requirements for the source of annotations arise: First of all, the information should allow to draw relevant biological conclusions from microarray data sets. Here annotations on aliases, clone information or sequence information are not the optimal choice. In this context, information on functionality, pathway membership or transcription factor binding sites are more promising. As a second requirement for annotation data, the data structure should allow for machine-processing of the annotations as well as human-readability. Only two larger sets of the annotation sources listed in Table 2.1 meet this second criterion, namely the Gene Ontology and the KEGG database. KEGG is a powerful source to associate subsets of genes or even complete experimental conditions to pathways, but is not optimal when trying to increase the level of abstraction. This, however, often is inevitable in order to identify ’general themes’ that account for the observed differences between experimental conditions.

The structure of the Gene Ontology, on the other hand, allows to analyze annotation data at any level of abstraction: from as general concepts as ’*signal transducer activity*’ down to specific annotations such as ’*thyrotropin releasing hormone and secretagogue-like receptors*

Gene annotation data on:	Reference to Source / Database
Sequence (genomic,mRNA)	EMBL [37], Genbank [38], DDBJ [39], dbEST [40], dbSTS [41], RefSeq [42], Ensembl [43]
Localization (cellular / tissue / organ - level)	GO [44], Human Protein Atlas [45]
Expression patterns	GEO [46], ArrayExpress [47]
Functionality	GO [44], BRENDA [48]
Transcription factor binding sites	TRANSFAC [49], DBTBS [50]
Interaction partners (genomic & proteomic level)	CTD [51], BIND [52], STRING [53]
Proteins	SWISSPROT [54], PROSITE [55]
Post-translational modifications	DSDBASE [56], PhosphoBase [57], RESID [58]
Pathway-membership	KEGG [59], EMP [60]
Relevance as disease target	GeneDis [61], KMDB [62]
Protein 3-d structure	PDB [63], SWISS-Model [64], ModBase [65]
CpG-island & status of methylation	MethDB [66, 67]

Table 2.1: Various types of gene-annotation data and their sources. This non-comprehensive listing on gene-annotation data demonstrates not only the wealth of information that being available, but also demonstrate the heterogeneity in data structures. Not all kinds of annotations are, however, of immediate relevance for the functional interpretation of high-throughput data set like microarray data.

Gene Ontology Overview	
total annotation terms	19977
biological process	10745
cellular component	1799
molecular function	7433
different organisms	30
overall annotated gene products	≈ 1 600 000
annotated by electronic means	≈1 450 000
curated annotations	≈ 150 000

Table 2.2: Overview of the basic statistics from the Gene Ontology. The provided numbers represent the status of the Gene Ontology Project in August 2006.

activity'. Being structured as an ontology it allows computer-based processing of the captured information, for more details on the structure and their benefits please refer to 2.1.2. Moreover in the recent years GO has developed to a *de facto* standard for the annotation of gene products. This is not only demonstrated by the rapid increase of publications making use of GO (more than 900 as of August 2006) or tools that have been developed to exploit the ontology (for a selection of these please refer to [34]), but also by the tremendous growth of available annotation terms as well as annotated gene products for a wide range of species. A snapshot (as of August 2006) of these numbers is provided in Table 2.2.

Up to now the GO has been mainly used in the analysis of microarray data, but due to the ever growing popularity and information content, it is also being used in a broader context, such as gene function prediction [68] or construction and analysis of cellular pathways. Moreover the Gene Ontology is currently used in computer science to, for instance, test description logic to build consistent ontologies [69, 70, 71].

2.1 Gene Ontology

2.1.1 Overview

The Gene Ontology was initiated in 1998 by members of the labs associated with the three model organism databases, namely the *Saccharomyces* Genome Database (SGD), the *Drosophila* genome database Flybase and the Mouse Genome Informatics Databases (MGD). The GO Consortium is comprised of members of these labs and was joined in 2000 by the *Arabidopsis* Information Resource (TAIR) and the *Caenorhabditis elegans* group, completing the group of fully sequenced model organisms.

With complete genomic sequences at hand it is very appealing to compare the genomes of the different organisms. Comparative analysis between model organisms shows that especially for genes involved in biological core processes such as DNA replication or translation, a high degree of conservation can be expected [72]. One opportunity that arises from this finding, is the possibility to transfer biological annotations between organisms based on, for instance, sequence homology. While there are methods for sequence comparison available [73, 74], the transfer of the annotation data poses a problem due to the incompatibility of annotations terms between the different databases. This incompatibility arises not only from the different terms being used, but also from the differing definition of the terms (if some unambiguous definitions are provided at all).

One major goal for the GO Consortium therefore was ’to develop cross-species biological vocabularies that are used by multiple databases to annotate genes and gene products in a consistent way’ [75]. One key prerequisite that will allow interoperability between databases (or research groups) is the use of controlled vocabulary. This means that the terms allowed to describe properties of gene products are restricted to those concepts available in the ontology. Moreover to ensure a consistent usage of these concepts between research-groups or even different fields of research an unambiguous definition has to be provided for each concept. An additional aim for GO is to keep the concepts (terms) of the ontology as organism-unspecific as possible, allowing for an applicability across species .

2.1.2 The structure of GO

The GO originally has been divided into three main ontologies focusing on different properties of the gene products:

1. **Molecular Function:** describes what a gene product does at a biochemical level, without specifying localisation or the broader context of the function. E.g. ’*ligase*’ or ’*sugar-transporter*’ are valid concepts of Molecular Function ontology.
2. **Biological Process:** describes the broader biological context the individual gene products contribute to. A biological process could be perceived as the results of an (ordered) sequence of molecular functions. E.g. ’*mitosis*’ or ’*cell differentiation*’ are valid concepts of Biological Process ontology.
3. **Cellular Component:** describes the cellular localisation of gene products. This ontology encompasses concepts like ’*membrane*’ as well as ’*chaperone complex*’.

Although an ontology tries to capture all existing information in a certain field of research, GO limits itself to these three aspects of gene properties. According to the consortium this subset (of ontologies) has been chosen to ’initially focus on three precise terms that are of immediate and exceptional utility and that span our various organismal domains’ [75].

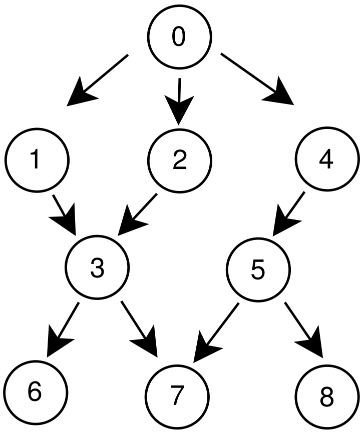


Figure 2.1: Schematic example of a directed acyclic graph (DAG). Every vertex (depicted as circles) is connected to at least one other vertex, without any directed cycles. In contrast to hierarchies a DAG allows for each vertex having more than one parental vertex, here vertex three and seven are examples of this. The vertex having no parental vertex is called the root of the graph (here vertex 0). Those vertices having no children vertices are called leaves of the graph (here vertices 6-8).

The concepts in the ontology are structured as a directed acyclic graph (DAG), which is a directed graph with no directed circles. In other words, for every vertex V there is no non-empty directed path starting and ending in V. In contrast to hierarchies DAGs allow for each vertex having multiple parents, an example of which is provided in Figure 2.1.

In an ontology not only the concepts but also the relations connecting them are defined. Originally two types of relations have been defined in the GO, namely the 'is a' and 'part of' relations. The former is applicable if a concept is an instance of its father (e.g., 'fructose-transporter-activity' is a 'transporter-activity'), the latter is appropriate if the child represents a component of the father (e.g. 'mitochondrium' is part of 'cytoplasm').

Each term (concept) in the ontology has a unique identifier, by which it is accessible. This ID consists of 'GO:' followed by seven digits, e.g. 'GO:0006260' is the ID for the concept of 'DNA replication'. This accessibility by unique IDs fulfills a further requirement for interoperability between databases and research groups.

The complete ontology is available for download from [76]. Just recently a subset of the GO called 'GO slim' became available for download. These terms are selected to focus on more general biological processes, omitting concepts describing very specific and detailed aspects of gene products.

2.1.3 Annotation of gene products

The annotation of gene products with concepts is independent from the development and definition of the concepts of the ontology. A central requirement to consistently annotate gene

Evidence Code	meaning
IMP	inferred from mutant phenotype
IGI	inferred from genetic interaction
IPI	inferred from physical interaction
ISS	inferred from sequence similarity
IDA	inferred from direct assay
IEP	inferred from expression pattern
IEA	inferred from electronic annotation
TAS	traceable author statement
NAS	non-traceable author statement
ND	no biological data available
RCA	inferred from reviewed computational analysis
IC	inferred by curator

Table 2.3: Evidence codes for annotation of gene products. Each annotation of a gene product has to be provided with an evidence code, corresponding to the source of information the annotation is based on. All available codes and their meaning are listed here.

products are the definitions provided with each concept. Since the annotation of gene products is done and maintained to a large extend by the corresponding databases (e.g. SGD for *S. cerevisiae*), the strict application of definitions, which sometimes is referred to as ’commitment to the ontology’, is crucial.

For each annotation several pieces of information are mandatory: firstly a reference to the source (publication, database or computational analysis), secondly information on what kind of evidence the annotation is based has to be provided. This is done by choosing the appropriate term from a small controlled vocabulary (Table 2.3). Detailed definitions of these codes are given in the GO website [77] to ensure consistent usage of the codes.

The annotation of a gene product to one of the three ontologies is independent of its annotation to the other ontologies. Nevertheless, as it already has been noticed by the GO Consortium [75] that relationships between the ontologies exist, especially between ’Molecular Function’ and ’Biological Process’. Recently tools have been developed that exploit these relationships and thereby assist the annotation process [78].

2.1.4 Exploiting the ’true path rule’ - i.e. how to associate genes to GO terms

One property of the GO is the so called ’true path rule’. This states that the parental vertices of each individual vertex in the ontology have to be true for this vertex as well [75], i.e. all parental concepts of any concept have to describe valid properties of the associated gene product(s). In the annotation process the curator tries to capture the complete knowledge that is available for any given gene product, such that gene products are annotated with the most specific concept possible. Even though this ensures that all information about the gene product is captured, in many cases this leads to sparsely ’populated’ concepts. E.g. ’*glutamate-cysteine ligase activity*’ is only associated to a single gene product in *S. cerevisiae* and thereby rendering the term next to unusable for functional interpretation in a larger context than a single gene.

The flat files distributed by the GO consortium contain only these ’most-specific gene annotations’. Amongst other information an association between a GO-accession ID (e.g. GO:006536, for ’glutamate metabolism’) and an appropriate organism specific gene-identifier are provided. By using only this most specific association one would loose valuable information (i.e. all the parental concepts being true for the associated gene product). Therefore I decided to ’blow up’ these flat files by annotating each gene with each parental term of its initial concept as well. These father-child associations can be extracted directly by the SQL query provided in 7.1. The increased amount of storage/memory used by these ’blown up’ annotations don’t pose a problem. As a result the concept of ’*glutamate metabolism*’ is associated with 40 different gene products, compared to a single gene in the provided flat-file.

These expanded annotations exhibit improvements over the distributed flat files in two main aspects: Firstly, they increase the number of associated gene products for many of the concepts and thereby providing a better statistical basis for any conclusions drawn from the concepts. Moreover only by these expanded associations more general terms become applicable at all (as shown for ’*glutamate metabolism*’). Secondly it allows to perform an analysis at any given level of abstraction or specificity (i.e. distance from root concept in the ontology) without loosing any information (i.e. associated gene products).

2.2 Integration of gene annotation data in Correspondence Analysis

2.2.1 Correspondence Analysis

Correspondence Analysis (CA) is a exploratory method to analyze two-way as well as multi-way tables. CA can project the information captured in a data matrix into a lower dimensional subspace, commonly 2- or 3-dimensional, with minimal loss of information. In contrast to Principal Component Analysis (PCA), CA can represent row and column variables in the same space, allowing to identify associations not only among rows or columns, but also between them.

The following paragraph gives a concise overview of the calculations performed in a CA: Let N denote the original data matrix, being composed of I rows (here genes) and J columns (here hybridizations). The grand total of N is represented by n. The correspondence matrix P is derived from dividing the elements of N by the grand total: $p_{ij} = n_{ij}/n$. The row masses of P are defined as $r_i = \sum_j p_{ij},\ i = 1,...,I$ and analogously the column masses are defined as $c_j = \sum_i p_{ij},\ j = 1,...,J$. Based on this matrix S is computed by $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$. S is submitted to singular value decomposition: $S = U\Lambda V^T$. Λ is a diagonal matrix with its diagonal elements being the singular values of S. These are sorted in descending order and denoted by λ_k. The resulting row-coordinates in the new subspace are calculated by: $f_{ik} = \lambda_k u_{ik}/\sqrt{r_i}$. The column coordinate(s) are calculated by: $g_{jk} = \lambda_k v_{jk}/\sqrt{c_j}$. More details on CA are provided in [79, 80].

2.2.2 Interpretation of Correspondence Analysis plots

Based on the original data matrix CA calculates the χ^2- statistics as a measure of association between rows and columns, whereas higher values indicate an existing association. In the biplot the points are depicted as such that the sum of their distance to the centroid equals the total interia of the matrix, which in turn is the χ^2 divided by the grand total of the matrix (n). The greater the distance between a point and the centroid, the higher is its row contribution to the χ^2- statistic, which means that the larger the distance, the larger is the difference between that point's profile and the average profile. In turn, if two rows exhibit a similar profile, they will be plotted in close vicinity to each other.

Along with points representing the rows of the data matrix, CA visualizes the columns using the same criterion. Commonly the row points are plotted along with the columns in the same plot, i.e. a symmetric map where rows and columns are scaled in principal coordinates, as opposed to an asymmetric map where rows are scaled in principal coordinates and columns in standard coordinates (or *vice versa*). This implies that in case of symmetric maps the distance between row and column points as such, can not be used as a measure of association. To facilitate interpretation of association among rows (genes) and columns (hybridizations), the standard coordinates for each hybridization (or experimental condition) are plotted. These coordinates equal a row profile with all its mass concentrated in a single column (i.e. hybridization). These are artificial points that exhibit the strongest possible association of a row to column, such that a combined display of these artificial points with 'real' data would result in a shrinkage of the real data to small area around the centroid. To circumvent this, lines are drawn from the centroid to these standard coordinates which indicate the direction and thus can be used as a guide lines to evaluate associations among rows and columns.

Figure 2.2 gives a CA plot of an artificial data set to demonstrate the interpretation. Here genes are depicted as black dots, hybridizations as colored squares:

- Genes and hybridizations which have a similar profile are plotted in close vicinity to each other. E.g. the three genes on the upper left hand side of the plot are both upregulated in 'blue'.

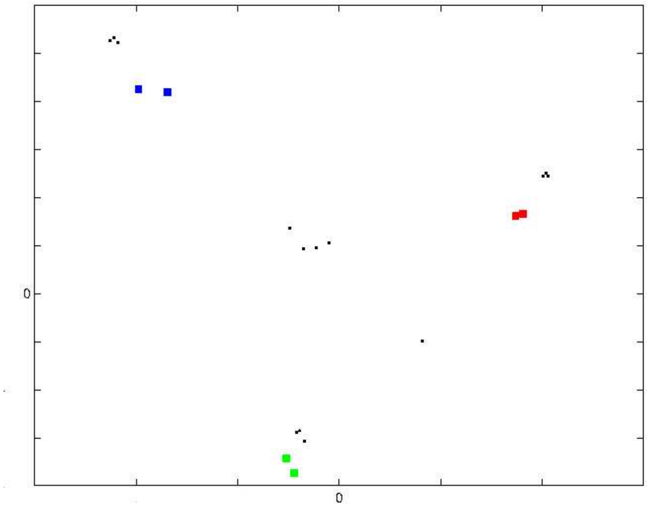


Figure 2.2: CA plot of an artificial data set. Genes are depicted as black dots, hybridizations as colored squares. Genes having similar expression profiles are plotted in close vicinity to each other. If no significant change in the intensities is present, the genes are plotted near the center of the map (i.e. the centroid). Genes being upregulated in the 'blue' condition are plotted in the direction of the blue squares. Genes being downregulated in blue (and up in red and green) they are plotted on the opposite site of the centroid.

- Genes being particularly upregulated in one experimental conditions are plotted in the direction of this condition.
- Genes being particularly down-regulated in an experimental condition are plotted on the opposite site of the centroid of that experimental condition. The gene in the lower right corner in the plot is down-regulated in blue and is thus plotted on the opposite site of the centroid as the 'blue' condition.
- Genes which do not exhibit any significant change in their expression intensities over the experimental conditions are plotted near the center of plot (i.e. the centroid) - regardless of the intensity level. In turn, this means that genes that show a significant change in their expression profile are plotted at larger distances to the centroid, such that the genes being plotted at the margins are the most differential ones.

One should keep in mind, however, that a projection of high dimensional data in a two-dimensional subspace is always accompanied with loss of information (except for cases of $J <= 3$, where all of the variance can be captured by two-dimensional maps). Thus it is advisable to check the percentage of variance that is explained by the first two dimensions. Moreover, since the percentages only are summary values for each dimension, the quality of display for individual data points should be assessed as well the before the selection of genes as candidate genes (see 2.2.3).

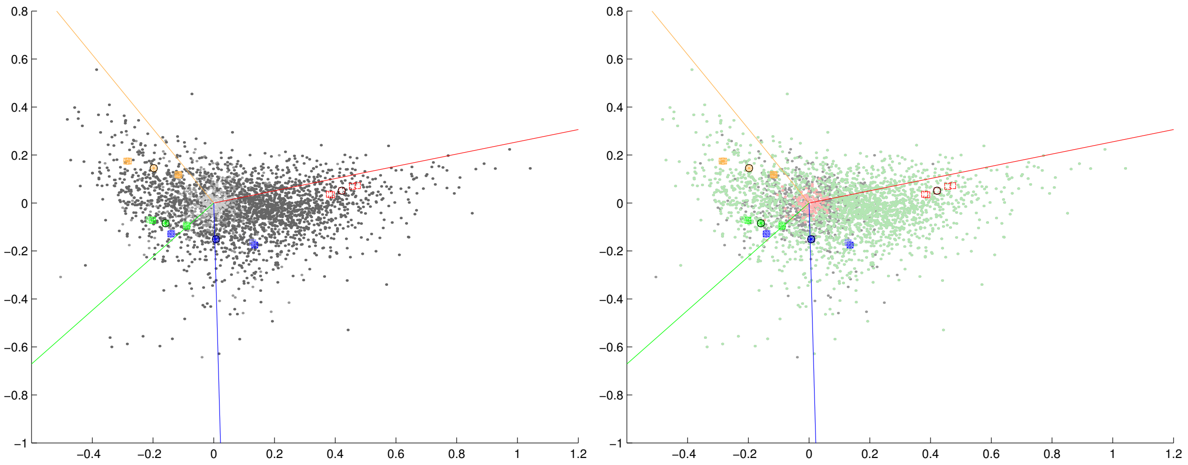
2.2.3 Visualizing the quality of display in CA

In order to analyze high-dimensional data (such as microarray data) CA projects the data in a lower dimensional plane such that the loss of information is minimal. For a more detailed description on CA please refer to 2.2.1. This projection implies that some data points are displayed well, whilst others are not. The overall information content of a particular two dimensional plot can be assessed by the percentage of the total variance captured by the two first principal axes. This does not provide any information on the quality of display for each individual data point. This, however, is valuable information, especially when staring to build biological hypothesis on the positioning of individual genes. Here one should ensure that the gene's position in the two dimensional plot is capturing a sufficient amount of that gene's total variance. To this end I implemented the display of this information in CA, by calculating the percentage of variance that is explained by the first two principal axis for each data point (i.e. gene).

The different levels of quality are categorized into three groups: genes of high quality (more than 80% of their total variance explained), genes of low quality (less than 30% of their total variance explained) and genes of medium display quality in between the previously defined borders. The categorization is color-coded in the analysis plot, providing a grey-scale and color mode. Whereas the grey-scale is useful when generating overview plots with large numbers of displayed genes, while it is rather unfeasible to distinguish shades of grey on smaller numbers of genes and thus the color-mode is preferable.

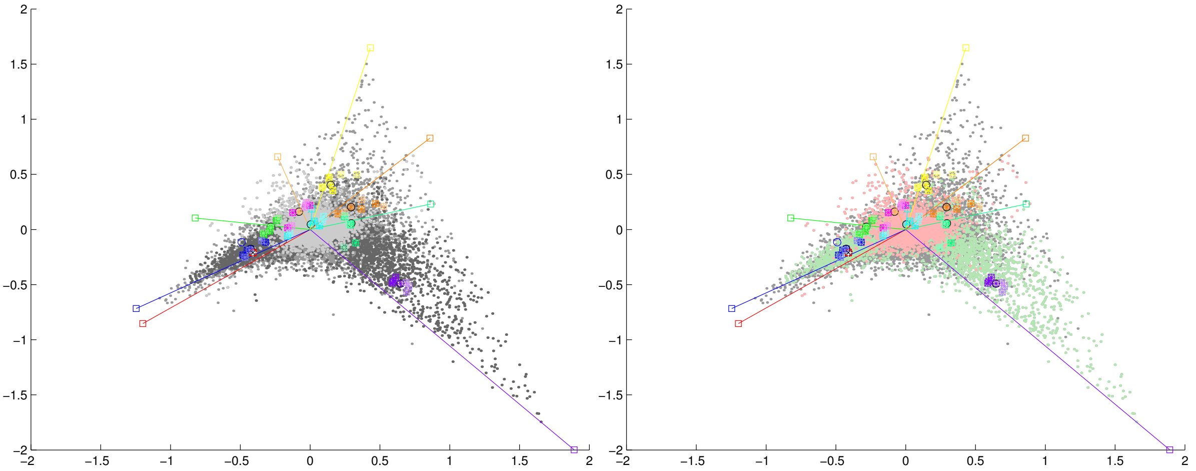
Figure 2.3 shows the dataset described in 2.4.1 additionally displaying the gene-by-gene quality information. In this setting the majority of annotations being plotted near the margins of the plot is of high quality. This is not surprising since more than 90% of the total variance in the dataset is explained by the first two axis. Some spots of minor quality of display can be found near the centroid of the map.

The importance of displaying this information becomes more apparent, when data sets with larger number of genes and, more importantly, larger numbers of experimental conditions are analyzed. Figure 2.4 provides an example of a microarray study analyzing the developmental process in *Drosophila melanogaster* [81]. In this analysis 14551 genes are displayed with 8 corresponding experimental conditions (i.e. developmental stages of Drosophila). In this more complex setting two important points become apparent: First, compared to the previously shown data set, the overall number of the low-quality genes has increased dramatically. Secondly the position of these low-quality genes is not restricted to a small area around the centroid, but these can also be found near the margins of the plot and thus being potential candidates for further analysis. These two points clearly demonstrate that the necessity to assess the quality of display for each gene individually, before subsequent analysis or even the development of a biological hypothesis based on the gene's position in the CA plot.



(a) (b)

Figure 2.3: Quality of display for individual genes. CA can project the data points on a two dimensional plane such that the loss of information is minimal, nevertheless it is inevitable to have information on the quality of display for individual data points. To this end the genes are categorized into three groups: genes of high quality (more than 80% of total variance explained), genes of low quality (less than 30% of the total variance explained) and genes of medium display quality in between the previously defined borders. In Fig.(a) high quality is depicted by black, medium by grey and low quality by light grey. In Fig. (b) the categories are depicted by different colors: high quality genes are represented by green points, medium by grey and low quality by red points. The data-set shown here is described in 2.4.1.



(a)

(b)

Figure 2.4: Quality of display - *Drosophila melanogaster* data-set. In these Figures the quality of display for individual genes is coded in grey-scale (a) and three different colors (b). In Fig.(a) high quality is depicted by black, medium by grey and low quality by light grey. In Fig. (b) the categories are depicted by different colors: high quality genes are represented by green points, medium by grey and low quality by red points. The data-set analyzes the developmental stages of Drosophila melanogaster. The developmental process is reflected by the positioning of the experimental conditions in a u-shape, which the embryonic stage being placed on the left site of the centroid (i.e. colored in red) and the adult fly on the opposite site (i.e. colored in purple).

2.2.4 Gene annotation data in CA

In CA it is common practise to diminish the influence of identified outliers but plotting them without mass (i.e. as supplementary points). Besides this application the addition of supplementary variables to the analysis [82, 83] is a well known method to enhance the interpretability of a CA plot. Moreover [84, 85, 86] have already shown the applicability of CA to analyze Boolean matrices.

2.2.4.1 Boolean implementation

The association of gene products with annotations can be considered a Boolean variable: annotations for a specific gene product are either available or not. Accordingly, in this implementation, each annotation is represented by a 0/1-vector (the length of which is given by the number of features on the array), being 1 for each gene that is associated to the particular annotation (refer to Table 3.1 for an example of the encoding). These annotations vectors are added as supplementary columns to the data-matrix. Since annotations are implemented supplementary (here as supplementary columns), they do not contribute to the computation of the principal axes (i.e. are plotted without mass [83, 87].

In the resulting plot (Fig. 2.5) the annotations (depicted as cyan squares) are predominant over all other aspects of the plot. In case of Boolean vectors the individual elements can only be one of two values. This results in maximal relative changes, which are displayed by CA. Thus the annotations will be positioned in the outer margins of the plot. Compared to these 'artifical' Boolean vectors the relative changes derived from the 'natural' transcription intensities are rather small and thus the corresponding points are concentrated in the center of the plot. As can be seen in Fig. 2.5 the genes (grey dots) as well as the experimental conditions (colored full squares) occupy only a rather small area in the center of the plot.

To assess the usefulness of integration of GO annotations, the displayed annotations are analyzed in context of the experimental setting (for an in depth analysis refer to 2.4.1). Here only two annotations (labeled with corresponding GO-IDs in 2.5) will be mentioned exemplarily : '*carbohydrate transporter activity*' (GO:0015144) and '*tricarboxylic acid cycle'* (GO:0006099). The first dimension of this CA plot distinguishes between experimental condition(s) in which *S. cerevisiae* was grown in media containing glucose (left-hand side of the plot) and *S. cerevisiae* grown in media with no glucose (right-hand side). Since the selected annotations (GO:0015144, GO:0006099) are projected on opposite sides of the centroid (along the first dimension), they can be considered as candidates for characterizing the difference between these two main groups of experimental settings (i.e. glucose vs. non-glucose).

With the annotation '*carbohydrate transporter activity*' being placed on the left-hand side of the plot, it is considered as being associated with conditions of *S. cerevisiae* being grown in the presence of glucose. This makes immediate sense, since glucose (if available) is used as the major carbon source for yeast and thus needs to be transported in the cell [88]. In absence

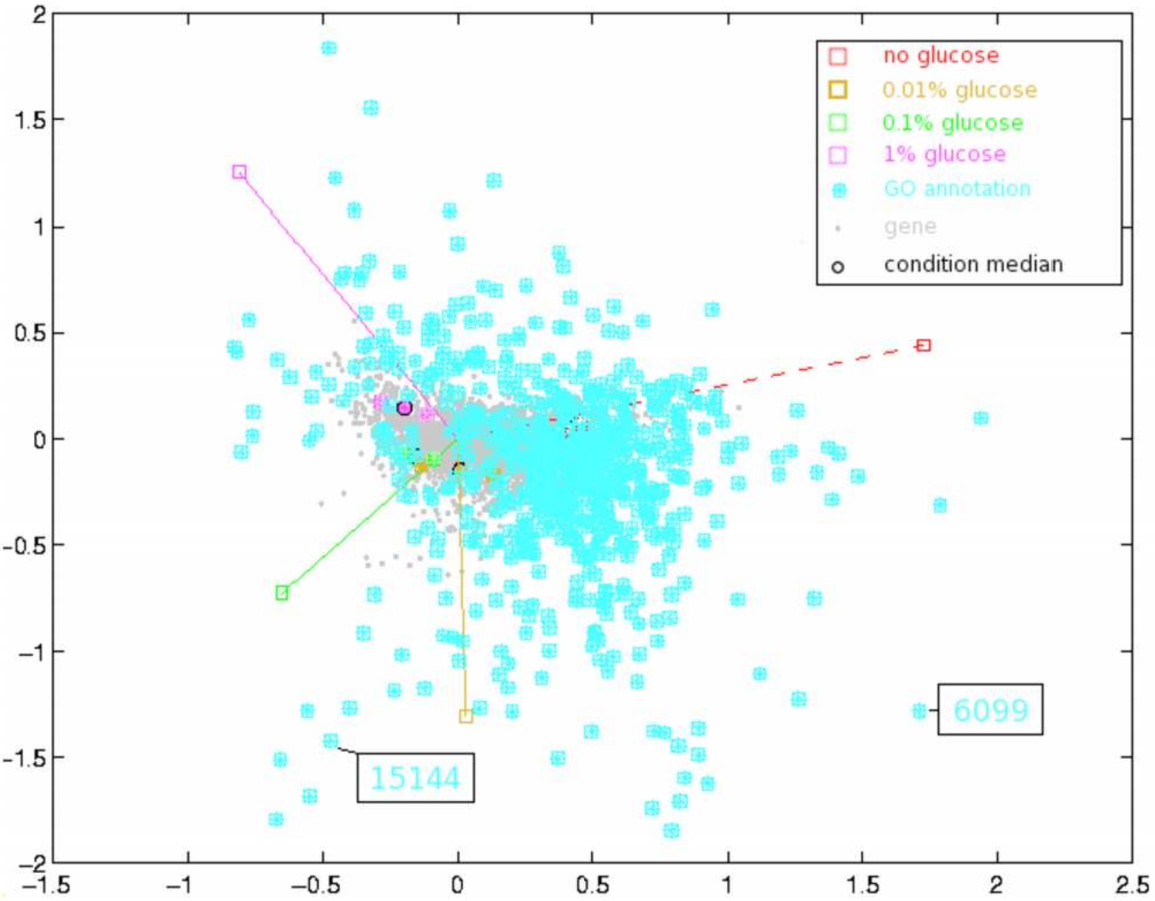


Figure 2.5: CA with GO annotations coded as Boolean columns. GO annotations are added to the data matrix as supplementary columns and are displayed as cyan squares. Genes are marked as grey dots, experiments as full squares, color coded according to the experimental conditions they belong to (see legend in upper right corners). Numbers adjacent to the annotations (if provided) reflect the GO-IDs (truncated of 'GO:' and leading zeros). Representations for each condition are depicted in standard coordinates as colored empty squares terminating the lines in the plot (e.g. the green square terminating the green line denotes the standard coordinate of the condition with 0.1% glucose).

of glucose, not only different carbon sources have to be used, but also a distinct set of pathways, like the tricarboxylic acid cycle (i.e. GO:0006099), is being activated for the utilization of non-glucose energy sources. Moreover it is known [89] that the TCA-cycle is repressed in the presence of glucose, being in concordance with the positioning of the corresponding annotation in Figure 2.5.

