Development of Computational Methods for Metabolic Network Analysis based on Metabolomics Data

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.) im Fach Bioinformatik
eingereicht im Jahre 2008 an der
Naturwissenschaftlich-Technischen Fakultät I
der Universität des Saarlandes

Von

Priti Talwar

Saarbrücken
2008

Priti Talwar: *Development of Computational Methods for Metabolic Network Analysis based on Metabolomics Data.* Ph.D. thesis for obtaining the academic degree of a doctor of the natural sciences in bioinformatics, submitted to the Faculty I of Natural Sciences and Technology (mathematics and computer science) of the Saarland University, Saarbrücken, Germany, 2008.

Dekan (Dean of the Faculty): Prof. Dr. Joachim Weickert

Einreichung (Submission): 15 September 2008

Kolloquium (Colloquium): 26 November 2008

Vorsitzender (Chairman): Prof. Hans-Peter Lenhof

Gutachter (Reviewers): 1. Prof. Dr. Thomas Lengauer, Ph.D.
2. Prof. Dr. Jörg Rahnenfuhrer

Protokollant (Minute taker): Dr. Mario Albrecht

Summary

The baker's yeast, *Saccharomyces cerevisiae*, is a simple eukaryotic organism with approximately 6000 genes. *Saccharomyces cerevisiae* is an ideal model organism for large-scale functional studies and provides a system in which genes can be systematically inactivated by way of gene-knockout methods. A substantial fraction of the 6000 genes in *Saccharomyces cerevisiae* encode proteins for which currently we do not know any confirmed or putative function. Prediction of the functional role of these proteins is a challenging problem in systems biology, especially as many of these genes have no overt phenotypes. In our study, we aim at a better understanding of the underlying functional relationships between genes working across diverse metabolic pathways using intracellular metabolite profiling studies. We applied bioinformatics methods and statistical analysis techniques in combination with metabolic profiling to understand the function and the regulatory mechanisms of specific genes involved in central carbon metabolism and amino acid biosynthesis. The experimental work was carried out by the group of Prof. Elmar Heinzle (Biochemical Engineering, Saarland University), our collaboration partner. ^{13}C stable isotope substrates can be used as tracers to generate detailed metabolic profiles of gene knockouts. Detailed and quantitative information on the physiological cellular states is measured by ^{13}C -metabolic profiling of cultures grown on novel high throughput oxygen sensor microtiter plates. In this dissertation, we worked towards developing systematic approaches for study of *Saccharomyces cerevisiae* genes of unknown function based on the metabolic profiles of knockout mutants under varied environmental conditions. In the first step, we have developed a software tool called CalSpec for automation of Gas Chromatography Mass Spectrometry data acquisition and analysis routine, as this is a bottleneck in the metabolic profiling studies. In the next step, we worked on large scale statistical analysis of metabolic profiling data. We applied various algorithms for finding closely related mutants which show similar metabolic profiles. According to our hypothesis, similarity in the metabolic profiles can be used to find functionally linked genes. *Saccharomyces cerevisiae* is known to be robust to majority of genetic perturbations. In these cases where the mutants show no overt

phenotypes, we developed a sensitive outlier detection method to detect those subsets of metabolic profile features which are most differentiating (outliers) for all mutants. The second part of this dissertation involves developing computational tools for metabolic pathway analysis on the basis of genome scale metabolic models, as well as integration of various newly emerging experimental techniques. In recent years, genome scale metabolic models have been and are continuing to be assembled for various organisms. In the year 2003, first comprehensive genome scale metabolic model for yeast became publicly available. With the emergence of system biology area of research, diverse computational approaches have been developed. In this work, we developed a new webserver called MetaModel, for analysis of genome scale metabolic networks of eukaryotic organisms.

Acknowledgement

I wish to sincerely thank Prof. Thomas Lengauer for his excellent supervision and giving me a chance to work on the present project. Prof. Lengauer has introduced me to the area of “computational systems biology” and has spearheaded my present work. I am grateful for his supervision and continuous support. I would also like to express my sincere thanks to Prof. Jörg Rahnenfuhrer (TU Dortmund) who has helped me throughout my thesis work and given me invaluable ideas. He has been a great friend and mentor.

The present project has been performed in collaboration with group of Prof. Elmar Heinzle, Technical Biochemistry department at University of Saarland. Together with Prof. Heinzle, Prof. Christoph Wittmann (now Uni. Münster) has helped me in understanding the problem of automation of mass spectrometric data analysis, as well as many aspects of yeast metabolomics. I wish to thank Prof. Elmar Heinzle, Prof. Christoph Wittmann, Vidya Velagapudi and other members of Technical Biochemistry department for providing me the yeast growth and fractional labeling datasets.

There has been multitude of people who made my Ph.D. years truly memorable. I especially like to convey my gratitude to my dear friends and colleagues Jochen Maydt, Dr. Ingolf Sommer, Kirsten Roomp, Oliver Sanders, Frederik Gwinner, Dr. Francisco Domingues, Dr. Mario Albrecht, Lars Kunert, Prof. Iris Antes, Christoph Welsch, Dr. Andreas Kaemper, Dr. Gabriele Mayr and all the D-3 group members. They have been a source of constant support and encouragement during my stay in the group.

I am grateful to Dr. Joachim Büch for all the help and support during the project. Another person who did me innumerable favors and helped me in settling down is Ruth Schneppen-Christmann. Her ever smiling face and timely help has eased many things for me. I wish to also thank my German language teacher, Simone Schulze. What ever German I know, I owe it to Simone who was ever so patient with us.

Also, I wish to thank my teachers, colleagues and friends from Saarbrücken and around the world, especially Prof. V. Sundararajan (CDAC), Prof. Vidyasagar (Uni. Pune), Shince Thomas, Sanjay Londhe (DIC Pune), Pintu & Pinky Das, Nivedita Jena, Naresh C. Pal, Meenakshi Jha, Kavitha Telikepalli and Ventakesh Srinivasan. I also extend my sincere thanks to my mentors and colleagues at my new workplace (BIMR, San Diego), especially Andrei Ostermann for his scientific enthusiasm and motivation.

I also wish to thank my husband, Dr. Ravanan Palaniyandi for his understanding and constant support. What ever I am today is only because of love, determination and endless sacrifices by my mother Veena Talwar and my elder brother Pradeep Talwar (Minti). Minti bhaiya, thanks for being a great elder brother.

In the end, I wish to again thank Prof. Thomas Lengauer for teaching me the “good science” and encouraging me during the entire project. You are a true mentor and a dynamic teacher.

Contents

Part I Introduction

Part II Statistical analysis of the data

Part III Web server for metabolic network analysis

Part I: Introduction

Chapter 1

1. *Saccharomyces cerevisiae* biology

1.1. *S.cerevisiae* biochemistry

Yeast is a collective term for unicellular basidiomycetous and ascomycetous fungi. These two types of fungi differ only in the way they produce spores.

Basidiomycetous fungi bears sexually produced spores on a “basidium” which is like a club-shaped structure whereas the ascomycetous fungi bear the spores inside a sac-like structure known as “ascus”.

S. cerevisiae is commonly known as Baker’s yeast. *S. cerevisiae* is an organism of choice in large-scale functional analysis. The *S. cerevisiae* genome encompasses 16 chromosomes and is 12-megabases (Mb) in size. With an average of 1 gene per 2kb of genomic sequence, yeast genome roughly encodes 6,200 genes [Goffeau1996]. Another important characteristic, in contrast to higher eukaryotes, is that only 263 *S. cerevisiae* genes possess intronic regions [Costanzo2000]. This makes computational methods for gene identification in yeast very simple. *S. cerevisiae* is known to be stable in both haploid and diploid states. The stability of the mutants makes yeast very attractive for mutational studies and gene function prediction methods. Large scale comparative studies of yeast and human genes have emerged as a powerful approach for human gene function prediction. This is due to the fact that nearly 50% of human genes responsible for genetic diseases have yeast homologues.

1.1.1. *S.cerevisiae*: A model organism

In the last few years, genetically modified organisms have been extensively employed as a functional genomics tool for predicting the role of genes and their protein products [Kumar2001]. Nevertheless, few models express the expected phenotype thought to be associated with the gene or protein. There is thus a need to further define the phenotype resultant from a genetic modification in order to understand how the transcriptional or proteomic network may accomplish altering the metabolism and bringing forth the expected phenotype. *Saccharomyces cerevisiae* is the most widely studied type of yeast and is an extensively used eukaryotic organism for experimentation in biological research. *S. cerevisiae* also has the distinction of being the first eukaryotic organism with a completely sequenced genome. In fact, its genome sequence was completed in 1996 [Dujon1996 and Goffeau1996]. An important characteristic of this organism is that it is amenable to genetic modification, and it is therefore possible to engineer the metabolism and thereby exploit the organism as a host, for example in industrial production of many different chemicals. Yeast has been increasingly used in the area of metabolomics. As currently defined, "metabolome" approaches stand for the approaches which use the complete pool of cellular metabolites. This includes the whole range of molecules and intermediates which are subjected to biochemical conversions through metabolic pathways, for generation of chemical blocks and energy for growth and for maintenance of cellular functions. "Metabolic profiling" of cellular concentrations of metabolites provides the detailed snapshot of the cell's phenotype. Metabolic profiling has been extensively applied to yeast and other organisms [Raamsdonk2001, Adams2003, Trethewey1999, Fiehn2002, Watkins 2002, and Castrillo2003]. These large-scale metabolic profiling studies also lead to development of sensitive, large-scale high-throughput methods for "metabolite screening" [Oliver2002]. Another approach called "metabolite footprinting" uses metabolite profiling methods for analyzing specific metabolites which are released into the medium. Kell et al. used this footprinting approach for the classification of yeast knockout mutants [Kell2000]. In recent years, new and powerful methods have been developed for yeast knockout analysis. Among these are metabolic control analysis, strategies for the elucidation of the function of new genes and metabolic pathways as well as the role of amino acids and other specific metabolites in controlling gene expression, metabolome analysis, metabolic profiling approaches for biomarker discovery, and drug target screening [Teusink1998, Raamsdonk2001, Trethewey2001, Fuente2002, Weckwerth2002, Fafournoux2000, Hansen2000, So2000, Griffin2001, and Watkins2002].

1.1.2. Physiology: diauxic growth

The effect of glucose on a variety of cellular processes has been extensively studied in *Saccharomyces cerevisiae*. Some of these include glucose repression of genes used in growth on alternative carbon sources and the induction of genes needed for glucose transport and protein synthesis [Carlson1999, Gancedo1998, Johnston1999, Warner1999, and Newcomb2003]. Glucose is known to have a profound effect on the transcription of yeast genes. It is known that switch from anaerobic growth to aerobic respiration upon depletion of glucose is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage [Johnston1992]. This shift from anaerobic fermentation of glucose to aerobic respiration of ethanol is called the "diauxic shift". The “Diauxic shift” is known to involve major changes in gene expression [DeRisi1997]. It is known that genes encoding glycolytic enzymes are down-regulated as glucose gets exhausted, whereas the expression levels of genes involved in oxidative metabolism increase.

1.2. Carbohydrate metabolism

Yeast cells have evolved to undergo a variety of metabolic changes in response to fluctuating nutrient levels in the environment, many of which are coordinated by proteins such as TOR, Sch9, and PKA [Kaeberlein2007]. TOR proteins are known to be highly conserved from yeast to humans and regulate multiple cellular processes in response to nutrients, including cell size, autophagy, ribosome biogenesis and translation, stress response, actin organization, carbohydrate and amino acid metabolism [Schmelzle2000]. Sch9 and Protein kinase A (PKA) are nutrient-responsive protein kinases that modulate replicative aging in yeast [Lin2000, Fabrizio2004]. It is well known that yeast responds to decreasing glucose levels by shifting growth behavior from one that favors fermentation to one that favors respiration [see section 1.2.1].

1.2.1. Central carbon metabolism

The salient feature of eukaryotic central carbon metabolism is its dissection into cytosolic and mitochondrial subpathways, connected by intercompartmental transport of metabolites [Michal1998, Rose1989, Strathern1982, and Zimmermann1997]. Tricarboxylic acid (TCA) cycle operates in the mitochondria, and the glycolysis and the pentose phosphate pathway (Pen*P*p) are located in the cytosol [Fraenkel1982, Gancedo1989, and Pronk1996]. It is known that oxaloacetate (OxAc), pyruvate (Prv) and acetyl-CoA (AcCoA) are present in both mitochondria and cytosol. In addition, systems for their transport across the mitochondrial membrane have been identified [Pronk1996, Kispal1993, Roermund1999 Palmieri1999]. Hence, these three intermediates are key metabolites to distinguish cytosolic (cyt) and mitochondrial (mt) pools.

In yeast, OxAc is produced both in TCA cycle, and in cytosol by the action of pyruvate carboxylase [Fraenkel1982, Gancedo1989, Rohde1991, van Urk1989]. Cytosolic OxAc is actively transported by the proton motive force at the inner mitochondrial membrane. Cytosolic Prv is produced in the cytosol by glycolysis and a fraction of it is transported into the mitochondria yielding mitochondrial Prv [Gancedo1989, Pronk1996]. In addition, mitochondrial Prv is also synthesized from malate by malic enzyme. The transport is actively driven by the mitochondrial proton motive force suggesting a largely unidirectional transport from the cytosol into the mitochondria. Mitochondrial AcCoA and cytosolic AcCoA can be derived from mitochondrial Prv and cytosolic Prv, respectively, either by the pyruvate dehydrogenase complex in the mitochondria, or via a cytosolic 'by-pass pathway' [Haarasilta1977]. AcCoA can cross the inner mitochondrial membrane via the 'carnitine shuttle'. This shuttle consists of the carnitine O-acetyltransferase serves to balance the cytosolic and mitochondrial AcCoA pools by facilitated diffusion.

1.2.2. Alternative carbon sources: glucose, fructose and galactose

Saccharomyces cerevisiae can grow with a variety of carbon sources, but glucose and fructose are the preferred carbon sources. Presence of glucose and fructose lead to the down-regulation of the synthesis of those enzymes which are required for the catabolism of other alternative carbon sources [Gancedo1998]. This is also known as “catabolite repression”. *S.cerevisiae* encodes for GAL genes which are required for the galactose catabolism. The galactose metabolism is subject to dual control namely via the induction of the GAL genes by galactose and via the repression of the GAL genes by the presence of glucose [Johnston1992]. While translational control by glucose is rare, glucose triggers inactivation and/or proteolysis of a number of proteins. Glucose is also known to be involved in another phenomenon known as “catabolite inactivation” in which glucose triggers inactivation and/or proteolysis of a variety of proteins [Holzer1976]. Glucose is known to cause rapid phosphorylation of Fructose 1, 6-bisphosphate and proteolytic degradation of this enzyme [Funayama1980, Müller1981].

1.3. *S. cerevisiae* gene knockout library

In the present work, we used yeast haploid knockout strains for studying the physiological growth profiles associated with individual gene knockouts.. Section 1.3.1 and 1.3.2 describes the reference strain and the detailed description of the knockout strains used in the study, respectively.