Even though the resulting biplots are not optimal in the display of the combined information (due to the predominance of the annotation in the plot), these results already demonstrate the usefulness of combining gene annotation and transcription data. In Figure 2.5 GO annotations plotted in the outer margins of the plot already provide a ’functional overview’ of the biological processes being relevant in the experimental settings under investigation.

2.2.4.2 Intensity based implementation

In contrast to the Boolean approach (2.2.4.1), in which the genes have been perceived as being ’attributes of the annotations’, annotations can be considered as ’attributes of the genes’ as well. Whilst in the Boolean approach the annotations are represented by a column-vector, here they are represented by a row-vector, having a length equivalent to the number of experimental conditions.

When representing annotations as row-vectors a Boolean encoding of the association between genes and GO annotations is not possible and since there commonly exist more than one gene per GO annotation, this information has to be integrated. Here Kurt Fellenberg developed the idea to calculate ’representatives’ for each available annotation, based on the annotated genes (i.e. their expression intensities). For each annotation, this representative expression profile is calculated by the row-wise sum of the annotated genes: let x_{ij} be the normalized expression intensities for gene $i = 1..n$, in condition $j = 1..m$; $A_k \subset \{1..n\}$ denote the set of genes annotated to GO term k. $\sum_{i_k \in A_k} x_{i_k j}$ is used as a representative gene profile for term k. These vectors are added as supplementary rows to the data-matrix (for an example of the encoding please refer to Table 3.1). As with supplementary columns, supplementary rows do not contribute to the computation of the principal axes [83]. Summation of the expression intensities of the annotated genes places the corresponding annotation in the center of gravity of these genes, e.g. the position of annotation GO:0006099 in Fig. 2.18 represents the center of gravity (centroid) of the annotated genes (tagged by red circles).

Figure 2.6 shows the same data set as in Figure 2.5 (Boolean implementation), besides that the representatives of the annotations are added as supplementary rows and are depicted as blue dots. In other words the positions of the genes, as well as the experimental conditions are identical to those in Fig. 2.5. One obvious difference between the plots is the spread of the annotations around the centroid. In the intensity-based approach the annotations are more tightly clustered around the centroid. This can be also seen on the different scale that is necessary to display the annotations: In the Boolean encoding 4 units of the first dimension

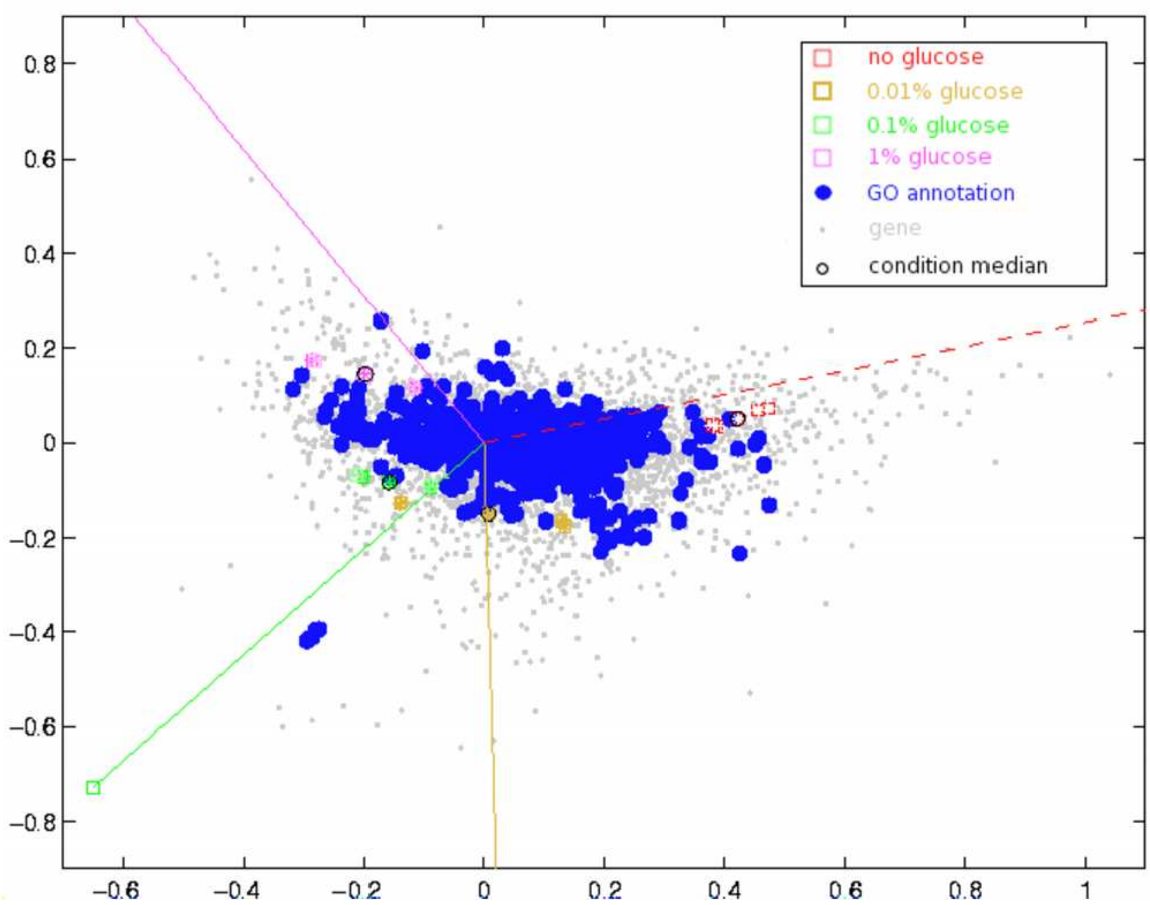


Figure 2.6: CA with GO annotations coded as 'sum of genes' rows. Annotations are added as supplementary rows (after summing up the profiles of the annotated genes) to the data matrix and are depicted as solid blue circles. Genes are marked as grey dots, experiments as full squares, color coded according to the experimental conditions they belong to (see legend in upper right corners). Representations for each condition are depicted in standard coordinates as colored empty squares terminating the lines in the plot (e.g. the green square terminating the green line denotes the standard coordinate of the condition with 0.1% glucose).

compared to ~1.3 units in the intensity-based coding have to be displayed to account for the spread of the annotations.

Moreover, this implementation allows to detect more characteristics in the data: For instance the clearly distinguishable cluster of annotations in the lower left corner of Fig. 2.6, is not distinguishable in the Boolean implementation (Fig.2.5) and thus probably would have been missed. The annotations comprised in the cluster are listed in Table 2.7 and mainly describe the biological process of transportation of sugar (i.e. glucose) into the cell. This clearly is a prerequisite for the utilization of glucose as an energy source and thus the cluster provides valuable information about the regulated processes. For an in-depth interpretation of the results refer to 2.4.1 or [90].

2.2.4.3 How to assign genes to a single, best-fitting GO term

The motivation to assign a gene to a single, unique GO term is mainly based on two objectives: The first one being, that this reduction could be used as an annotation filter, reducing the number of displayed annotations and thereby enhancing the interpretability of the CA-plot (for detailed discussion of annotation filters please refer to 2.3). The second one is based on the intention to analyze the annotations not only as supplementary information, but to give them mass in the analysis (i.e. calculate the principal axes based on the annotations rather than the genes). Here one has to ensure that each gene is accounted for the same number of times in the data-matrix (e.g. a doubling of the complete data-matrix would be valid), which holds not true in the provided files.

This is due to two reasons, one of which is the different number of annotations being initially provided per gene. Genes that have been studied thoroughly are associated to a higher number of annotations than others. The second reason being, that these annotations reside at varying levels of the ontology, which leads to large discrepancies when exploiting the 'true-path-rule' (see 2.1.4). For instance gene products being annotated with '*dihydropyridine-sensitive calcium channel activity*' (GO:0015270) would be annotated with nine parental terms (this annotation has a distance of 9 (vertices) from the root-vertex of the ontology), whereas gene products being annotated with '*chaperone regulator activity*' (GO:0030188) would be annotated to only 3 parental terms. Thus genes annotated with the former annotation would be represented three times more in the data matrix compared to those of the latter annotation. More generally speaking, the larger the distance of the annotation term to the root-vertex, the higher the number of representations of the annotated gene products. An immediate solution for the latter reason would be to only use the annotation-files as distributed by the GO consortium (http://www.geneontology.org/GO.current.annotations.shtml). In these, the gene products are only annotated with the most specific annotation term, reflecting the current state of knowledge.

This approach however, besides not fully solving the problem, poses another difficulty: as already mentioned in 2.1.4, the utilization of the most specific annotation terms available, will often result in terms being associated to only one or two genes. This low number is insufficient to base any statistically sound conclusions on it. Moreover since the concepts are commonly very specific - to capture all available information - they are not useful for analyzing the data from a more abstract point of view. However, this generalization is of particular importance for the functional interpretation of microarray data, since complex phenotypes often are not appropriately summarized by these specific concepts. Furthermore researches are commonly interested in broader biological process /objectives that are effected in the chosen experimental setting.

In order to represent each feature the same number of times two options arise: Firstly, one could determine the highest number of occurrences for a gene and then represent each gene as many times (this would be equivalent of multiplying the whole matrix with this factor). Secondly, Stefan Winter developed the idea to reduce the number of annotations for a each gene to a single one. I decided to follow the second approach mainly for three reasons. First, multiplying the data matrix with the largest factor, will blow up the matrix to an extent that makes it computationally unfeasible. Additionally a multiplication would not solve the problem how to distribute the genes to their associated annotation term. If a gene is annotated with a small number of concepts, the high number of repetitions of this gene would be concentrated to these few annotations and outweigh the influence of a gene being annotated with larger numbers of concepts. Finally, a reduction in the number of displayed annotations will increase the interpretability of the plot, by circumventing a massive overlay of annotation, as seen in Figure 2.6.

The assignment of gene products to single annotation term is a reduction of information. It is obvious that an arbitrary selection of unique assignments will most likely not reduce the information in a sensible way. Hence it would be preferable to base these unique assignment on a criterion which can identify annotations that are of potential importance in the selected experimental setting. The comparison of different types of measures that could be used to achieve this goal and the selection of the best suitable is discussed in depth in 2.3.2.

As a result use the mean of all pairwise Speaman's correlation coefficients of the annotated genes of a concept to calculate as a measure, based on which the unique assignment of genes to a concept can be done.

The most straightforward approach in selection of the best-fitting association is an iterative process, which is sketched in Figure 2.4 on the next page. In the initial step a filter score (in this case mean of pairwise correlation coefficients, 3rd column of Table 2.4) for all annotations is calculated. At this stage, a gene can be annotated with multiple GO terms, as can be seen in the first column of the same Table. Here gene 'a' is annotated with the IDs '23' , '54' and '346' (first, second and last row). In a second step the annotation with the highest filter score is selected, in this case annotation '23' and the gene products being annotated with this term are now uniquely associated to this annotation. In other words, the genes 'a b c d e' are subtracted from all other annotations. For example, in case of gene 'a', it would be subtracted from annotations '54' and '346'.

Gene products	Annotation - ID	Filter score
a b c d e	23	0.98
a c u z k	54	0.95
i o l j t	6423	0.91
s j f d b	53547	0.89
l o w m e	45375	0.86
f i a u c l k	346	0.85

Table 2.4: Unique assignment of gene products to the best fitting GO term. The left-hand column lists the gene products associated to a specific annotation-ID (listed in the middle column). The individual genes are here exemplarily denoted by lower-case characters, i.e. the genes ’a, b ,c, d and e’ are annotated with the annotation having ID 23 (first row). The last column represents the chosen filter-criterion, based on which the association is done.

The subsequent step is the recalculation of the filter score. This is necessary since the set of annotated genes has changed for some annotations due to the removal of the genes being associated to the annotation having the largest score value. For example in Table 2.4, after the removal of the first row, the score for all of the remaining annotations (except ’6423’) has to be recalculated, because each annotation’s set of genes has been altered (i.e. reduced). Again the annotation having highest score is selected and if the score is above a user defined threshold the gene products are uniquely associated to this annotation and the procedure starts from the beginning. If the score is below the threshold, the process terminates. The overall work flow is sketched in Figure 2.7.

This unique assignment has been applied to the *S. cerevisiae* dataset described in 2.4.1. Here only genes that have a maximal normalized intensity value equal to or larger than 10 in any experimental condition (intensity filter) were submitted to analysis, resulting in a total of 2898 genes. Based on this set of genes 1814 GO annotations are comprised of two or more genes. After the unique assignment of genes to GOs (with a threshold of 0.7) this set of annotations was reduced to 118 annotations with a total of 257 different genes - a comprehensive list of resulting annotations is given in 7.6.

Figure 2.8(a) shows the positions of the original set of annotations (depicted as grey crosses) as supplementary rows and the unique set (blue dots) with mass. Besides having been reduced in numbers and thus enhancing the clarity of the plot the average distance to the centroid of the unique annotations tends to be larger as compared to the original set. To demonstrate the change in the position between unique and original set Figure 2.8(b) shows only those annotations being present in both sets. For reasons of clarity the change of position is highlighted by a line connecting the annotation from both sets for a small subset of annotations.

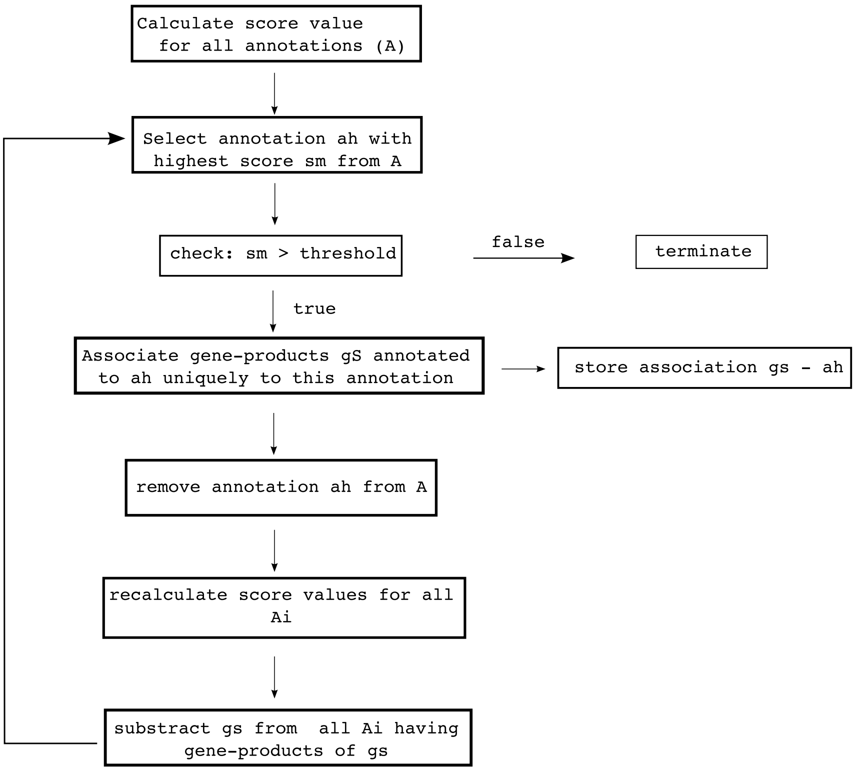
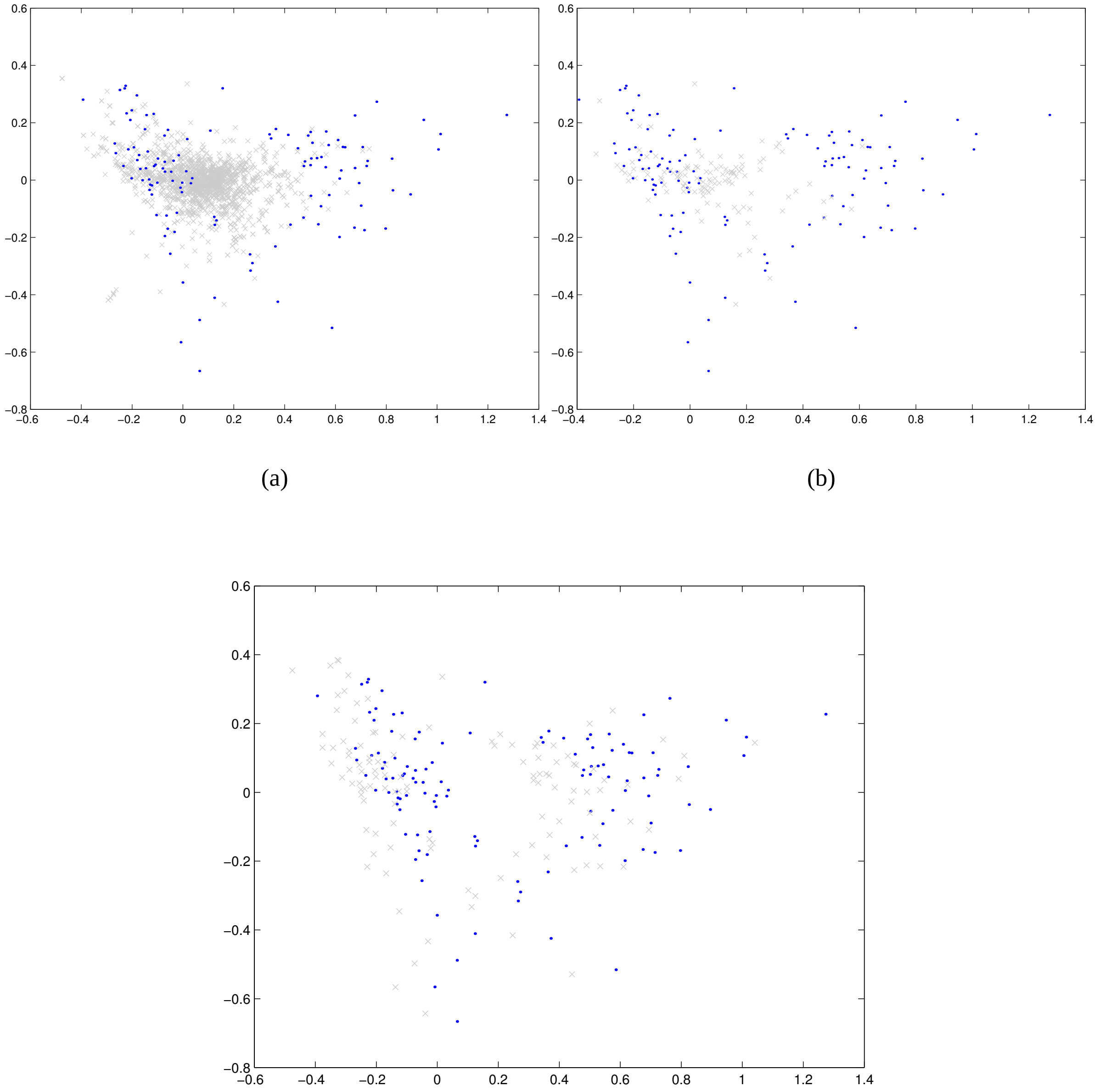

Calculate score value for all annotations (A)
Select annotation ah with highest score sm from A
check: sm > threshold
false
terminate
true
Associate gene-products gS annotated to ah uniquely to this annotation
store association gs - ah
remove annotation ah from A
recalculate score values for all Ai
substract gs from all Ai having gene-products of gs

Figure 2.7: Work flow for unique assignment of gene products to annotations. In the initial step the selection score is calculated, from which the annotation with the largest score is chosen. The gene products annotated with this term are uniquely associated to it, by subtracting (i.e. deleting) them from all remaining annotations. The selected annotation is removed from the set and (in combination with the annotated genes) stored. The scores values of the annotations whose set of annotated genes has been altered, is recalculated in the subsequent step. From this point on, again the annotation with the highest score is selected and if the score is above a user-defined threshold, the next unique assignment is created.



(a)

(b)

(c)

Figure 2.8: Unique assignment of genes to GOs. In all plots the annotations based on the unique assignment of genes are depicted as blue dots. In Figures a-b the grey crosses represent the original set of annotations, whereas in (c) they represent the unique set of GOs plotted as supplementary rows in contrast to GOs with mass (blue dots). The unique assignment results in 118 distinct annotations, these are plotted (with mass) in (a), where all the original annotations are plotted as supplementary rows. Figure (b) shows the same set, with the original set now being reduced to those annotations being present in the unique set as well. Figure (c) shows the difference in position when plotting the unique set of GOs with mass (blue dots) and as supplementary rows (grey circles). For clarity genes and experimental conditions are not plotted.

2.2.4.4 Data as supplementary points vs. points with mass

The display of additional information in correspondence analysis is commonly done by adding this data as supplementary points. These data points, in contrast to data points ’with mass’, are not relevant in the selection of the optimal subspace, in which the data is going to be projected. In other words, commonly the subspace is calculated based on the expression profiles of the filtered genes such that these are displayed in an optimal way. In order to display the annotations in an optimal way these can be plotted with mass as well.

Here I analyzed the *S. cerevisiae* Glucose-data set described in detail in 2.4.1. Initially, to account for multiple occurrences of genes in the annotation files, genes were assigned to a single, best-fitting annotation (as described in 2.2.4.3), resulting in 118 different annotations (listed in Table 7.2) associated to 257 distinct genes. Otherwise the data set is identical to the one presented in Figure 2.6.

Figure 2.8(c) shows the resulting CA plot when this unique set of annotations is plotted with mass (depicted as blue dots). In accordance with 2.2.4.2, each annotation is represented by a row-profile which corresponds to the row-wise sum of the expression-profiles of the annotated genes. In order to asses the impact on the annotation’s position, when plotting them with mass, the positions of the same annotations without mass are depicted by grey crosses in the same plot. Here no major difference in the resulting positions is visible. Even though none of the absolute positions are identical, the distances and the ordering between them, i.e. their relative positions with respect to each other, has not changed for the majority of annotations. Here the positions of the annotations with mass are shifted towards larger positive values along the first dimension.

2.2.4.5 Representation of a single gene by multiple features

The algorithms discussed here are integrated in the M-CHiPS analysis system [33]. This uses a database for the storage of the transcription data along with various data for experimental- and gene-annotations. To be able to analyze this data the individual spots are are associated to unique numerical ids. Since in the M-CHiPS-system the most basic point of reference that is used, is the location of the spotted material in the original microwell plates, each position in the plate can be identified (and accessed) via an unique id. This non-gene centered approach allows for easy access of transcription intensities and also could be used to relatively easy retrieve positional information (with respect to the position in the plates) and identify technical bias of the transcription intensities that could arrive from, for instance, dried out wells in the outer regions of the spotting-plate. Nevertheless it is not possible to immediately identify spots/features on the array that represent the same gene. This mainly is due to two reasons: First of all the exact same material (e.g. PCR-product or oligo) is contained in multiple wells of the spotting plates. Or, secondly, the same gene could be represented by different PCR products or oligos. E.g. some spots may represent the complete sequence of the gene, whilst others only part of the gene, or some spots may even represent non-overlapping sequences of the same gene.

For conventional analysis, this does not pose a major problem, since these genes will then represented by multiple points in the analysis. Ideally these points should be plotted in close vicinity to each other, i.e. exhibiting similar expression profiles. If this being the case, they even can serve to increase the confidence in the observed profile. If a grouping or categorization based on genes is applied however, this multiplicity turns out to be problematic.

As defined in 2.2.4.2 for each annotation a representative row profile is calculated by summing up the expression profiles of the annotated genes, more precisely it is the sum of annotated features. Since the annotation data naturally is gene-centered data, the summation based on the features is not optimal. Given the situation that a gene is represented by multiple features on the array multiple expression profiles of the same gene will contribute to the representative. Speaking in terms of CA, this gene will have a higher mass than the others. Due to this the position of the representative in the plot can be unduly strong shifted to the 'overrepresented' gene. This effect is especially profound if the overrepresented gene's profile strongly differs from the profiles of the other annotated genes.

To eliminate this effect, the transcription profiles of the genes being represented by multiple features can be summarized as well. A prerequisite for this is to have an identifier available in the database, which allows the identification of these candidates. Commonly the identifier that is used to create the association of the features to GO annotations can be utilized, in case of *S. cerevisiae* this would be the systematic name (e.g. 'YPR173C') or in case of *homo sapiens* the RefSeq-IDs could be used (e.g. 'NM_002746').

Having identified multiple representations of the same gene on the array I summarize the gene-wise median of the expression profiles. Figure 2.9 shows the difference in position of annotations, when plotting them with a) not accounting for multiplicity (depicted by purple dots) and b) calculating the median (depicted by blue dots).

2.3 Filtering of GO annotation terms

In case of the *S. cerevisiae* data set (2.4.1) 98.9% of the genes on the yeast-array could be assigned to at least one GO annotation. While in the original association file the most descriptive annotation (having the greatest depth in the tree) is recorded, I also assigned all father-nodes to the corresponding gene (see 2.1.4), resulting in a data set of 156674 tuples containing 3138 different annotations. These numbers alone strongly suggest the use of some filter-criteria as well as the fact that only very few of these annotations are descriptive for a condition and/or a gene-cluster and are thus worth to be judged by eye.

In an initial analysis of microarray data in context of gene-annotations however, it is not advisably to drastically reduce the amount of available information by only selecting a small

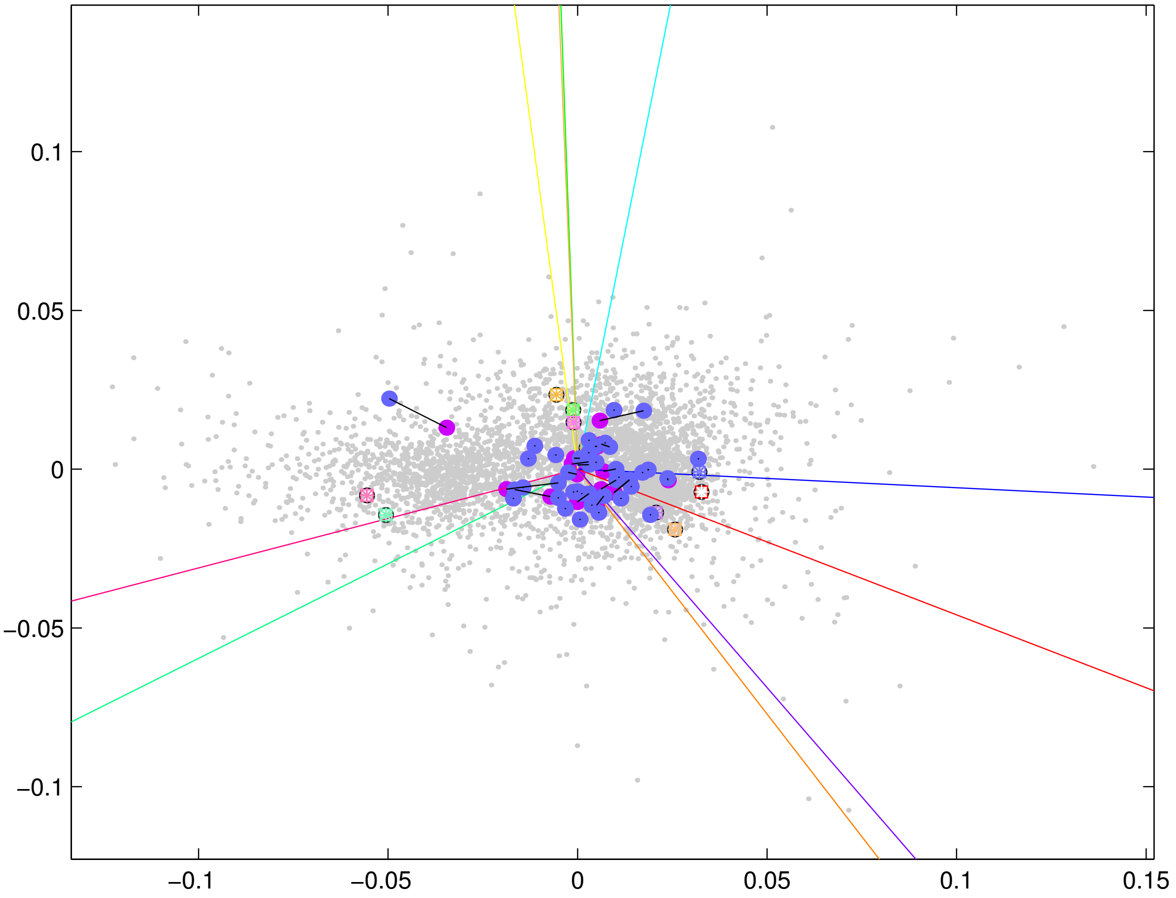


Figure 2.9: CA plot showing the difference in taking mean of doubles to the original position. The plotted dataset is described in 2.5.4.2. Here the focus is on the change in position if genes that are represented by multiple clones on the array are accounted for as a.) individual rows (purple circles) or b.) are summarized by calculating the gene-wise median (blue circles). Points representing the same TF are connected via a black line. The rest of the figure follows the outline of the previews ones, i.e. genes are represented by grey dots, experimental conditions by colored squares.

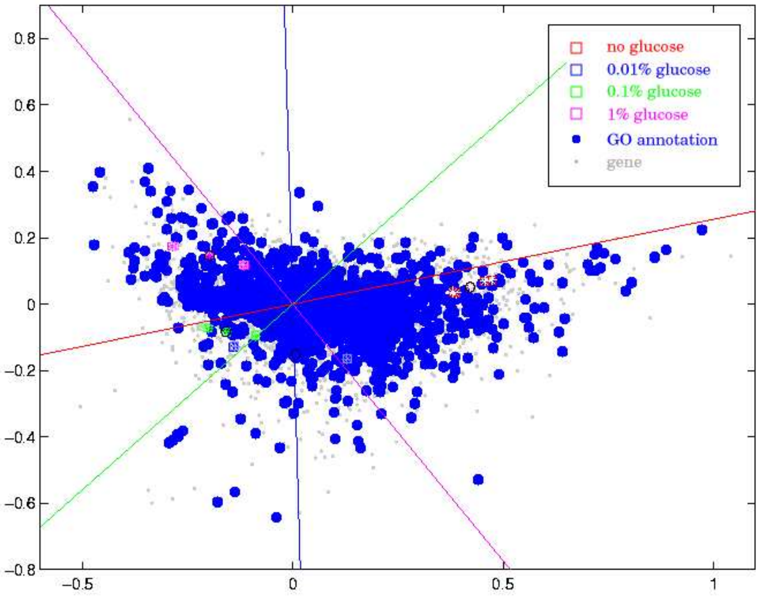


Figure 2.10: No filter for GO annotations applied. Here the same data as in Fig. 2.6 is plotted, except that no annotation filter has been applied. This means, in this plot all available annotations for this set of genes are shown, summing up to 2797 different annotations.

subset of annotations. The number of available annotations is mainly dependent on two variables: The first one being the knowledge about the organism that is captured by the ontology and the second one being the number of genes spotted on the array, or more specifically, the genes submitted to the analysis. In general an analysis without any kind of annotation filter is not advisable either. To demonstrate this, the same data set as in Fig. 2.6 is plotted, now displaying all annotations that could be associated to the genes on the array (Fig. 2.10).

The number of displayed annotations sums up to 2797, transforming the center of the plot to blue mass, due to the massive overlay of annotations. Noticeably the number of annotations being displayed near the margins of the plot is larger, as compared to Fig. 2.6 on page 23, potentially rendering the former plot ’more interesting’. Here one has to realize that the majority of the annotations placed in the outer regions are associated to small numbers of genes, commonly between one and three. If the development of biological hypothesis about the experimental setting is indented, one has to be cautious to base any conclusion on the position of annotations being associated to low numbers of genes. With these low numbers it is hard to statistically validate, whether there the calculated position for the annotation is relevant in the experimental context, or the small number of genes exhibits these similar profile by mere chance.

From this follows that filtering of annotations is inevitable, not only to limit the analysis to annotations comprised of a minimal number of genes, in order to increase the confidence in the observed relations, but also to reduce the overall number of annotations down to a level that is feasible to be analyzed by eye.

2.3.1 Based on ontological characteristics

The measures that can be used to decrease the number of annotations can be roughly divided in two groups: criteria based on features of the annotated genes (see 2.3.2) and characteristics of the underlying ontology, which will be discussed in this section.

The structure of the ontology, i.e. the separation into three main sub-trees (namely ’Molecular Function’, Biological Process’ and ’Cellular Component’) offers the most intuitive starting point for filtering of annotations. By removing one of these sub-trees the number of annotations can be reduced not only very effectively, but also in a sensible way, since in some experimental settings the main focus of the analysis might be on the localisation of the gene products (i.e. only the children from the term ’cellular component’ will be analyzed), whilst in other settings this is of only minor importance, and thus only the two other sub-trees are submitted to the analysis.

Another way to restrict the numbers of annotations, is to provide a user defined list of terms that are considered interesting in the given context. Here two ways are possible, one that could be referred to as ’opt-out’, where a list of annotations not to be displayed in the analysis is provided and another method, ’opt-in’, in which a list of annotations which have to be displayed is provided. In the practical situations the latter approach will be the most common one. While this approach commonly generates very well interpretable plots, with low number of annotations, it highly depends on the knowledge and diligence of the scientist assembling this list. Moreover the probability of discovering completely novel relationships between annotations and experimental conditions is rather low, since in these unexpected annotations will very likely not be part of the assembled list. Hence this way should not solely be used to reduce the number of annotations, but rather as a way of positively filtering the annotations, i.e. the user defined set of annotations is always added to the analysis, even though they might not fulfill other filter-criteria.

A further criterion is the level of specificity (or the other way round, level of generalization) an annotation should have in the ontology, in order to be displayed in analysis. This level is roughly reflected by the distance of a term to the root node of the ontology. In general it holds true that the higher the distance, the more specific, i.e. the more information is captured by the term. It should be realized however, that the distance measure can not be used to quantify the amount of information. So it is not valid to state that a term with a distance of four from the root node carries half the amount of information compared to a node of distance eight. In some cases the information content that is captured by concepts at the same level of the ontology can differ drastically, for example: ’*glutathione dehydrogenase (ascorbate) activity*’ (GO:0045174) and ’*enzyme activator activity*’ (GO:0008047) both have a distance of four to the root node, but the specificity (i.e. information content) of the terms clearly is not the same. One major factor for the specificity of terms, or in other words, the depth of the ontology in

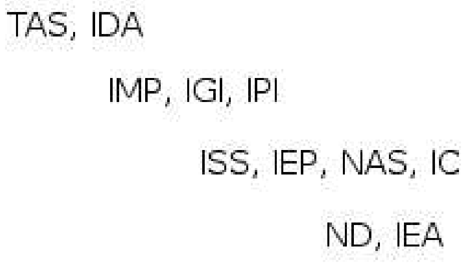


Figure 2.11: Ordering of evidence codes based on quality of annotation. Here a proposed ordering of the evidence codes is given. With TAS providing the most trustworthy annotations and at the other end of the scale IEA, being the most unreliable form of annotation.

specific areas largely depends on the amount of knowledge that is available in this field. If for a specific area a lot of research has been conducted, the depth of this subsection will be larger and the annotations tend to be more complex.

Yet another approach for the selection of annotations is, not only to define single annotations that should be integrated in the analysis, but also incorporate their parental- (or child-) terms as well. This can be fairly easily done by the query in 7.1. Extracting the parental notes is commonly used when very specific information about the gene product is available and the dataset is to be analyzed with respect to some broader concepts. Extraction of child-terms is applied when some general ideas about the effected mechanism (for instance pathways) exists, but with insufficient detail. This specification implies that this information is provided in the gene annotation files.

Additionally the evidence codes, that are provided with each annotation of a gene product provide valuable information for filtering. These codes represent the quality/reliability of the given annotation. Hence annotations based on ’traceable author statement’ (TAS) can be considered to be more reliable as compared to ’inferred from electronic annotation’ (IEA), to compare the two most extreme codes. Restricting the displayed annotations to those with TAS-code will increase the reliability of annotations significantly, but on the other hand the number of annotations that are displayed are so small (an again with only low numbers of annotated genes as well), that the generation of new hypothesis will become difficult. Fig. 2.11 provides a proposed rough ordering of the evidence codes based on the underlying quality of annotation, and could be used to define a compromise between reliability of annotation and sufficient numbers of annotation and genes in the analysis. Please note that the evidence codes in the corresponding user-interface (Fig. 7.2 on page 100) are listed in descending order of quality.

2.3.2 Based on gene characteristics

One fairly obvious but nevertheless crucial filter is the number of annotated gene products per annotation term. By this the minimal and maximal numbers of annotated genes can be defined, in order to qualify it for the display in the CA plot. The upper boundary (i.e. maximal number of genes) is suitable to discard annotations that are not very likely to provide any functional information for the experimental setting. That is, terms annotated with very high numbers of genes (e.g. a couple of hundreds) commonly refer to very general concepts like '*development*' or '*response to stimulus*'. These terms carry only little useful information for the functional interpretation and thus are futile to be displayed in the CA map. Moreover, as described in 2.2.1 on page 14, in CA points being closest to the margins of the plot, are the most 'interesting' ones (i.e. highest difference to average profile). However, in the majority of cases, annotations with large number of genes will be positioned near the centroid of the map - marking them as non-interesting for further analysis.

Yet another gene-characteristic could be used for filtering: one can argue that annotations, whose annotated genes show a homogeneous expression profile are of particular interest/relevance in the given experimental context and thus should be further analyzed. One measure that accounts for this characteristic is the correlation coefficient. The application of this and subsequent testing is described in detail in 2.3.3.2.

2.3.3 Receiver Operating Characteristic curves to evaluate filter performance

Having multiple filters for the annotations at hand, one needs to evaluate their performance, in order to find a set of standard parameters (or even a combination of different filter measure), that could be applied to the data. A standard methodology to evaluate performance of classifiers are ROC curves. ROC curves have been developed in the 1950 when trying to identify true signals in the highly noisy radio signals. Nowadays it has become a common method in the medical field to evaluate the performance of different cut-off values in diagnostic tests [91, 92].

ROC-curves allow to assess the accuracy of predictions, e.g. in medical tests, a threshold has to be chosen above (or below) which the patient is considered diseased. In ideal situations the two populations (i.e. diseased and healthy) could be distinguished at a certain threshold (or range of thresholds). In real-life situations this commonly is not the case, since there is an overlap between both populations at a certain value-range of cut-offs (Fig. 2.12). In the most simple case ROC curves assess the accuracy of a binary predictor. For clarity all possible outcomes for such a binary predictor are listed in 2.5. Some cases will be correctly identified as positives (True Positive, TP), whereas other positive cases will falsely classified negative (False Negative, FN) - due to the overlap of populations (Fig. 2.12). Accordingly, cases being negative, can be classified as negative (True Negative, TN) or positive (False Positive, FP). In

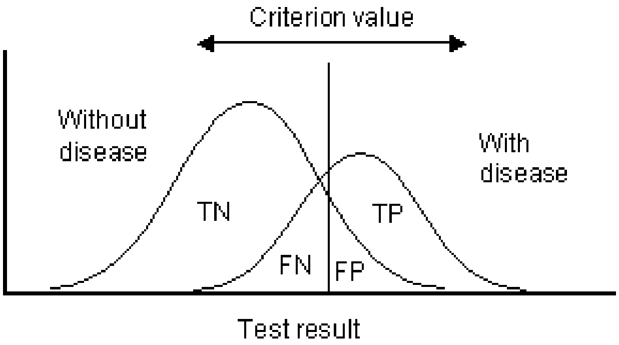


Figure 2.12: Distribution of populations in a binary system. Given the situation that patients have to be classified as diseased or normal based on a test. In the ideal situation there would be a cut-off value (or a range of values) that clearly separates both populations. In real-life situations however it is common to observe an overlap of both populations in a certain cut-off range. Figure reproduced from [93]

case of binary predictors the effects on true (or false) positives and true (or false) negative are commonly reported separately. This accounts for the fact, that the consequences of missing a positive (i.e. false negative) might have different impact than classifying a negative as positive (i.e. false positive).

Thus the values can reported as false positive and false negative rates. Whereas the false positive rate (FPR) is calculated by $FPR = \frac{FP}{FP+TN}$ and the false negative rate (FNR) by $FNR = \frac{FN}{FN+TP}$. The rates of the correctly classified samples can be obtained accordingly: the true positive rate (TPR) by $TPR = 1 - FPR$, or $TPR = \frac{TP}{TP+FN}$ and the true negative rate (TNR) by $TNR = 1 - FNR$, or $TNR = \frac{TN}{TN+FP}$.

In literature the TPR is often referred to as *sensitivity*, i.e. the probability that a case with its true class being positive will be correctly identified as positive out off all cases being classified positive [94]. Whereas the TNR is often referred to as the *specificity*, i.e. the probability that a case with its true class being negative will be correctly identified as negative out off all cases being classified negative. As already mentioned this is easily calculated for a binary predictor, but becomes is not possible for a continuous predictor as for instance, the homogeneity criterion (i.e. correlation coefficient) for the GO annotations. To assess the performance of a continuous predictor, it has to be transformed to binary form by analysing the results at different thresholds. The performance is commonly evaluated based on the resulting sensitivity and specificity. This can be done by calculating and displaying these values for individual thresholds separately. This however, becomes less and less informative with increasing numbers of analyzed thresholds. ROC curves allow the overall performance of a classifier by representing the results in a 2 dimensional plot (Fig. 2.13).

Typically ROC curves display the trade off between false negative and false positive rates

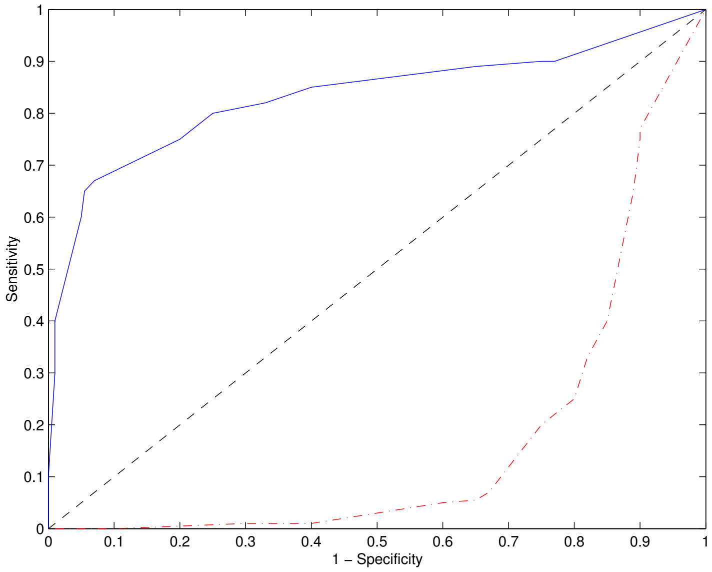


Figure 2.13: ROC curve with artificial data. The solid line depicts a ROC curve derived from artificial data. The dashed line represents the ROC curve for a random classifier, whereas the dash-dotted line is derived from a classifier which performs worse than random. In these cases it is advisable to inverse the prediction of the classifier, which (in this case) would result in the solid line.

for different cut-off values. The resulting curve always goes through two points (0,0) and (1,1). Whereas (0,0) represents the state where no case is classified as being positive and (1,1) represents the cut-off at which every possible case is considered as being positive. This means that on the one hand the classifier identifies all true positive cases correctly, but on the other hand all true negative cases are incorrectly classified. This means the sensitivity is 1 (all true positive cases are classified positive) and the specificity 0 because all true negative cases have been classified as positive. Since the x-axis displays 1-specificity this case will result in the coordinates (1,1).

A random classifier would generate a curve close to the diagonal connecting (0,0) and (1,1) (e.g. dashed line in Fig. 2.13). In some cases the resulting curve of a classifier is beneath the diagonal, this indicates that the performance is worse than random (dash-dotted line in the same Figure). Here an inversion of the prediction will result in an above average classifier.

The quality of a predictor can be assessed from the curve. The perfect predictor would be represented by a single point in the plot, namely (1,0). According to this, the steeper the slope, the better the predictor. This behaviour can be summarized by calculating the area under the curve (AUC). For an ideal classifier the AUC will be 1 (having 100% sensitivity and 100% specificity), for a random classifier the AUC is 0.5 (dashed line in Fig. 2.13). Any ROC analysis is based on a set of samples for which the true state is known, sometimes refereed to as the ’golden standard’, the definition of this for the gene annotations is described in the following section(2.3.3.1).

True class	Predicted class		total
	p	n	
P	True Positive (TP)	False Negative (FN)	$TP+FN$
N	False Positive (FP)	True Negative (TN)	$FP+TN$
Total	$TP+FP$	$FN+TN$	

Table 2.5: Comprehensive overview of outcomes for a binary predictor. When using a (binary) classifier for the prediction of classes the following results are possible: Cases that are correctly identified as positives (True Positive, TP), whereas other positive cases will falsely classified negative (False Negative, FN). Accordingly, cases being negative, can be classified as negative (True Negative, TN) or positive (False Positive, FP).

2.3.3.1 Definition of ’standard annotations’

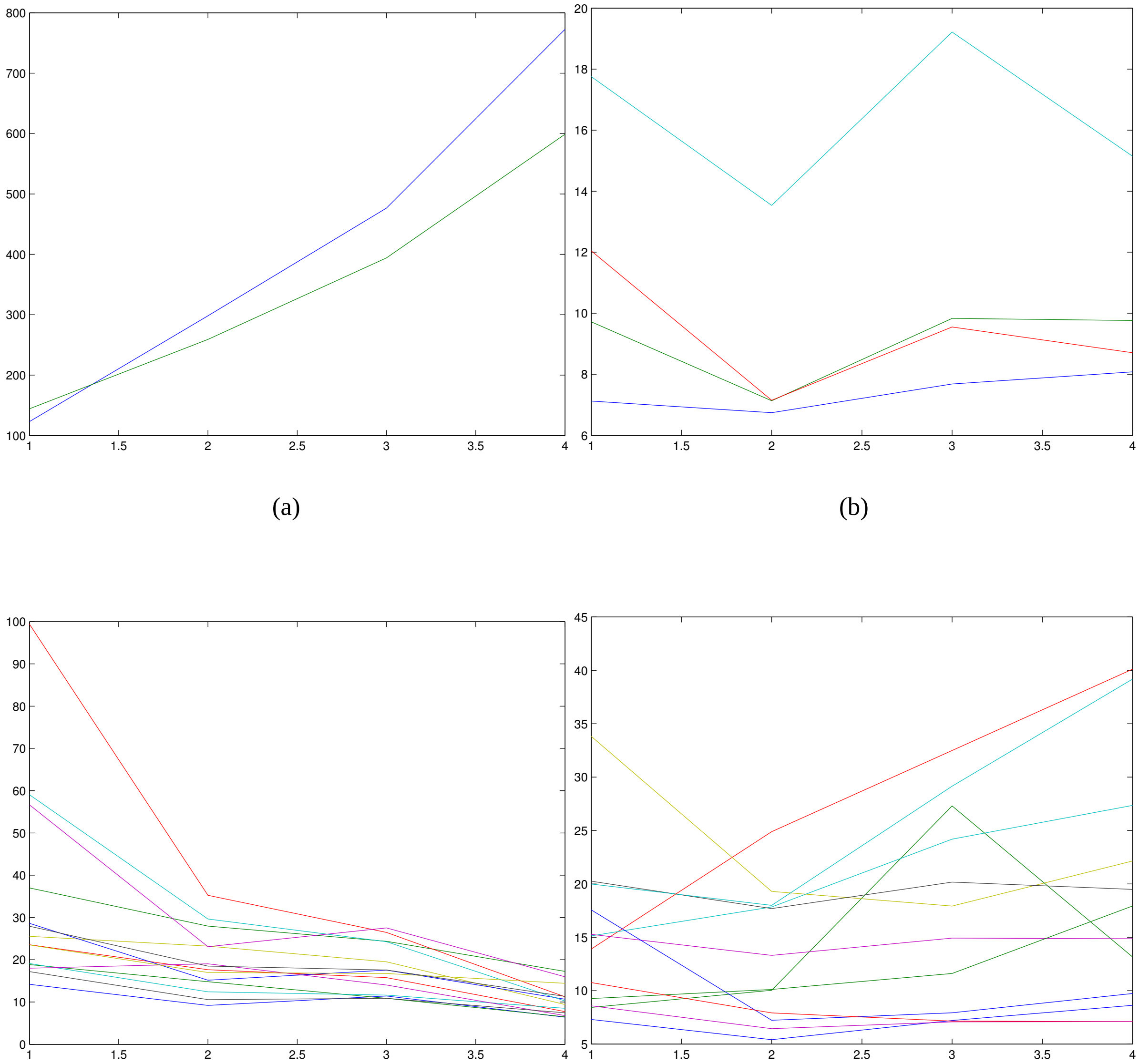
In order to perform a ROC analysis it is necessary to have a ’golden standard’, i.e. to have cases for which the true class is known (see Table 2.5). In case of the GO annotations, a set of ’standard annotations’ has been defined for the *S. cerevisiae* data set (2.4.1).

As explained in 2.3.3.2 the homogeneity of the expression profiles of the annotated genes can be used as a filter criterion. Based on this, a set of 65 annotations has been manually selected and classified into two groups, namely ’homogeneous’ (i.e. positive) and ’non-homogeneous’ profiles (i.e. negative). In Figure 2.14 examples of the expression profiles for these different groups are provided. In the selection of the standard annotations care was taken not to over-represent annotations with small numbers of annotated genes in the positive groups.

2.3.3.2 Identification of optimal measure of homogeneity

As a measure for similarity in expression I used the absolute values of all pairwise correlation coefficients (R) of the genes associated to one GO-Annotation. To calculate R the condition-medians of repeated hybridizations for one experimental condition have been used. Since I consider an annotation containing anti-correlated genes to be descriptive as well, I work with the absolute values of R. The pairwise R for each annotation were condensed to one number by calculating the minimum, median, 75percentile and percentage of genes with $R > 0.8$.

The overall goal for the annotation filter is to reduce the number of displayed annotations to a set which is feasible to analyze/interpret manually. One possible criteria on which annotations can be selected is the homogeneity of the expression-profiles of the annotated genes. This filter combined with an appropriate gene-filter (e.g. one that selects for differentially expressed genes) will very likely select annotations that are relevant in the given experimental context. A further benefit of a homogeneity filter is that annotations being too unspecific will very likely be filtered out, due to the comparably large number of annotated genes.



(a) (b)

(c) (d)

Figure 2.14: Examples for the 'Golden Standard' in ROC analysis. Figure (a)-(c) show examples of annotations that have been categorized as having homogeneous expression-profiles (i.e. they are classified as positive). Whilst the classification for(a) is obvious, the increasing number of genes in (b) and (c) make make it more and more difficult to make a clear call. (d) shows an example of an annotation that has been classified as non-homogeneous (i.e negative). The x-axis represents the different experimental conditions, whereas the y-axis depicts the normalized intensities.

To evaluate the homogeneity of expression profiles correlation coefficients are a useful method. Here, the most common one is Pearson's Correlation Coefficient, which basically represents the quality of a least squares fit to the data. However the following assumptions have to be met to apply this correlation coefficient:

- linear relationship between x and y
- continuous random variables
- both variables must be normally distributed
- x and y must be independent of each other.

In the case of microarray data the intensity distribution commonly is heavily skewed to the right [95, 96], thus not meeting the assumption of an normal distribution. Moreover in the case of GO annotations it is questionable if the genes annotated to the same GO term can be considered as being independent of each other.

To circumvent these problems non-parametric correlation coefficients can be used to compute a measure for the strength of association between the annotated genes. The most prominent non-parametric coefficient is 'Spearman's Rank Correlation Coefficient' , which is defined by:

$$\rho = 1 - 6 \sum \frac{d^2}{N(N^2 - 1)}$$

, where d represents the difference in statistical ranks of the two variables (i.e. genes) and N is the number of pairs of values. The independence from the assumption of normality distribution is achieved by calculating the correlation not on the raw numbers (i.e. intensities) but on the ranks. Another correlation coefficient that is based on the ranks, is the one from Kendall (τ), which is defined by:

$$\tau = \frac{2p}{\frac{1}{2}n(n-1)} - 1$$

, where n represents the number of pairs and p the sum over all items. In the following the performance of both coefficients based on the previously defined 'golden standard' has been compared.

Since the correlation coefficient measures the association between two variables (i.e. here two genes), multiple values are generated for a annotation having more than two genes. To be able to apply an annotation filter a single descriptive value for each of the annotations has to be provided. Again there are multiple ways (e.g. mean, median, minimal, percentiles, etc.) by which to 'summarize' the pairwise correlation coefficients of the genes. Since there is

	Area under curve (AUC)			
	mean	median	minimal	75 percentile
Kendall's τ	0.80	0.76	0.19	0.71
Spearman's ρ	0.78	0.78	0.19	0.71

Table 2.6: Area under curve for different correlation coefficients. Here the performance of the different correlation coefficients and the different ways to summarize them is evaluated based on the 'area under curve'. This area is calculated based on the trapezoid method.

no best choice that could be deducted from theoretical considerations, the different ways for summarizing have been compared by ROC analysis as well.

To assess the performance of the filters quantitatively the corresponding AUCs have been calculated and are given in Table 2.6. One obvious fact is that taking the minimal coefficient out off all pairwise will result in a poor classification. The remaining methods are in the same range whilst mean and median slightly outperform the 75 percentile. For clarity the ROC curves of taking the mean and median of both Rs are plotted in Figure 2.15. Here, the performance of both Rs summarized by the mean differs only marginally, with both showing adequate performance.

While the AUC is providing a measure for the performance of the classifier over the complete range of filter values, the value range commonly used for correlation coefficients is between 0.6 and 1. Thus the performance of the different summarizing methods at fixed cut-offs has been plotted in Figure 2.16.

Again the performance of both correlation coefficients is highly similar, the overall highest sensitivity, however, is accomplished when using Spearman's ρ. Moreover at high cut-off values Spearman's ρ outperforms Kendall's τ in terms of sensitivity while showing the specificity. Thus, the mean of all pairwise ρ's is being used in all subsequent analysis as the measure for homogeneity of the annotated gene-profiles.

2.4 Biological validation of algorithms

Whilst the development of new methodology to analyze high-throughput data could be entirely restricted to theoretical considerations and testing on artificial data, it is crucial to evaluate the performance in the context of 'real' biological data. To this end I applied the integration of annotation data to various datasets, two of which are analyzed here exemplarily. The initial evaluation will be done on data from the model organism *Saccharomyces cerevisiae* in an experimental setup which investigates well known pathways (2.4.1). In a subsequent step, the method is applied to a more complex organism and experimental setting, by analyzing different tumor samples from *homo sapiens* (2.4.2), these are results are also available from [90].

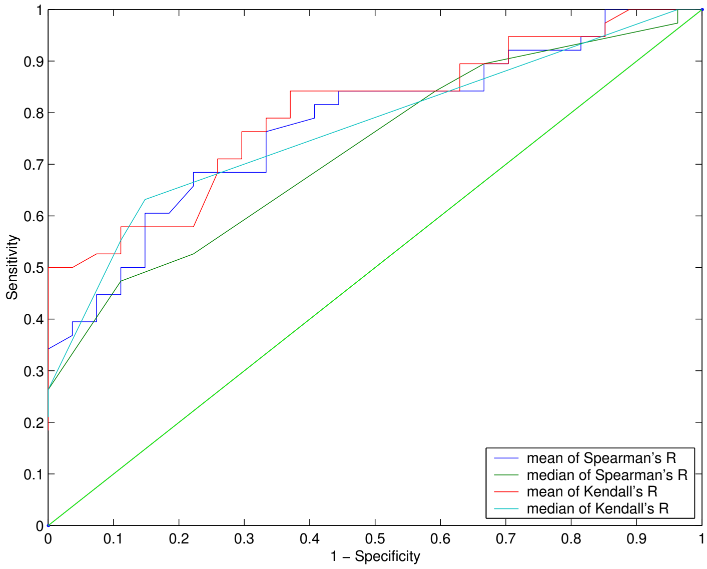


Figure 2.15: ROC analysis of correlation coefficients. Here the corresponding ROC curves for the different Rs are shown. Whereas the in one setting the pairwise Rs have been summarized by calculating the mean (blue and red line), in the other by the median (cyan and dark green line). The light green diagonal in the plot depicts the curve for a random classifier.

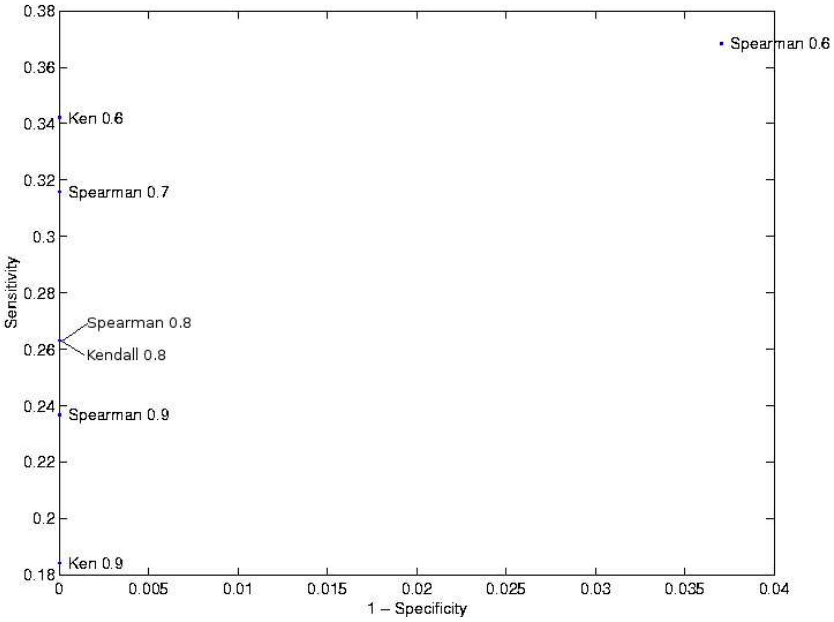


Figure 2.16: Performance of filter parameters at fixed cut-off values. Here specificity and 1-sensitivity values are plotted for practically relevant thresholds for correlation coefficients (e.g 0.6 - 0.9).

Prior to any analysis shown in the following a general filtering of annotations was performed: the root term (Gene Ontology) and the three main category terms (biological process, cellular component and molecular function) were deleted, as well as the annotation 'unknown' from each category. Since these terms do not carry any useful information for functional interpretation of gene-clusters or experimental conditions plotting these would be futile.

2.4.1 *Saccharomyces cerevisiae* - glucose data set

2.4.1.1 Experimental setup

To demonstrate the applicability of this approach I analyze a dataset of the model-organism *Saccharomyces cerevisiae* focusing on the well-studied glucose pathway [88]. The relevant data is publicly available from [97].

In this experiment, the arrays consist of 6103 yeast-genes [98]. Mapping of the genes to GO terms was carried out using the systematic gene-name (e.g.: YBR166C) and the common name (TYR1) in the association file provided by SGD. A total of 3060 distinct GO annotations was found, which annotate 5506 (90.22%) genes on the array. In Figures 2.5 and 2.6 only genes that have a maximal normalized intensity value equal to or larger than 10 in any experimental condition (intensity filter), resulting in a total of 2898 genes, were submitted to analysis.

In this experimental setting *Saccharomyces cerevisiae* had been grown in media containing different amounts of glucose (0%, 0.01%, 0.1%, 1%). For each of these conditions, RNA had been isolated, processed and hybridized to microarrays [98]. The data had been normalized and filtered [99, 87] such that the data-matrix being submitted to CA holds normalized transcription intensities, rows depicting the genes, columns the experimental conditions. Conditions are represented by the gene-wise median of repeatedly performed hybridizations. The gene annotations were filtered such that only GO terms containing a minimum of 5 genes are displayed in analysis.

2.4.1.2 Intensity based coding

The resulting CA plot of this dataset is provided in Figure 2.5 on page 21. The experimental conditions are ordered in a clockwise orientation following increasing levels of glucose. The condition with no glucose is the only one on the right hand side of the centroid, in other words the x-axis (1st principal axis, which accounts for the largest variance in the data set) explains the separation of glucose from non-glucose conditions. This implies, that the difference in expression profiles between glucose vs. non-glucose states is larger than the changes due to varying concentrations of glucose. This is in agreement with current knowledge, since in the absence of glucose as an energy source, different carbon sources have to be accessed for the utilization of energy, resulting in major changes of activated pathways.

GO accession	(Main Category) GO term
GO:0008643	(P) carbohydrate transport
GO:0015749	(P) monosaccharide transport
GO:0008645	(P) hexose transport
GO:0015144	(F) carbohydrate transporter activity
GO:0015145	(F) monosaccharide transporter activity
GO:0015149	(F) hexose transporter activity
GO:0015578	(F) mannose transporter activity
GO:0005353	(F) fructose transporter activity
GO:0005355	(F) glucose transporter activity

Table 2.7: GO annotations forming the cluster in Fig. 2.6. Cluster members are listed with their corresponding GO accession id, main category (P= Biological Process, F= Molecular Function) and GO term. These annotations are linked to a set of 13 genes, all of which belong to more than one annotation of the cluster (Table 2.8).