1.3.1. Reference strain

Saccharomyces cerevisiae deletion mutants with a parental genotype of BY4742 $MAT\alpha$ with his, leu, lys and ura auxotrophy were obtained from Open Biosystems (Heidelberg, Germany). This mutant library was used for the entire experimental work by our collaboration partners, the group of Dr. Elmar Heinzle (Technical biochemistry, Saarland University, Germany). The above mutants also possess antibiotic Geneticin resistance, which is used as a marker. From this collection, the parental strain, which was used as the reference strain, and a set of deletion mutants, where genes are known to be involved in central carbon metabolism, and few strains with unknown function were chosen for further analysis.

1.3.2. Description of selected knockout mutants

The Saccharomyces Genome Database (SGD) and the Gene Ontology (GO) provide rich and up-to-date resources for annotation concerning the unique and multiple functions of the *S. cerevisiae* genes [Cherry1998, sgd, Harris2004, go, go2000, Dwight2002, Hong2007]. GO provides a rich, precise and structured controlled vocabulary for describing the cellular role of genes and gene products in a given organism. SGD collects and organizes biological information about the chromosomal features and gene products of the budding yeast *Saccharomyces cerevisiae*. In the scenario of ever increasing and changing biological knowledge of cellular roles of gene products, SGD and GO provide media for organizing and querying biological annotations for individual genes and gene products at various stages of completion and for deciphering probable or predicted links between cellular roles of two or more genes in the same or different organisms. This is made possible by inclusion of evidence coming from high-throughput experiments and computational predictions, in the absence of published experimental data [Hong2007].

The preliminary set of genes which are known or putative regulators of central carbon metabolism in yeast, were selected using two basic criteria: 1) the deletion must be viable, 2) the deletion must be active in more than three cellular processes depicted by the GO classification [Figure 1]. The hypothesis is that if an ORF is involved in multiple cellular processes, it might result in a more explicit phenotype in knockout experiments.

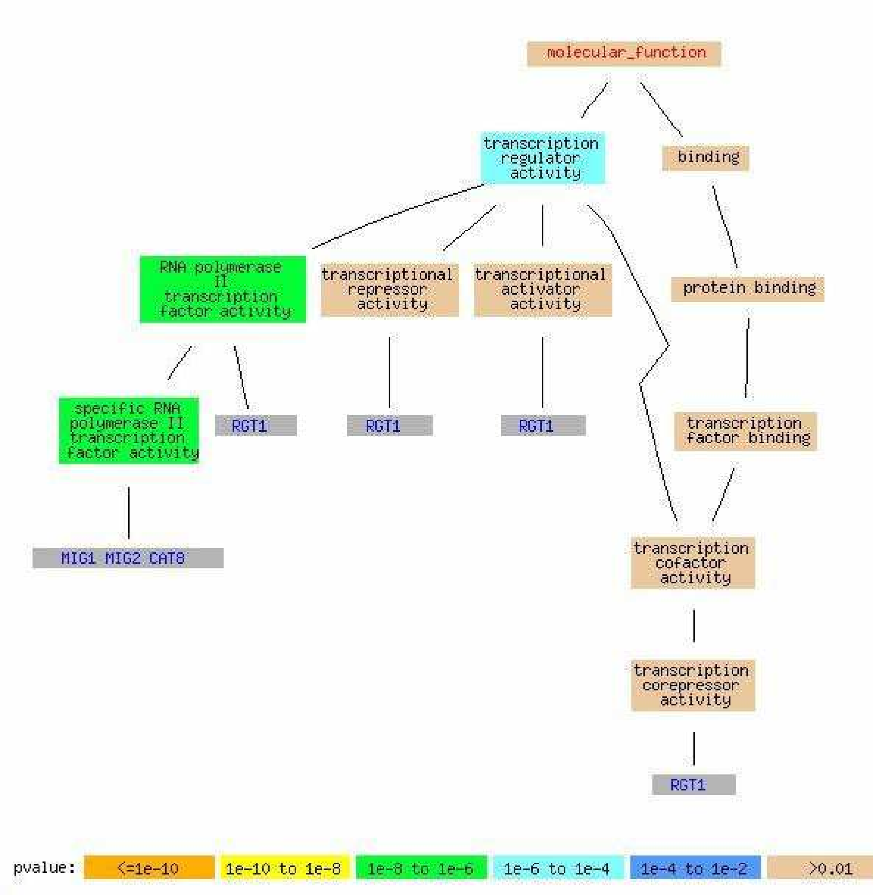

molecular_function
transcription regulator activity
binding
RNA polymerase II transcription factor activity
transcriptional repressor activity
transcriptional activator activity
protein binding
specific RNA polymerase II transcription factor activity
RGT1
RGT1
RGT1
transcription factor binding
MIG1 MIG2 CAT8
transcription cofactor activity
transcription corepressor activity
RGT1
pvalue:
<=1e-10
1e-10 to 1e-8
1e-8 to 1e-6
1e-6 to 1e-4
1e-4 to 1e-2
>0.01

Figure 1 *Gene Ontology function terms associated with the transcription factor RGT1*

1.4. GO annotations for selected knockout mutants

Table 1 lists the molecular function, biological processes and location GO terms associated with the entire mutant set under study.

ORF	Accession	Mutant Id	Biological process	Molecular function	Cellular component
ABZ1	YNR033W	53	para-aminobenzoic acid metabolism	4-amino-4-deoxychorismate synthase activity	Unknown
ABZ1	YNR033W	62	para-aminobenzoic acid metabolism	4-amino-4-deoxychorismate synthase activity	Unknown
ACE2	YLR131C	17	G1-specific transcription in mitotic cell cycle	transcriptional activator activity	Nucleus
ADR1	YDR216W	25	Transcription	transcription factor activity	nucleus
CAT8	YMR280C	6	regulation of transcription from Pol II promoter	specific RNA polII transcription factor activity	Nucleus
CYB2	YML054C	1	electron transport	L-lactate dehydrogenase	Mitochondrial intermembrane space
DLD2	YDL174C	46	aerobic respiration	D-lactate dehydrogenase (cytochrome) activity	Mitochondrial inner membrane
FBP1	YLR377C	61	Gluconeogenesis	fructose-bisphosphatase activity	Cytosol
FBP26	YJL155C	22	Gluconeogenesis	fructose-2,6-bisphosphate 2-phosphatase activity	Cytosol
FOX2	YKR009C	34	fatty acid beta-oxidation	3-hydroxyacyl-CoA dehydrogenase activity	Peroxisomal matrix
FUM1	YPL262W	10	tricarboxylic acid cycle	fumarate hydratase activity	Cytosol
GAD1	YMR250W	5	response to oxidative stress	glutamate decarboxylase activity	Cytoplasm
GAL10	YBR019C	50	galactose metabolism	unknown	Unknown
GAL11	YOL051W	9	transcription from Pol II promoter	RNA polymerase II transcription mediator activity	mediator complex
GAL4	YPL248C	11	regulation of transcription	DNA-dependent transcriptional activator activity	Nucleus
GAL7	YBR018C	49	galactose metabolism	UTP-hexose-1-phosphate uridylyltransferase activity	Cytoplasm
GAL80	YML051W	2	regulation of transcription	DNA-dependent transcription	Nucleus
GCR2	YNL199C	33	regulation of transcription from Pol II promoter	transcriptional activator activity	Nucleus
GLK1	YCL040W	16	carbohydrate metabolism	glucokinase activity	cytosol
GLO1	YML004C	3	glutathione metabolism	lactoylglutathione lyase activity	Unknown
HXK2	YGL253W	27	fructose metabolism	hexokinase activity	Nucleus
IMP2	YIL154C	64	DNA repair	transcription co-activator activity	Unknown
KGD1	YIL125W	55	tricarboxylic acid cycle	oxoglutarate dehydrogenase (lipoamide) activity	Mitochondrial matrix
KGD2	YDR148C	23	tricarboxylic acid cycle	unknown	Mitochondrial matrix
LAT1	YNL071W	57	pyruvate metabolism	dihydrolipoamide S-acetyltransferase activity	Mitochondrion
LEU4	YNL104C	58	leucine biosynthesis	2-isopropylmalate synthase activity	Cytoplasm
MAE1	YKL029C	18	pyruvate metabolism	malate dehydrogenase (oxaloacetate decarboxylating) activity	Mitochondrion
MAL33	YBR297W	42	regulation of transcription	DNA-dependent* transcription factor activity	Nucleus
MIG1	YGL035C	31	regulation of transcription from Pol II promoter	RNA pol II transcription factor activity	Nucleus
MIG2	YGL209W	26	regulation of transcription from Pol II promoter	RNA pol II transcription factor activity	Nucleus
MSN4	YKL062W	20	response to stress	transcription factor activity	Nucleus
NGG1	YDR176W	24	histone acetylation	transcription cofactor activity	SAGA complex
NRG1	YDR043C	41	regulation of transcription from Pol II promoter	DNA binding activity	Nucleus
NRG2	YBR066C	51	invasive growth	transcriptional repressor activity	Nucleus

ORF	Accession	Mutant Id	Biological process	Molecular function	cellular component
PCK1	YKR097W	39	gluconeogenesis	phosphoenolpyruvate carboxykinase (ATP) activity	cytosol
PFK1	YGR240C	30	glycolysis	6-phosphofructokinase activity	Cytoplasm
PFK2	YMR205C	4	glycolysis	6-phosphofructokinase activity	Cytoplasm
PFK26	YIL107C	54	fructose 2,6-bisphosphate metabolism	6-phosphofructo-2-kinase activity	Cytoplasm
PFK27	YOL136C	36	fructose 2,6-bisphosphate metabolism	6-phosphofructo-2-kinase activity	Cytoplasm
PGU1	yjr153w	40	pseudohyphal growth	polygalacturonase activity	Extracellular
RBK1	YCR036W	43	ribose metabolism	ATP binding activity	Unknown
RGT1	YKL038W	19	glucose metabolism	DNA binding activity	nucleus
RPE1	YJL121C	38	pentose-phosphate shunt	ribulose-phosphate 3-epimerase activity	Cytosol
RTG3	YBL103C	47	transcription initiation from Pol II promoter	specific RNA pol II transcription factor activity	Nucleus
SFA1	YDL168W	45	formaldehyde assimilation	formaldehyde dehydrogenase (glutathione) activity	Unknown
SIN4	YNL236W	63	transcription from Pol II promoter	RNA pol II transcription mediator activity	mediator complex
SIP3	YNL257C	7	transcription initiation from polII promotor	transcription cofactor activity	nucleus
SNF11	YDR073W	14	chromatin modeling	RNA pol II transcription factor activity	Nucleosome remodeling complex
SNF2	YOR290C	65	chromatin modeling	RNA pol II transcription factor activity	Nucleosome remodeling complex
SNF5	YBR289W	59	chromatin modeling	RNA pol II transcription factor activity	Nucleosome remodeling complex
SNF6	YHL025W	37	chromatin modeling	RNA polymerase II transcription factor activity	Nucleosome remodeling complex
SRB8	YCR081W	44	negative regulation of transcription from Pol II promoter	RNA polII transcription mediator activity	transcription factor complex
SSN2	YDR443C	60	negative regulation of transcription from Pol II promoter	RNA pol II transcription factor activity	transcription factor complex
SSN3	YPL042C	29	protein amino acid phosphorylation	RNA pol II transcription factor activity	transcription factor complex
SSN8	YNL025C	52	meiosis	RNA pol II transcription factor activity	transcription factor complex
SUC2	yil162w	56	sucrose catabolism	beta-fructofuranosidase activity	Cytoplasm
SWI3	YJL176C	21	chromatin modeling	general RNA pol II transcription factor activity	Nucleosome remodeling complex
TAF14	YPL129W	12	transcription initiation from Pol II promoter	general RNA pol II transcription factor activity	Nucleosome remodeling complex
TYE7	YOR344C	8	transcription	transcription factor activity	nucleus
UGA1	YGR019W	28	nitrogen utilization	4-aminobutyrate aminotransferase activity	Intracellular
UGA2	YBR006W	48	response to oxidative stress	succinate-semialdehyde dehydrogenase (NAD(P)+) activity	Unknown
XKS1	YGR194C	15	xylulose catabolism	xylulokinase activity	Unknown
YBR184W	YBR184W	13	unknown	unknown	Unknown
YDR248C	YDR248C	35	unknown	unknown	Unknown
ZWF1	YNL241C	32	pentose-phosphate shunt	glucose-6-phosphate 1-dehydrogenase activity	Cytoplasm

Table 1 *GO terms associated with the selected mutants.*

Chapter 2

2. Quantitative high-throughput techniques

In the last twenty five years, the desire to combine high-throughput technology, provided by high-performance liquid chromatography (HPLC) and, in particular, by capillary gas chromatography (GC), with the exquisite accuracy and sensitivity provided by mass spectrometry (MS) has led to the development of the most efficient analytical technologies presently available, i.e., LC-MS and GC-MS and variations of them, e.g., LC-tandem MS and GC-tandem MS [Dimitrios2001]. Initially, MS and GC and, later, liquid chromatography (LC) have developed independently. LC-MS is used to analyze polar, thermally labile and high-molecular-mass compounds, such as peptides and proteins. On the other hand, GC-MS is a method of choice for analyzing low-molecular-mass compounds. As a rule, these compounds are mostly polar in nature, and their analysis by GC-MS requires chemical conversion, i.e., derivatization of the compounds into non-polar, volatile and thermally stable derivatives amenable to GC analysis. In the present work, we use GC-MS technique for physiological profiling.

2.1. Gas chromatography-mass spectrometry

2.1.1. Technique

Since metabolites are chemical compounds, these are amenable to analysis techniques like molecular spectroscopy and MS. The selectivity, sensitivity and resolution of these spectroscopy techniques are enhanced by GC or liquid chromatography (LC) processes. The method of choice depends on the type of the metabolite sample to be analyzed [Goodacre2004]. The basic requirements for a substance to be analyzed using GC-MS instrumentation technique are 1) thermal stability and 2) volatility. Hence, GC-MS techniques are best suited for relatively low molecular weight compounds. Additionally, the analyte is subjected to chemical modification using various derivatizing compounds, like TBDMS, to overcome various absorption effects that might lead to inaccurate quantification.

2.1.2. Estimation of intracellular amino acid pool labeling

The introduction of isotope labeling provided the basis for determination of intracellular amino acid pools and pathway activity (flux) in a metabolic network [Noronha2000, Park1997]. In labeling experiments, cells are fed with a labeled substrate (in our studies-^{13}C labeled substrate), and the labeling patterns of certain intracellular metabolites are measured. Any given metabolite can exhibit numerous labeling patterns depending on its chemical compositions and the number of reactions it participates in. Each individual labeling pattern of a given metabolite/intermediate can be regarded as a "labeling state" of that metabolite/intermediate. Since the measurements of labeling patterns of the intracellular metabolites are usually difficult to perform due to their small pool sizes, analysis of the amino acids has been a widely used approach for elucidating the labeling states of intermediates in the central metabolism. Additionally, these labeling measurement data provide independent constraints on the intracellular fluxes and thus enable the determination of the fluxes that are unobservable by the conventional flux analysis using only metabolite balances. Nuclear magnetic resonance (NMR) spectroscopy is extensively used for labeling pattern measurement [Sauer1999, Dieuaide-Noubhani1995, and Zupke1995].