With the experimental conditions being placed at sensible positions, the subsequent step is to analyze the displayed annotations. Here the majority (depicted as blue dots) is concentrated around the centroid of the map. In this setting, however, a distinct cluster of annotations can be observed, which was not detectable in the Boolean approach (Fig. 2.6 on page 23). This cluster consists of 9 different annotations, whose GO accessions and corresponding terms are listed in Table 2.7. The annotations are linked to a set of 13 genes, all of which belong to more than one annotation of the cluster (Table 2.8). All annotations describe, at different levels of detail, the activation of carbohydrate-transport into the cell.

The position of the cluster indicates negative association with the control condition (no glucose in medium) and positive association with the remaining conditions with stronger association, i.e. up-regulation, in response to low glucose signals (0.01% and 0.1% glucose). Indicating that at low levels of glucose higher concentrations of transporters have to be present in the membrane, to efficiently uptake glucose in the cell. This is consistent with prior findings identifying HXT1 to HXT7 as key enzymes for the uptake of glucose with HXT2, HXT6 and HXT7 being important for growth on 0.1% glucose [88].

As a subsequent step the stringency of the gene-filter is further increased to validate the positions of the annotations based on the most reliable genes in the data set. Therefore genes were filtered out, whose transcription intensities remain below 30 and/or show a minmax-separation (a quality filter that assesses how well repeatedly measured genes are separated under two different conditions [99]) of less than 0.3. In resulting two figures (2.17 and 2.18) based on this gene-set, the y-coordinates of the data-points were multiplied by -1 (i.e. mirrored at the x-axis) for better interpretability.

From Figure 2.17 it is obvious that the reduction in gene numbers also resulted in a reduction of displayed annotations. Nevertheless the previously identified annotation cluster (describing

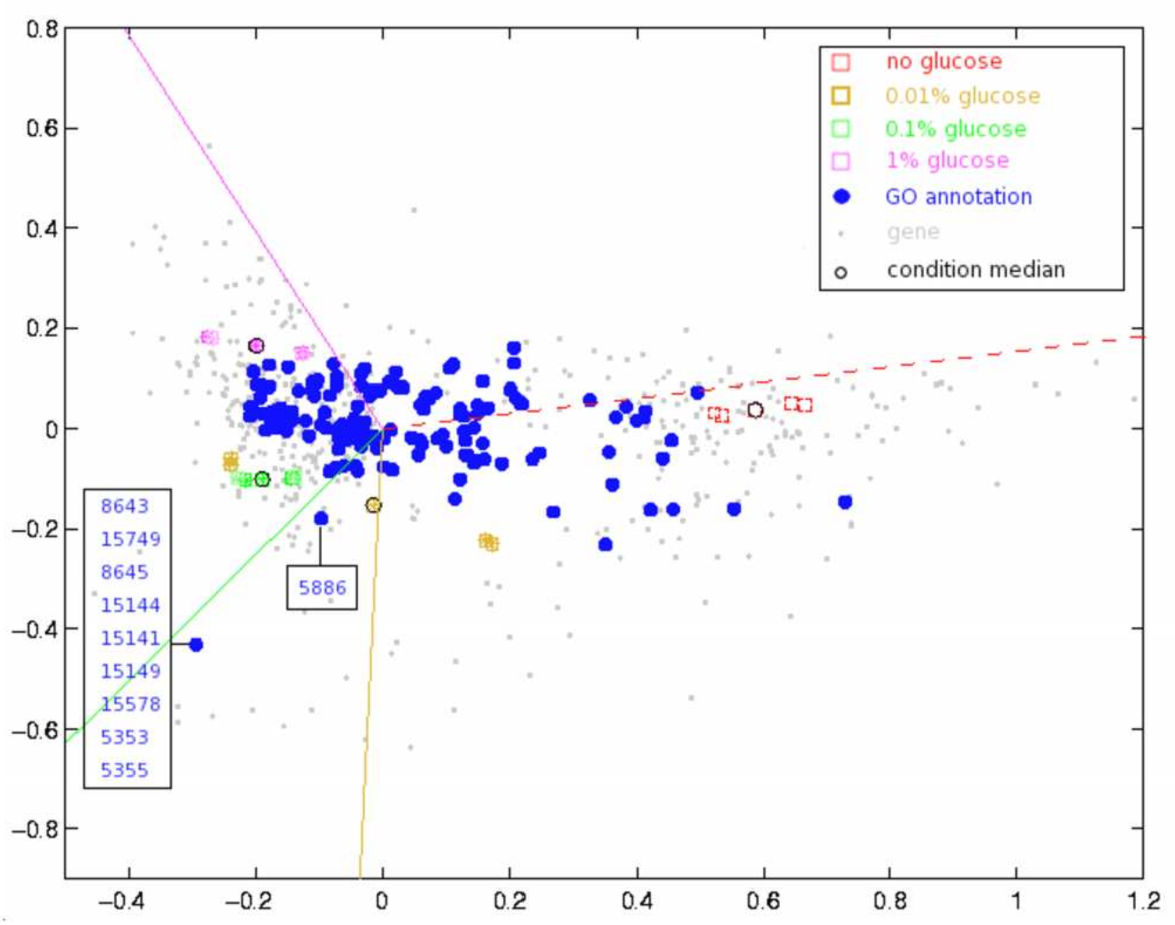


Figure 2.17: CA-Map with stringent gene filter (444 genes remaining). GO annotations are added to the data-matrix as supplementary rows and are depicted as solid blue circles. GO IDs have been truncated as in the previous figures (see also legend in upper right corner). For better readability, annotations forming the transporter cluster are listed in the adjacent box. In this plot ≈93% of the total variance in the data-set is explained by shown two first principal axis.

Genes	GO identifiers (truncated of ’GO:’ and trailing zeros)
YPL026C (SKS1)	8643, 8645, 15749
YDL194W (SNF3)	5355, 15144, 15145, 15149
YPL244C (HUT1)	8643, 15144
YGL225W (GOG5)	8643, 15144
YHR094C (HXT1)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YMR011W (HXT2)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR345C (HXT3)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YHR092C (HXT4)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YHR096C (HXT5)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR343C (HXT6)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YDR342C (HXT7)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YJL214W (HXT8)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749
YJR158W (HXT16)	5353, 5355, 8643, 8645, 15144, 15145, 15149, 15578, 15749

Table 2.8: Genes annotated to GO terms in transporter cluster (Fig. 2.6). The first column is comprised of the systematic gene name and the common name is given in brackets. The second column lists the GO annotations of the genes. Genes marked in bold fulfill stringent gene filter criteria and their corresponding annotations are displayed in Fig. 2.17.

the transportation of glucose into the cell) is still distinguishable with this reduced set of genes. The annotations comprised in the cluster remain constant (as shown in Table2.7). Only the number of distinct genes is reduced to six (marked bold in Table 2.8).

Due to the reduction of displayed annotations a further annotation (’plasma membrane’ - GO:0005886) becomes apparent, being associated with condition containing 0.1% glucose in the medium, i.e. being located in the same direction as the ’transporter-cluster’. This annotation is a member of the ’cellular localization’ ontology and thus describes the positioning of the annotated gene product in the cell. Since the genes encoding for the transporter proteins are among those annotated to this term, the positioning of this annotation makes sense, since the localization of transporter proteins in the plasma membrane is inevitable to enable transport of glucose into the cell and thus is also in agreement with the literature [89].

2.4.1.3 Application of Spearman-filter

Through the reduction in number of genes the number of displayed annotations was reduced as well, but still resulting in too large numbers to be thoroughly analyzed by eye. Thus annotation filters as described in 2.3.3.2 were applied. In the resulting plot (Fig. 2.18) annotation terms with less than three annotated genes and a mean Spearman correlation coefficient of less than 0.8 were discarded, leaving a total of 15 annotations to be displayed, further enhancing the clarity of the plot.

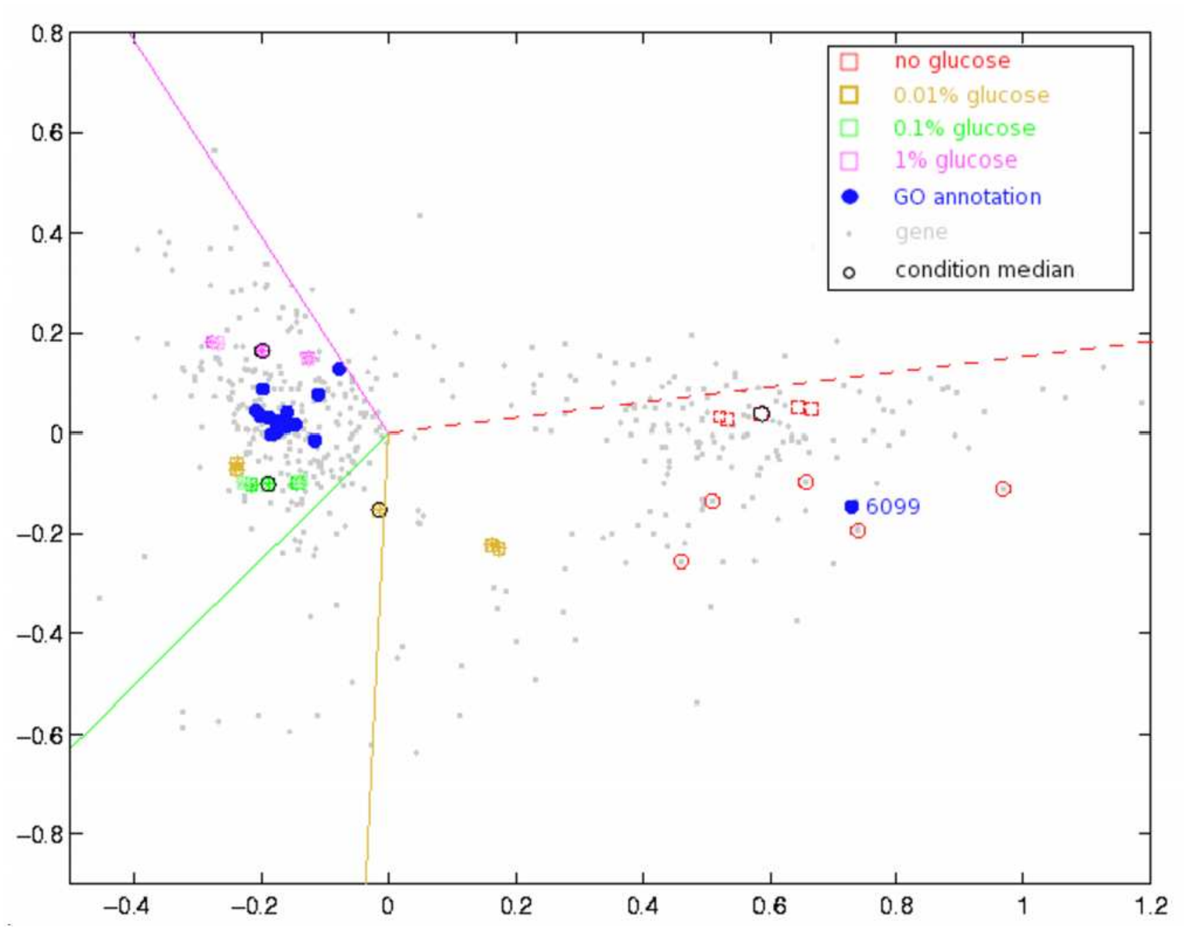


Figure 2.18: CA-Map with filtered set of GO annotations. To filter out GO terms annotating inhomogeneously transcribed gene sets, the mean of Spearmans correlation coefficient was applied. The plot follows the layout of the previous figures, see legend in upper right corner. Annotations forming the cluster left of the centroid are listed in Table 2.9. Genes annotated as 'tricarboxylic acid cycle' (GO:0006099) are encircled red. Note that this annotation represents the center of gravity of the corresponding genes.

GO identifier	(Main Category) GO term
GO:0000027	(P) ribosomal large subunit assembly and maintenance
GO:0000028	(P) ribosomal small subunit assembly and maintenance
GO:0042257	(P) ribosomal subunit assembly
GO:0005840	(C) ribosome
GO:0005830	(C) cytosolic ribosome (sensu Eukarya)
GO:0005842	(C) cytosolic large ribosomal subunit (sensu Eukarya)
GO:0005843	(C) cytosolic small ribosomal subunit (sensu Eukarya)
GO:0030529	(C) ribonucleoprotein complex
GO:0004553	(F) hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0015926	(F) glucosidase activity
GO:0006096	(P) glycolysis
GO:0006099	**(P) tricarboxylic acid cycle**
GO:0006445	(P) regulation of translation
GO:0006450	(P) regulation of translational fidelity
GO:0008652	(P) amino acid biosynthesis

Table 2.9: GO annotations displayed in Fig. 2.18, are listed with their corresponding GO-accession, main category (P= Biological Process, F= Molecular Function) and GO term. The annotation marked in bold is positively associated to the control condition. Remaining annotations are negatively associated to the control condition.

Here, the annotation ’tricarboxylic acid cycle’ (GO:0006099, lower right quarter of the plot) is associated with the control condition, suggesting that the annotated genes are repressed in the presence of glucose. Examplariy for this annotation, the corresponding genes are encircled red in the plot. Please note that the position of the annotation is the center of gravity of the annotated genes.

The genes are comprised of YKL085W (MDH1, malate dehydrogenase), YDR148C (KGD2, 2-oxoglutarate dehydrogenase), YLR304C (ACO1, aconitase), YNR001C (CIT1, citrate synthase) and YIL125W (KGD1, alpha-ketoglutarate dehydrogenase). All of which are known to be involved in the TCA cycle. For example MDH1, catalyzes interconversion of malate and oxaloacetate [100], whereas KDG1 and KGD2 are components of the mitochondrial alpha-ketoglutarate dehydrogenase complex, which catalyze a step in the tricarboxylic acid (TCA) cycle, namely the oxidative decarboxylation of alpha-ketoglutarate to succinyl-CoA [101, 102]. Whereas CIT1, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate and also being the rate-limiting enzyme of the TCA cycle [103, 104].

In the opposite direction of the centroid, a cluster of annotations can be found, whose members are listed in Table 2.9. In the presence of glucose in the media, it is used as the primary energy source for *S. cerevisiae*, such that pathways utilizing this source are expected to be upregulated. The activation of these pathways is described by the association of the corresponding

GO annotations, like ’glycolysis’, ’glucosidase activity’ and ’hydrolase activity, hydrolyzing O-glycosyl compounds’ with the glucose containing conditions.

Out of the total of 10 genes being annotated with the term ’glycolysis’, some are discussed examplarily : YCR012W (PGK1), which is a 3-phosphoglycerate kinase, catalyzing the transfer of high-energy phosphoryl groups from the acyl phosphate of 1,3-bisphosphoglycerate to ADP to produce ATP and thus is a key enzyme in glycolysis and gluconeogenesis [105, 106]. Furthermore the gene YKL060C (FBA1) is annotated thereto as well, which is a fructose 1,6-bisphosphate aldolase, a cytosolic enzyme required for glycolysis and gluconeogenesis. FBA1 catalyzes the conversion of fructose 1,6 bisphosphate into two 3-carbon products, namely glyceraldehyde-3-phosphate and dihydroxyacetone phosphate [107, 108, 109]. Finally YMR205C (PFK2), beta-subunit of heterooctameric phosphofructokinase, is involved in glycolysis which is indispensable for anaerobic growth and activated by fructose-2,6-bisphosphate and AMP. Mutation in this gene inhibits glucose induction of cell cycle-related genes [110, 111, 112].

Moreover apart from annotations describing energy metabolism the reminder of annotations is mainly comprised of terms referring to the ribosome . Their positions in the CA-map indicate up-regulation of the corresponding genes at 0.1% and 1% glucose, consistent with [88]. In the presence of sufficient amounts of glucose, yeast cells invest energy in the production of ribosomes to enable rapid growth and reproduction. This is also reflected by the up-regulation of genes responsible for ’amino acid biosynthesis’ (GO:0008652), which is essential for prolonged growth.

2.4.2 *Homo sapiens*

Whilst the usefulness of the integration of annotations in CA for biological interpretation has been shown for lower eukaryotes (2.4.1), performance in a more complex organism as well as a more complex experimental set-up remains to be demonstrated. To this end I analyzed microarray data studying different subtypes of human cancer [113].

RNA from human ductal adenocarcinomas, cystic tumors and normal pancreas tissue was extracted, labeled and hybridized to a cDNA microarray. The resulting data has been processed analogous to the previous *S. cerevisiae* data set. Annotations were added as supplementary rows to the data matrix (intensity based implementation, 2.2.4.2) and filtered by Spearman’s correlation coefficient such that the number of displayed annotations was reduced to 35 (Fig. 2.19).

The separation of normal tissue from cancer samples is clearly visible along the x-axis, whereas the potential new tumor entity [113] separates from ductal and cystic tumors along the y-axis. Three clusters of annotations can be distinguished: annotations associated to normal tissue (A), associated to ductal and cystic (B) and generally tumor associated (C). The complete list of displayed annotations is given in Table 7.1.

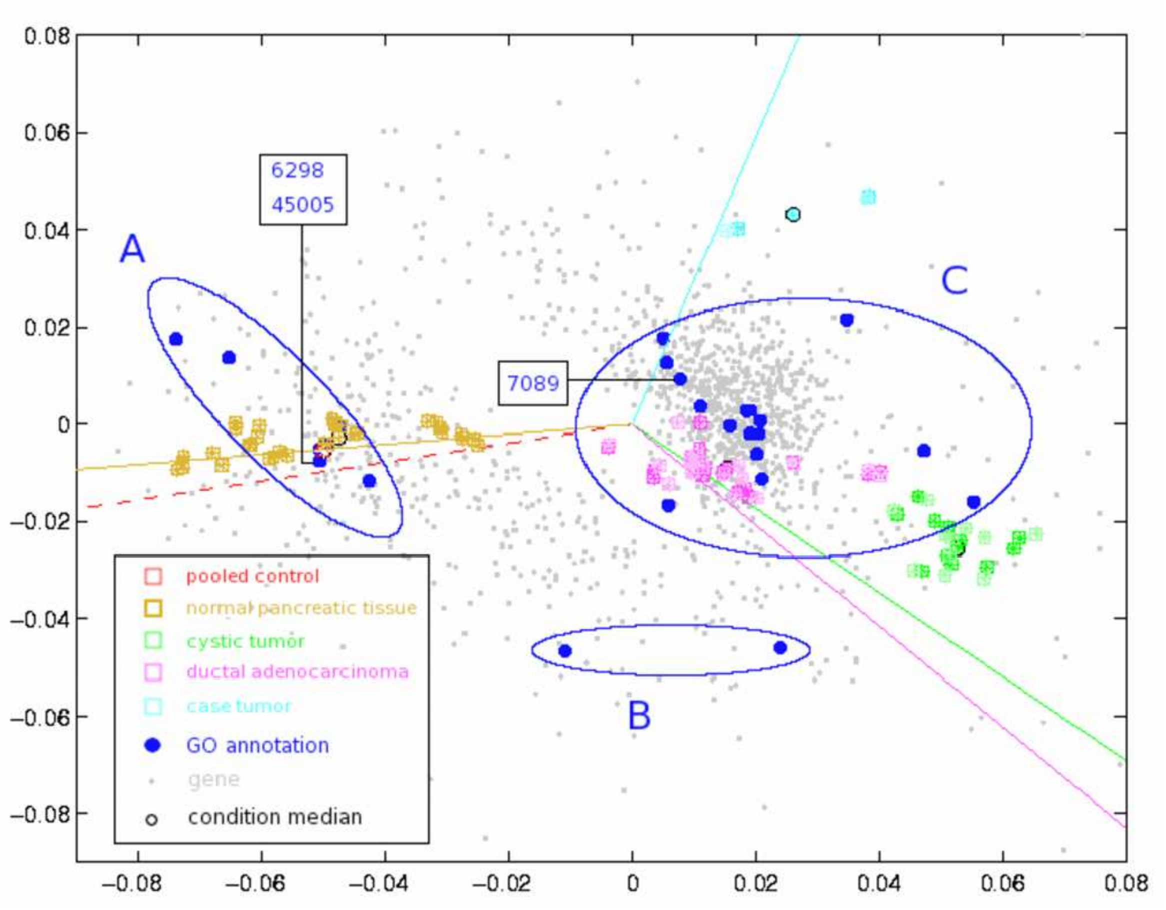


Figure 2.19: CA-Map of human pancreatic cancer study. Comparison of ductal adenocarcinomas (pink), cystic tumors (green) and normal tissue (yellow). A potential new tumor entity is colored cyan. The plot follows the layout of the previous figures: see legend in lower left corner. GO annotations are grouped in 3 different clusters named A, B and C. A comprehensive listing of all displayed annotations is provided in Table 7.1.

Amongst others, the annotations ’mismatch repair’ (GO:0006298) and ’maintenance of fidelity during DNA-dependent DNA replication’ (GO:0045005) are contained in the normal-associated cluster A. The genes annotated to them are MSH2, MSH3 and MLH1. These are DNA mismatch repair enzymes. It is known that mutations in the MSH2 cosegregate with susceptibility to different types of cancer [114, 115]. Moreover, high frequencies of functional inactivation of MLH1 by hypermethylation of the promoter region were found in colorectal cancers [116]. Additionally it has been reported that a loss of function of these repair enzymes could be associated with invasive bladder cancer [117] and that these genes could be potential prognostic factors in colorectal cancers [118, 119].

Annotations in cluster B should be evaluated cautiously. Even though there are several different features (i.e. clones) associated to each of the annotations, effectively there is only 1 gene per annotation, since the distinct annotated clones contain identical genes. Nevertheless, these annotations describe a mechanism, namely DNA methlyation, which just recently has been recognized as playing an important role in the regulation of gene expression. Moreover it already has been shown that abnormal methylation patterns in the promotor regions of genes can lead to the over-/under- expression of the corresponding gene. Various cases of abnormal methylation patterns have been identified in several tumor types [120, 121, 122, 123]. The steadily growing importance of this area of research is clearly demonstrated by the demand for a ’human epigenome project’ [124].

In the tumor associated-cluster C, the annotation ’traversing start control point of mitotic cell cycle’ (GO:0007089) can be found. The comprised genes are CDK10 (represented by 3 different clones), CDC2 (alias CDK1) and CDC25C. It is well known that cyclin dependant kinases play a crucial role in controlling the cell cycle [125], with CDK10 having a potential role in regulating the G2/M phase [126]. The progression through the cell cycle, is a highly regulated ordered series of events. Alteration in the genetic control of the cell cycle can lead to unrestrained cell proliferation. The corresponding mutations occur mainly in two classes of genes: proto-oncogenes and tumor-suppressor genes.

Activation of CDK is regulated through dephosphorylation by members of the Cdc25 phosphatase family. Cdc25A plays an important role at the G1/S-phase transition, Cdc25B undergoes activation during S-phase and Cdc25C activates CDK1-cyclin B during entry into mitosis. Deregulation or overexpression of Cdc25 allows for unscheduled activation of CDK-cyclins and can be associated with tumour formation. Cdc25A and Cdc25B are potential human oncogenes [127]. Cdc25B is overexpressed in 32% of primary breast cancers. Transcription of Cdc25A and Cdc25B genes is activated by c-Myc, an oncogene found to be frequently mutated in human cancers [128]. Raf, a kinase downstream of the frequently mutated Ras oncogene, is able to bind, activate and deregulate Cdc25 protein [129].

2.5 Analysis of transcription factors in CA

GO annotations provide information on regulated functional process, localization of gene products or broader biological objectives - this information is useful to identify general aspects of the experimental condition(s), however it lacks the specificity to identify potential key regulatory elements. One group of genes that are frequently found to exhibit this role of key regulatory genes (proteins) are transcription factors (TFs). In short, transcription factors are proteins that bind to defined DNA sequence-motifs, which commonly are located in the promoter or enhancer region of a gene and thereby regulate its transcription.

2.5.1 Transfac

TRANSFAC is a database storing information on ’eukaryotic transcription regulating sequence elements and the transcription factors binding to and acting through them’ [130]. The ’main’ information can be considered as being the TFs themselves and the sequence motifs they bind to, both of which are stored in separate tables (namely ’Factor’ and ’Site’ respectively). These are connected by a many-to-many relation, since the majority of TFs bind to more than a single site. Besides the actual nucleotide-sequence of the site, information on the corresponding gene, genomic location and a short description of the regulated gene are given in the ’site’ table. The table ’Factor’ provides information on the actual TFs, namely the encoding gene, homologs, organism and known interacting factors. Additionally the Table ’matrix’ stores information on the nucleotide distribution of the binding sites, thus enabling reconstruction of, for instance, a consensus sequence. The ’gene’ table provides information on the regulated genes. The analysis presented here is based on TRANSFAC version 9.3, which is comprised of 16819 distinct sites, 7668 distinct factors and 11910 distinct gene entries.

2.5.2 Integration of transcription factors in CA

The most straightforward approach in the identification of relevant TFs in microarray studies would be to analyze the TF’s expression profiles to find significant regulation. This strategy might not be optimal for several reasons: First of which, the changes in concentration at the mRNA level of the TF might be so small that they are not necessarily identified as significant by microarray analysis, yet have an impact on the expression levels of the target genes. Moreover, it is known that major control mechanisms for the activity of TFs are not based on a change in the mRNA level, but are the result of modifications such as phosphorylation [131], acetylation [132] or ligand binding just to name a few - in other words any postranslational modifications can have an influence on the activity of the TF. These levels of control can’t be assessed with a microarray experiment when analyzing the expression level of the TF. Thus I decided to focus the analysis on the behaviour of the TF’s target genes.

The presence of a TF binding site can be considered as a property of a gene, which would allow for representing the TFs in a CA as supplementary columns as well as rows. In case of the GO annotations the coding of annotation data as supplementary rows has proven superior to supplementary columns and thus the TFs are integrated analogously. In short, for each TF a representative row-profile is calculated, by the row-wise summation of the expression profiles of the genes having a binding-site of this TF: let x_{ij} be the normalized expression intensities for gene $i = 1..n$, in condition $j = 1..m$; $A_k \subset \{1..n\}$ denote the set of genes with a binding site for TF k.$\sum_{i_k \in A_k} x_{i_k j}$ is used as a representative gene profile for TF k. These vectors are added as supplementary rows to the data-matrix (for an example of the encoding please refer to Table 3.1). As with supplementary columns, supplementary rows do not contribute to the computation of the principal axes [83]. The biological validation of this method is described in 2.5.4.1.

2.5.3 Incorporation of ChIP-chip data

Chromatin immunoprecipitation (ChIP) is a method which allows to analyze whether a particular protein binds to a specific DNA sequence *in vivo*. In 2000 Ren et al. [133] combined the ChIP- and microarray-methodology and developed a method for the systematic analysis of protein-DNA interactions. In this approach, the protein of interest (for instance a TF) is immunoprecipitated along with the genomic fragments it is bound to. These fragments are isolated, labeled and hybridized to a microarray (e.g. tiling-arrays). The resulting data identifies the genomic sequences of the binding sites and thereby downstream target genes being potentially regulated by that particular TF. Slight modifications in the ChIP protocol allow to answer different biological questions: if, for instance, polymerases are used for the precipitation areas of active transcription in the genome can be identified.

The TRANSFAC database (version 9.3) provides a table (’fragment’) in which the results of some ChIP-chip experiments are stored: Besides basic information like, the analyzed TF and the corresponding binding sequence, data on the corresponding publications and most importantly on effected genes are provided. Based on this TFs have been matched to the human array described in 2.5.4.2. In Figure 2.22 the TFs derived from TRANSFAC (purple dots) are plotted along with those derived from ChIP-chip experiments (blue dots), a comprehensive listing of these is given in Table 2.13.

Noticeably, all TFs from ChIP-chip data are plotted on the left-hand side of the centroid, with Sp1 (T00759) and RelA (T00594) being closest to the margins of the plot, i.e. being the most differential ones. It is known that Sp1 binds to GC box promoter elements and is activated by, for instance, insulin [134]. The transcription of genes like calmodulin and collagen type I alpha I [135] are regulated by Sp1. RelA (alias p65) is a subunit of the NFkappa-b complex and it has been reported that RelA is constitutively activated in human pancreatic adenocarcinomas cells [136]. Moreover the inhibition of RelA functionality can result in inhibition of tumor cell growth *in vitro* and *in vivo* [137].

Thus it is surprising to find RelA being located in the direction of the ’normal phenotype’ in Figure 2.22. One should note however that the number of associated genes per TF ranges from 103 to 4890 (see Table 2.13), these numbers are very high and it is questionable if all these genes are under regulatory control of a single TF.

2.5.4 Biological validation

Analogous to the GO annotations, the biological validation of the method is a crucial step and thus initial validation of the integration of TFs was based on transcription-data from the model organism *Saccharomyces cerevisiae*. As a subsequent step, to assess the performance in more complex settings, I evaluated the method in context of human microarray data studying the treatment effects on different malignant cell lines.

2.5.4.1 *Saccharomyces cerevisiae*

The data set used in this analysis, is described in detail in 2.4.1 and has been preprocessed analogously. In short, *S. cerevisiae* had been grown in media containing different amounts of glucose (0%, 0.01%, 0.1%, 1%) [98]. The data has been filtered such that, genes were filtered out, whose transcription intensities remain below 10 and/or show a minmax-separation (a quality filter that assesses how well repeatedly measured genes are separated under two different conditions [99]) of less than 0.05 . This results in a total of 1580 genes displayed in the plot. The transcription factors have been filtered such that only those with a minimum of 4 associated genes were submitted to analysis, resulting in 10 different TFs, as shown in Figure 2.20.

In this CA map the predominant variance is between glucose and non-glucose conditions separated along the first principal component, i.e. the non-glucose condition is positioned on the right-hand site of the plot while the rest is found on the opposite site. The glucose containing conditions are ordered in a clockwise orientation of ascending glucose concentration along the second principal component. Here the TFs are depicted as purple circles and for some of the most interesting ones, i.e. those being closest to the margins of the plot, their corresponding TRANSFAC accession numbers are given in the Figure. A comprehensive listing of all plotted TFs is provided in Table 2.11. Two TFs are associated with the non-glucose condition, namely ’T03538’ and ’T03227’. The most differential one of those, which corresponds to the transcription factor CAT8, will be discussed in the following.

CAT8 encodes a zinc-finger cluster protein that mediates derepression of a number of genes during the diauxic shift, which is the transition between fermentative and nonfermentative metabolism [138]. Genomic studies have shown that least 30 genes, encoding proteins involved in gluconeogenesis, ethanol utilization, and the glyoxylate cycle, being regulated by Cat8p [139, 140].

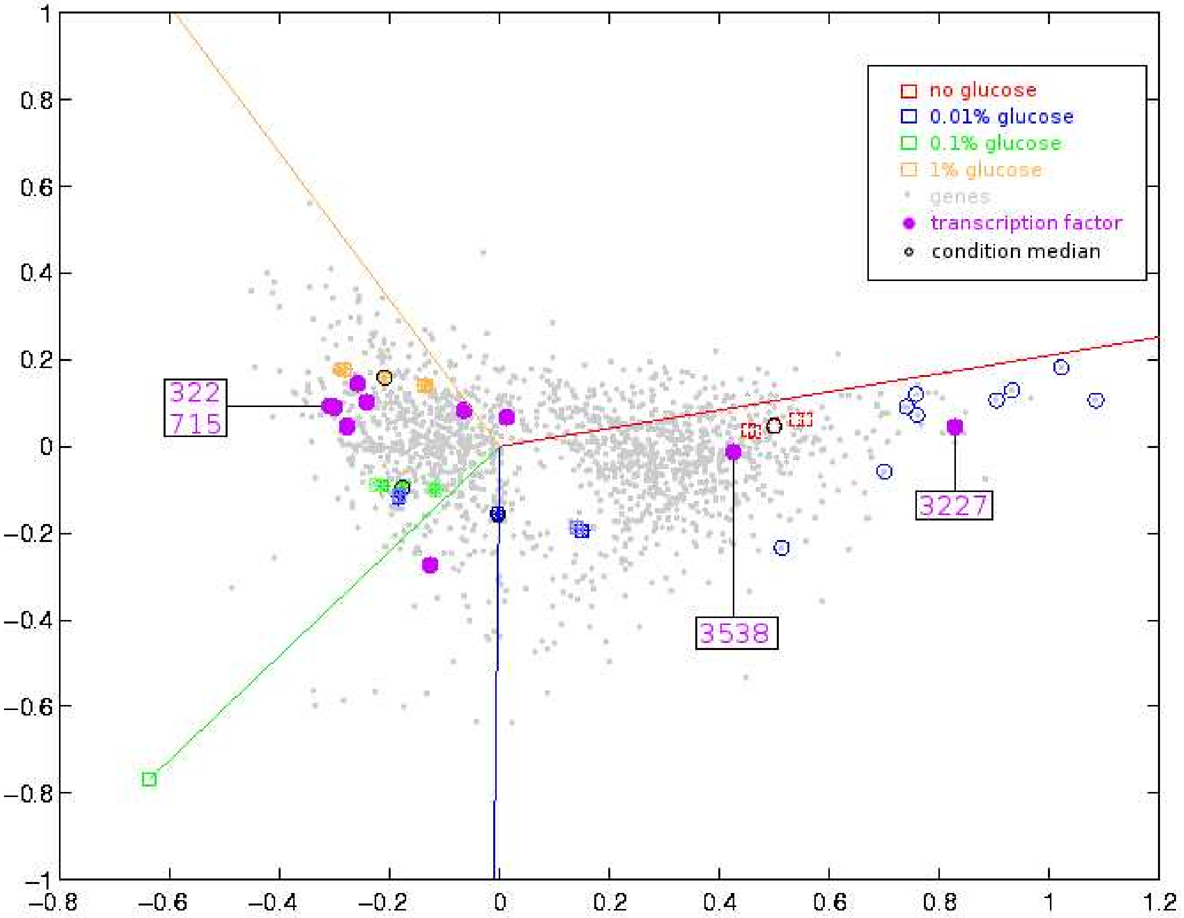


Figure 2.20: CA plot of glucose data-set with transcription factors. TFs are added to the data-matrix as supplementary rows and are depicted as solid purple circles (see also legend in upper right corner). For some of the most prominent TFs the corresponding TRANSFAC-IDs are provided in an adjacent box (the IDs have been truncated of 'T' and trailing zeros). A comprehensive listing of annotations with their corresponding association can be found in Table 2.11. For TF '3227' the annotated genes are marked with blue circles exemplarily, please note that the position of the TF is the center of gravity of these annotated genes. In this plot ≈94% of the total variance in the data set is explained by the two first principal axis.

Under experimental conditions where glucose is available in abundance, expression of CAT8 is repressed by the DNA binding protein Mig1p, which recruits the repressor complex Ssn6p-Tup1p and binds to a site in the CAT8 promoter [138, 141]. When the concentration of glucose in the medium decreases, Mig1p is phosphorylated and transported to the cytoplasm, relieving repression of Cat8p and likely recruiting a transcription activator of CAT8 expression as well [142]. Cat8p functions to derepress the transcription of target genes by binding to the carbon source-responsive element (CSRE) upstream of these genes [141, 143]. While glucose regulates transcription of CAT8, it also appears to regulate Cat8p activity; Cat8p is phosphorylated in derepressed cells and addition of glucose triggers dephosphorylation [141].

The genes that are associated to CAT8 are exemplarily encircles in blue in Figure 2.20, please note that the position of the TF is the center of gravity of the corresponding genes. Table 2.10 provides an overview of the corresponding genes.

For example, the gene YLR377C (FBP1) is known to be involved in gluconeogenesis which is the process whereby glucose is synthesized from non-carbohydrate precursors. Gluconeogenesis mediates the conversion of pyruvate to glucose, whereas the opposite pathway, the formation of pyruvate from glucose, is known as glycolysis. FBP1 catalyzes the reaction from fructose-1,6-bisphosphate to fructose-6-phosphate [144]. Moreover it is known that the transcriptional regulation is effected through consensus sequences in the FBP1 promoter region for the repressor Mig1p and the derepressing Cat8p [145, 144]. Both of these enzymes are TFs and are displayed in Figure 2.20, whereas MIG1 is associated with the glucose containing conditions, and thus being in agreement with literature [145].

Moreover, on the opposite site of the centroid TFs like ABF1 and GCR1 can be found, indicating an upregulation in the presence of glucose. ABF1 is known to be a a site-specific DNA-binding protein, that binds to the consensus sequence: 5-TnnCGTnnnnnnTGAT-3 [161]. The genes that are transcriptionally regulated by ABF1 are involved in diverse cellular processes, one major of which being carbon source regulation. Examples of genes regulated by ABF1 are ADH1, CDC19, PGK1, ENO1 and ENO2 [162, 163, 164], all being relevant in the process of glycolysis.

Accession number	Name	Short description	Association	Number of features
T00056	ABF1	Genes regulated by ABF1 are involved in a diverse range of cellular processes including carbon source regulation, nitrogen utilization, sporulation, meiosis, and ribosomal function.	glucose	6 (90)
T00322	GCR1	Transcriptional activator of genes involved in glycolysis. DNA-binding protein that interacts and functions with the transcriptional activator Gcr2p	glucose	4 (9)

Accession number	Name	Short description	Association	Number of features
T00509	MIG1	Essential TF involved in glucose repression, by repression of HXT2 and HXT4 in presence of glucose. MIG1 is a C2H2 zinc finger protein similar to mammalian Egr and Wilms tumor proteins.	glucose	6 (-)
T00715	RAP1	Is involved in the transcription activation of genes encoding for ribosomal proteins and glycolytic enzymes [165, 166].	glucose	9 (13)
T00725	REB1	RNA polymerase I enhancer binding protein.	-	5 (26)
T00726	CAR1 repressor	CAR1 is an arginase which is responsible for arginine degradation, its expression responds to both induction by arginine and nitrogen catabolite repression [167, 168]	glucose	6 (20)
T01286	ROX1	Heme-dependent repressor of hypoxic genes; contains an HMG domain that is responsible for DNA bending activity [169, 170]	glucose	4 (10)
T03227	CAT8	Is required for positive regulation of gluconeogenesis, for detailed discussion please refer to text.	non glucose	9 (5)
T03491	MED8	Subunit of the RNA polymerase II mediator complex. It associates with core polymerase subunits to form the RNA polymerase II holoenzyme and is essential for transcriptional regulation [171]	low glucose	5 (-)
T03538	RCS1	Involved in iron utilization and homeostasis. Mutants exhibit growth defect on a non-fermenTable (respiratory) carbon source [172]	no glucose	4 (-)

Table 2.11: Comprehensive listing of transcription factors shown in Figure 2.20. Their corresponding Transfac accession number, name, association to experimental condition, as well as the number of associated gene in the given filter setting are listed. The number in given in brackets in the last columns depicts the number of associated genes when using an algorithm for the prediction of TFBS, as described in .

GCR1 binds to the consensus sequence 5-CTTCC-3 and is known to interact with GCR2 [173, 174], both of which have been shown to be transcriptional activators of glycolytic en-

SGD ID	Gene name	short description
YKR097W	PCK1	phosphoenolpyruvate carboxykinase, is a key enzyme in gluconeogenesis, which catalyzes an early reaction in carbohydrate biosynthesis. Glucose represses transcription and accelerates mRNA degradation. PCK1 is regulated by Mcm1p and Cat8p and located in the cytosol [146, 147, 139]
YNL117W	MLS1	malate synthase, is an enzyme of the glyoxylate cycle, which is involved in the utilization of non-fermenTable carbon sources. Its expression is subject to carbon catabolite repression and localizes in peroxisomes during growth in oleic acid medium [148, 149].
YKL217W	JEN1	lactate transporter,which is required for uptake of lactate and pyruvate. Its expression is derepressed by transcriptional activator Cat8p under nonfermentative growth conditions, and repressed in the presence of glucose, fructose and mannose [150, 151, 139]
YLR174W	IDP2	isocitrate dehydrogenase; catalyzes the oxidation of isocitrate to alpha-ketoglutarate. Enzyme levels are elevated during growth on non-fermenTable carbon sources and reduced during growth on glucose [152, 153].
YAL054C	ACS1	acetyl-CoA synthetase, catalyzes the formation of acetyl-CoA from acetate and CoA and is expressed during growth on nonfermenTable carbon sources and under aerobic conditions [154, 155]
YJR095W	SFC1	mitochondrial succinate-fumarate carrier, transports succinate into and fumarate out of the mitochondrion and is required for ethanol and acetate utilization [156, 157]
YOL126C	MDH2	malate dehydrogenase is one of the three isozymes that catalyze interconversion of malate and oxaloacetate. It is involved in gluconeogenesis during growth on ethanol or acetate as carbon source - interacts with Pck1p and Fbp1p [158, 159].
YER065C	ICL1	isocitrate lyase, catalyzes a key reaction of the glyoxylate cycle, namely the formation of succinate and glyoxylate from isocitrate. The expression of ICL1 is induced by growth on ethanol and repressed by growth on glucose [160]
YLR377C	FBP1	fructose-1,6-bisphosphatase, is key regulatory enzyme in the gluconeogenesis pathway, for detailed discussion please refer to text.

Table 2.10: Comprehensive list of genes associated to TF CAT8. The concise gene description was mainly derived from the Saccharomyces Genome Database (SGD).

zymes in *S. cerevisiae* [175, 176]. Null and point mutants have decreased levels of most glycolytic enzymes [177, 178, 179]. The GCR1 mutant of yeast grows at near wild-type rates on nonfermentable carbon sources but exhibits a severe growth defect when grown in the presence of glucose, even when nonfermentable carbon sources are available [179], indicating the importance of this TF in the presence of glucose and thus supporting the positioning of the factor in the CA plot.

Up to now the TFs displayed in CA were solely based on association data stored in TRANSFAC, since this data, even though constantly growing, is yet not abundant, the number of associated genes per TF are rather low (as can bee seen in Table 2.11). This problem becomes more and more severe the more stringent the corresponding gene filters are selected, resulting in fewer and fewer genes to be displayed in the analysis. One way to circumvent this problem is by integration of predicted TF-binding sites. The algorithm used for the prediction is explained in 7.7. Figure 2.21 shows the same data set as in Fig. 2.20, but here the association of TFs and genes is based on the predicted binding sites. As before, only TFs with more than 3 associated genes are shown, resulting in a total of 24 TFs. A list of TFs that are now displayed but have not been listed in Table 2.11 can be found in Table 2.12.

The initial step is to compare the positions of the TFs that are present in both Figures, i.e. that are based on TRANSFAC annotations compared to those based on predicted conserved binding sites in the promoter region. It is apparent, that the predicted TFs are more densely concentrated around the centroid of the map, this is mainly due to the increased number of annotated genes. With larger numbers of genes per TF chances that genes exhibiting non-homogeneous expression profiles will be annotated to the same TF are higher, resulting in a position of the TF closer to the centroid. Nevertheless, for none of the TFs a major change in position (e.g. swap to the different site of the centroid) could be observed. Exemplarily three TFs shared between both Figures have been encircled in black (Fig. 2.21), corresponding to TF ’T03227’ on the right-hand side and ’T00322’ and ’T00715’ on the left-hand side of the plot.

The proximity of resulting positions of TFs based on TRANSFAC and the predicted BS, not only indicate the usefulness of the prediction algorithm but also increases confidence levels of the positioning of the TFs in the plot, by basing it on a larger number of genes. An analysis solely based on the predicted ones would also describe the major regulated processes in this experimental setting. Moreover an additional set of factor becomes submitted to analysis (as listed in Table 2.12), some of which provide an more in-depth picture of the regulated processes.

One example of which is the TF ’T00321’ (GCN4), its positioning in the upper-left part indicates an association to conditions containing high-levels of glucose in the medium. From what is known about GCN4, it is one of the key factors for the general control of aminoacid biosynthesis, by being involved in the derepression of genes from 19 out of the 20 amino acid biosynthetic pathways [180, 181]. Moreover results indicate that it might contribute to processes like purine biosynthesis, organelle biosynthesis, autophagy and glycogen homeostasis

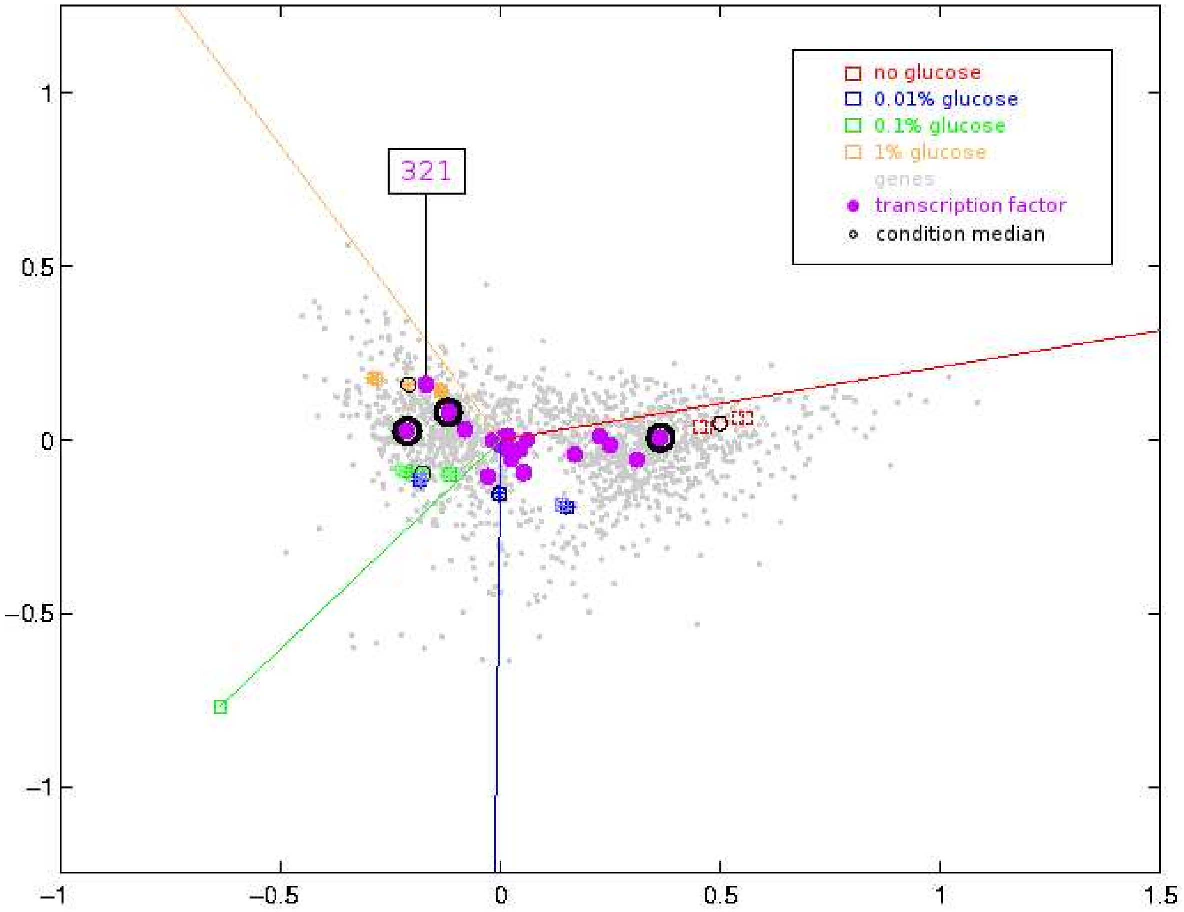


Figure 2.21: CA plot with predicted TFs. Here the displayed TFs are based on the prediction algorithms described in 7.7. TFs are added to the data-matrix as supplementary rows and are depicted as solid purple circles (see also legend in upper right corner). TFs encircled in black are examples of shared TFs between TRANSFAC and prediction based analysis. A listing of additional displayed annotations can be found in Table 2.12.

TRANSFAC accession	Factor name	Number of associated features
T00011	ADR1	157
T00321	GCN4	4
T00346	HAP1	16
T00349	HAP2	5
T00350	HAP3	5
T00351	HAP4	5
T00385	HSF1	545
T00487	MATalpha2	294
T00488	MATa1	28
T00500	MCM1	33
T00772	STE12	143
T00778	TAF	23
T00798	TBP	461
T01257	MSN2	18
T01258	MSN4	18
T01286	ROX1	10
T03525	PDR3	10
T03707	XBP1	63

Table 2.12: List of TFs that are displayed in Fig. 2.21 in addition to those listed in Table 2.11.

[182]. An upregulation of this factor, as indicated by the plot, is also supported by the fact, that in the presence of sufficient amounts of glucose yeast cell invest energy for rapid growth and reproduction, for both of which an upregulated amino-acid-biosynthesis is inevitable.

2.5.4.2 *Homo sapiens*

The human gene expression study comprised samples of the keratinocyte cell line HaCaT which can be genetically and phenotypically divided into five different subtypes. The original nontumorigenic HaCaT (Tetra) cell line [183] was transfected with the Ha-ras oncogene and the cells were injected in mice for tumor growth resulting in benign tumorigenic cells (HaCaT-A5 and HaCaT-I7), and malignant cells that grow locally invasive (HaCaT-II4) or metastasize (HaCaT-A5RT3) [184]. The H-Ras expression is different in the transformed cell lines, low in HaCaT-A5, moderate in HaCaT-I7 and -II4 and very strong in HaCaT-A5RT3. The normalized expression data were filtered with respect to signal intensity and 7289 genes were selected for CA by two-way Anova analysis (P value < 0.05) including both parameters, HaCaT variant and treatment. In the resulting CA plot, the grouping of the experimental conditions corresponds corresponds very well to the intensity of the Ha-Ras expression of the respective cell line, e.g. not or weakly expressed (Tetra, A5), moderately expressed (I7, II4) and strongly expressed (A5RT3, Figure 2.22). In this analysis, a total of 9 different TFs are displayed in the CA plot and listed in Table 2.13. Three TFs, TP53, NF-IL6-2 and NF-kappaB, which had the strongest association to any cell line phenotype are discussed in-depth. The TF most strongly anti-correlated to the parental HaCaT (Tetra) line is NF-IL6-2 (’T00581’), that is positioned on the opposite site of the centroid between the cell lines I7, II4 and A5RT3. The common character of these cell lines is the transformation with the oncogene Ha-Ras and the stronger Ha-Ras expression. Interestingly, NF-IL6-2 (C/EBP) is a known leucine zipper transcription factor CCAAT/enhancer binding protein which is well described to be mediated by Ha-Ras oncogene in keratinocytes and skin carcinomas [185, 186]. The elevated expression level of associated target genes like ICAM1, IL1-beta and SAA2 are strongly associated with tumorigenesis, tumor invasion and cell survival. The second transcription factor NF-kappa B (’T00590’) was strongly associated to the malignant cell line A5RT3 which harbors the potential to metastasize. NF-kappa B induces wound-responsive and inflammatory response genes, and regulates anti-apoptotic processes. Increasing evidence supports the hypothesis that secondary inflammatory responses could be responsible for the invasive potential of cancer cells [187, 188, 189]. The activation of the nuclear factor NF-kappa B was observed in many cancer diseases like breast and ovarian tumors [190, 191]. In the initial phase of tumorigenesis in the epidermis, Ras-initiated keratinocytes progress to a cancerous state if NF-kappa B is blocked [192]. The NF-kappa B inhibition seems to be changed in advanced tumor progression. For example, in squamous cell carcinoma the activation of NF-kappa B along with genes involved in proliferation, angiogenesis and metastasis was found to be dependant on Wnt-signaling in a mouse model [193]. The third described TF is the well known tumor suppressor gene TP53 (’T00671’). In line with the described functionality of the gene, this TF was shown to be closest related to the parental HaCaT cells and HaCaT-A5 cells representing the nontumorigenic and one of the benign phenotype. In response to DNA damage or cellular stress, TP53 signal transduction enhances arrest of cell cycle progression or induces apoptosis [194]. Although, mutations in TP53 were found in all variants of HaCaT, TP53 independent mechanism are known to activate relevant downstream targets like p21 which results in cell cycle arrest and apoptosis [195, 196]. Moreover, the finding supports the early observation, that in the HaCaT cell line, the Ha-Ras-transformation and TP53 mutations are not sufficient to induce a malignant phenotype [184, 197]. Further genetic events, like the changes in WNT-signaling, TGF-signaling and NF-kappa B signaling are necessary to result in malignant and metastasizing phenotypes.

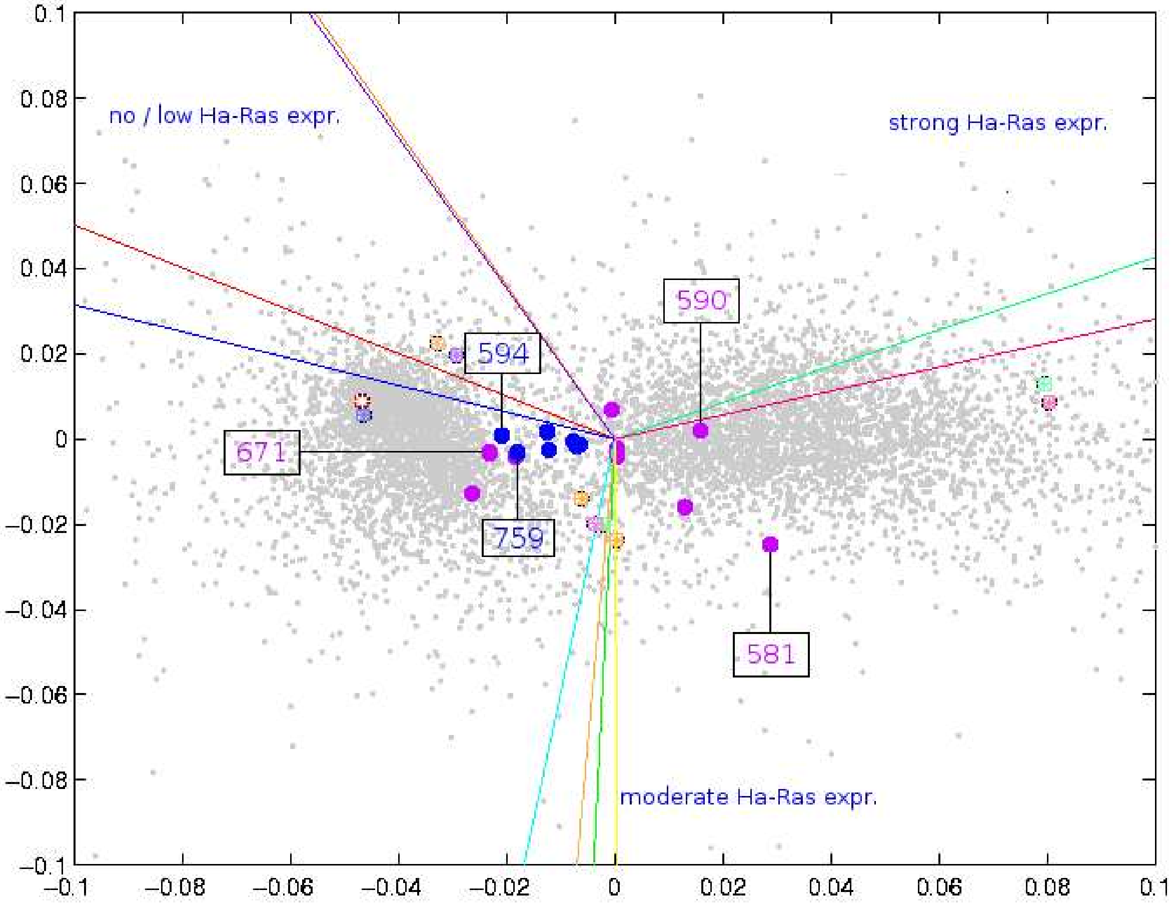


Figure 2.22: CA plot of human cancer data set with TFs. The experimental conditions can be roughly divided into three groups based on their Ha-Ras expression: weak, moderate, strong. This phenotypical grouping corresponds to the clustering of conditions that can be seen in the plot. Here TFs have been added to the data-matrix as supplementary rows. TFs derived from TRANSFAC are depicted as purple colored dots, while TFs derived from ChIP-Chip experiments are depicted as blue circles. Exemplarily corresponding TRANSFAC accession IDs are provided (being cut of from trailing zeros and 'T') for individual TFs. In the following the cell-lines corresponding to the displayed conditions will be given in a clockwise orientation, starting with the blue line: HaCaT(Tetra-DMSO), HaCaT(Tetra), HaCat(A5), HaCat(A5-DMSO), HaCaT(A5RT3), HaCaT(A5RT3-DMSO), HaCaT(II4-DMSO), HaCaT(I7), HaCaT(I7-DMSO), HaCaT(II4).

TRANSFAC accession	Factor name	Number of associated features
T00029	AP-1	7
T00163	CREB	8
T00261	ER-alpha	7
T00581	NF-IL6-2	5
T00590	NF-kappaB	10
T00593	p50	5
T00594	RelA	5
T00671	p53	15
T00759	Sp1	23
T00368	*HNF-1alpha-A*	*732*
T03286	*HNF-6alpha*	*825*
T03828	*HNF-4alpha*	*4890*
T00140	*c-Myc*	*451*
T00759	*Sp1*	*311*
T00163	*CREB*	*103*
T00594	*RelA*	*124*

Table 2.13: Comprehensive List of TFs display in Fig. 2.22. TFs are listed with their TRANSFAC accession, name and number of associated features in the chosen filter setting. The last TFs (written in italics) are derived from ChIPChip-data.

3 Discussion

3.1 Gene annotation data

In recent years microarrays have become a standard technique to analyze complete transcriptomes, resulting in large amounts of data. Consequently a wealth of methods has been developed to identify significantly regulated genes. Correspondence Analysis, a projection method, has already been successfully applied in the visualization of microarray data and extraction of relevant genes [198, 199]. The ultimate goal of any experiment, however, is to develop new or validate existing hypothesis. This step currently poses a major bottleneck in microarray data analysis.

In order to deduce any biological hypothesis from microarray data, functional annotations for genes have to be gathered and analyzed. Since the majority of available methods presents the set of regulated genes in long lists, the subsequent functional interpretation is commonly done by eye in a spread-sheet like data format, rendering this step not only time-consuming but also prone to errors.

The analysis shown in this thesis are based on an in-house analysis software called M-CHiPS in which the expression profiles and experimental annotations are stored in an underlying database. Even though columns that hold gene annotation data exist, these are in general sparsely populated. Moreover, since no constrains for these columns are defined, the annotation data are stored as free text, interfering with a statistical analysis of gene properties.

In this work Gene Ontology was chosen as the major source of gene annotations for several reasons: First of all, being structured as an ontology the individual annotation terms (or concepts) are predefined. In other words the concepts available for annotation are based on a set of controlled vocabularies. Since for each of these concepts an unambiguous definition is provided, the risk of misinterpretation of the meaning of concepts is minimized and thus allowing for cross-laboratory annotation process. Moreover, each concept can be accessed by a unique identifier, which provides the basis for a statistical analysis of gene annotation data.

This annotation data is distributed through flat files from the GO consortium. These files as such are not optimal for immediate functional interpretation. In the annotation process each gene product is associated to the term describing its functionality most specifically and it is this most specific association(s) that are provided in the files. If solely these associations were used, the information that is stored in the hierarchy of the ontology would be ignored. Thus I associate each gene to all parental concepts of its original annotation as well, thereby fully exploiting the structure of the ontology. Only by expanding the associations in this way, more general concepts become applicable for the analysis by increasing the number of associated genes to a statistically relevant level.

There are, however, areas in which GO should be improved: Smith et al. [200] reported examples of semantic errors in the ontology, which are likely to become more frequent as the ontology becomes more complex. These errors could result from inappropriate construction of parts of the ontology or from imprecise definitions of the concepts and their relations. To ensure semantic consistency of an already complex and still growing ontology, the use of software solutions is inevitable [201]. Another critical aspect is the varying quality of the definitions of concepts and relations. Ambiguous or even circular definitions can result in non-consistent usage of annotations terms - software solutions that address this problem are being developed [202].

Overall, however, GO has proven to be an excellent source of annotation data, well suitable for functional interpretation of high-throughput data. At this point it already has become a *de facto* standard for the annotation of gene products, which is well documented by more than 900 publications on the development or usage of GO (as of August 2006). With growing numbers of annotation terms as well as annotated genes, the usefulness of GO will still increase.

In summary, using GO as a source of information for the annotation of microarray data combines the benefits of an extensible human-curated, cross-species database with the capability to statistically analyze the annotations. This renders it the currently best choice for functional interpretation of microarray data.

3.2 Comparison of implementations

In the course of this work different ways of integrating functional annotations in CA have been implemented and compared. In an initial approach the annotations have been coded as Boolean variables and added to the data matrix as supplementary columns. Here the annotation vector holds a ’1’ if the annotation is associated to the corresponding gene and a ’0’ if not (an example of the encoding is given in Table 3.1).

A CA plot based on this encoding is shown in Figure 2.5, in which the annotations are predominant over all other aspects of the plot. Here the genes and experimental conditions are restricted to a rather small area around the centroid, leaving the plot not well interpretable. The reason for this is the Boolean nature of the (annotation-) vector. With only two possible values for each position the relative changes between the state ’annotated’ and ’not annotated’ will always be maximal. Since CA is sensitive for relative changes, these extreme differences will result in positions close to the standard coordinates (i.e. the positions with the maximal association of a column to row - or *vice versa*). In case of ’normal’ transcription intensities the relative changes between the experimental conditions will be much smaller. Thus genes and hypridizations are plotted closer around the centroid, resulting in the predominance of the Boolean vectors.