In recent years, two-dimensional ^{1}H–^{13}C NMR spectroscopy has been used for analyzing the labeling states of proteinogenic amino acids [Szyperski1995]. This technique has been used to estimate the flux distribution in *Escherichia coli* [Schmidt1999_a], *Aspergillus niger* [Schmidt1999_b] and *Bacillus subtilis* [Sauer1997]. In last few years, the gas chromatography–mass spectrometry method (GC-MS) has been used as an alternative to NMR [Dauner2000, Donato1993]. GC–MS analysis is much more sensitive than the NMR method and can thus provide labeling measurements with higher precision.

2.2. CalSpec

In the past, intensive research has been carried out concerning the quantitative investigation of metabolic networks as the basis for understanding metabolic functioning and regulation machinery of specific metabolic systems [Bailey1991, Bailey1998, Cameron D.C.1997]. A powerful approach to quantifying metabolic fluxes is based on tracer studies with ^{13}C-labeled substrates combined with mass spectrometry (MS) measurement of ^{13}C labeling patterns of biomass constituents [Christensen1999, Wittmann1999, Wittmann2002]. In these tracer studies, the measured labeling pattern reflects the metabolic state of the cell and is used to calculate intracellular flux parameters. We developed CalSpec software tool for automatic processing of labeling data from MS spectra.

2.2.1. Mass Isotopomer Distribution

The ^{13}C isotope labeling studies basically target a set of metabolites which can be quantitatively measured in a given experimental setup. The unfragmented metabolite fraction is called the parent fraction. The metabolite fragments that arise from the fragmentation of the parent fractions are called partial fractions. A fraction, *x*, without any label (^{13}C), and, having a molecular weight *m*, is denoted as x_m. When any/all of the carbon atoms of the fraction *x,* has ^{13}C label incorporated, then this is denoted as x_{m+i}, where *i* is the number of ^{13}C labeled carbon atoms present in the respective fragment.

Mass isotopomer distributions (MID) are vectors of abundances of various mass isotopomers of the parent and partial fraction of any metabolite. CalSpec calculates the MIDs according to equation 1.

$$\mathrm{MID} = \begin{bmatrix} x_m \\ x_{m+1} \\ x_{m+2} \\ .. \\ .. \\ x_{m+i} \end{bmatrix}$$

Equation 1

2.2.2. Automation of calculation of Mass Isotopomer Distribution

Metabolic flux analysis is useful when applied in comparative studies. Thus experimental and computational tools for efficient metabolic flux analysis on a broad level are highly desired. Efficient flux analysis on a broad level, however, requires a straightforward approach that can be parallelized and automated for all steps involved. A time-consuming and error prone step in the whole procedure of metabolic flux analysis is the extraction of labeling patterns from mass spectra, which is has been done manually, so far. Manual processing of these MS data sets is highly time-consuming and subject to error. The CalSpec software tool addresses this problem by automatic processing of MS spectra. It has been tested and applied to gas chromatography/mass spectrometry (GC-MS) analysis of t-butyldimethylsiloxy (TBDMS)-derivatized amino acids. Amino acids have been shown to provide valuable labeling information for flux calculations in ^{13}C tracer experiments [Wittmann1999, Daumer2000]. CalSpec is especially useful for routinely analyzing samples derived from protein hydrolysates or cultivation supernatants, for example. It should be noted that care has to be taken regarding isobaric interference of the target analytes with other compounds, which might occur in highly complex mixtures such as cell extracts, for example. CalSpec could not detect the interference of the target analytes with other compound, unless they exhibit low abundances compared to the total fragment abundance which can be detected as a bad peak signal.

2.2.3. Implementation

CalSpec automatically performs identification of specified analytes in the MS spectrum and the subsequent quantification of labeling patterns. An overview on the steps involved in the data processing by CalSpec is given in Figure 2.

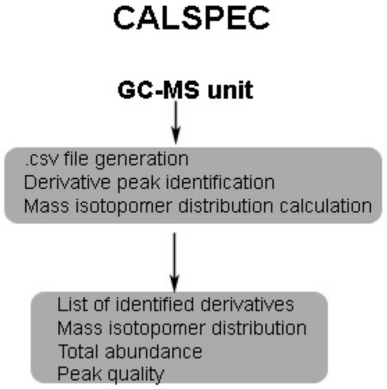

CALSPEC
GC-MS unit
.csv file generation
Derivative peak identification
Mass isotopomer distribution calculation
List of identified derivatives
Mass isotopomer distribution
Total abundance
Peak quality

Figure 2 *CalSpec data processing schema.*

The software module is flexible and can be easily adapted to variants of MS measurements. The module was developed and tested for GC/MS analysis of TBDMS-derivatized amino acids using a GC with HP-5MS capillary column (5% phenylmethylsiloxane diphenylpolysiloxane, 30 m × 250 µm, electron impact ionisation at 70 V and a quadrupole detector; Agilent Technologies, Waldbronn, Germany) as described previously [Wittmann2002]. The instrument was equipped with MSD Productivity ChemStation software (G1701C, Rev. Code 00.01; Agilent Technologies) generating specific data files (*.ms files). These data files are first converted into comma-separated value (*.csv) files using a macro that was supplied by Agilent Technologies and further modified by us. Corresponding macros must be developed to use the program with non-Agilent systems.

The tool has an initial step for conversion of the *.ms file format originating from the GC-MS system into a *.csv file. In this step, a widely used platform independent file format is generated that can be further processed.

In the next step, identification of the amino acid TBDMS-derivatized fragments present in the sample is carried out using the presence of typical mass-to-charge (m/z) signals observed in the spectra. In this way, the sample can be checked for the presence of specified analytes and thus, the preceding experimental protocol can be evaluated. For identification, the user should modify (i) the elution time (T_e) of an analyte in the GC run, and (ii) the m/z values of corresponding specific ion clusters to be observed, in the **param.txt** parameter file according to the format. Table 2 lists the standard parameters used for individual amino acid fragments, in CalSpec. If the user has a different set of parameters, then it should be defined in the param.txt before running the CalSpec program. Parameters are defined in comma separated .txt file where the first element in a row is amino acid derivative elution time and the subsequent elements in the row are the molecular weight of individual fragments coming from that amino acid, after fragmentation steps (see table 2). These parameters are used to check for the presence of an analyte. Identification of peaks is currently performed in a window of (T_e - 0.25) to (T_e + 0.25) min for all analytes. In each time window, the spectrum is scanned at the specified m/z values, whereby the user can define a threshold for each signal that has to be exceeded to indicate presence of the corresponding analyte. In this way, signals with low abundance can be excluded that are subjected to interference with background noise and therefore should not be considered for flux estimation [Daumer2000].

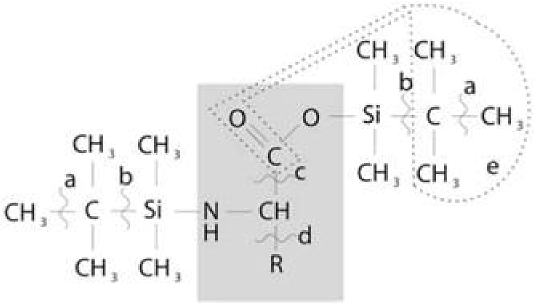


Figure 3 *Fragmentation of TBDMS-derivatized amino acids.* The amino acid with its specific side chain (R) is in gray. Fragmenting at the denoted positions leads to the following fragments: M = Molecular weight of parent fragment (a) methyl group dissociation; (b) tert-butyl group dissociation; (c) C(O)O-TBDMS ion dissociation; (d) the double silylated fragment and the side chain (sc)+, consisting of R and possibly further TBDMS groups dissociation; (e) CO of the amino acid and a tert-butyl group (grouped within the dashed line) dissociation.(Nanchen 2006)

To ensure the presence of an analyte, the specific m/z values, correlating to typical fragments such as [M0], [M1], or [M2] for TBDMS-derivatized amino acids in the time window should exceed the threshold set.

```
! Elution time, M0, M1, M2, M3,
12.18,158,232,260
12.83,218,246
15.40,186,260,288
16.55,200,274,302
17.37,200,274,302
18.20,184,258,286
23.20,218,292,320
23.91,288,362,390
24.59,302,376,404,159
26.09,234,308,336,302
27.71,316,390,418,302
30.24,330,432
31.00,302,315,417
32.45,329,431,488
34.77,414,442
36.95,196,338,440
37.20,517,545
37.57,364,438,466,302
*
```

Table 2 *Parameter file “ param.txt”. The first element of a row is the elution time for a particular amino acid fragment in gc-ms equipment and the rest of the elements of the row stands for m/z ratios of unlabeled fragments for a given amino acid (like [M0]+0, [M0]+0, or [M0]+0)*

[M0], [M1], [M2] etc are the typical fragments obtained after cleavage of the parent fragment by dissociation of side chains of varying molecular weights like 57 D(tert-butyl group), 159 D(C(O)O-TBDMS), 302 D(double silyl fragment) etc. [Figure 3]. Table 3 lists the amino acid fragments and all of their labeled subfragments identified.

Amino acid fragment Mol. Wt.	*Fragments identified*	*Amino acid fragment*	*Fragments identified*
Ala 158	M[0]+[0,1,2]	Asp 316	M[0]+[0,..,3]
Ala 232	M[1]+[0,1,2]	Asp 390	M[1]+[0,..,3]
Ala 260	M[2]+[0,..,3]	Asp 418	M[2]+[0,..,4]
Gly 218	M[0]+[0,1]	Asp 302	M[3]+[0,1,2]
Gly 246	M[1]+[0,1,2]	Glu 330	M[0]+[0,..,4]
Val 186	M[0]+[0,..,4]	Glu 432	M[1]+[0,..,5]
Val 260	M[1]+[0,..,4]	Asn 302	M[0]+[0,1,2]
Val 288	M[2]+[0,..,5]	Asn 315	M[1]+[0,..,3]
Leu 200	M[0]+[0,..,5]	Asn 417	M[2]+[0,..,4]
Leu 274	M[1]+[0,..,5]	Lys 329	M[0]+[0,..,5]
Leu 302	M[2]+[0,..,6]	Lys 431	M[1]+[0,..,6]
Ile 200	M[0]+[0,..,5]	Lys 488	M[2]+[0,..,6]
Ile 274	M[1]+[0,..,5]	Arg 414	M[0]+[0,..,5]
Ile 302	M[2]+[0,..,6]	Arg 442	M[1]+[0,..,6]
Pro 184	M[0]+[0,..,4]	His 196	M[0]+[0,..,4]
Pro 258	M[1]+[0,..,4]	His 338	M[1]+[0,..,5]
Pro 286	M[2]+[0,..,5]	His 440	M[2]+[0,..,6]
Met 218	M[0]+[0,..,4]	Gln 517	M[0]+[0,..,4]
Met 292	M[1]+[0,..,4]	Gln 545	M[1]+[0,..,5]
Met 320	M[2]+[0,..,5]	Tyr 364	M[0]+[0,..,8]
Ser 288	M[0]+[0,1,2]	Tyr 438	M[1]+[0,..,8]
Ser 362	M[1]+[0,1,2]	Tyr 466	M[2]+[0,..,9]
Ser 390	M[2]+[0,..,3]	Tyr 306	M[3]+[0,1,2]
Thr 302	M[0]+[0,1,2]	Phe 234	M[0]+[0,..,8]
Thr 376	M[1]+[0,..,3]	Phe 308	M[1]+[0,..,8]
Thr 404	M[2]+[0,..,4]	Phe 336	M[2]+[0,..,9]
Thr 159	M[3]+[0,1,2]	Phe 302	M[3]+[0,1,2]

Table 3 *Amino acid fragments identified by CalSpec***.**

For example, for the amino acid fragment ala_158 (alanine with mol.wt. 158), CalSpec searches for the unlabeled (i.e. ([M0+0)/([M1]+0)/([M2]+0)) fragment, 1 carbon labeled ([M0]+1) and 2 carbons labeled ([M0]+2) [Figure 3].This is denoted in an abbreviated manner in table 3, as M[0]+[0,1,2]. First, abundance levels of different mass fractions of the analytes are calculated. It is known that the high resolution of GC separation can lead to isotope fractionation, i.e. gradients for the relative abundance of different mass isotopomers over a peak [Daumer2000]. To correctly extract labeling information from a peak, all mass scans performed by the MS detector during the elution of the peak have to be taken into account. CalSpec therefore integrates the different m/z signals, by calculating mean abundances for all mass isotopomer fractions over the entire peak. The automated specification of the time window ensures that the same signals are considered in every measurement. By contrast, manual integration is error prone and tedious. The output file is generated, which contains a list of the specified analytes, information about their presence, and the abundance of mass isotopomer fractions. This file has *.xls format and therefore can be easily imported into any text editing application.

2.2.4. Results

The developed software tool, CalSpec, is useful for efficient processing of ^{13}C labeling data from MS measurements in ^{13}C flux analysis [Talwar2003].

```
CalSpec result file
Input spectra file: sample_input.CSV
Elements in row:
Element 1:  amino acid derivative type, Element 2: mass/charge, Element 3:Molecular weight of
the fragment, Element 4: abundance[mx]/abundance[m0], Element 5: specific ion abundance
Element 6: total fragment abundance, Element 7: peak quality warning
Element 8: detector limit warning abundance>10^8
 met    320    M[2]+0     1.000000    2877.111111
 met    321    M[2]+1     1.914343    5507.777778
 met    322    M[2]+2     1.326794    3817.333333
 met    323    M[2]+3     0.478605    1377.000000
 met    324    M[2]+4     0.759816    2186.074074
 met    325    M[2]+5     0.235885    678.666667     16443.962891    *Bad
Peak*
```

Table 4 *Typical result file generated after a CalSpec run.*

These MS data sets are generated in huge numbers due to (i) replicate measurements of one sample to assess the confidence in the measured values and estimation of error; (ii) replicate measurements of one experiment to check for isotopic steady-state; or (iii) different measurements of one sample with different protocols to obtain additional labeling information via alternative fragments. Data processing by CalSpec takes only a few seconds per spectrum, whereas the same task requires up to 30 min or more if done manually. Table 4 is an example of a part of one such output file generated after CalSpec execution.

Part II: Statistical Analysis of the Data

Chapter 3

3. *Introduction to the statistical analysis of data*

Metabolic profiles can be mined using a range of pattern recognition techniques, including hierarchical cluster analysis, principal components analysis, partial least squares and neural networks. Furthermore, the metabolic perturbations generated due to gene knockouts can be used to classify strains by grouping mutants in clusters if they arose from deletions, which are involved in identical or related cellular functions. This suggests that a process of defining a phenotype through the global changes induced in metabolism could be used to predict the function of genes deleted in a given system [Griffin2004]. Raamsdonk et.al. found that yeast mutants involving the deletion of one of two genes encoding the same enzyme, 6-phosphofructo-2-kinase, produced identical metabolic phenotype, and deletions involving oxidative phosphorylation also clustered together. Thus, such a process of defining a phenotype through the global changes induced in metabolism may be used to predict the function of genes deleted or up regulated in a given system through comparative metabolomics [Raamsdonk2001, Gareth2004]. In the present study, we have applied various statistical techniques for the analysis of metabolic profiles and for understanding the effects of gene knockouts on the metabolic network functionality of wild type yeast.