In a subsequent approach annotations have been represented as row-vectors: for each annotation a representative row profile is calculated by the rowwise sum of the expression intensities of the annotated genes. The last rows of Table 3.1 show this encoding exemplarily. Here an annotation is represented based on ’natural’ expression intensities, such that the resulting relative changes will be in the same order of magnitude as for the rest of the data. Figure 2.6 gives an example of a CA plot based on this encoding, where neither of the variables is predominant. Here the majority of annotations is plotted around the centroid of the map. This concentration is to be expected in such an experimental setting: in general, only few functional processes are differentially regulated between the experimental conditions, consequently only few annotations should be positioned near the margins of the plot. Exceptions from this are settings in which complete developmental cycles of organisms (e.g. *Drosophila melanogaster* [203]) are analyzed. In these settings large numbers of genes are temporarily activated.

Having established a suitable encoding, this has been validated for its biological applicability on two data sets of different complexity: the first one analyzes the functional changes in the transcriptome of *S. cerevisiae* when grown in media with different concentrations of glucose (2.4.1), whereas the second focuses on differences between human tumor types (2.4.2).

In the resulting CA plot of the yeast data set (Figure 2.6 on page 23) a cluster of annotations describing the transportation of glucose into the cell is clearly distinguishable. This already points out a major functional difference between the experimental conditions, since in the absence of glucose the corresponding transporters do not need to be expressed at high levels. After filtering of the annotations further terms describing relevant biological processes become apparent. These include concepts like ’glucosidase activity’, ’glycolysis’ or ’TCA cycle’, which describe the different pathways for energy-utilization of *S. cerevisiae* being alternatively activated, depending on the availability of glucose. These results already indicate the usefulness of analyzing microarray data in context of GO, by providing researchers with an immediate characterization of the major pathways being differentially regulated between these experimental conditions.

To further validate this approach, GO annotations were used in the analysis of a more complex organism and experimental setting, namely a human cancer study. Here different tumor samples where compared to normal tissue. In the resulting CA plot (Figure 2.19 on page 50) an upregulation of genes involved in DNA mismatch repair in the normal tissue can be immediately identified. It is well known that their upregulation, or more precisely the lack of expression in cancerous samples, is one of the key events in the development of cancer. Moreover annotations pointing to mechanisms of enhanced cell proliferation and inflammatory-like responses are associated with the tumor samples. Unrestrained cell proliferation is a well known characteristic of tumors and chronic inflammation has been reported as a risk factor in various cancers [204, 205].

Additionally annotations describing transcriptional regulation due to changes in methylation

Gene	Exp. cond.1	Exp. cond. 2	Exp. cond. 3	Exp. cond. 4	Term 1	Term 2	.
A	13	300	23	432	1	0	.
B	457	398	355	932	0	1	.
C	24	458	44	364	1	1	.
D	324	245	98	34	0	0	.
E	478	928	293	99	0	1	.
F	38	485	21	375	1	0	.
.	.	.	.	.	.	.	.
Term 1	75	1243	88	1171			
Term 2	959	1784	692	1395			
.	.	.	.	.			

Table 3.1: Ways of adding supplementary information to the data matrix. Here a schema of an artificial data-matrix for a simplified (i.e. only one repetition per exp. condition) microarray experiment is given. The last columns examplarily represent two annotations added as supplementary columns with the individual cells holding Boolean values. The last rows are examples of the same two annotations, now being added as supplementary rows to the data matrix. The representative row-profile is calculated by summing up the expression intensities of the annotated genes.

patterns were associated to the most aggressive types of tumors in this study. Just in the recent years the methylation status of CpG-islands in the promotor regions of genes has been identified as an important regulatory mechansim of the transcriptional activity of the associated gene(s). In subsequent studies it could be shown that aberrant methylation patterns are often associated with the development and/or occurrence of tumors [120, 121, 122, 123]. Hence the association of this annotation to the aggressive tumor type indicates potential relevance of these mechanisms.

In summary, integration of annotation data as supplementary rows in CA, generates well interpreteable plots in which relevant functional processes can be immediately identified. Based on this the researcher can deduce functional hypothesis in a fast and intuitive way, without the need for comparing long lists of annotations.

3.3 Applicability of annotation filters

For most of the genomes the numbers of available annotations are already so high that an unrestricted display will result in a massive overlay of annotations, as demonstrated for the *S. cerevisiae* data set (Figure 2.10 on page 32).

In order to increase the clarity of the plot and display only relevant annotations different filters have been tested. The first and probably most intuitive way of filtering is based on the categorization of the GO in three main ontologies, namely ’Molecular Function’, ’Biological

Process’ and ’Cellular Component’. In experimental settings where the localization of the gene products is not of relevance, the number of annotations can be reduced by deselection of the corresponding category by approximately 9%. This corresponds to the overall percentage of concepts in the ’Cellular Component’ ontology. Larger reductions of displayed annotations can be achieved by removing one of the remaining ontologies (i.e. ≈37% in case of ’Molecular Function’ and ≈54% for ’Biological Process’).

The number of associated genes per annotation term can be used as a filter criterion as well. Here annotations having less than a minimal or more than a maximal threshold will be discarded from the analysis. I recommend to use at least the threshold defining the minimal number of genes per annotation: If not applied large numbers of annotations will be only associated to one or two genes. Since the position of these annotations is calculated based on one or two transcription profiles these annotations tend to be plotted near the margins of the plot, which normally would indicate relevance in the experimental context. These annotations, however, are generally not suitable for functional interpretation for mainly two reasons: First of all, annotations with low numbers of associated genes tend to be rather specific and thus do not have the necessary level of abstraction to provide a functional overview. More importantly, however, these low numbers are a poor statistical basis to derive any functional hypothesis from. The filter for the maximal number of genes is useful to discard annotations describing too general processes like ’response to stimulus’, for which the number of genes can be up to several hundreds. Since, these annotations are likely to be plotted near the centroid of the map and thus are marked as ’non-interesting’ they should be filtered out beforehand to further enhance the clarity of the plot.

The evidence codes provided along with each annotation of a gene product can be used as a further filter criterion. These codes describe the kind of data (i.e. evidence) the annotation is based on. Since the evidence mainly varies in the level of confidence of the experimental data, they could be perceived as a measure of quality for the annotation. Since there is no unified quality-based ordering of these codes a rough ranking is proposed as shown in Figure 2.11. While the quality-differences between ’traceable author statement’ and ’inferred from electronic annotation’ are obvious, a ranking of ’inferred from genetic interaction’ and ’inferred from mutant phenotype’ is less intuitive. These corresponding filter option has been implemented in the user-interface (see Figure 7.2 on page 100) with the codes being listed in descending order of quality. Each of the codes can be individually (de-)selected allowing to choose any combination of codes and thus accounting for any possible (user-defined) ranking. One of the most efficient filters to decrease the number of terms, is to discard annotations that are ’inferred from electronic annotation’. In case of *homo sapiens* this results in a reduction of more than 34% of the available annotations. While this filter setting improves the confidence for the remaining annotations, it is advisable to perform at least one analysis including all annotations, to fully exploit the available data. Here one should keep in mind, however, that a thorough validation of the functional properties is inevitable.

Another filter criterion is based on the structure of the ontology, it accounts for the distance of an annotation to the root node of the ontology. With this filter the abstraction level of the plotted terms can be controlled. Generally speaking, the higher this distance is, the more detailed the resulting annotations are. One should be aware, however, that the distance itself can not be used as an absolute measure for information content, in the sense that terms having a distance of six carry twice the amount of information compared to those at a distance of three. Even the content of information that is captured by concepts at the same level of the ontology can differ drastically, for example: ’*glutathione dehydrogenase (ascorbate) activity*’ (GO:0045174) and ’*enzyme activator activity*’ (GO:0008047) both have a distance of four to the root node, but the specificity (i.e. information content) of the terms clearly is not the same. The level of specificity at a certain depth in the ontology is influenced to a great extent by the information available in that area. When applying this filter, the heterogeneity of the ontology should always be kept in mind, especially when used in combination with the ’minimal number of genes’ filter.

Finally a filter based on the homogeneity of the expression profiles of the annotated genes is presented. Here correlation coefficients along with different ways two summarize them were applied and tested. Their performance was assessed by ROC curves based on a predefined set of ’standard annotations’, the best overall performance was achieved by calculating the mean of all pairwise Spearman’s correlation coefficients (ρ) of the expression profiles of the annotated genes. Since Spearman’s ρ is based on the ranks, rather than the actual intensities, it does not assume normal distribution of the data, making it applicable to microarray data which is known to be heavily right-skewed [95, 96]. In this setup annotations comprised of anti-correlated genes will be filtered out due to negative coefficients. Since I consider anti-correlation as potentially interesting as well, I account for this by taking the mean of the absolute values of the pairwise ρ. One should note, however, in cases where half of the genes of an annotation are anti-correlated to the other half, this annotation will be plotted near the centroid of the map, despite high values of ρ. The application of this filter results in a large decrease of displayed annotations, especially at high thresholds (0.85 - 1). Here one should be aware that the resulting set of annotations is biased towards annotations with low numbers of genes, commonly two to four. This is due to the fact that the higher the number of annotated genes, the higher the probability for an annotated gene having a deviating expression profile, which consequently results in a lower ρ. As a side effect the annotations selected by this filter tend to be rather specific in the information content (i.e. high distance from the root node), such that reasonable thresholds (0.6 - 0.8) in combination with a sensible ’number of genes filter’ give the most promising results. The usefulness of this filter could be demonstrated in the analysis of yeast as well as human data sets. In both cases the application of the correlation coefficient filter resulted in well interpretable plots displaying reasonable numbers of annotations, but only after application of this filter, annotations describing relevant functional properties became apparent.

3.4 Integration of transcription factors

The integration of annotation data in CA has proven to generate well interpretable plots in which the associations of annotations to experimental conditions or clusters of genes can be used to describe their functional characteristics. Even though this approach is very powerful in identifying common properties, these commonly relate to more general concepts and processes. The identification of key regulatory elements being responsible for the observed differences in expression is not feasible by this approach.

Here, transcription factors (TFs) represent a group of genes which often act as key regulatory elements. TFs are proteins that bind to specific DNA sequence motifs and thereby regulate the transcription activity of a nearby gene. Theoretically the relevance of a TF in an experimental setting can be assessed by two approaches. One of which focuses on the behavior of the TF itself, the second one on the changes in the TF's target genes. The former approach poses several problems: small changes in the expression level of the TF, which are undetectable by microarray technology, can still effect the expression of the target genes. More importantly, however, the functionality of TFs is often dependent on post-translational modifications, such as phosphorylation, acetylation or the binding of a co-factor. These effects are not detectable at the mRNA level and thus I decided to assess the relevance of a TF by the transcriptional changes of it's target genes.

To this end the TRANSFAC database has been used as a source for the association of transcription factor binding sites (TFBS) to genes. The structure of this association is similar to those derived from the GO and thus the TFs are integrated analogous: For each TF a representative expression profile is calculated based on the annotated genes and added to the data matrix as supplementary rows. This approach was subsequently validated for its usefulness on a *S. cerevisiae* and a human cancer data set.

In case of the yeast data, which compares experimental conditions with varying concentrations of glucose, TFs such as CAT8, ABF1 and GCR1 were identified as being differentially regulated. It is known that CAT8 is responsible for the derepression of multiple genes during the diauxic shift. Target genes include FBP1 and PCK1, both of which are known to be key enzymes in the gluconeogenesis pathway [144, 147], such that the observed upregulation in the 'non-glucose' condition is in agreement with literature. Furthermore ABF1 and GCR1 were upregulated in glucose containing conditions, this is supported by prior findings, reporting both TFs as activators of glycolytic enzymes [162, 163, 164, 175, 176].

The analyzed human cancer data, compares HaCaT cell lines that were transfected with the Ha-Ras oncogene. Here target genes of TFs such as C/EBP and NF-kappa B were identified as being differentially expressed. C/EBP (NF-IL6-2) is known to be mediated by Ha-Ras in keratinocytes [185, 186]. Interestingly NF-kappa B is associated to the cell line A5RT3 which harbors the potential to metastasize. This indicates that an initial NF-kappa B inhibition [192] could be changed in advanced tumor progression. The activation of NF-kappa B can be observed in many cancer diseases like breast and ovarian tumors [190, 191].

For a number of TFs, however, only a few validated target genes are available, which is a non-optimal statistical basis to draw any conclusions from their positioning in a CA plot. To circumvent this the binding matrices provided in the TRANSFAC database were used to predict potential binding sites in the upstream regions of genes (7.7). With this approach the number of associated genes/TF could be increased, at the cost of decreasing the level of confidence. In case of the *S. cerevisiae* integration of the predicted sites, not only supported the prior findings (solely based on TRANSFAC data) but also allowed to identify previously undetected TFs being relevant in the given experimental context. In case of the human data set, however, the positioning of the TFs based the predicted binding sites was not in every case in accordance with TRANSFAC data. Moreover, some of the plotted associations of TFs to experimental conditions could not be explained with current knowledge. Both findings indicate the need for improvement in case of the human data set. Here a refinement of the prediction algorithm or the application of quality filters for the predicted prior to visualization could render this approach applicable.

In 2000 Ren et al. [133] published the Chromatin-immunoprecipitation-chip (ChIP-chip) that allows to identify DNA-protein interactions on a genome wide scale. With this the target sequence(s) of a DNA binding protein are identified along with downstream genes, that are potentially regulated by the corresponding protein. One of the first class of proteins to be analyzed by this method are transcription factors. I integrated the ChIP-chip data analogous to the data from TRANSFAC, such that each TF is represented by a summary profiles of its annotated genes. In this data set all TFs were plotted on one side of the centroid, even though, from a biological point of view, this association does not make sense for all TFs. This positioning is very likely due to the very large numbers of genes being associated with each TF, in this setting up to ≈4300 genes/TF. Since the position of the TF represents the centroid (i.e. weighted average) of the annotated genes, it is plotted in the direction of the majority of the genes. This is very likely the reason why all TFs, based on ChIP-chip data, can be found on the left-hand side of the centroid in Figure 2.22, regardless of their biological function. Since it is questionable, that over 4000 genes are under the control of a single TF I rather expect a high false-positive rate in the ChIP-chip data. Thus it is advisable to further validate these gene-TF associations before submitting them to subsequent analysis.

In summary, the presented method provides a mean to visualize microarray data in context of transcription factor binding sites. The resulting plot allows for an intuitive identification of TFs being relevant for the transcriptional changes between the chosen experimental conditions.

3.5 Future prospects

As the integration of gene annotation data in CA has proven as a powerful method to identify relevant functional processes (in case of GO) or even target individual genes as key regulatory elements (in case of integrated TFs), a subsequent step is to expand this approach to other data sources as well. Here data stored in KEGG could be of immediate use in identifying regulated pathways, even though this will exhibit some overlap with the data provided by GO, I expect that some processes will be represented more precisely due to the higher resolution of KEGG. One example of this would be the highly interrelated processes of glycolysis and gluconeogenesis: in both pathways the majority of reactions is catalyzed by the same set of genes (either in forward or reverse direction), only for a few thermodynamically irreversible reactions a different enzyme is utilized. Due to this large overlap, it will be hard to differentiate the two processes based on GO categories, but could prove feasible based on data from KEGG.

Furthermore, information on genomic localization can help to identify genes being co-regulated based on their vicinity in the genome or even to identify potential chromosomal deletions. Besides the mere localization, complete sequences from the upstream regions of genes (or selected motifs thereof) could be submitted to the analysis. From a medical/pharmaceutical point of view it is most interesting to integrate data on the disease-relevance of genes. In general almost every gene annotation data available could be added to a CA plot to facilitate deduction of biological hypothesis.

Besides the visualization of annotation data as an aid in functional interpretation of microarray data, integration of external data sources provides a basis for more accurate reconstruction of regulatory networks as compared to microarray data alone. To this end database structures and analysis methods should be developed that can store and integrate data on, for instance, protein levels (i.e. derived from two-dimensional gels or the upcoming protein arrays). To further facilitate the development of sound network topologies protein-protein interaction data should be integrated as well.

4 Summary

DNA microarrays have become a standard technique to assess the mRNA levels for complete genomes. To identify significantly regulated genes from these large amounts of data a wealth of methods has been developed. Despite this, the functional interpretation (i.e. deducing biological hypothesis from the data) still remains a major bottleneck in microarray data analysis. Most available methods display the set of significant genes in long lists, from which common functional properties have to be extracted. This is not only a tedious and time-consuming task, which becomes less and less feasible with increasing numbers of experimental conditions, but is also prone to errors, since it is commonly done by eye.

In the course of this work methods have been developed and tested, that allow for a computer-based analysis of functional properties being relevant in the given experimental setting. To this end the Gene Ontology was chosen as an appropriate source of annotation data, because it combines human-readability with computer-accessibility of the annotations term and thus allows for a statistical analysis of functional properties.

Here the gene-annotations are integrated in a Correspondence Analysis which allows to visualize genes, hybridizations and functional categories in a single plot. Due to the increasing amounts of available annotations and the fact that in most settings only few functional processes are differentially regulated, several filter criteria have been developed to reduce the number of displayed annotations to a set being relevant in the given experimental setting.

The applicability of the presented visualization and filtering have both been validated on datasets of varying complexity. Starting from the well studied glucose-pathway in *S. cerevisiae* up to the comparison of different tumor types in human. In both settings the method generated well interpretable plots, which allowed for an immediate identification of the major functional differences between the experimental conditions [90].

While the integration of annotation data like GO facilitates functional interpretation, it lacks the capability to identify key regulatory elements. To facilitate such an analysis, the occurrence of transcription factor binding sites in upstream regions of genes has been integrated to the analysis as well. Again this methodology was biologically validated on *S. cerevisiae* as well human cancer data sets. In both settings TFs known to exhibit central roles for the observed transcriptional changes were plotted in marked positions and thus could be immediately identified [206].

In essence, integration of supplementary information in Correspondence Analysis visualizes genes, hybridizations and annotation data in a single, well interpretable plot. This allows for an intuitive identification of relevant annotations even in complex experimental settings. The presented approach is not limited to the shown types of data, but is generalizable to account for the majority of the available annotation data.

5 Zusammenfassung

DNS-Chips (’Microarrays’) haben sich zu einer der Standardmethoden zur Erstellung von genomweiten Expressionsstudien entwickelt. Mittlerweile wurden dazu eine Vielzahl von Methoden zur Identifizierung von differentiell regulierten Genen veröffentlicht. Ungeachtet dessen stellt die abschliessende funktionelle Interpretation der Ergebnisse einen der Engpässe in der Analyse von Chip-Daten dar. Die Mehrzahl der Analysemethoden stellt die signifikant regulierten Gene in Listen dar, aus denen in einem weiteren Schritt gemeinsame funktionelle Eigenschaften abgeleitet werden müssen. Dies stellt nicht nur eine arbeitsintensive Arbeit dar, die mit steigender Anzahl an experimentellen Konditionen immer weniger praktikabel wird, sondern ist auch fehleranfällig, da diese Auswertung im allgemeinen auf dem visuellen Vergleich von Listen beruht.

In der vorliegenden Arbeit wurden Methoden für eine rechnergestützte Auswertung von funktionellen Geneigenschaften entwickelt und validiert. Hierzu wurde die ’Gene Ontology’ als Quelle für die Annotationsdaten ausgewählt, da hier die Daten in einem Format gespeichert sind, das sowohl eine leichte menschliche Interaktion sowie die statistische Analyse der Annotationen ermöglicht.

Diese Genannotation wurden als Zusatzinformationen in die Korrespondenzanalyse integriert, welches eine simultane Darstellung von Genen, Hybridisierungen und funktionellen Kategorien in einer Grafik ermöglicht. Aufgrund der ständig wachsenden Anzahl an verfügbaren Annotationen und der Tatsache, daß zwischen den meisten experimentellen Bedingungen nur wenige funktionelle Prozesse differentiell reguliert sind, wurden Filter entwickelt, die die Anzahl der dargestellten Annotationen auf eine im gegebenen experimentellen Kontext relevante Gruppe reduzieren.

Die Anwendbarkeit der Visualisierung und der Filter wurde auf Datensätzen unterschiedlicher Komplexität getestet: beginnend mit dem gut verstandenen Glukosestoffwechsel im Modellorganismus *S. cerevisiae,* bis hin zum Vergleich unterschiedlicher Tumortypen im Menschen. In beiden Fällen generierte die Methode gut zu interpretierende Grafiken, in denen die funktionellen Hauptunterschiede durch die dargestellten Annotationen gut beschrieben werden [90].

Während die Integration von Annotationsdaten wie GO die funktionelle Interpretation vereinfacht, fehlt die Möglichkeit zur Identifikation einzelner relevanter Schlüsselgene. Um eine solche Analyse zu ermöglichen, wurden Daten zum Vorkommen von Transskriptionsfaktorbindestellen in den 5’-Bereichen von Genen integriert. Auch diese Methode wurde an Datensätzen von *S. cerevisiae* und vergleichenden Studien von humanen Krebszelllinien validiert.

In beiden Fällen konnten Transkriptionsfaktoren identifiziert werden, die für die beobachteten transkriptionellen Unterschiede von entscheidender Bedeutung sind [206].

Zusammenfassend, ermöglicht die Integration von Zusatzinformationen in die Korrespondenzanalyse eine simultane Visualisierung von Genen, Hybridisierungen und Annotationsdaten in einer einzigen, gut zu interpretierenden Grafik. Dies erlaubt auch in komplexen experimentellen Bedingungen eine intuitive Identifizierung von relevanten Annotationen. Der hier vorgestellte Ansatz, ist nicht auf die gezeigten Datenstrukturen beschränkt, sondern kann auf die Mehrzahl der verfügbaren Annotationsdaten angewendet werden.

6 Publications

Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data.
Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD, Fellenberg K.
Bioinformatics. 2005 May 15;21(10):2424-9. Epub 2005 Mar 3

Visualizing expression patterns in context of transcription factor binding sites using Correspondence Analysis
Busold CH, Beissbarth T, Kuner R, Hauser NC, Boukamp P, Schäfer R, Sers C, Lund P, Sültmann H, Poustka A, Hoheisel JD, Fellenberg K.
submitted -BMC Bioinformatics

Expression profiling of glial genes during Drosophila embryogenesis.
Altenhein B, Becker A, Busold C, Beckmann B, Hoheisel JD, Technau GM.
Dev Biol. 2006 May 5; [Epub ahead of print]

Systematic Interpretation of Microarray Data using Experiment Annotations
Fellenberg K, Busold C, Witt O, Bauer A., Beckmann B., Hauser NC, Frohme M, Winter S, Dippon J, and Hoheisel J.
submitted, 2006

The transcriptomes of Trypanosoma brucei Lister 427 and TREU927 bloodstream and procyclic trypomastigotes.
Brems S, Guilbride DL, Gundlesdodjir-Planck D, Busold C, Luu VD, Schanne M, Hoheisel JD, Clayton C.
Mol Biochem Parasitol. 2005 Feb;139(2):163-72.

An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome.
Hild M, Beckmann B, Haas SA, Koch B, Solovyev V, Busold C, Fellenberg K, Boutros M, Vingron M, Sauer F, Hoheisel JD, Paro R.
Genome Biol. 2003;5(1):R3. Epub 2003 Dec 22.

Use of complex DNA- and antibody-microarrays as tools in functional analyses.
Bauer A, Beckmann B, Busold C, Brandt O, Kusnezow W, Pullat J, Aign V, Fellenberg K, Fleischer R, Jacob A, Frohme M, Hoheisel JD.
Comp. Funct. Genom, vol. 4, pp. 520-524, 2003

Bibliography

[1] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[2] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[3] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.*, 93:10614–10619, 1999.

[4] D. Shalon, S. J. Smith, and P. O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, 6:639–645, 1996.

[5] M. Beier and J.D. Hoheisel. Production by quantitative photolithographic synthesis of individually quality checked DNA microarrays. *Nucleic Acids Res*, 28(4):E11, 2000.

[6] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–4, 1999.

[7] http://www.genomic.ch/techno_array.php.

[8] G.K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4): 265–73, 2003.

[9] T. Bammler, R.P. Beyer, S. Bhattacharya, G.A. Boorman, A. Boyles, B.U. Bradford, R.E. Bumgarner, P.R. Bushel, K. Chaturvedi, D. Choi, M.L. Cunningham, S. Deng, H.K. Dressman, R.D. Fannin, F.M. Farin, J.H. Freedman, R.C. Fry, A. Harper, M.C. Humble, P. Hurban, T.J. Kavanagh, W.K. Kaufmann, K.F. Kerr, L. Jing, J.A. Lapidus, M.R. Lasarev, J. Li, Y.J. Li, E.K. Lobenhofer, X. Lu, R.L. Malek, S. Milton, S.R. Nagalla, J.P. O'malley, V.S. Palmer, P. Pattee, R.S. Paules, C.M. Perou, K. Phillips, L.X. Qin, Y. Qiu, S.D. Quigley, M. Rodland, I. Rusyn, L.D. Samson, D.A. Schwartz, Y. Shi, J.L. Shin, S.O. Sieber, S. Slifer, M.C. Speer, P.S. Spencer, D.I. Sproles, J.A. Swenberg, W.A. Suk, R.C. Sullivan, R. Tian, R.W. Tennant, S.A. Todd, C.J. Tucker, B. Van Houten, B.K. Weis, S. Xuan, and H. Zarbl. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, 2(5):351–6, 2005.

[10] S.W. Chua, P. Vijayakumar, P.M. Nissom, C.Y. Yam, V.V. Wong, and H. Yang. A novel normalization method for effective removal of systematic variation in microarray data. *Nucleic Acids Res*, 34(5):e38, 2006.

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, 1998.

[12] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6): 418–27, 2001.

[13] JA Hartigan and MA Wong. A k-means clustering algorithm. *Applied Statistics*, 28: 100–108, 1979.

[14] T. Kohonen. Analysis of a simple self-organizing process. *Biological Cybernetics*, 43: 59–69, 1982.

[15] T. Kohonen. *Self Organizing Maps*. Berlin: Springer-Verlag, 1989.

[16] Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley, 1977.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, CA, 1984.

[18] TM Cover and PE Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967.

[19] RM Neal. *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[20] BD Ripley. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.

[21] V Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.

[22] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–74, 2001.

[23] JB Kruskal. Multidimensional sclaing by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[24] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*, page 223. Academic Press, London, 1st edition, 1984.

[25] M. J. Greenacre. *Correspondence Analysis in Practice*, pages 181–183 and 36. Academic Press, London, 1st edition, 1993.

[26] H Burkhardt and B Smith. *Handbook of Metaphysics and Ontology*. Munich: Philosophia Verlag, 1991.

[27] D.B. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38:33–48, 1995.

[28] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*, 6(5):R44, 2005.

[29] J.D. Westbrook and P.E. Bourne. STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics*, 16(2):159–68, 2000.

[30] P.L. Whetzel, H. Parkinson, H.C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S.A. Sansone, C. Taylor, J. White, and C.J. Stoeckert, Jr. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–73, 2006.

[31] B. Trombert-Paviot, J.M. Rodrigues, J.E. Rogers, R. Baud, E. van der Haring, A.M. Rassinoux, V. Abrial, L. Clavel, and H. Idir. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *Int J Med Inform*, 58-59:71–85, 2000.

[32] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 1993.

[33] K. Fellenberg, N.C. Hauser, B. Brors, J.D. Hoheisel, and M. Vingron. Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. *Bioinformatics*, 18(3):423–33, 2002.

[34] http://www.geneontology.org/GO.tools.microarray.shtml.

[35] G. Dennis, Jr, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(5):P3, 2003.

[36] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4): 578–80, 2004.

[37] http://srs.ebi.ac.uk/.

[38] http://www.ncbi.nlm.nih.gov/.

[39] http://www.ddbj.nig.ac.jp/.

[40] http://www.ncbi.nlm.nih.gov/dbEST/.

[41] http://www.ncbi.nlm.nih.gov/dbSTS/.

[42] http://www.ncbi.nlm.nih.gov/RefSeq/.

[43] http://www.ensembl.org/index.html.

[44] http://www.geneontology.org/.

[45] http://www.hpr.se/.

[46] http://www.ncbi.nlm.nih.gov/geo/.

[47] http://www.ebi.ac.uk/arrayexpress/.

[48] http://www.brenda.uni koeln.de/.

[49] http://www.gene regulation.com/pub/databases.html.

[50] http://dbtbs.hgc.jp/.

[51] http://ctd.mdibl.org/.

[52] http://www.bind.ca/.

[53] http://string.embl.de/.

[54] http://www.expasy.org/sprot/.

[55] http://www.expasy.org/prosite/.

[56] http://caps.ncbs.res.in/dsdbase//dsdbase.html.

[57] http://www.cbs.dtu.dk/databases/PhosphoBase/.

[58] http://www.ebi.ac.uk/RESID/.

[59] http://www.genome.jp/kegg/.

[60] http://empproject.com/.

[61] http://life2.tau.ac.il/GeneDis/.

[62] http://mutview.dmb.med.keio.ac.jp/MutationView/jsp/index.jsp.

[63] http://www.rcsb.org/pdb/Welcome.do.

[64] http://swissmodel.expasy.org/SWISS MODEL.html.

[65] http://modbase.compbio.ucsf.edu/.

[66] C. Grunau, E. Renault, A. Rosenthal, and G. Roizes. MethDB–a public database for DNA methylation data. *Nucleic Acids Res*, 29(1):270–4, 2001.

[67] C. Amoreira, W. Hindermann, and C. Grunau. An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res*, 31(1):75–7, 2003.

[68] O.D. King, R.E. Foulger, S.S. Dwight, J.V. White, and F.P. Roth. Predicting gene function from patterns of annotation. *Genome Res*, 13(5):896–904, 2003.

[69] R. Stevens, C. Wroe, S. Bechhofer, P. Lord, A. Rector, and C. Goble. Building ontologies in dAML + oIL. *Comp. Funct. Genomics*, 4(1):133–141, 2003).

[70] P. Lord, R. Stevens, C. Goble, and J. Horrocks. *Description logics: OWL and DAML + OIL*. John Wiley & Sons, 2005.

[71] C.J. Wroe, R. Stevens, C.A. Goble, and M. Ashburner. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput*, pages 624–35, 2003.

[72] G.M. Rubin, M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, and S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–15, 2000.

[73] W. Fleischmann, S. Moller, A. Gateau, and R. Apweiler. A novel method for automatic functional annotation of proteins. *Bioinformatics*, 15(3):228–33, 1999.

[74] M.A. Andrade, N.P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis, and C. Sander. Automated genome sequence analysis and annotation. *Bioinformatics*, 15(5):391–412, 1999.

[75] GO Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33, 2001.

[76] http://www.geneontology.org/GO.downloads.shtml#ont.

[77] http://www.geneontology.org/GO.evidence.shtml.