3.1. Description of the data

Sections 3.1.1-3.1.3 describes the data used in the statistical analysis. We have used the physiological profiles, amino acid labeling profiles and the transcript co-response data for elucidating the interrelationships in various gene knockouts under study.

3.1.1. Physiological Growth (PG) profiles

The physiological growth (PG) profile of a mutant refers to the vector of physiological features like yield coefficients, which characterize the growth behaviour of a gene-knockout mutant. In the present study, we used a data set I comprising 59 single gene knockout mutants. Out of these 59 mutants, 37 are grown under conditions in which ^{13}C-labeled glucose is the sole carbon source, 41 mutants are grown under conditions in which ^{13}C-labeled fructose is the sole carbon source and, analogously, 24 mutants are grown with ^{13}C-labeled galactose as the sole provider of carbon atoms. These mutant subsets were selected for further computational analysis as they do not exhibit severe growth defects under the supply of the specific carbon source and the PG profiles are available from our collaboration partners.

The physiological profile PG_i for mutant i consists of the growth rate μ_i, the biomass yield Yxs_i, the ethanol yield Yp_i, the rate of biomass production Q_i and the rate of ethanol production Qp_i. A dissimilarity matrix is generated using the PG_i for all three growth conditions, using the Euclidean distance as a dissimilarity measure.

3.1.2. GC-MS amino acid fragment fractional labeling (FL) profiles

The GC-MS amino acid fragment fractional labeling profile (FL profile) for a mutant is a vector of fractional labelings of selected TBDMS [Tert-butyldimethylsilyl] amino acid derivatives and the respective mass fractions of these amino acid derivatives. For example alanine has two fractions namely ala_260 and ala_232. Other amino acid fractions which are quantified using the GC-MS spectra are gly_246, val_288, val_260, val_186, ile_200, pro_286, ser_390, ser_362, ser_288, thr_404, thr_376, phe_336, asp_418, glu_432, arg_442, arg_414 and tyr_466. For an amino acid fraction with three carbon atoms, the fractional labeling is calculated as follows:

$$FL_i = [0*(m+0)+1*(m+1)+2*(m+2)+3*(m+3)]/3 \qquad \textbf{Equation 2}$$

Where i denotes the index number of the fraction, m stands for the molecular weight of the i^{th} fraction (like [M-57], [M-85], [M-159] etc.) and $(m+0)$ denotes the intensity of the gc-ms peak (i.e. count associated with the signal) (known as abundance henceforth) of the unlabeled i^{th} amino acid fraction. Similarly $(m+1)$ is the abundance of the amino acid fraction where exactly one carbon is labeled (^{13}C).

3.1.3. Transcript co-response data

An important advance in the area of reconstruction of function relationships among genes is the elucidation of transcriptional units which are characterized by correlated changes in the mRNA expression levels. Transcript co-response profiles are the basis for the attempt to deduce functional relationships between genes from correlations in the corresponding mRNA expression levels. CSB.DB is a publicly accessible systems biology database for the analysis of large-scale transcript co-response data [Steinhauser2004a]. We downloaded from the CSB.DB database the correlation coefficient (ρ) for all-against-all pair combinations of the genes in the mutant set I under study. The implicit assumption here is that common transcriptional control of genes is reflected in corresponding, synchronous changes in transcript levels [Steinhauser2004].

3.2. Analysis of metabolic profiling data using clustering algorithms

Once quantitative datasets are obtained using high throughput techniques, there is a wide spectrum of data-analysis strategies that can be pursued with metabolite profiles [Roessner2001, Roessner-Tunali2003, Sauter1988, Allen2003, Brindle2002, Huhman2002]. The fundamental approach is to simply compare the abundance of a metabolite between an experimental and a control sample, and to use standard statistics to assess the significance of differences. These approaches can then be extended by one dimension in order to look at correlations in abundance of individual metabolites across different samples [Kose2001]. In recent past, a lot of interest has been focused on grouping approaches for whole metabolite profiles [Roessner2001, Fiehn2000, Huhman2002, and Kose2001].

3.2.1. Methods

In general, clustering algorithms aggregate observations into groups, henceforth called clusters, such that the pairwise dissimilarities between observations in the same cluster are lower than those of observations assigned to different clusters [Jain1999]. Generally, various clustering methods fall under either partitional or hierarchical clustering technique [Figure 4]. A partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique.

There are three types of clustering algorithms combinatorial, mixture modelling and mode seeking. Combinatorial algorithms do not assume any probability model in the reference data whereas mixture modelling algorithms treat each data point as a sample from some population described by a probability density function [Hastie2001]. Mode seeking algorithms are nonparametric in nature and directly estimate distinct modes of the probability density function. The squared error technique belongs to the partitional algorithms category, and minimizes the squared error for a clustering m of a observation set l (containing k clusters) is

$$e^2(l,m)=\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|x_i^{(j)}-c_j\right\|^2$$ **Equation 3**

where x_i^j is the i^{th} observation belonging to the j^{th} cluster and c_j is the centroid of the j^{th} cluster [Jain99].

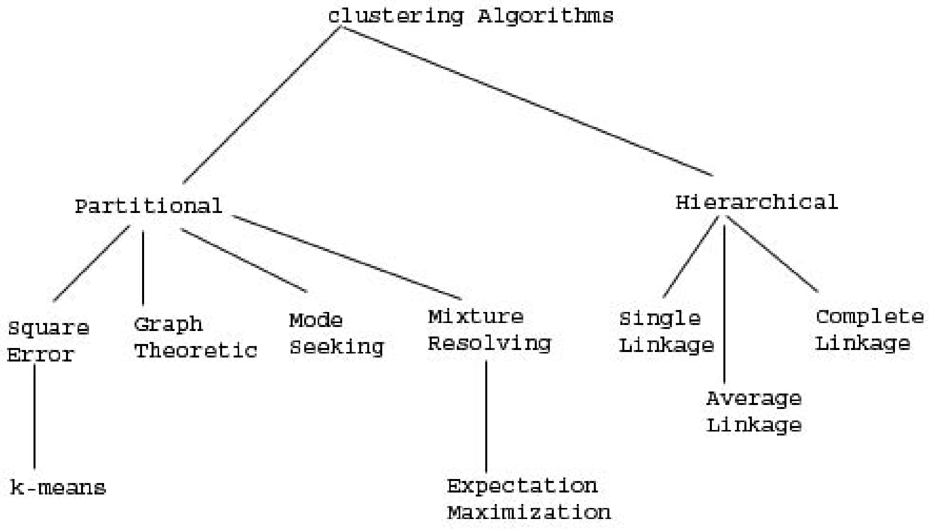


Figure 4 *Clustering algorithms [Jain99]*

The k-means clustering algorithm is the simplest and commonly used algorithm employing a squared error criterion [McQueen 1967]. It starts by partitioning the input points into k initial sets (either randomly, or using some heuristic), followed by the calculation of mean point (centroid) of each set. The algorithm follows the following two steps iteratively. In the first step, it calculates new partitions by associating each input point to the closest centroid. In the second step, the cluster centers are recalculated using the new partitions. This is iterated until convergence, e.g. until there is no reassignment of any pattern from one cluster to another, or until the squared error ceases to decrease significantly after some number of iterations.

3.2.1.1. Supervised and unsupervised learning algorithms

Learning algorithms can be generally divided into two classes namely supervised learning algorithms and unsupervised learning algorithms. Supervised learning algorithms use the response variable to guide the learning process whereas unsupervised learning algorithms look as how the original data is clustered with out any knowledge of response variables.

Supervised methods are powerful methods that can be applied if one has some previous information about which genes are expected to cluster together. In these cases by selecting an initial number of cluster (k), one could perform k–way classification. The Support Vector Machine (SVM) method is one such popular example of supervised learning methods [Brown2000, Quackenbush2001]. SVMs map the input vector x into high- dimensional feature space Z through some nonlinear mapping, chosen *a priori*. In this space, an optimal separating hyperplane is constructed for data classification.

Unsupervised learning methods work with the observed patterns Y_i. Each pattern is usually regarded as an independent sample coming from the underlying unknown probability density function $P(Y)$. For example, density estimation methods like Bayesian networks, and feature selection techniques try to directly identify statistical regularities/irregularities in the input data [Grenander1976, Barlow1989, and Nowlan1990]. In the present thesis, we have applied the Partitioning Around Medoids (PAM) algorithm (sec. 3.2.1.2) and the hierarchical clustering algorithm (3.2.1.3), both of which belong to unsupervised clustering techniques.

3.2.1.2. Partitioning Around Medoids algorithm

Partitioning around medoids (PAM) was originally introduced by Kaufman and Rousseeuw [Kaufman1990]. The general idea of the PAM algorithm is based on the search for k representative objects or medoids among the observations of the dataset. These observations should represent the structure of the data. After identification of a set of k medoids, k clusters are constructed by assigning each observation to the nearest medoid. The objective is to find k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object. We use the ***PAM*** function as defined in the statistical programming environment R [R2005].

For an arbitrary dissimilarity matrix, ***PAM*** aims at minimizing the sum over all objects of the distances to the closest of *k* prototypes [Kaufman1990]. This objective functions is locally optimized in two steps. In the BUILD phase, initial prototypes are chosen. In the SWAP phase, potential single replacements of prototypes with other data points are considered iteratively. Out of all pairs of objects, in which one is a prototype and the other is not, the swap (if any) that decreases the objective function most, is made. The algorithm is well suited for metabolic profiling datasets since it combines the flexibility of hierarchical clustering regarding arbitrary similarity matrices with the optimization approach of *k*-means.

3.2.1.3. Hierarchical clustering algorithm

In hierarchical agglomerative clustering, each mutant is initially assigned to a separate singleton cluster [Jain1999]. Then, iteratively, the two closest clusters in terms of the distance are joined, forming a new node of the clustering tree. The similarity matrix is updated with this new node replacing the two joined clusters. This process is repeated until only a single cluster remains. In each repeat step, the updated similarity matrix is calculated using the mutant dissimilarity between the mutants from the two joined clusters. The average linkage uses the average distance, single linkage the smallest and the complete linkage the largest distance. Hierarchical clustering is the most popular clustering algorithm in diverse areas like DNA microarray analysis due to the easy visualization of the cluster through a dendrogram [Figure 6]. In such a plot, a line connects clusters when they are joined. The height of this line denotes the distance between the clusters. The cluster with the smaller variation is plotted on the left-hand side. Another advantage is that this procedure provides a hierarchy of clusterings with the number of clusters ranging from one to the number of objects [Rahnenfuehrer2006].

3.2.1.4. Silhouette width

We used the silhouette width as the measure of quality of clustering.

The silhouette width is a quantitative measure of the quality of a clustering. Equation 4 displays the formula for the silhouette width, $s(i)$, for a data point *i* in cluster *x*.

$$s(i) = \frac{y(i) - x(i)}{\max(x(i), y(i))} \qquad \textbf{Equation 4}$$

where mutant i belongs to cluster x; $x(i)$ is the average dissimilarity of object i to all members in its cluster x and $y(i)$ is the average dissimilarity of object i to all members of the nearest neighbouring cluster y (i.e. the cluster which has the minimum dissimilarity for the data points in cluster x). The "average silhouette width" for a cluster is calculated by calculating the mean of all $s(i)$ values for that cluster. The average silhoutte width over the entire mutant set is denoted by *s*. It is calculated as an average of the “average silhouette width” for all the clusters.

3.2.2. Results of clustering methods

Unsupervised learning methods have been extensively applied in studies on high throughput DNA microarray data but have not been systematically applied to PG and FL profiles. Here, we present the results from an investigation of metabolite profiling data with PAM and HC clustering algorithms [Hastie2001]. The objective is to partition metabolic profiles corresponding to different mutant sets into groups with higher similarities among mutants within a group than between mutants from different groups. This approach provides a means of estimating the discriminatory power of physiological mutant data including growth rate, biomass production, rate of ethanol production and rate of product formation. In the present analysis framework, we performed a stepwise analysis on four subsets of mutants derived from dataset I, namely: Mutant set M_{total} refers to the set of 59 mutants which is used for calculation of FL profiles from GC-MS spectra; mutant set M_{glu} is the set of 37 mutants which are grown under glucose conditions; mutant set M_{fru} is the set of 41 mutants which are grown under fructose conditions, and mutant set M_{gal} is the set of 24 mutants which are grown under galactose conditions.

We first start with performing large scale analysis of the FL profiles alone.

3.2.2.1. FL profile analysis

We apply following steps, namely step 1 to 3 to the M_{total} mutant dataset. Thereafter, we also present the results that we obtained for individual analysis steps [1-3] in the current section.

Step1: Hierarchical clustering of mutant set M_{total}, using FL profiles. In this step we study whether the overlap in the biosynthetic pathways of amino acids would result in the overlap in the labeling pattern of fragments originating from these biosynthetically linked amino acids. Also, we study whether this association between FL profiles of biosynthetically linked amino acids could be used in clustering of related mutants.

Step 2: Analysis of biosynthetically linked amino-acid fragments using corresponding FL profiles. In this step, we analyze whether the amino acid fragments which are linked by precursor-product relationship, also show similar FL profiles or not.

Step 3: Estimation of the optimal number of clusters in mutant sets: M_{glu}, M_{fru} and M_{gal}, using PAM and HC; In this step, we apply PAM and HC algorithms on the mutants grown with diverse carbon sources. The basic idea is to identify whether the mutants which cluster together, also have some functional relationship, or whether we can make certain hypothesis about the possible functional relationships.

Analysis of step 1 results: Hierarchical clustering of the mutant set M$_{total}$, using FL profiles.

FL profiles show low variance among functionally closely related mutants. We analyzed the complete FL profile for each mutant for the mutant set M_{total} (Table 1). We found that the complete FL profile does not show statistically significant differences among the profiles obtained for the reference strain and for the functionally closely related mutant set M_{total}. This observation is plausible because all the mutants in this set are knock-outs of known or putative regulators of central carbon metabolism in *S. cerevisiae*. Figure 5 represents the box plot of FL profiles for entire mutant set of 59 mutants. This is in agreement with the results presented in a related publication by Gombert et.al. [Gombert2001].