[78] J.B. Lee, J.J. Kim, and J.C. Park. Automatic extension of Gene Ontology with flexible identification of candidate terms. *Bioinformatics*, 22(6):665–70, 2006.

[79] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1st edition, 1984.

[80] M. J. Greenacre. *Correspondence Analysis in Practice*. Academic Press, London, 1st edition, 1993.

[81] M. Hild, B. Beckmann, S.A. Haas, B. Koch, V. Solovyev, C. Busold, K. Fellenberg, M. Boutros, M. Vingron, F. Sauer, J.D. Hoheisel, and R. Paro. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. *Genome Biol*, 5(1):R3, 2003.

[82] R. Micciolo, I. Vantini, G. Cavallini, W. Piubello, G. Talamini, L. Benini, and L.A. Scuro. Correspondence analysis in a study of the clinical evolution of uncomplicated chronic relapsing alcoholic pancreatitis. *Stat Med*, 4(3):303–9, 1985.

[83] M. J. Greenacre. *Correspondence Analysis in Practice*, chapter 12, pages 95–102. Academic Press, London, 1st edition, 1993.

[84] D.L. Hoffman and G.R. Franke. Correspondence analysis graphical representation of categorical data in marketing research. *Journal of Marketing Research*, XXIII:213–27, 1986.

[85] B. Charnomordic and S. Holmes. Correspondence Analysis with R. *Stat. Comp. and Stat. Graph. Newsletter*, 12:19–25, 2001.

[86] C. Dieterich, R. Herwig, and M. Vingron. Exploring potential target genes of signaling pathways by predicting conserved transcription factor binding sites. *Bioinformatics*, 19 Suppl 2:II50–II56, 2003.

[87] K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, and M. Vingron. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, 98: 10781–10786, 2001.

[88] Z. Yin, S. Wilson, N.C. Hauser, H. Tournu, J.D. Hoheisel, and A.J. Brown. Glucose triggers different global responses in yeast, depending on the strength of the signal, and transiently stabilizes ribosomal protein mRNAs. *Mol Microbiol*, 48(3):713–24, 2003.

[89] E. Boles and C. P. Hollenberg. The molecular genetics of hexose transport in yeasts. *FEMS Microbiol Rev*, 21(1):85–111, Aug 1997.

[90] C.H. Busold, S. Winter, N. Hauser, A. Bauer, J. Dippon, J.D. Hoheisel, and K. Fellenberg. Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data. *Bioinformatics*, 21(10):2424–9, 2005.

[91] J.A. Swets. Measuring the accuracs of diagnostic systems. *Science*, (240):1285–1293, 1988.

[92] M.H. Zweig and Campell. G. Reciever Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, 39:561–77, 1993.

[93] http://www.medcalc.be/manual/mpage06 13a.php.

[94] John A. Swets and Ronald M. Pickett. *Evaluation of diagnostic systems : methods from signal detection theory*. Academic Press : New York, 1982.

[95] L. Hunter, R.C. Taylor, S.M. Leach, and R. Simon. GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, 17 Suppl 1:S115–22, 2001.

[96] Y. Zhao and W. Pan. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 19(9):1046–54, 2003.

[97] http://mips.gsf.de/proj/eurofan/eurofan_2/b2/dkfz/results_mce37.txt.

[98] N. C. Hauser, M. Vingron, M. Scheideler, B. Krems, K. Hellmuth, K. D. Entian, and J. D. Hoheisel. Transcriptional profiling of all open reading frames of *Saccharomyces cerevisiae*. *Yeast*, 14:1209–1221, 1998.

[99] T. Beißbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J. M. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schütz, A. Poustka, and M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16:1014–1022, 2000.

[100] L. McAlister-Henn and L.M. Thompson. Isolation and expression of the gene encoding yeast mitochondrial malate dehydrogenase. *J Bacteriol*, 169(11):5157–66, 1987.

[101] B. Repetto and A. Tzagoloff. Structure and regulation of KGD1, the structural gene for yeast alpha-ketoglutarate dehydrogenase. *Mol Cell Biol*, 9(6):2695–705, 1989.

[102] B. Repetto and A. Tzagoloff. Structure and regulation of KGD2, the structural gene for yeast dihydrolipoyl transsuccinylase. *Mol Cell Biol*, 10(8):4221–32, 1990.

[103] M. Suissa, K. Suda, and G. Schatz. Isolation of the nuclear yeast genes for citrate synthase and fifteen other mitochondrial proteins by a new screening method. *EMBO J*, 3(8):1773–81, 1984.

[104] K.S. Kim, M.S. Rosenkrantz, and L. Guarente. Saccharomyces cerevisiae contains two functional citrate synthase genes. *Mol Cell Biol*, 6(6):1936–42, 1986.

[105] R.A. Hitzeman, L. Clarke, and J. Carbon. Isolation and characterization of the yeast 3-phosphoglycerokinase gene (PGK) by an immunological screening technique. *J Biol Chem*, 255(24):12073–80, 1980.

[106] C.C. Blake and D.W. Rice. Phosphoglycerate kinase. *Philos Trans R Soc Lond B Biol Sci*, 293(1063):93–104, 1981.

[107] Z. Lobo. Saccharomyces cerevisiae aldolase mutants. *J Bacteriol*, 160(1):222–6, 1984.

[108] H.G. Schwelberger, S.D. Kohlwein, and F. Paltauf. Molecular cloning, primary structure and disruption of the structural gene of aldolase from Saccharomyces cerevisiae. *Eur J Biochem*, 180(2):301–8, 1989.

[109] A. Kumar, S. Agarwal, J.A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K.H. Cheung, P. Miller, M. Gerstein, G.S. Roeder, and M. Snyder. Subcellular localization of the yeast proteome. *Genes Dev*, 16(6): 707–19, 2002.

[110] T. Kriegel, J. Behlke, and G. Kopperschlager. Hydrodynamic studies on the quaternary structure of reacting yeast phosphofructokinase. *Biomed Biochim Acta*, 46(5):349–55, 1987.

[111] A. Arvanitidis and J.J. Heinisch. Studies on the function of yeast phosphofructokinase subunits by in vitro mutagenesis. *J Biol Chem*, 269(12):8911–8, 1994.

[112] L.L. Newcomb, J.A. Diderich, M.G. Slattery, and W. Heideman. Glucose regulation of Saccharomyces cerevisiae cell cycle genes. *Eukaryot Cell*, 2(1):143–9, 2003.

[113] I. Esposito, A. Bauer, J.D. Hoheisel, J. Kleeff, H. Friess, F. Bergmann, R.J. Rieker, H.F. Otto, G. Kloppel, and R. Penzel. Microcystic tubulopapillary carcinoma of the pancreas: a new tumor entity? *Virchows Arch*, 444(5):447–53, 2004.

[114] K. Orth, J. Hung, A. Gazdar, A. Bowcock, J.M. Mathis, and J. Sambrook. Genetic instability in human ovarian cancer cell lines. *Proc Natl Acad Sci U S A*, 91(20):9495–9, 1994.

[115] T.A. Brentnall, C.E. Rubin, D.A. Crispin, A. Stevens, R.H. Batchelor, R.C. Haggitt, M.P. Bronner, J.P. Evans, L.E. McCahill, N. Bilir, and et al. A germline substitution in the human MSH2 gene is associated with high-grade dysplasia and cancer in ulcerative colitis. *Gastroenterology*, 109(1):151–5, 1995.

[116] J.M. Wheeler, N.E. Beck, H.C. Kim, I.P. Tomlinson, N.J. Mortensen, and W.F. Bodmer. Mechanisms of inactivation of mismatch repair genes in human colorectal cancer cell lines: the predominant role of hMLH1. *Proc Natl Acad Sci U S A*, 96(18):10296–301, 1999.

[117] T. Thykjaer, M. Christensen, A.B. Clark, L.R. Hansen, T.A. Kunkel, and T.F. Orntoft. Functional analysis of the mismatch repair system in bladder cancer. *Br J Cancer*, 85 (4):568–75, 2001.

[118] A. Jansson, G. Arbman, H. Zhang, and X.F. Sun. Combined deficiency of hMLH1, hMSH2, hMSH3 and hMSH6 is an independent prognostic factor in colorectal cancer. *Int J Oncol*, 22(1):41–9, 2003.

[119] J. Plaschke, S. Kruger, B. Jeske, F. Theissig, F.R. Kreuz, S. Pistorius, H.D. Saeger, I. Iaccarino, G. Marra, and H.K. Schackert. Loss of MSH3 protein expression is frequent in MLH1-deficient colorectal cancer and is associated with disease progression. *Cancer Res*, 64(3):864–70, 2004.

[120] P.A. Jones and P.W. Laird. Cancer epigenetics comes of age. *Nat Genet*, 21(2):163–7, 1999.

[121] J.G. Herman and S.B. Baylin. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med*, 349(21):2042–54, 2003.

[122] S.B. Baylin. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol*, 2 Suppl 1:S4–11, 2005.

[123] S.B. Baylin and J.E. Ohm. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer*, 6(2):107–16, 2006.

[124] M. Esteller. The necessity of a Human Epigenome Project. *Carcinogenesis*, 2006.

[125] A.W. Murray. Recycling the cell cycle: cyclins revisited. *Cell*, 116(2):221–34, 2004.

[126] M. Kasten and A. Giordano. Cdk10, a Cdc2-related kinase, associates with the Ets2 transcription factor and modulates its transactivation activity. *Oncogene*, 20(15):1832–8, 2001.

[127] I. Nilsson and I. Hoffmann. Cell cycle regulation by the Cdc25 phosphatase family. *Prog Cell Cycle Res*, 4:107–14, 2000.

[128] K. Galaktionov, X. Chen, and D. Beach. Cdc25 cell-cycle phosphatase as a target of c-myc. *Nature*, 382(6591):511–7, 1996.

[129] K. Galaktionov, C. Jessus, and D. Beach. Raf1 interaction with Cdc25 phosphatase ties mitogenic signal transduction to cell cycle activation. *Genes Dev*, 9(9):1046–58, 1995.

[130] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24(1):238–41, 1996.

[131] T. Hunter and M. Karin. The regulation of transcription by phosphorylation. *Cell*, 70 (3):375–87, 1992.

[132] S. Mukherjee, G. Keitany, Y. Li, Y. Wang, H.L. Ball, E.J. Goldsmith, and K. Orth. Yersinia YopJ acetylates and inhibits kinase activation by blocking phosphorylation. *Science*, 312(5777):1211–4, 2006.

[133] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500): 2306–9, 2000.

[134] X. Pan, S.S. Solomon, D.M. Borromeo, A. Martinez-Hernandez, and R. Raghow. Insulin deprivation leads to deficiency of Sp1 transcription factor in H-411E hepatoma cells and in streptozotocin-induced diabetic ketoacidosis in the rat. *Endocrinology*, 142 (4):1635–42, 2001.

[135] A.C. Poncelet and H.W. Schnaper. Sp1 and Smad proteins cooperate to mediate transforming growth factor-beta 1-induced alpha 2(I) collagen expression in human glomerular mesangial cells. *J Biol Chem*, 276(10):6983–92, 2001.

[136] W. Wang, J.L. Abbruzzese, D.B. Evans, L. Larry, K.R. Cleary, and P.J. Chiao. The nuclear factor-kappa B RelA transcription factor is constitutively activated in human pancreatic adenocarcinoma cells. *Clin Cancer Res*, 5(1):119–27, 1999.

[137] H.W. Sharma and R. Narayanan. The NF-kappaB transcription factor in oncogenesis. *Anticancer Res*, 16(2):589–96, 1996.

[138] D. Hedges, M. Proft, and K.D. Entian. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast Saccharomyces cerevisiae. *Mol Cell Biol*, 15(4):1915–22, 1995.

[139] V. Haurie, M. Perrot, T. Mini, P. Jeno, F. Sagliocco, and H. Boucherie. The transcriptional activator Cat8p provides a major contribution to the reprogramming of carbon metabolism during the diauxic shift in Saccharomyces cerevisiae. *J Biol Chem*, 276(1): 76–85, 2001.

[140] C. Tachibana, J.Y. Yoo, J.B. Tagne, N. Kacherovsky, T.I. Lee, and E.T. Young. Combined global localization analysis and transcriptome data identify genes that are directly coregulated by Adr1 and Cat8. *Mol Cell Biol*, 25(6):2138–46, 2005.

[141] F. Randez-Gil, N. Bojunga, M. Proft, and K.D. Entian. Glucose derepression of gluconeogenic enzymes in Saccharomyces cerevisiae correlates with phosphorylation of the gene activator Cat8p. *Mol Cell Biol*, 17(5):2502–10, 1997.

[142] M.J. De Vit, J.A. Waddle, and M. Johnston. Regulated nuclear translocation of the Mig1 glucose repressor. *Mol Biol Cell*, 8(8):1603–18, 1997.

[143] S. Roth, J. Kumme, and H.J. Schuller. Transcriptional activators Cat8 and Sip4 discriminate between sequence variants of the carbon source-responsive promoter element in the yeast Saccharomyces cerevisiae. *Curr Genet*, 45(3):121–8, 2004.

[144] C.J. Klein, L. Olsson, and J. Nielsen. Glucose control in Saccharomyces cerevisiae: the role of Mig1 in metabolic functions. *Microbiology*, 144 (Pt 1):13–24, 1998.

[145] J.J. Mercado and J.M. Gancedo. Regulatory regions in the yeast FBP1 and PCK1 genes. *FEBS Lett*, 311(2):110–4, 1992.

[146] M.D. Valdes-Hevia, R. de la Guerra, and C. Gancedo. Isolation and characterization of the gene encoding phosphoenolpyruvate carboxykinase from Saccharomyces cerevisiae. *FEBS Lett*, 258(2):313–6, 1989.

[147] M. Proft, D. Grzesitza, and K.D. Entian. Identification and characterization of regulatory elements in the phosphoenolpyruvate carboxykinase gene PCK1 of Saccharomyces cerevisiae. *Mol Gen Genet*, 246(3):367–73, 1995.

[148] A. Hartig, M.M. Simon, T. Schuster, J.R. Daugherty, H.S. Yoo, and T.G. Cooper. Differentially regulated malate synthase genes participate in carbon and nitrogen metabolism of S. cerevisiae. *Nucleic Acids Res*, 20(21):5677–86, 1992.

[149] M. Kunze, F. Kragler, M. Binder, A. Hartig, and A. Gurvitz. Targeting of malate synthase 1 to the peroxisomes of Saccharomyces cerevisiae cells depends on growth on oleic acid medium. *Eur J Biochem*, 269(3):915–22, 2002.

[150] M. Casal, S. Paiva, R.P. Andrade, C. Gancedo, and C. Leao. The lactate-proton symport of Saccharomyces cerevisiae is encoded by JEN1. *J Bacteriol*, 181(8):2620–3, 1999.

[151] O. Akita, C. Nishimori, T. Shimamoto, T. Fujii, and H. Iefuji. Transport of pyruvate in Saccharomyces cerevisiae and cloning of the gene encoded pyruvate permease. *Biosci Biotechnol Biochem*, 64(5):980–4, 2000.

[152] R.J. Haselbeck and L. McAlister-Henn. Function and expression of yeast mitochondrial NAD- and NADP-specific isocitrate dehydrogenases. *J Biol Chem*, 268(16):12116–22, 1993.

[153] T.M. Loftus, L.V. Hall, S.L. Anderson, and L. McAlister-Henn. Isolation, characterization, and disruption of the yeast gene encoding cytosolic NADP-specific isocitrate dehydrogenase. *Biochemistry*, 33(32):9661–7, 1994.

[154] C. De Virgilio, N. Burckert, G. Barth, J.M. Neuhaus, T. Boller, and A. Wiemken. Cloning and disruption of a gene required for growth on acetate but not on ethanol: the acetyl-coenzyme A synthetase gene of Saccharomyces cerevisiae. *Yeast*, 8(12): 1043–51, 1992.

[155] M.A. van den Berg, P. de Jong-Gubbels, C.J. Kortland, J.P. van Dijken, J.T. Pronk, and H.Y. Steensma. The two acetyl-coenzyme A synthetases of Saccharomyces cerevisiae differ with respect to kinetic properties and transcriptional regulation. *J Biol Chem*, 271 (46):28953–9, 1996.

[156] M. Fernandez, E. Fernandez, and R. Rodicio. ACR1, a gene encoding a protein related to mitochondrial carriers, is essential for acetyl-CoA synthetase activity in Saccharomyces cerevisiae. *Mol Gen Genet*, 242(6):727–35, 1994.

[157] L. Palmieri, F.M. Lasorsa, A. Vozza, G. Agrimi, G. Fiermonte, M.J. Runswick, J.E. Walker, and F. Palmieri. Identification and functions of new transporters in yeast mitochondria. *Biochim Biophys Acta*, 1459(2-3):363–9, 2000.

[158] K.I. Minard and L. McAlister-Henn. Isolation, nucleotide sequence analysis, and disruption of the MDH2 gene from Saccharomyces cerevisiae: evidence for three isozymes of yeast malate dehydrogenase. *Mol Cell Biol*, 11(1):370–80, 1991.

[159] N. Gibson and L. McAlister-Henn. Physical and genetic interactions of cytosolic malate dehydrogenase with other gluconeogenic enzymes. *J Biol Chem*, 278(28):25628–36, 2003.

[160] E. Fernandez, F. Moreno, and R. Rodicio. The ICL1 gene from Saccharomyces cerevisiae. *Eur J Biochem*, 204(3):983–90, 1992.

[161] R. Beinoraviciute-Kellner, G. Lipps, and G. Krauss. In vitro selection of DNA binding sites for ABF1 protein from Saccharomyces cerevisiae. *FEBS Lett*, 579(20):4535–40, 2005.

[162] P.K. Brindle, J.P. Holland, C.E. Willett, M.A. Innis, and M.J. Holland. Multiple factors bind the upstream activation sites of the yeast enolase genes ENO1 and ENO2: ABFI protein, like repressor activator protein RAP1, binds cis-acting sequences which modulate repression or activation of transcription. *Mol Cell Biol*, 10(9):4872–85, 1990.

[163] A. Chambers, C. Stanway, J.S. Tsang, Y. Henry, A.J. Kingsman, and S.M. Kingsman. ARS binding factor 1 binds adjacent to RAP1 at the UASs of the yeast glycolytic genes PGK and PYK1. *Nucleic Acids Res*, 18(18):5393–9, 1990.

[164] H.Y. Yoo, S.Y. Jung, Y.H. Kim, J. Kim, G. Jung, and H.M. Rho. Transcriptional control of the Saccharomyces cerevisiae ADH1 gene by autonomously replicating sequence binding factor 1. *Curr Microbiol*, 31(3):163–8, 1995.

[165] D. Shore and K. Nasmyth. Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. *Cell*, 51(5):721–32, 1987.

[166] J.D. Lieb, X. Liu, D. Botstein, and P.O. Brown. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet*, 28(4):327–34, 2001.

[167] R.A. Sumrada and T.G. Cooper. Ubiquitous upstream repression sequences control activation of the inducible arginase gene in yeast. *Proc Natl Acad Sci U S A*, 84(12): 3997–4001, 1987.

[168] L. Kovari, R. Sumrada, I. Kovari, and T.G. Cooper. Multiple positive and negative cis-acting elements mediate induced arginase (CAR1) gene expression in Saccharomyces cerevisiae. *Mol Cell Biol*, 10(10):5087–97, 1990.

[169] T. Keng. HAP1 and ROX1 form a regulatory pathway in the repression of HEM13 transcription in Saccharomyces cerevisiae. *Mol Cell Biol*, 12(6):2616–23, 1992.

[170] J. Deckert, R. Perini, B. Balasubramanian, and R.S. Zitomer. Multiple elements and auto-repression regulate Rox1, a repressor of hypoxic genes in Saccharomyces cerevisiae. *Genetics*, 139(3):1149–58, 1995.

[171] L.C. Myers, C.M. Gustafsson, D.A. Bushnell, M. Lui, H. Erdjument-Bromage, P. Tempst, and R.D. Kornberg. The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes Dev*, 12(1):45–54, 1998.

[172] R. Gil, J. Zueco, R. Sentandreu, and E. Herrero. RCS1, a gene involved in controlling cell size in Saccharomyces cerevisiae. *Yeast*, 7(1):1–14, 1991.

[173] H.V. Baker. GCR1 of Saccharomyces cerevisiae encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC sequence motif. *Proc Natl Acad Sci U S A*, 88(21):9443–7, 1991.

[174] H. Uemura and Y. Jigami. Role of GCR2 in transcriptional activation of yeast glycolytic genes. *Mol Cell Biol*, 12(9):3834–42, 1992.

[175] A. Chambers, E.A. Packham, and I.R. Graham. Control of glycolytic gene expression in the budding yeast (Saccharomyces cerevisiae). *Curr Genet*, 29(1):1–9, 1995.

[176] H. Uemura and D.G. Fraenkel. Glucose metabolism in gcr mutants of Saccharomyces cerevisiae. *J Bacteriol*, 181(15):4719–23, 1999.

[177] M.J. Holland, T. Yokoi, J.P. Holland, K. Myambo, and M.A. Innis. The GCR1 gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in Saccharomyces cerevisiae. *Mol Cell Biol*, 7(2):813–20, 1987.

[178] H. Uemura and D.G. Fraenkel. gcr2, a new mutation affecting glycolytic gene expression in Saccharomyces cerevisiae. *Mol Cell Biol*, 10(12):6389–96, 1990.

[179] M.C. Lopez and H.V. Baker. Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns. *J Bacteriol*, 182(17):4970–8, 2000.

[180] A.G. Hinnebusch and G.R. Fink. Positive regulation in the general amino acid control of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 80(17):5374–8, 1983.

[181] A.G. Hinnebusch and K. Natarajan. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryot Cell*, 1(1):22–32, 2002.

[182] K. Natarajan, M.R. Meyer, B.M. Jackson, D. Slade, C. Roberts, A.G. Hinnebusch, and M.J. Marton. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol*, 21(13):4347–68, 2001.

[183] P. Boukamp, R.T. Petrussevska, D. Breitkreutz, J. Hornung, A. Markham, and N.E. Fusenig. Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. *J Cell Biol*, 106(3):761–71, 1988.

[184] P. Boukamp, E.J. Stanbridge, D.Y. Foo, P.A. Cerutti, and N.E. Fusenig. c-Ha-ras oncogene expression in immortalized human keratinocytes (HaCaT) alters growth potential in vivo but lacks correlation with malignancy. *Cancer Res*, 50(9):2840–7, 1990.

[185] S. Zhu, K. Yoon, E. Sterneck, P.F. Johnson, and R.C. Smart. CCAAT/enhancer binding protein-beta is a mediator of keratinocyte survival and skin tumorigenesis involving oncogenic Ras signaling. *Proc Natl Acad Sci U S A*, 99(1):207–12, 2002.

[186] M. Shim, K.L. Powers, S.J. Ewing, S. Zhu, and R.C. Smart. Diminished expression of C/EBPalpha in skin carcinomas is linked to oncogenic Ras and reexpression of C/EBPalpha in carcinoma cells inhibits proliferation. *Cancer Res*, 65(3):861–7, 2005.

[187] F. Balkwill and L.M. Coussens. Cancer: an inflammatory link. *Nature*, 431(7007): 405–6, 2004.

[188] F.R. Greten, L. Eckmann, T.F. Greten, J.M. Park, Z.W. Li, L.J. Egan, M.F. Kagnoff, and M. Karin. IKKbeta links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell*, 118(3):285–96, 2004.

[189] A. Mantovani. Cancer: inflammation by remote control. *Nature*, 435(7043):752–3, 2005.

[190] H. Kobayashi, T. Yagyu, T. Kondo, N. Kurita, K. Inagaki, S. Haruta, R. Kawaguchi, T. Kitanaka, Y. Sakamoto, Y. Yamada, N. Kanayama, and T. Terao. Suppression of urokinase receptor expression by thalidomide is associated with inhibition of nuclear factor kappaB activation and subsequently suppressed ovarian cancer dissemination. *Cancer Res*, 65(22):10464–71, 2005.

[191] B.B. Aggarwal, S. Shishodia, Y. Takada, S. Banerjee, R.A. Newman, C.E. Bueso-Ramos, and J.E. Price. Curcumin suppresses the paclitaxel-induced nuclear factor-kappaB pathway in breast cancer cells and inhibits lung metastasis of human breast cancer in nude mice. *Clin Cancer Res*, 11(20):7490–8, 2005.

[192] M. Dajee, M. Lazarov, J.Y. Zhang, T. Cai, C.L. Green, A.J. Russell, M.P. Marinkovich, S. Tao, Q. Lin, Y. Kubo, and P.A. Khavari. NF-kappaB blockade and oncogenic Ras trigger invasive human epidermal neoplasia. *Nature*, 421(6923):639–43, 2003.

[193] A. Kobielak and E. Fuchs. Links between alpha-catenin, NF-kappaB, and squamous cell carcinoma in skin. *Proc Natl Acad Sci U S A*, 103(7):2322–7, 2006.

[194] L.J. Ko and C. Prives. p53: puzzle and paradigm. *Genes Dev*, 10(9):1054–72, 1996.

[195] T.A. Lehman, R. Modali, P. Boukamp, J. Stanek, W.P. Bennett, J.A. Welsh, R.A. Metcalf, M.R. Stampfer, N. Fusenig, E.M. Rogan, and et al. p53 mutations in human immortalized epithelial cell lines. *Carcinogenesis*, 14(5):833–9, 1993.

[196] M.B. Datto, Y. Li, J.F. Panus, D.J. Howe, Y. Xiong, and X.F. Wang. Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. *Proc Natl Acad Sci U S A*, 92(12):5545–9, 1995.

[197] P. Boukamp, W. Peter, U. Pascheberg, S. Altmeier, C. Fasching, E.J. Stanbridge, and N.E. Fusenig. Step-wise progression in human skin carcinogenesis in vitro involves mutational inactivation of p53, rasH oncogene activation and additional chromosome loss. *Oncogene*, 11(5):961–9, 1995.

[198] H. Kishino and P.J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform Ser Workshop Genome Inform*, 11:83–95, 2000.

[199] K. Fellenberg, N.C. Hauser, B. Brors, A. Neutzner, J.D. Hoheisel, and M. Vingron. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A*, 98(19): 10781–6, 2001.

[200] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. *AMIA Annu Symp Proc*, pages 609–13, 2003.

[201] I. Yeh, P.D. Karp, N.F. Noy, and R.B. Altman. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, 19(2): 241–8, 2003.

[202] J. Kohler, K. Munn, A. Ruegg, A. Skusa, and B. Smith. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics*, 7:212, 2006.

[203] B. Altenhein, A. Becker, C. Busold, B. Beckmann, J. Hoheisel, and M. Technau. Expression profiling of glial genes during drosophila embryogenesis. *(in press) Developmental Biology*, 2006.

[204] M.A. Dobrovolskaia and S.V. Kozlov. Inflammation and cancer: when NF-kappaB amalgamates the perilous partnership. *Curr Cancer Drug Targets*, 5(5):325–44, 2005.

[205] H. Lu, W. Ouyang, and C. Huang. Inflammation, a key event in cancer development. *Mol Cancer Res*, 4(4):221–33, 2006.

[206] C.H. Busold, T. Beissbarth, R. Kunert, N.C. Hauser, P. Boukamp, R. Schäfer, C. Sers, P. Lund, H. Sültmann, A. Poustka, J. Hoheisel, and K. Fellenberg. Visualizing expression patterns in context of transcription factor binding sites using Correspondence Analysis. *BMC Bioinformatics - submitted*, 2006.

[207] A. Buness, W. Huber, K. Steiner, H. Sultmann, and A. Poustka. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics*, 21(4): 554–6, 2005.

[208] A.E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–9, 2003.

7 Appendix

7.1 SQL query to extract father-child relations in GO

select distinct p.id from graph_path inner join term as t (t.id = graph_path.term2_id) inner join term as p on (p.id = graph_path.term1_id) where t.acc =’?’;

This query will extract all parental note from the GO term specified by ’?’, given the database is in the format as described in 7.1. Since the GO accesion (e.g. GO:0000001) is used as an identfier, this query tends to be slow for large amounts of data, since the used accession implies that the corresponding column in the database needs to be of the type ’CHAR’ or ’Text’. Retrieval of parental terms can be significantly sped up by an inital extraction of the internal GO-ID (as provided in the ’term’-Table), which in turn can be used as follows to retrieve all parental terms:

select distinct term1_id from graph_path where term2_id = ?;

Analogous to the previous query the ’?’ is to be replaced by the corresponding GO ID. Since this ID, now is of the type ’Integer’ and the respective column is indexed, even large queries will extract all parental terms in a feasible time frame.

7.2 Experimental procedures for human cancer study

Glass slides used for this study carried 37,530 cDNA clones selected from the Human UniGene 3.1 clone set (German Resource Center for Genome Research, Berlin, Germany). The cDNA array is submitted to Gene expression omnibus database GEO including manufacture protocol (GEO accession GPL3050). HaCaT is a human keratinocyte cell line. The original spontaneous immortalized, benign HaCaT-Tetra cell line has been Ha-Ras-transformed and the cells were injected in mice for tumor growth resulting in one additionally benign cell line HaCaT-A5, in two malignant cell lines HaCaT-I7 and HaCaT-II4, and in one malignant cell line HaCaT-A5RT3 with the potential to metastasize. HaCaT cells were routinely cultured in DMEM (Sigma), supplemented with 10% fetal calf serum and antibiotics at 37°C, 5% CO2 and 95% air. Cells were subcultured once a week at a 1:10 dilution. For the treatment with the MEK inhibitor U0126 (Promega), cells were seeded at 8x105 per plate in 10 cm culture dishes and treated with 10 μM U0126 or the solvent DMSO for 48 hours and finally collected for RNA preparation. Total RNA was prepared using the RNeasy-Midi-Kit (Qiagen) following the manufactures instructions. The isolated RNA was quantified by UV-spectroscopy and quality controlled using the Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA). All of the samples yielded high-quality RNA (28S/18S rRNA and E260/E280 ratio larger than 1.8). Samples were stored at 80°C until further use. Linear amplification was performed using the MessageAmpTM aRNA Kit (Ambion) according to the manufactures instructions. Amplified RNA samples were labeled with Cy3 and a common reference aRNA (Stratagene) with Cy5. Cy3- and Cy5-labeled samples were purified with Microcon YM-30 columns (Milipore, Bedford, MA, USA), combined and resuspended in 50 μl 1x DIG-Easy hybridization buffer (Roche Diagnostics), containing 10x Denhardt's solution and 2ng/μl Cot1-DNA (Invitrogen). Hybridisations (in duplicate) and washing were done as previously described. The hybridized arrays were scanned with the GenePix 4000B microarray scanner (Axon Instruments), and analyzed using GenePix Pro 4.1 software. The normalized expression data were filtered with respect to signal intensity and 7289 genes were selected for analysis in Correspondence Analysis by two-way Anova analysis (P value < 0.05) including both parameters, HaCaT variant and treatment. The displayed transcription factors for this experimental dataset are based on the association data from TRANSFAC database. Only these TFs were displayed which are associated to a minimum of 5 gene hits in the expression set. The TFs are added as supplementary rows to the data matrix in the same manner like in the yeast dataset analysis. Preprocessing and most of the statistical analysis were done using R (www.r-project.org) and Bioconductor (www.bioconductor.org). After quality control, all cDNA microarray data were normalized using arrayMagic [207].

7.3 UML schema of the GO database

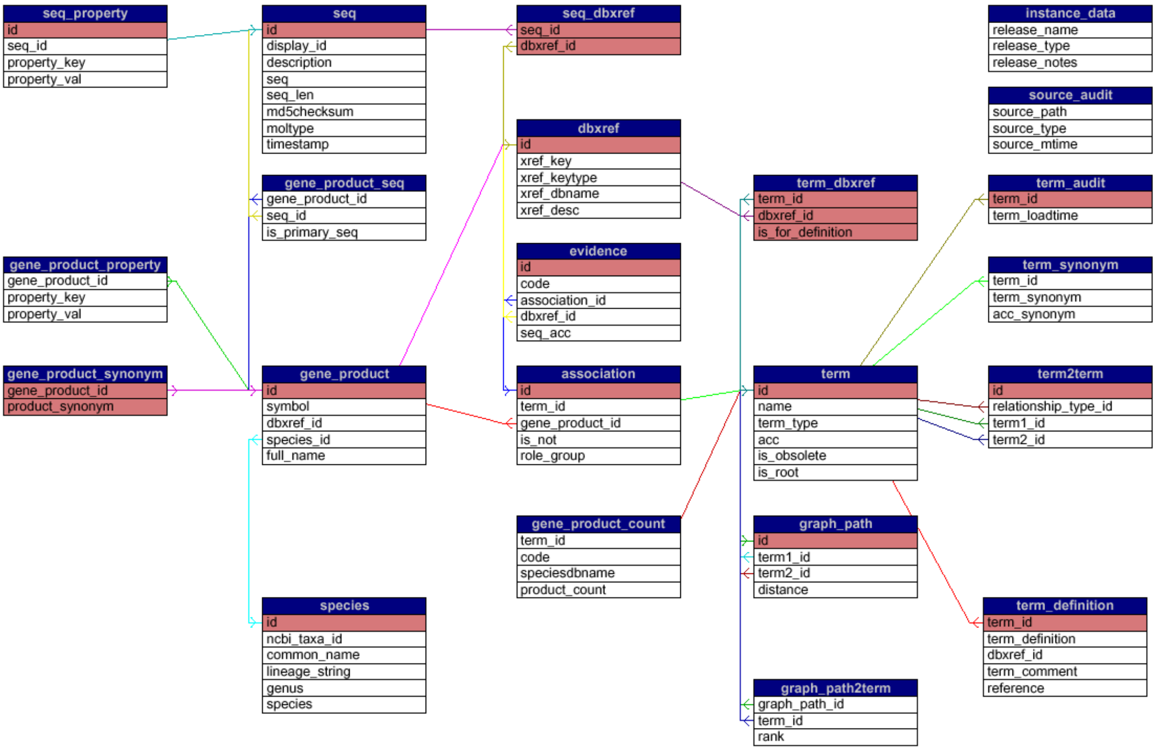
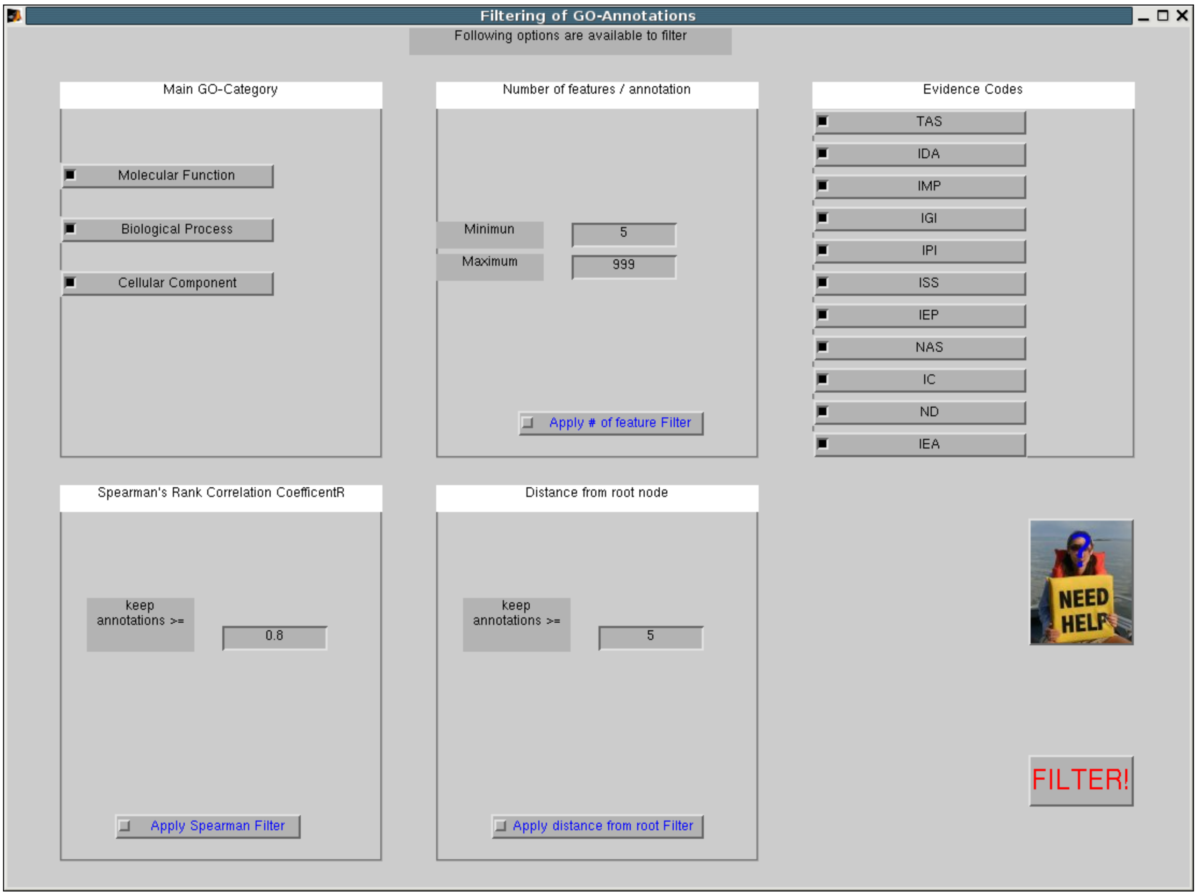


Figure 7.1: UML schema of the GO database. Reproduced from http://www.godatabase.org/dev/sql/doc/diagrams.html.

7.4 Snapshots of GUI



Filtering of GO-Annotations

Following options are available to filter

Main GO-Category

☒ Molecular Function

☒ Biological Process

☒ Cellular Component

Number of features / annotation

Minimun 5

Maximum 999

☐ Apply # of feature Filter

Evidence Codes

☒ TAS

☒ IDA

☒ IMP

☒ IGI

☒ IPI

☒ ISS

☒ IEP

☒ NAS

☒ IC

☒ ND

☒ IEA

Spearman's Rank Correlation CoefficentR

keep annotations >= 0.8

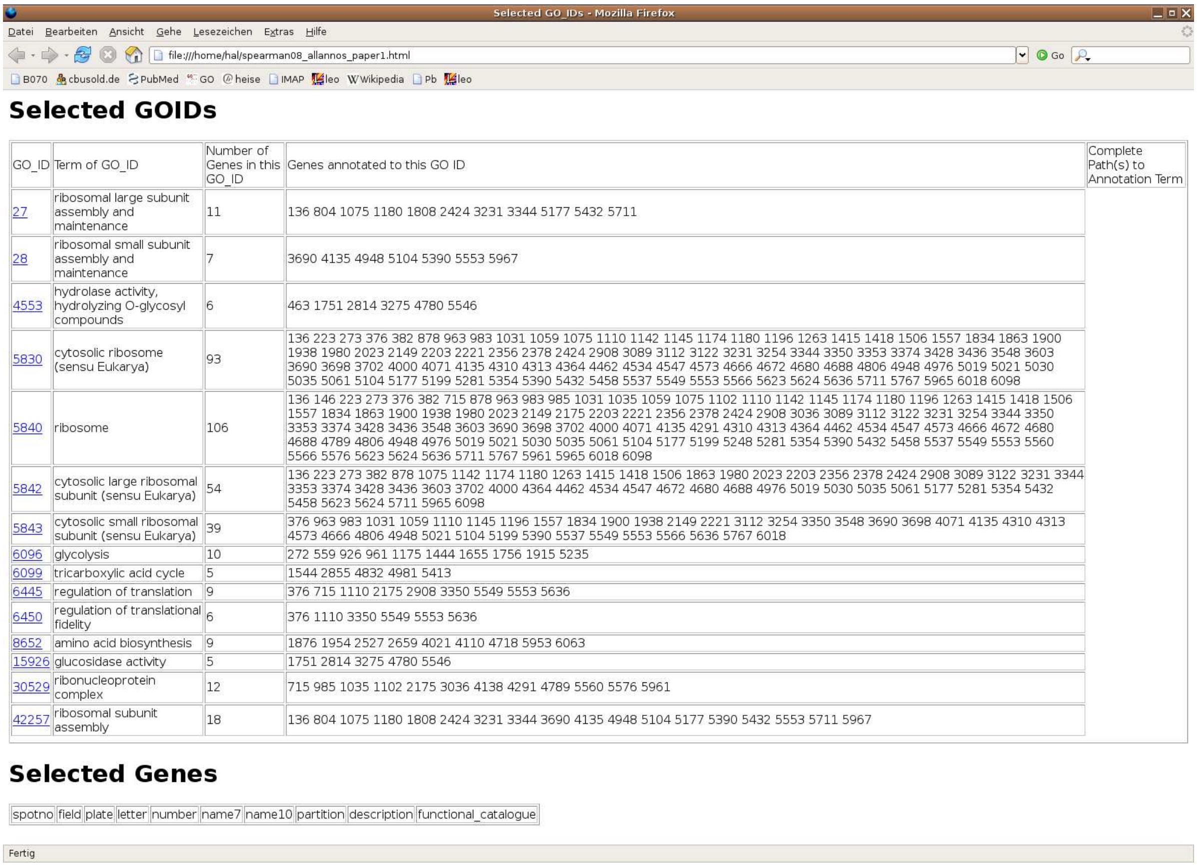
☐ Apply Spearman Filter

Distance from root node

keep annotations >= 5

☐ Apply distance from root Filter

Figure 7.2: User interface for filtering of GO annotations. In the first panel any combination of subsets of the ontology can be selected. The second panel lets the researcher define minimal and maximal boundaries, for the number of genes being associated to an annotation. In the last panel in the first row the annotations, more precisely the annotated gene products, can be filtered based on the evidence codes. Here the codes are roughly ordered by the quality of annotation, i.e. to top-most code (TAS, 'traceable author statement') has the highest quality, whereas IEA ('inferred from electronic annotation') at the bottom, the lowest. This is not a standard ordering based on some criterion, but a suggestion from the author. The first panel in the second rows allows the specification for the threshold of the correlation coefficient below which the annotations should be not displayed. The last panel allows to select annotations based on their position in the ontology, i.e. their distance from the root node. The filtering options and their effects on the data are discussed in 2.3 on page 30.


Selected GO_IDs - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
file:///home/hal/spearman08_allannos_paper1.html
Go
B070 cbusold.de PubMed GO heise IMAP leo Wikipedia Pb leo
Fertig

Selected GOIDs

GO_ID	Term of GO_ID	Number of Genes in this GO_ID	Genes annotated to this GO ID	Complete Path(s) to Annotation Term
27	ribosomal large subunit assembly and maintenance	11	136 804 1075 1180 1808 2424 3231 3344 5177 5432 5711	
28	ribosomal small subunit assembly and maintenance	7	3690 4135 4948 5104 5390 5553 5967	
4553	hydrolase activity, hydrolyzing O-glycosyl compounds	6	463 1751 2814 3275 4780 5546	
5830	cytosolic ribosome (sensu Eukarya)	93	136 223 273 376 382 878 963 983 1031 1059 1075 1110 1142 1145 1174 1180 1196 1263 1415 1418 1506 1557 1834 1863 1900 1938 1980 2023 2149 2203 2221 2356 2378 2424 2908 3089 3112 3122 3231 3254 3344 3350 3353 3374 3428 3436 3548 3603 3690 3698 3702 4000 4071 4135 4310 4313 4364 4462 4534 4547 4573 4666 4672 4680 4688 4806 4948 4976 5019 5021 5030 5035 5061 5104 5177 5199 5281 5354 5390 5432 5458 5537 5549 5553 5566 5623 5624 5636 5711 5767 5965 6018 6098	
5840	ribosome	106	136 146 223 273 376 382 715 878 963 983 985 1031 1035 1059 1075 1102 1110 1142 1145 1174 1180 1196 1263 1415 1418 1506 1557 1834 1863 1900 1938 1980 2023 2149 2175 2203 2221 2356 2378 2424 2908 3036 3089 3112 3122 3231 3254 3344 3350 3353 3374 3428 3436 3548 3603 3690 3698 3702 4000 4071 4135 4291 4310 4313 4364 4462 4534 4547 4573 4666 4672 4680 4688 4789 4806 4948 4976 5019 5021 5030 5035 5061 5104 5177 5199 5248 5281 5354 5390 5432 5458 5537 5549 5553 5560 5566 5576 5623 5624 5636 5711 5767 5961 5965 6018 6098	
5842	cytosolic large ribosomal subunit (sensu Eukarya)	54	136 223 273 382 878 1075 1142 1174 1180 1263 1415 1418 1506 1863 1980 2023 2203 2356 2378 2424 2908 3089 3122 3231 3344 3353 3374 3428 3436 3603 3702 4000 4364 4462 4534 4547 4672 4680 4688 4976 5019 5030 5035 5061 5177 5281 5354 5432 5458 5623 5624 5711 5965 6098	
5843	cytosolic small ribosomal subunit (sensu Eukarya)	39	376 963 983 1031 1059 1110 1145 1196 1557 1834 1900 1938 2149 2221 3112 3254 3350 3548 3690 3698 4071 4135 4310 4313 4573 4666 4806 4948 5021 5104 5199 5390 5537 5549 5553 5566 5636 5767 6018	
6096	glycolysis	10	272 559 926 961 1175 1444 1655 1756 1915 5235	
6099	tricarboxylic acid cycle	5	1544 2855 4832 4981 5413	
6445	regulation of translation	9	376 715 1110 2175 2908 3350 5549 5553 5636	
6450	regulation of translational fidelity	6	376 1110 3350 5549 5553 5636	
8652	amino acid biosynthesis	9	1876 1954 2527 2659 4021 4110 4718 5953 6063	
15926	glucosidase activity	5	1751 2814 3275 4780 5546	
30529	ribonucleoprotein complex	12	715 985 1035 1102 2175 3036 4138 4291 4789 5560 5576 5961	
42257	ribosomal subunit assembly	18	136 804 1075 1180 1808 2424 3231 3344 3690 4135 4948 5104 5177 5390 5432 5553 5711 5967	

Selected Genes

spotno	field	plate	letter	number	name7	name10	partition	description	functional_catalogue

Figure 7.3: Snapshot of results output. Here a snapshot of the format in which the GO terms that are selected in a CA plot are presented to the users. The first column represents the GO accession id being truncated of 'GO:' and trailing zeros. This ID is a link to the GO website, providing more detailed information on it (e.g. it's position in the ontology - amoungst other), the second column gives the name of the term, the third the number of genes being associated to this term in the current analysis (i.e. after applying gene- and annotations filters) and in the last column spotnos of the associated genes are provided. Finally all the genes being associated to one of the selected annotations are listed at the bottom of the page along with all available information on them that is stored in the database.

7.5 Comprehensive listing of annotations displayed in human cancer data set

Cluster	GO id	GO term
Tumor associated		
	GO:0009611	(P) response to wounding
	GO:0009888	(P) histogenesis
	GO:0004722	(F) protein serine/threonine phosphatase activity
	GO:0004725	(F) protein tyrosine phosphatase activity
	GO:0004842	(F) ubiquitin-protein ligase activity
	GO:0005001	(F) transmembrane receptor protein tyrosine phosphatase activity
	GO:0006310	(P) DNA recombination
	GO:0006470	(P) protein amino acid dephosphorylation
	GO:0008287	(C) protein serine/threonine phosphatase complex
	GO:0016879	(F) ligase activity, forming carbon-nitrogen bonds
	GO:0016881	(F) acid-D-amino acid ligase activity
	GO:0019208	(F) phosphatase regulator activity
	GO:0019888	(F) protein phosphatase regulator activity
	GO:0045595	(P) regulation of cell differentiation
	GO:0004693	(F) cyclin-dependent protein kinase activity
	GO:0007089	(P) traversing start control point of mitotic cell cycle
	GO:0007172	(P) signal complex formation
	GO:0007265	(P) RAS protein signal transduction
	GO:0007254	(P) JNK cascade
Ductal / Cystic associated		
	GO:0006306	(P) DNA methylation
	GO:0040029	(P) regulation of gene expression, epigenetic
	GO:0012502	(P) induction of programmed cell death
	GO:0043067	(P) regulation of programmed cell death
	GO:0043068	(P) positive regulation of programmed cell death
Normal associated		
	GO:0003735	(C) structural constituent of ribosome
	GO:0005006	(F) epidermal growth factor receptor activity
	GO:0005216	(F) ion channel activity

Cluster	GO id	GO term
	GO:0005261	(F) cation channel activity
	GO:0005830	(C) cytosolic ribosome (sensu Eukarya)
	GO:0005840	(C) ribosome
	GO:0005843	(C) cytosolic small ribosomal subunit (sensu Eukarya)
	GO:0006298	(P) mismatch repair
	GO:0007173	(P) epidermal growth factor receptor signaling pathway
	GO:0016055	(P) Wnt receptor signaling pathway
	GO:0045005	(P) maintenance of fidelity during DNA-dependent DNA replication

Table 7.1: Comprehensive list of GO annotations displayed in Figure 2.19. Annotations are grouped according to the clusters in Fig. and their corresponding GO-id, main category (P= Biological Process, F= Molecular Function) and GO term are given.

7.6 Unique assignments of gene products

GO accession	GO term
GO:0006099	tricarboxylic acid cycle
GO:0000022	mitotic spindle elongation
GO:0000027	ribosomal large subunit assembly and maintenance
GO:0000041	transition metal ion transport
GO:0000070	mitotic sister chromatid segregation
GO:0000096	sulfur amino acid metabolism
GO:0000132	establishment of mitotic spindle orientation
GO:0000274	proton-transporting ATP synthase, stator stalk (sensu Eukaryota)
GO:0000275	proton-transporting ATP synthase complex, catalytic core F(1) (sensu Eukaryota)
GO:0000329	vacuolar membrane (sensu Fungi)
GO:0000902	cellular morphogenesis
GO:0003924	GTPase activity
GO:0003969	RNA editase activity
GO:0004028	3-chloroallyl aldehyde dehydrogenase activity
GO:0004175	endopeptidase activity
GO:0004478	methionine adenosyltransferase activity
GO:0004672	protein kinase activity

GO accession	GO term
GO:0006099	tricarboxylic acid cycle
GO:0004674	protein serine/threonine kinase activity
GO:0004824	lysine-tRNA ligase activity
GO:0005047	signal recognition particle binding
GO:0005489	electron transporter activity
GO:0005667	transcription factor complex
GO:0005672	transcription factor TFIIA complex
GO:0005740	mitochondrial envelope
GO:0005743	mitochondrial inner membrane
GO:0005746	mitochondrial electron transport chain
GO:0005774	vacuolar membrane
GO:0005816	spindle pole body
GO:0005819	spindle
GO:0005823	central plaque of spindle pole body
GO:0005830	cytosolic ribosome (sensu Eukaryota)
GO:0005843	cytosolic small ribosomal subunit (sensu Eukaryota)
GO:0005854	nascent polypeptide-associated complex
GO:0005887	integral to plasma membrane
GO:0005905	coated pit
GO:0005934	bud tip
GO:0005935	bud neck
GO:0005976	polysaccharide metabolism
GO:0006007	glucose catabolism
GO:0006067	ethanol metabolism
GO:0006071	glycerol metabolism
GO:0006100	tricarboxylic acid cycle intermediate metabolism
GO:0006103	2-oxoglutarate metabolism
GO:0006163	purine nucleotide metabolism
GO:0006260	DNA replication
GO:0006325	establishment and/or maintenance of chromatin architecture
GO:0006365	35S primary transcript processing
GO:0006417	regulation of protein biosynthesis
GO:0006450	regulation of translational fidelity
GO:0006457	protein folding
GO:0006461	protein complex assembly

GO accession	GO term
GO:0006099	tricarboxylic acid cycle
GO:0006486	protein amino acid glycosylation
GO:0006493	protein amino acid O-linked glycosylation
GO:0006497	protein amino acid lipidation
GO:0006508	proteolysis
GO:0006528	asparagine metabolism
GO:0006536	glutamate metabolism
GO:0006555	methionine metabolism
GO:0006566	threonine metabolism
GO:0006631	fatty acid metabolism
GO:0006643	membrane lipid metabolism
GO:0006696	ergosterol biosynthesis
GO:0006800	oxygen and reactive oxygen species metabolism
GO:0006820	anion transport
GO:0006839	mitochondrial transport
GO:0006869	lipid transport
GO:0006873	cell ion homeostasis
GO:0006885	regulation of pH
GO:0006892	post-Golgi vesicle-mediated transport
GO:0006970	response to osmotic stress
GO:0006972	hyperosmotic response
GO:0006996	organelle organization and biogenesis
GO:0007015	actin filament organization
GO:0007020	microtubule nucleation
GO:0007088	regulation of mitosis
GO:0007163	establishment and/or maintenance of cell polarity
GO:0007568	aging
GO:0008028	monocarboxylic acid transporter activity
GO:0008144	drug binding
GO:0008443	phosphofructokinase activity
GO:0008526	phosphatidylinositol transporter activity
GO:0008553	hydrogen-exporting ATPase activity, phosphorylative mechanism
GO:0008610	lipid biosynthesis
GO:0009057	macromolecule catabolism
GO:0009063	amino acid catabolism

GO accession	GO term
GO:0006099	tricarboxylic acid cycle
GO:0009064	glutamine family amino acid metabolism
GO:0009072	aromatic amino acid family metabolism
GO:0009084	glutamine family amino acid biosynthesis
GO:0009267	cellular response to starvation
GO:0009310	amine catabolism
GO:0009408	response to heat
GO:0010035	response to inorganic substance
GO:0015893	drug transport
GO:0015926	glucosidase activity
GO:0015934	large ribosomal subunit
GO:0016615	malate dehydrogenase activity
GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0016757	transferase activity, transferring glycosyl groups
GO:0016773	phosphotransferase activity, alcohol group as acceptor
GO:0016791	phosphoric monoester hydrolase activity
GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
GO:0016866	intramolecular transferase activity
GO:0017171	serine hydrolase activity
GO:0019202	amino acid kinase activity
GO:0019207	kinase regulator activity
GO:0019725	cell homeostasis
GO:0030001	metal ion transport
GO:0030003	cation homeostasis
GO:0030120	vesicle coat
GO:0030478	actin cap
GO:0030528	transcription regulator activity
GO:0042123	glucanosyltransferase activity
GO:0042625	ATPase activity, coupled to transmembrane movement of ions
GO:0042645	mitochondrial nucleoid
GO:0045333	cellular respiration
GO:0046037	GMP metabolism

GO accession	GO term
GO:0006099	tricarboxylic acid cycle
GO:0050794	regulation of cellular process

Table 7.2: Comprehensive listing of uniquely assigned gene products. Here all annotations are provided which have been selected when assigning each annotated gene product to a single annotation, as described in 2.2.4.3

7.7 Prediction of TF binding sites

The yeast genome was searched for potential predicted TFBS using the 56 matrices for 44 different TFs designed from Funghi from the TRANSFAC database. 500bp upstream regions were extracted from all genes in the yeast genome from the version of the genome at UCSC (sacCer1, http://genome.ucsc.edu). The program Match [208] was used to scan for potential binding sites. All potential binding sites that fulfilled "minimize false positives" cutoff criteria that are suggested for each matrix in the TRANSFAC database were used. Altogether 19808 potential TFBS sites were predicted for 5542 yeast genes and 56 different TF matrices giving 15236 unique gene/site combinations.

Furthermore the human genome was scanned for conserved predicted TFBS using the 565 matrices designed for 386 different TFs from Vertebrates from the TRANSFAC database. 10kb upstream regions were extracted from all Refseq genes in the human, mouse and rat genomes from at UCSC (http://genome.ucsc.edu). The homologene database was used to determine the orthologous gene pairs and tripples. The pars and triples were aligned using the program MUSCLE. Each of the sequences was scanned for potential TFBS using the program Match [208]. All potential binding sites that fulfilled "minimize false positives" cutoff criteria that are suggested for each matrix in the TRANSFAC database were used. The positions of the TFBS in the aligned (gapped) sequences were computed and all TFBS kept that appeared at least in two of the aligned sequences in corresponding positions (+/-5 bp). Both of the predictions were carried out by Tim Beissbarth.

7.8 Software used

- PostgreSQL (Version 6.5.3 and 7.3.15): A powerful open source relational database system, which is publically available from 'http://www.postgresql.org/download/'.
- Matlab (Version 6.0) incl. Statistics Toolbox: Interpreted numerical programming environment. MathWorks Inc. MA, USA.

- Perl (Version 5.8.7): free, platform-independant, interpreted programming language, which is suited for regular expression searches in e.g. text files. The software is available from 'http://www.perl.com/download.csp'.
- R (Version 2.1.1): a freely available language and environment for statistical computing and graphics ('http://cran.r-project.org/index.html').

Abbreviations

AUC - area under curve

CA - Correspondence Analysis

ChIP-chip - chromatin immunoprecipitation-chip

cond. - condition

DAG - directed acyclic graph

DMSO - di-methyl-sulfoxide

DNA - desoxy-ribonucleicacid

exp. - experiment(al)

FN(R) - false negative (rate)

FP(R) - false positive (rate)

GEO - gene expression omnibus

GO - gene ontology

ID - identifier

KEGG - Kyoto Encyclopedia of Genes and Genomes

M-CHiPS - Multi-Conditional Hybridization Intensity Processing System

MDS - multi dimensional scaling

MGD - mouse genome database

MGED - Microarray Gene Expression Data

MIAME - minimal information about microarray experiments

mRNA - messenger ribonucleicacid

OBO - open biological ontologies

PCA - principal component analysis

ROC - receiver operating characterisics

RZPD - German Resource Center for Genome Research

SGD - Saccharomyces Genome Database

SOM - self-organizing maps

SQL - structured querying language

SVM - support vector machines

TAIR - Arabidopsis Information Resource

TCA - tricarboxylic acid

TF(BS) - transcription factor (binding site)

TN(R) - true negative (rate)

TP(R) - true positive (rate)

XML - extended markup language

Acknowledgements

First of all, I want to thank the supervisors of this work Prof. Dr. Thomas Dandekar, Dr. Kurt Fellenberg and Dr. Jörg Hoheisel for their guidance and help. I am also grateful to Prof. Dr. Wiemann for co-correcting this thesis.

Many thanks to Stefan Winter and Prof. Dr. Jürgen Dippon for mathematical enlightenment. Dr. Tim Beissbarth for good discussions and critically reading the manuscript. Dr. Nicole Hauser, Dr. Boris Beckmann, Dr. Andrea Bauer and Dr. Ruprecht Kuner who provided me with the microarray data presented in this thesis.

I would like to thank all members of the ’Functional Genome Analysis’ departement for a wonderful working atmosphere and all the spare-time activities I enjoyed sharing with you. Here special thanks to Jorge, Rafa, Michi, Andrea, Verena, Linda, Tim, Marcus and Jörg. Marc and Christoph for readily joining in and supporting the ’linux-crusade’.

I want to thank Florian Hauger and Robert Kirmse not only for correcting the manuscript, but also for the good times on Rubika and the many fruitful discussions and observations in the Botanik and Marstall.

A big thanks to the other ’Fischköpp’: Anette, Marc, Svenja, Tobi and Jan - the south will never understand...

Special thanks to Nicole Juling, not only for introducing me to the ’after-work-life’ in HD and the good times in the KBhf, but also for being a friend.

Last but not least, I owe the biggest ’thank you’ to my parents Friedlinde and Klaus Busold and my ’little’ sister Dorothee for their unlimitted support and encouragment throughout all the different stages of my PhD. Without you all this would not have been possible - DANKE!

Curriculum vitae

Personal Information

Date of birth	31.03.1975
Place of birth	Hameln, Germany
Citizenship	German
Familiy status	single

Education

1987 - 1994	Otto-Hahn-Gymnasium Springe
1995 - 2001	Study of Biology, University of Hannover Majors: Genetics, Biochemistry, Immunology
09.1997 - 02.1998	Research Assistant, Dept. Ertragsphysiologie
09.1998 - 09.1999	Exchange year; Northeastern University Boston, MA; DAAD scholarship
04.1999 - 06.1999	Research Assistant, Palmer Station, Antarctica
12.1999	Teaching Assistant, Dept. Biochemistry
09.2000 - 05.2001	Master Thesis at Max-Planck-Institute for Developmental Biology Tübingen, Dept. Genetics (Prof. C. Nüsslein-Volhard), Title: ’Etablierung von Methoden zur Genomkartierung beim Zebrafisch (*Danio rerio*) durch Microarrayhybridisierung’
01.01.2002 -	PhD-Position at German Cancer Research Center, Dept. Functional Genome Analysis (Dr. Jörg Hoheisel)

Ehrenwörtliche Erklärung

Gemäß §4 Absatz 3 Ziffern 3, 5 und 8 der Promotionsordnung der Fakultät für Biologie der Bayerischen Julius-Maximilians-Universität Würzburg Hiermit erkläre ich, Christian Busold, ehrenwörtlich, die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben. Die Dissertation wurde bisher weder vollständig noch teilweise einer anderen Hochschule mit dem Ziel einen akademischen Grad zu erwerben vorgelegt. Ich erkläre weiterhin, dass ich außer meines Diploms in Biologie an der Universität Hannover keine weiteren akademischen Grade erworben habe oder zu erwerben versucht habe. Die Arbeit wurde am Deutschen Krebsforschungszentrum (DKFZ) in Heidelberg angefertigt.

________________________ ____________________________

Ort, Datum Unterschrift