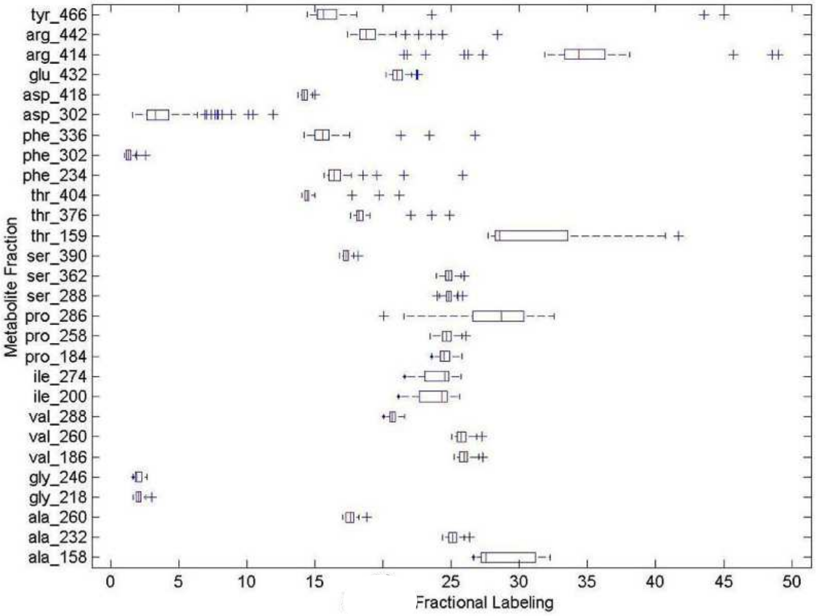


Figure 5 FL *Variance plot for 57 mutants grown under aerobic conditions*. The x-axis denotes a fractional labeling percentage labeling of individual fragments as $FL_i = [0*(m+0)+1*(m+1)+2*(m+2)+..+n*(m+n)]/N$, where n is the number of ^{13}C in a given fragment, N is the total number of carbon atoms in a fragment, $(m+n)$ refers to the intensity of GC-MS peak of the molecular weight= m+n. The y–axis denotes the amino acid fraction with the corresponding molecular weight (using the notation introduced in Section 2.2.3)

Analysis of step 2 results: Analysis of biosynthetically linked amino acid fragments using corresponding FL profiles.

Amino-acid fragments which are in a precursor–product relationship also show close correlation in the FL profiles. We found that amino-acid fragments which are linked biosynthetically also are clustered into close proximity when the FL profile for the mutant spectra is used for hierarchical clustering [Figure 6].

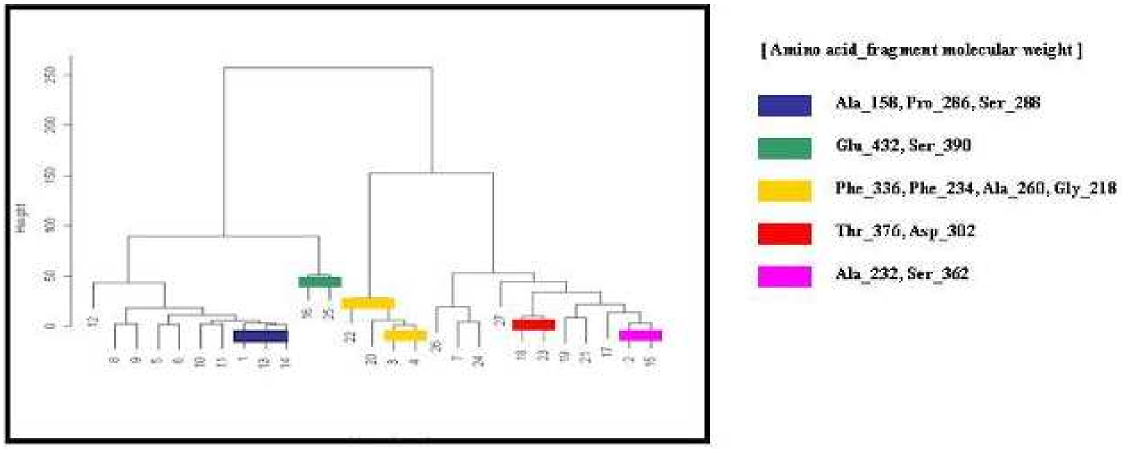


Figure 6 *Dendrogram of FL profiles of amino acid precursor- product network components*

Hierarchical clustering plot using amino acid *FL* profiles for 57 mutants. The y-axis denotes the separation among the clusters; the x-axis denotes the respective clusters from the smallest variation (left) to largest variation (right). Five sets of biosynthetically linked amino acids fragment denoted in the plot are: The first set (colored in blue) consists of Alanine fragment with molecular weight 158 D, represented as Ala_158, and Pro_286 , Ser_288. The other four sets are: II (colored in green): Glu_432, Ser_390; III (colored in yellow): Phe_336, Phe_234, Ala_260, Gly_218; IV (colored in red): Thr_376, Asp_302, V: Ala_232, Ser_362 (colored in pink).

This provides evidence for tight metabolic coupling of amino acids which are in a precursor-product relationship. For example, valine and alanine fragments are used for deciphering the compartmentation of pyruvate and acetyl-CoA by one such precursor-product relationship [Falco1985]. The quantitative information on the enzyme activities could also be estimated by studying the labeling profiles of corresponding amino acid fragments. It is already established that the activity of malic enzyme (Mae1p) can be estimated by quantitative analysis of the *FL* profiles of pyruvate and phosphoenolpyruvate [Boles1998].

Analysis of step 3 results: Estimation of the optimal number of clusters using PAM and HC.

In this step, we compute the optimal number of clusters in our datasets, using the clustering algorithms PAM and HC. Initially, we calculate a dissimilarity matrix, D using

FL profiles for each mutant. Then, we calculate the silhouette width s [see 3.2.1.4]. By construction, s lies in the interval [1, -1]. In our setting, clusters with high silhouette values have the property that the dissimilarity among mutants within the cluster is much lower than the dissimilarity between mutants belonging to different clusters. High silhouette widths (generally silhouette width in the range of 0.7-1) give us confidence in the assignment and elucidation of functional association, if any, among mutants calculated using solely the FL profiles.

Here we present the results for the cluster analysis of FL profiling data for three sets of mutants, namely M_{glu}, M_{fru}, and M_{gal}. There is overlap among the mutant sets whenever a mutant shows considerable growth behaviour in several different conditions. M_{gal} is the smallest mutant set mainly because the majority of the mutants were slow growers under galactose conditions.

M_{glu} leads to s values of 0.31 and 0.36 under PAM and HC, respectively. Hence the quality of this classification is rather weak and resulting clusterings contain many small clusters (9 with PAM and 10 with HC) and singletons [Figure 7].

Figure 7 presents the results of clustering of mutants grown with glucose as the sole carbon source. The graphs on the top show the average silhouette widths in dependence of the number of clusters. The histogram on the bottom shows the silhouette widths of knockouts in the best clustering. N : denotes the number of mutants, and C_j : the optimal number of clusters.

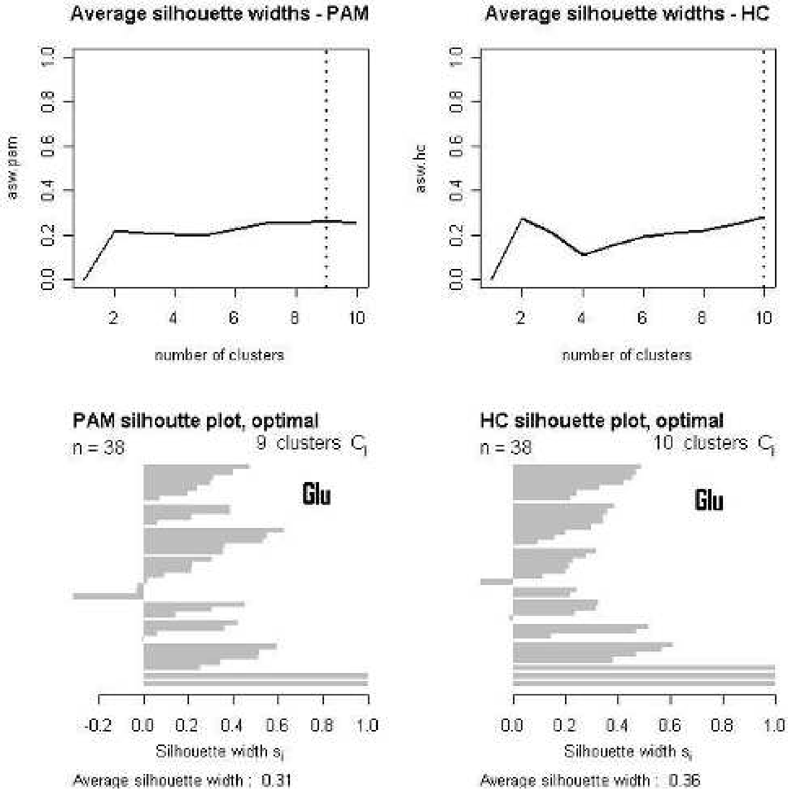


Figure 7 *FL profiles based differentiation of mutant set* M_{glu} .

The s values of 0.36 and 0.31 for PAM and HC for M_{fru} point towards a weak clustering among the mutants and lead to the conclusion that FL data is not sufficiently discriminative and gives the same results as for the glucose conditions [Figure 7].

Figure 8 presents the results of clustering of mutants grown with fructose as the sole carbon source. The graphs on the top show the average silhouette-widths in dependence of the number of clusters. The histogram on the bottom shows the silhouette widths of knockouts in optimal clustering. N : number of mutants; C_j : optimal number of clusters.

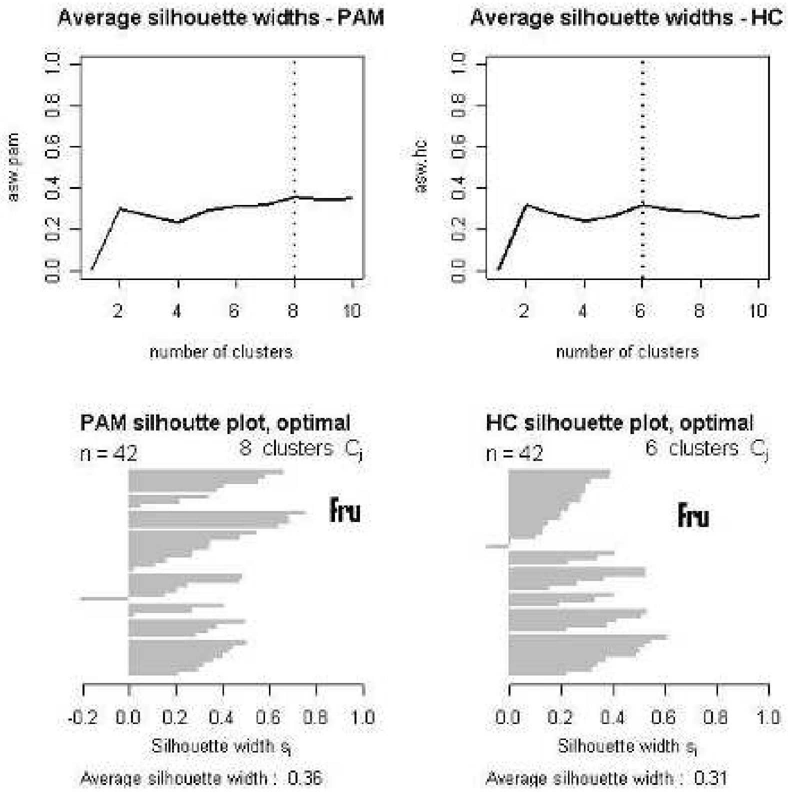


Figure 8 *FL profiles based differentiation of mutant set* M_{fru} .

Figure 9 presents the results of clustering of mutants grown with galactose as the sole carbon source. The graphs on the top show the average silhouette widths in dependence to number of clusters. The histogram on the bottom shows the silhoutte widths of knockouts in optimal clustering. N : number of mutants; C_j : optimal number of clusters. In the case of growth with galactose too, we found the s to be 0.33 and 0.42 for PAM and HC clustering respectively. The clustering results are indicative of the weak grouping as the silhouette width is in the order of 0.5 s units. The s value equal to or less than 0.5 indicates bad clustering. All three mutant sets grown under glucose, fructose and galactose show similar FL profiles. The above result points to a need for integrating complementary sources of data which strengthen the confidence in the predicted functional association among the mutants.

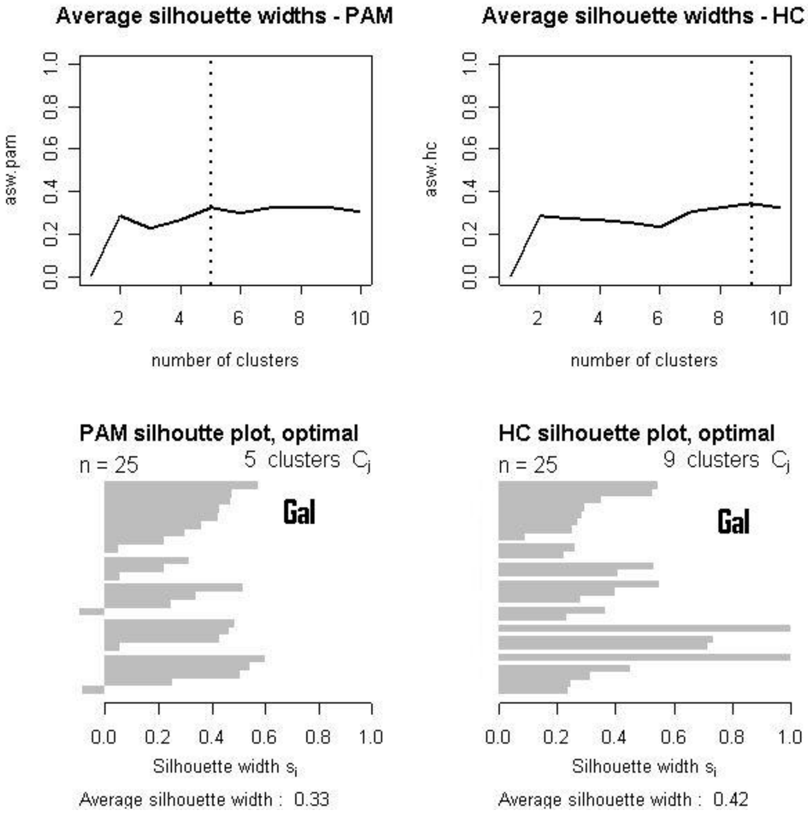


Figure 9 *FL profiles based differentiation of mutant set* M_{gal}

3.2.2.2. *PG* profiles

In this preliminary study, we apply the PAM and HC algorithms to the *PG* profiles of a dataset of 109 mutants [refer appendix 3]. The silhouette width with PAM and HC algorithm was found to be 0.45 and 0.44 respectively. The PG profiles under glucose and fructose conditions show higher similarity than with the PG profiles under galactose conditions [Figure 9]. *PG* profiles provide global features which must be complemented with other heterogeneous data types, for mutant differentiation. *PG* profiles alone were not sufficient for mutant differentiation. Hence, in the next step, we performed an integrated analysis of *PG* profiles and transcript co-response profiles under glucose, fructose and galactose conditions.

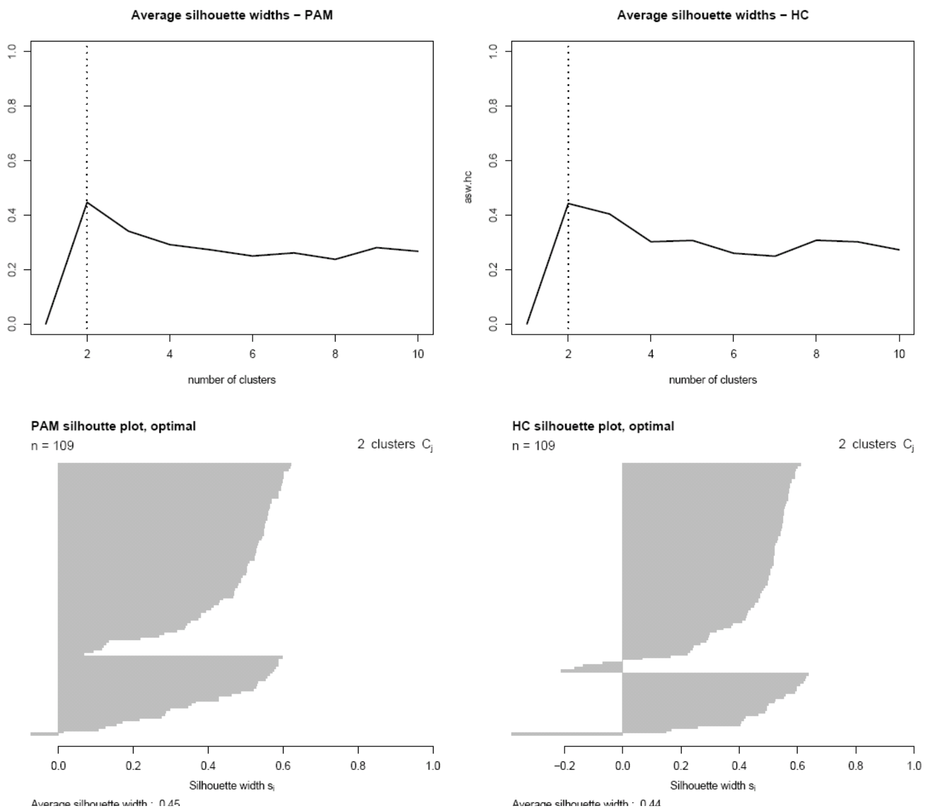


Figure 10 *PG profile based differentiation of mutant set* M_{glu}

3.2.2.3. Pairwise correlation analysis of *PG* profiles and transcript co-response correlation analysis

Here we investigated whether the similarity at the level of *PG* profiles for a set of mutant is also reflected in transcript co-response correlation between gene expression profiles of these mutants?

First we downloaded the transcript co-response data from CSB.DB for our mutant set. Next, we looked for only those ORF pairs where the co-response profile correlation values exceed 0.7. The mutants which show co-response profiles with correlation below 0.7 were not considered to be significant pairs. The value 0.7 was taken to be the cutoff to eliminate any weaker corresponding ORF pairs. Transcript co-response profiles on their own provide a preliminary indication of functional association among mutants. The transcript co-response profiles correlation is calculated using experimental data under culture conditions and hence there is a need to specifically test the accuracy of assignments.

We found that 33% of the highly correlated co-response profile mutant pairs also belong to the same cluster as given by PG profile analysis using the PAM and HC algorithms. It is important to see that we are able to find association, though weakly so, based on global features like fractional labeling and heterogeneous data like transcript co-response profiles. Out of 630 total pairs, 185 mutant pairs show correlation of at least 0.5 on the co-response profile level. Out of these, 91 pairs show a co-response profile correlation greater than or equal to 0.6 and there exist 29 mutant pairs which show co-response profile correlation of at least 0.7. These 29 mutant pairs were further studied as to whether there exists a strong correlation at the level of PG profiles, as well [Figure 11].

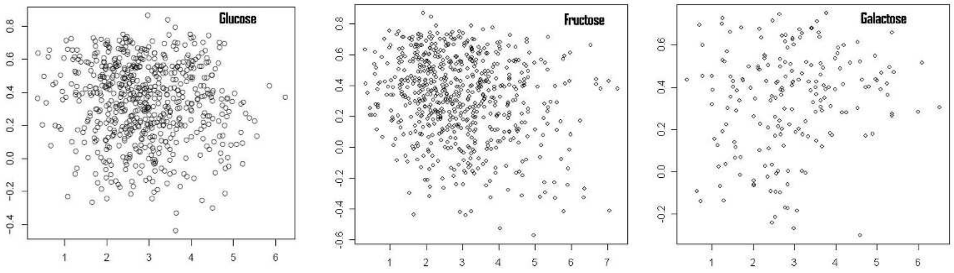


Figure 11 *Correlation of PG profile distances and transcript co-response profiles*

Figure 11 represents scatter plots of the PG profile vs. transcript co-response correlation under glucose, fructose and galactose conditions, respectively. The x-axis denotes pairwise distances for the respective mutant sets, using the PG profile; the y-axis denotes the transcript co-response profiles for all mutant pairs. Out of these 29 mutant pairs, 7 mutant pairs were also assigned to the same cluster using the PG profile analysis. We also corrected for multiple testing on the co-response profile data and the adjusted p-value was less than or equal to 4.347e-05 using the Bonferroni method. This signifies that these p-values relate to the hypothesis that there is a positive correlation between the co-response profile correlation values of mutant pairs.

We performed a Fisher test for the significance of the hypothesis that mutant pairs which show low values of Euclidean distance in the PG profile space show higher correlation in the co-response profile space. We specifically tested for the null hypothesis that PG profile correlation lower than mean (PG correlation) and co-response profile correlation higher than mean (co-response profile correlation) are independent.

We found statistically significant p values for all three sets M_{glu}, M_{fru}, and M_{gal} [Fig 10], confirming the dependence between PG profiles and co-response profiles.. The p-values are p=0.0455 for glucose, p=0.0017 for fructose, and p=0.0404 for galactose.

Table 5 enlists all the 9 pairs of mutant ORF sets which showed high correlation in co-response profiles and close correlation in PG profiles.

ORF Set	Confirmed biological relationship (+)	Suggested biological relationship (*)	No hypothesis on biological relationship (-)
YIL107C, YIL154C		*	
YOL136C, YIL107C	+		
YIL162W, YIL154C			-
YKL062W, YML054C			-
YKR097W, YBR018C	+		
YBR184W, YKL062W			-
YDR043C, YIL107C		*	
YDR073W, YGL035C		*	
YGR194C, YKR097W			-

Table 5 *ORF pairs showing high correlation in co-response profiles and PG profiles*

YIL107C [SGD: S000001369] and YIL154C [SGD: S000001416]

YIL107C is a knockout of the gene coding for PFK26 [Swiss-Prot: P40433] andYIL154C is a knockout of the gene coding for IMP2 [Swiss-Prot: P46972].

IMP2 is a known transcriptional coactivator. It is known that IMP2 [Swiss-Prot: P46972] is involved in glucose depression as well as in regulation of GAL genes.

The role of IMP2 in the galactose metabolism is predicted to be partially dependent on MIG1p along with NRG1p. However, disruption of MIG1p and NRG1p only partially relieves glucose repression of GAL genes, suggesting the existence of additional functional partners of IMP2. PFK26 is not needed to maintain adequate glycolytic activity but, rather, is concerned with maintaining the homeostasis of metabolite concentrations. It is probable that PFK26 activity is regulated by IMP2 for maintaining metabolite homeostasis.

YOL136C [SGD: S000005496] and YIL107C [SGD: S000001369]

Another pair which shows high correlation between co-response profile and PG profile is YIL107C and YOL136C. This pair is involved in catalysis of the same metabolic reaction namely the phosphorylation of fructose-6-phosphate to fructose-2, 6-bisphosphate using ATP. YOL136C encodes for PFK27 which has far less enzymatic activity than PFK26, but nevertheless, the same cellular role. This is a positive indication that the present methodology can detect meaningful and close functionally associated ORFs.

YIL162W [SGD: S000001424] and YIL154C [SGD: S000001416]

Also we found a close link among YIL162W coding for SUC2, and YIL154C. These might have a functional association but we do not have any experimental evidence yet confirming the finding.

YKL062W [SGD: S000001545] and YML054C [SGD: S000004518]

YKL062W is a transcriptional activator related to Msn2p and is activated in stress conditions. YML054C codes for a membrane protein active in mitochondrial intermembrane space. There is no known evidence of activity of YML054C in the MAPK signaling pathway. YKL062W is a poorly characterized gene and the gene product is involved in MAPk signaling pathway.

YKR097W [SGD: S000001805] and YBR018C [SGD: S000000222]

YKR097W is a key enzyme in gluconeogenesis and its transcription is repressed by glucose. YBR018C encodes Galactose-1-phosphate uridyl transferase, synthesizes glucose-1-phosphate and UDP-galactose from UDP-D-glucose and alpha-Dgalactose-1-phosphate in the second step of galactose catabolism.

YBR184W [SGD: S000000388] and YKL062W [SGD: S000001545]

YBR184W encodes a putative protein of unknown function. YKL062W is a transcriptional activator related to Msn2p. YBR184W shows physical interaction with Rad3p, Cnm67p and Jsn1p which show growth defects on fermentable carbon sources.

YDR043C [SGD: S000002450] and YIL107C [SGD: S000001369]

YDR043C mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response and is a known regulator of glucose-repressed genes.YIL107C plays a key role in transcriptional regulation involving protein kinase A.

YDR073W [SGD: S000002480] and YGL035C [SGD: S000003003]

The YGL035C knockout mutant leads to partial repression of glucose-regulated transcripts. YDR073W encodes for a subunit of the SWI/SNF chromatin remodeling complex involved in transcriptional regulation. YDR073W is known to have a functional interaction with the components SNF2p, SNF11p, and SNF12p of the SNF chromatin remodelling complex involved in transcriptional regulation. YGL035C gene product is regulated by SNF1p protein kinase by phosphorylation of MIG1 repressor. We propose a probable functional association among YGL035C and YDR073W via SNF1p.

The last mutant pair found were YGR194C [S000003426] and YKR097W [S000001805] and there is no conclusive evidence or probable functional relationship between these two mutants.

Out of 9 ORF sets, we were able to find positive confirmed biological relationship. For 3 ORF sets, we were able to suggest a probable functional relationship. For the rest 4 ORF sets, we were not able to find any biological relationship.

3.2.3. Discussion of clustering algorithms results

Using integrated analysis of co-response profiles and PG profiles, we found that high co-response profiles correlation tends to come with lower distance of the mutant PG profiles, for the present study. In the future, larger amounts of data could be used to further corroborate the finding. The mutant pairs which have high co-response profile correlation but are assigned to different clusters were not studied further since our method is directed towards mutant differentiation using combined analysis of PG profiles and transcript co-response profiles data. Additionally, we found that FL profiles were not sufficient to derive any functional associations among the mutant set under study. For the present set of mutant set, we could not study the FL profile and PG profiles together as the mutant dataset under study is different in these two cases.

High-throughput metabolic profiling studies are becoming increasingly useful for systematic analysis of cellular systems and provide a valuable means for quantification of cellular pathway activity. The present work provides a robust method for such studies. The present method comprises a procedure developed in-house for automation of GC-MS spectra analysis, quantification of summed fractional labeling of proteogenic amino-acid fragments in order to estimate metabolite concentrations which are vital indicators of state and extent of activity of certain subpathways and branch points in metabolic networks of *S. cerevisiae,* estimation of the extent of mutant association based on the global features growth rate μ ,biomass yield Y_{xs}, ethanol yield Y_p rate of biomass production Q_s and rate of ethanol production Q_p, followed by integration of transcript co-response profiles for mutant differentiation. In this framework, we have introduced a scheme for estimation of cluster quality in analysis of metabolic profiling data. This measure assesses whether the clustering is useful or is a mere weak assembly of distant mutant ORFs. We confirmed that the fractional labeling (FL) is a useful procedure for obtaining insights into the activity of a number of sub-pathways. In particular, we could uncover similarities among the FL profiles of those fragments which have biosynthetic linkages, such as precursor-product relationships. We show that by integrating transcript co-response profiles with PG profiles one can identify functionally related ORF sets and could use this to generate plausible hypotheses about the functional roles of genes involved in metabolism and regulation of *Saccharomyces cerevisiae* central carbon metabolism. We proved that by analysis of set of mutants involved in regulating the central carbon metabolism.

This framework can be extended by including *in-silico* flux estimates, in order to obtain greater insights into functional association among genes in eukaryotic organisms, using metabolic profiling data.

As described in the previous sections, we applied unsupervised learning techniques to identify characteristic features of a gene knockout under varied carbon sources. However, it was found that the unsupervised learning methods produced clustering with silhouette widths below 0.5, which indicates the absence of strong clusters [see section 3.2]. To overcome this limitation, we developed a novel approach, based on adaptive reweighted estimation of mean and covariance (ARW method), which could answer the following queries for a given large-scale gene knockout metabolic profiling datasets, even when these mutants show can only be clustered weakly:

1) Given a large-scale data set, which genes knockouts are most distinct (unlike/outliers) from the majority of the dataset?

2) For every gene knockout, what are the significant features which are characteristic of that knockout mutant?

These questions are routinely asked in large-scale analysis of data originating from high-throughput techniques like GC-MS, growth profiling, metabolomics etc. The following paragraph describes some of these large scale studies and their applications.

In the last few years, several large-scale profiling studies using yeast have been carried out. Many of these studies are done with the goals of predicting the modes of action of external metabolites and for characterizing genes of unknown function [Luesch2006]. The availability of genome-wide heterozygous/homozygous diploid and haploid gene deletion strains fuelled large-scale profiling studies. Fitness profiling on a genomic scale with numerous nutrients has resulted in the verification of target pathways such as those for lovastatin (*HMG*), hydroxyurea (small subunit of ribonucleotide reductase) and methotrexate (*DFR1*) [Lum2004, Giaever2004]. Another approach called drug-induced haplo-insufficiency (lowering the gene dosage of the gene encoding the drug target increases the susceptibility to the drug), is also used to study the deletion mutant fitness. Drug-induced haplo-insufficiency occurs when lowering the dosage of a single gene from two copies to one copy in diploid cells results in a heterozygote that displays increased drug sensitivity compared with wild-type strains [Baetz2004].

The basic idea is to determine the abundance of each deletion strain in the co-culture using a PCR to amplify the barcodes ("used for knockout identification") associated with each mutant. Giaever et al found the drug target of tunicamycin using drug-induced haplo-insufficiency [Giaever1999]. Another approach is a kind of fingerprinting/pattern matching strategy in which gene expression profiles of drug treated cells are compared with large scale expression profiles derived from deletion mutants [Hughes2000]. This basic assumption is that the "fingerprint" of an active compound on gene expression will show resemblance to the profile of mutant strains displaying defects in the targeted pathway or in which the target-encoding gene is knocked out.

In section 3.3, we study the outlier detection methods and their application to metabolic profiling data analysis. We developed an approach based on the adaptive reweighted estimation for mean and covariance (ARW) method and wrote a routine in the R programming environment which assigns p-values to all feature combinations in order to select the combination that is the most informative feature set for a given gene knockout.

3.3. Analysis of metabolic profiling data using an outlier detection method

Outliers are regarded as those observations which are found to come from a different underlying distribution than the distribution which encompasses rest of the data points in the dataset. Outliers are different from extreme values. This is because even when the extreme values are far away from the centre they still belong to the same distribution as rest of the dataset. The outliers can be either of univariate or multivariate in nature. The univariate outliers are usually a result of an experimental error and for their identification univariate approaches can be used. The multivariate outliers have a more complex nature and cannot be detected by univariate approaches [P.J. Rousseeuw 1987]. Identification of multivariate outliers requires multivariate techniques for example projection techniques. In a simple case, by projecting objects on one of the axes the outlier tends to be located far from the majority of the data, and thus, it can be easily detected. In figure 12, *b* is regarded as a multivariate outlier since none of the projections are sufficient to uncover the outlier because of its presence in the data cloud [Daszykowski2007]**.**

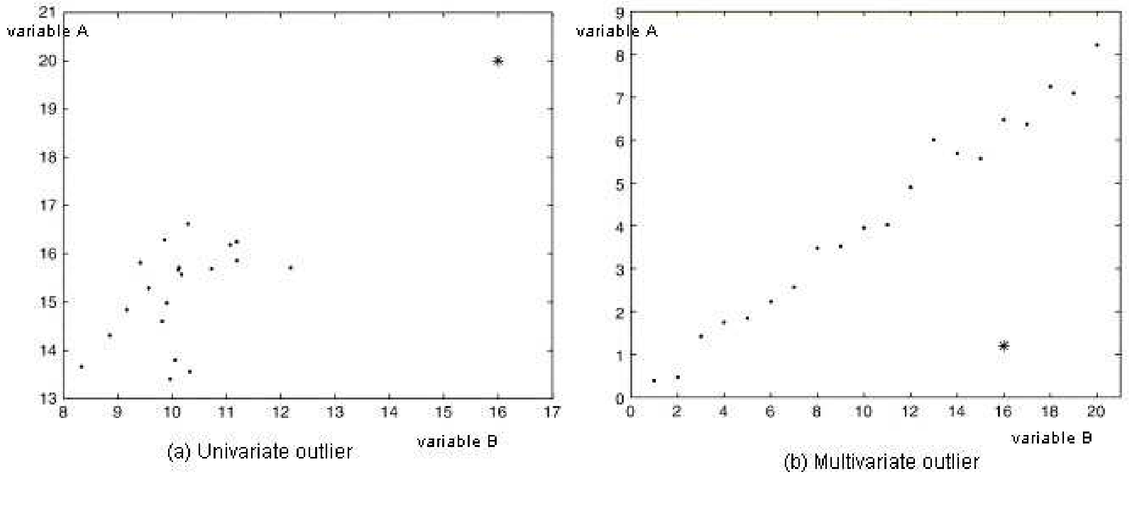


Figure 12 *Example of univariate and multivariate outliers (a) a univariate outlier (*) varies in terms of a single variable (b) a multivariate outlier (*) is an outlier which involves more than one variable.*

For a single variable x its mean, μ , is the sum of all elements divided by their number m :

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \textbf{Equation 5}$$

However, in the presence of outliers in the data, the mean is not a reliable estimate of the data location, and therefore it is said to be a non-robust estimator. Location estimators can be divided into two categories, robust and non-robust. The robust estimators aim to describe well the location of the majority of the data regardless of data contamination. The robustness of an estimator can be described by its breakdown point; a concept introduced by Hempel et al. [Hempel1986]. For a finite sample, the breakdown point of an estimator is the maximal fraction of outlying objects in the data, even in the presence of which the estimator yields acceptable estimates, other than in case of random estimator. For instance, the breakdown point of the mean estimator equals 0%, since a single outlier can bring the mean to an arbitrary value. There are different types of robust estimators. Parametric estimators assume a certain data distribution, for instance a normal distribution and thus such estimators simply eliminate outliers. Non-parametric estimators are robust in their nature because they do not require knowledge of the data distribution at hand.The lack of robustness of the mean estimator can be attributed to its least-squares nature. The mean of a random variable is the point minimizing the average Euclidean distance to all data objects. This condition is expressed as:

$$\overline{x} = \min_{\mu} \sum_{i=1}^{m} \left\| x_i - \mu(x) \right\| \qquad \textbf{Equation 6}$$

where ||…|| is the L2-Euclidean norm.

The median of the data is a robust alternative to the mean location estimator with a breakdown point of 50%, meaning that it takes contaminating 50% of the dataset to change the median value. The median of a variable is the middle element for an odd number of sorted elements. The median of a variable with an even number of sorted elements is the average of the two elements at the closest positions to the half-length of the variable.

In our study, where the data are multidimensional, i.e. the mutant profiles are described by several physico-chemical properties (variables), the data means and medians can be computed in a univariate manner, considering each data variable individually. This computation yields column means and column medians of the data (coordinate wise means and coordinate wise medians), respectively. It is also possible to consider the multidimensional nature of the data and the median as an estimate of a center of the multidimensional data cloud. The L1-median is a highly robust estimator of multivariate data location with a 50% breakdown point [Rousseeuw 1987]. The L1-median is a generalization of the univariate median. Although the L1-median seems to be the best-known multidimensional median, some other exists as well. Robust estimates of location as well as other robust estimates can be also derived applying the fuzzy set theory [Sârbu2001, Rajkó1994]. When outliers are present in the data, they can influence the data mean to a different degree depending on their distance from the data majority.

3.3.1. Methods

In the section 3.3.1.1., we describe the theory of multivariate outlier detection methods. Section 3.3.1.2 gives the theory behind the adaptive reweighted estimator of mean and covariance (ARW) method for outlier detection.

3.3.1.1. Multivariate outlier detection methods

Multivariate outlier detection methods can be grouped into two classes. One class comprises statistical methods that are based on estimated distribution parameters.

The second class comprises data mining methods that are typically parameter-free.

- **Statistical methods based on estimated distribution parameters**

Multivariate robust measures

The Mahalanobis distance is a widely used distance. It depends on the estimated parameters of the multivariate distribution. Given n observations from a p-dimensional dataset, $\overline{x}_n$ is the mean of the sample, and C_n denotes the covariance matrix**,**

$$C_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)(x_i - \overline{x}_n)^T \qquad \textbf{Equation 7}$$

The *Mahalanobis* distance (M) for each multivariate data point *i, i* = 1,…, *n* , is denoted by M_i and given by

$$M_i = \left(\sum_{i=1}^{n} (x_i - \overline{x}_n)^T C_n^{-1} (x_i - \overline{x}_n) \right)^{1/2}$$ **Equation 8**

The data points that have a large Mahalanobis distance are regarded as outliers. There are two effects namely masking and swamping effects that can affect the adequacy of the Mahalanobis distance in outlier detection.

Masking effect:

One outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance emerges as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

Swamping effect:

One outlier swamps a second observation, if the second observation can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers. Hadi et al. proposed a method to replace the mean vector by a vector of variable medians and to compute the covariance matrix for the subset of those observations with the smallest Mahalanobis distance [Hadi1992]. A modified version of Hadi's procedure is presented in [Penny2001]. Caussinus et al. proposed a robust estimate for the covariance matrix, which is based on weighted observations according to their distance from the center [Caussinus1990]. Other robust estimators of the location (centroid) and the shape (covariance matrix) include the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) [Rousseeuw1985, Rousseeuw1987 and Acuna2004].

- **Data-Mining Methods for Outlier Detection**

Data-mining methods are often non-parametric. Typically they do not assume an underlying generating model for the data. These methods can be classified into the following types: clustering methods, distance-based methods and spatial methods.

In clustering based methods, clusters of small size (including size 1) are regarded as the clustered outliers. PAM [section 3.2.1.2] and Clustering Large Applications (CLARA) fall under the category of clustering based methods [Kaufman1990].

CLARA essentially draws multiple samples from a dataset, applies PAM on each dataset and presents the best clustering as the output [Kaufmann1999]. CLARA has the advantage of being scalable to large datasets but at the same time it has disadvantages e.g. the efficiency of the algorithm depends on the sample size and also a good clustering using a sample will not be a good clustering for the entire dataset in case the sample is biased. Distance-based methods regard an observation as a distance-based outlier if at least a fraction β of the observations in the dataset are further than r from it [Knorr1997, Knorr1998]. These methods have drawbacks including the dependence on a parameter r and the lack of a ranking of the outliers. The methods usually have time complexity of the order of O (pn^2), where p is the number of features and n is the sample size.

In spatial outlier methods, an outlier is defined as a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood [Haining1993]. In other words, where an individual attribute value is not necessarily extreme in the distributional sense but is extreme in terms of the attribute values in adjacent areas.

3.3.1.2. Adaptive reweighted estimator for multivariate location and scatter: ARW algorithm

Multivariate outlier detection methods largely rely on the distance of the outlier from the centroid of the data as well as the shape of the dataset. The size and shape of multivariate data are quantified by the covariance matrix. In the majority of the methods, the *Mahalanobis* distance (M) is used as a distance measure.

For the normal distribution, the values of M_i^2 are approximately chi-square distributed with m degrees of freedom (χ_m^2). By setting the M_i^2 to certain quantiles of χ_m^2, it is possible to define ellipsoids that define sets of points having the same Mahalanobis distance [Gnanadesikan1977]. Thus all the points on a given ellipsoid have the same Mahalanobis distance to the centroid.

The presence of single extreme observations which are different from the main data cloud has a severe effect on Mahalanobis distance because of the sensitivity of the covariance matrix to outliers [Hampel1986, Maronna1998]. To overcome this problem, Rousseeuw and coworkers developed a method called minimum covariance determinant (MCD) estimator which is a robust estimator of the covariance matrix C and the mean t where both C and t are resistant to the presence of outliers [Rosseeuw1999, Rousseeuw1985].

Feature Index	Feature short descriptor	Feature long descriptor
1	mue	Growth rate
2	Qs	Rate of biomass production
3	Qp	Rate of ethanol production
4	QO2	Rate of biomass production on oxygen
5	ala_260	Alanine AAF M.W.=260
6	ala_232	Alanine AAF M.W.=232
7	gly_246	Glycine AAF M.W.=246
8	val_288	Valine AAF M.W.=288
9	val_260	Valine AAF M.W.=260
10	val_186	Valine AAF M.W.=186
11	ile_200	Isoleucine AAF M.W.=200
12	pro_286	Proline AAF M.W.=286
13	ser_390	Serine AAF M.W.=390
14	ser_362	Serine AAF M.W.=362
15	ser_288	Serine AAF M.W.=288
16	thr_404	Threonine AAF M.W.=404
17	thr_376	Threonine AAF M.W.=376
18	phe_336	Phenylalanine AAF M.W.=336
19	asp_418	Aspartic acid AAF M.W.=418
20	glu_432	Glutamic acid AAF M.W.=432
21	arg_442	Arginine AAF M.W.=442
22	arg_414	Arginine AAF M.W.=414
23	tyr_466	Tyrosine AAF M.W.=466

Table 6 *Feature index and long descriptors. Abbreviation: AAF stands for "Amino acid fragment"; M.W. stands for "Molecular weight"*

MCD essentially looks for a subset of h observations out of the total n observations such that the covariance matrix defined by the subset has the smallest determinant. Generally methods have a breakdown value of $n/(m+1)$ where n is the number of the observations and m are the number of dimensions [Donoho1982]. MCD looks for the ellipsoid with smallest volume that covers h data points where $n/2 \leq h < n.$, and has a breakdown value of $(n-h)/n$.

ARW is a powerful new method for multivariate outlier detection based on MCD which can distinguish between extreme values of a normal distribution and values originating from a different distribution (outliers) [Filzmoser2004]. It was originally applied for the analysis of geochemical data. The ARW method uses the MCD estimator with $h \approx 0.75n$. The location estimator is calculated as the average of these h points. The breakdown value with $h \approx 0.75n$ is approximately 25%. When the fraction of outliers exceeds 25% of the total observations, one would get completely biased estimates [Hampel1986]. M_i is calculated using the robust estimates of location and scatter and henceforth referred to as RD_i .

3.3.1.3. Implementation of the ARW algorithm in the R programming environment

INPUT:

- Physiological growth data and fractional labeling data for all mutants (Table 17 in appendix 1).
- Feature index refers to *nfeat* : 1, 2, 3……, 23(see Table 6).
- Maximum number of features to be used for all permutations of the parent dataset *nfeatsel*

OUTPUT:

- Most outlying feature set for each individual mutant
- P-values for all feature combinations from i=1, 2……, 8.

PROCEDURE:

- Drawing different combination of features from 1 to *nfeatsel* , from the parent dataset consisting of *nfeatsel* features using R function, "*combinations* "
- Applying function, "*my.arw*", to calculate the true theoretical chi square distribution followed by its comparison to the distribution coming from the permuted dataset.

- Calculation of the p-values for all feature combinations.
- Measurement of the feature combination (k) which gives the maximum of the distance from the rest of the dataset, for each mutant.

3.3.2. Results of outlier detection method (ARW)

In the present analysis, we were able to find highly significant feature combinations for each individual mutant present in the original dataset. This method proves to be a method for fast characterization for the metabolic profiling datasets for large scale knockout analysis.

We show that in the absence of strong phenotypic perturbations, for example in our case where the metabolic profiles prove not be sufficient in finding any underlying functional associations among majority of the mutant set, the ARW method can be used for a more granular analysis of each individual knockout mutant. Table 7 lists the most significant k - features for each individual mutant.

Mutant	**Min-pval**	**minF**	**Significant Feature Combination**
ACE2_gal	2.58387E-10	8	Mue, Qs, gly_246, ser_390, thr_404, phe_336, arg_414, tyr_466
ADR1_gal	4.24879E-11	8	gly_246, val_288, val_186, ser_390, ser_288, thr_404, thr_376, glu_432
CAT8_gal	0	2	QO2, ser_362
CYB2_gal	1.43743E-05	8	ala_260,val_288, ser_288, thr_404, phe_336, asp_418, glu_432 arg_442
DLD2_fru	1.12875E-11	4	QO2, val_260, val_186, phe_336
DLD2_gal	6.16483E-10	3	ser_390, ser_288, phe_336
FBP1_gal	1.89384E-11	5	Mue, Qs, QO2, ile_200, phe_336
FBP26_gal	0.002912756	3	gly_246, ser_390, thr_404
GAD1_gal	9.16376E-08	6	QO2, gly_246, val_288, val_260, pro_286, thr_404
GAL10_fru	1.68532E-13	8	Qs, Qp, ala_260, gly_246, val_260, phe_336, asp_418, arg_414
GAL10_glc	0.000804473	3	gly_246, ile_200, pro_286
GAL7_fru	4.32127E-11	3	QO2, phe_336, arg_414
GAL80_glc	1.42109E-14	1	tyr_466
GLK1_gal	6.64289E-07	8	ala_260, val_260, pro_286, ser_288, thr_404, asp_418, arg_414, tyr_466
HXK2_gal	0.000829674	8	val_288, ile_200, ser_362, thr_404, thr_376, phe_336, asp_418, glu_432
IMP2_fru	4.56749E-06	5	Qs, Qp, ile_200, ser_288, arg_442
IMP2_gal	7.11391E-08	4	QO2, ile_200, thr_404, phe_336
LEU4_fru	1.64293E-05	7	Mue, Qp, ala_260, val_260, thr_404, thr_376, arg_442
LEU4_gal	2.38705E-10	5	ala_260, val_288, ser_390, thr_404, arg_442

MAE1_gal	0.001224323	5	ala_260, ala_232, val_260, ser_362, thr_376
MAL33_fru	8.52123E-07	7	Qs, ala_232, val_288, ser_362, thr_404, thr_376, asp_418
MAL33_glc	0.01301821	4	Mue, Qp, gly_246, pro_286
MSN4_gal	3.33067E-16	8	Qp, ala_260, ala_232, gly_246, val_186, ile_200, ser_390, arg_442
PCK1_fru	1.18469E-07	8	QO2, val_288, val_260, ser_390, thr_376, phe_336, asp_418, arg_414
PCK1_gal	4.87388E-14	8	Mue, val_260, val_186, ser_390, ser_288, glu_432, arg_442, arg_414
PFK26_fru	1.76617E-05	8	ala_260, val_288, ser_362, ser_288, thr_376, phe_336, arg_442, arg_414
PFK26_gal	1.04222E-05	5	gly_246, val_186, pro_286, ser_362, arg_414
PFK27_gal	8.18322E-07	5	QO2, gly_246, ser_390, ser_362, ser_288
PGU1_fru	0	3	val_260, ile_200, arg_414
SFA1_fru	0.004536613	5	ala_260, ala_232, ser_390, glu_432, arg_414
SFA1_gal	0.002945495	3	QO2, ala_232, val_186
SFA1_glc	0.002293606	3	gly_246, ile_200, pro_286
SIP3_gal	9.08784E-12	8	Mue, ala_232, gly_246, val_288, val_260, pro_286, ser_362, arg_414
SNF11_gal	0	6	val_186, ile_200, ser_362, ser_288, thr_376, arg_414
SNF2_fru	3.18552E-10	8	ala_260, ala_232, val_288, val_186, pro_286, ser_288, thr_376, asp_418
SNF2_glc	0.002467652	2	ile_200, arg_442
SUC2_fru	1.02178E-06	6	ala_232, val_260, val_186, ser_390, ser_362, thr_376
SUC2_glc	0	1	tyr_466
TYE7_gal	1.89204E-08	8	Mue, ala_232, val_288, val_260, val_186, ile_200, ser_390, phe_336
UGA1_gal	0.003362803	3	ala_232, ser_288, thr_376
UGA2_fru	4.14022E-08	8	ala_232, val_288, val_260, val_186, ser_390, ser_362, thr_376, phe_336
UGA2_gal	0.000194988	3	val_288, ser_288, thr_404
YBR184W_gal	8.85803E-12	8	QO2, ala_232, val_260, pro_286, ser_362, ser_288, glu_432, arg_414
YDR248C_fru	1.24612E-07	8	QO2, gly_246, val_186, ser_390, thr_404, thr_376, phe_336, asp_418

Table 7 *List of most significant k - features for each mutant in the mutant set.* *min-pval* is the minimum of the pvalues obtained using the feature combination $i = 1,2,....,8$. minF is the feature combination which show *min-pval*. *Significant feature combination* is the feature combination which are most significant (*min-pval*) for a given mutant.

We calculate minimum of the pvalues *(min-pval*) for the feature combinations, $i = 1,2,..,8$. The *min-pval* is that lowest *p-value* obtained for any given mutant using any feature combinations namely, $i = 1,2,..,8$. Here *min-pval* denotes the feature combination which is the strongest outlier for a given mutant. Figure 13 denotes a plot of $\log_{10}$(*min-pval*) for mutants which are grown in three different experiments (MutFRU, MutGAL and MutGLC) namely differing in the type of carbon source used. It can be clearly seen from this plot that in general the MutGLC (Mutants set grown with glucose as carbon source) is less discriminatory than the other mutants grown in the other two conditions i.e. fructose and galactose.

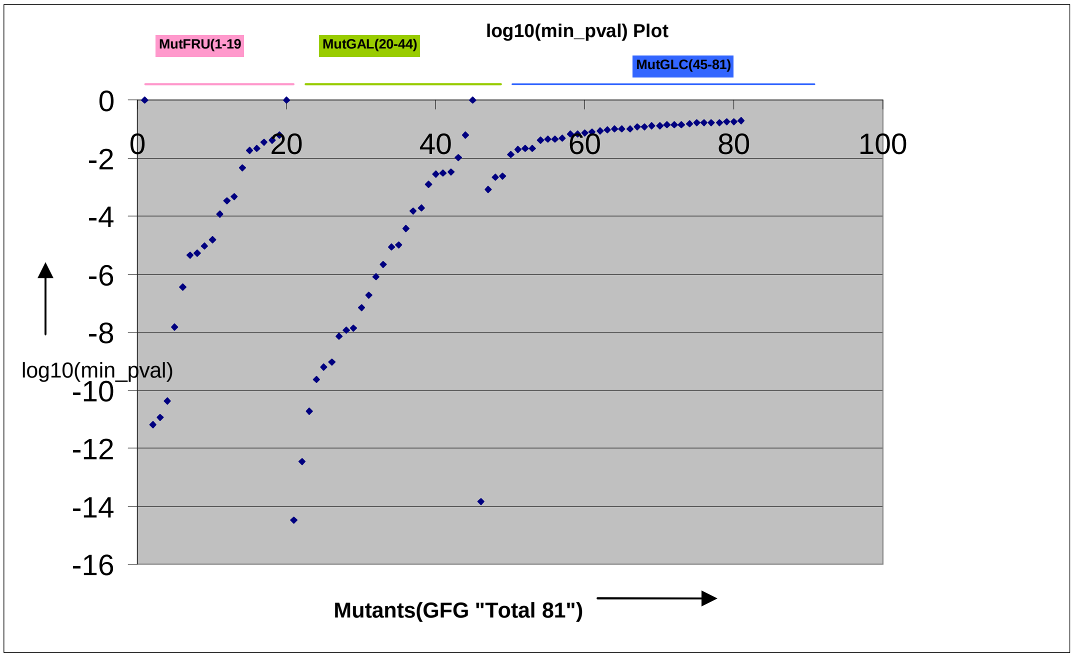


Figure 13 *Comparison of the* log_{10} *(min-pvalues) for the most significant* k *features combination for each mutant in the entire mutant set.*

3.3.3. Discussion of outlier detection method (ARW) results

The ARW method is an efficient method for identifying the most significant features for large scale metabolic profiling datasets as well as for comparison of knockout mutants under varying experimental conditions. In our analysis, we found that the knockout mutant of malic enzyme is a much more significant outlier when yeast is grown under glucose than when grown under galactose, thus pointing towards a differentiation of the metabolic phenotype. In the figure 14, it is evident that the metabolic profile of the malic enzyme when grown with glucose (Figure 14a, pvalue 0.06) shows the most differentiating outlier feature combination with 5 features whereas for the malic enzyme when grown with galactose show strong outlying behaviour with 3-5 features (Figure 14b, pvalue ~0.002). Preliminary investigation of the k- significant features for the malic enzyme knock-out in GLC and GAL conditions does reveal a different subset of amino acid fragments which are most discriminatory for the individual malic enzyme mutant.

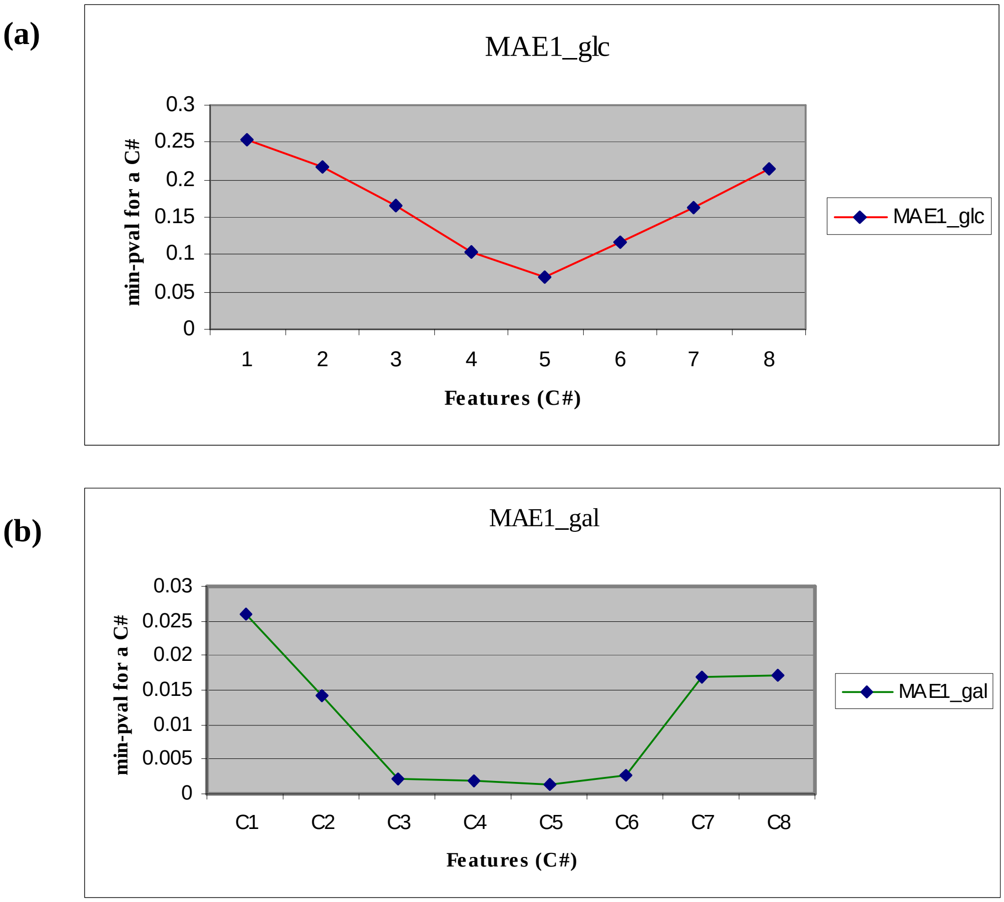


Figure 14 *Malic enzyme (MAE1):* plot of min-p values for the malic enzyme grown with glucose (a) and malic enzyme grown with galactose as carbon source (b). $C1, C2, ..., C8$ refers to the mutants set with 1, 2,….., 8 feature combinations, respectively.

It was also shown by Boles et al. that there is an alternate pathway for pyruvate metabolism.

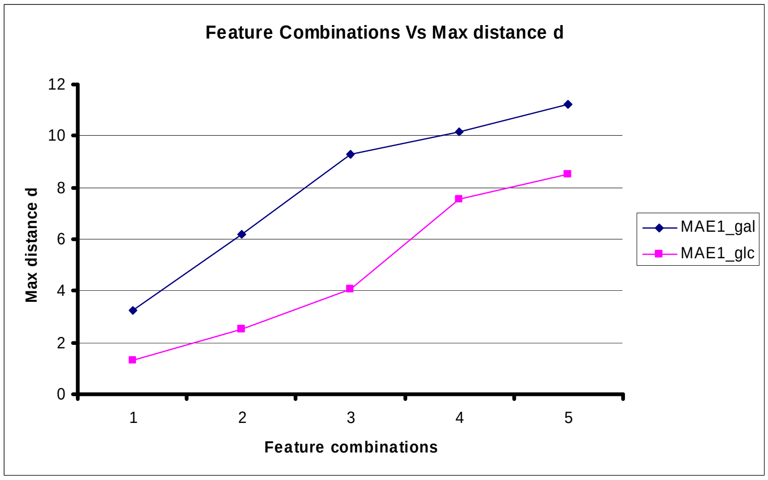


Figure 15 *Analysis of Malic enzyme knockout:* MAE_glc (grown with glucose) and MAE1_gal (grown with galactose)

They also reported that the malic enzyme shows much clearer phenotype with galactose compared to when grown with glucose (Figure 15) [Boles1998]. This is a novel method for in-depth and comparative analysis of large scale knockout datasets and is fairly fast in terms of computation time. For a set of 5 feature combinations, it takes about 5 hours on a single processor pentium P4 machine for calculation of the significant features of all knockouts and determination of the most outlying knockout.

Part III: Web server for metabolic network analysis

Chapter 4

4. Introduction to metabolic network analysis

4.1. Introduction

In living cells, the study of chemical transformations of substances is of central importance. On an organism level, these chemical transformations form the metabolic network specific for that organism. The term “metabolism” comprises of the large number of chemical reactions which convert one or more educts into one or more than one *product,* in the cellular environment. The area of metabolic network and pathway analysis has been vigorously researched in the last decade. With the emergence of system biology, diverse computational approaches have been developed. In this work, we developed a new webserver called MetaModel, for the analysis of genome-scale metabolic networks of eukaryotic organisms. Section 4.1.1-4.1.3 describes the general topological measures applied for network analysis and the KEGG pathway maps. Section 4.2 summarizes various mathematical approaches applied in pathway analysis. Sections 4.3-4.5 describe the theoretical basis and data used by MetaModel. In the current implementation, the server facilitates analysis of the *Saccharomyces cerevisiae* metabolic models iFF708 and iND750, and of user-defined custom models.

4.1.1. Types of biochemical networks

A graph (or network) representation can be used to define a system of genes and gene products that interact or regulate each other. These graph models can be directed or undirected, and can represent various biochemical relationships existing among the node members. For example, biochemical networks that capture mutual interactions like protein-protein binding can be best represented as undirected graph models, whereas directed graph models are suitable for representing biochemical reactions that transform a set of substrates into a set of products. These graphs can be augmented by labeling nodes and edges with additional information. Current incorporated information can be roughly classified into five categories namely 1) genomic information 2) transcriptomic information 3) proteomic information 4) metabolomic information and 5) interactomic information. Transcriptomic and metabolomic information can be used for indirect inference of molecular interactions. Except for the interactome, all other omics data generally provide information for labelling the nodes of a network.

4.1.2. Topological network parameters

In the literature, many graph theoretic topological measures have been used for studying biochemical networks. These measures can provide meaningful insights into functions and structural organization of biochemical networks [Christensen2007]. In the following we list the most widely used graph theoretic measures:

a) Degree and degree distribution

The degree of a node is defined as the number of edges incident to that node. In directed graphs, total degree of a node can be divided into an *out-degree* (# out-going edges) and an *in-degree* (# in-coming edges). In a graph in which edges have numerical weights, another measure called *node strength* can be defined. The strength of a node is the sum of the weights of the edges adjacent to that node. A global measure of network topology is called *degree distribution,* $P(k)$. $P(k)$ is defined as the probability of a randomly selected node to have degree k. $P(k)$ is a simple measure that is calculated but counting the number of nodes with k =1,2,….edges, and then dividing these numbers by the total number of nodes in the network. Recent studies have shown that the majority of cellular networks have a scale-free degree distribution [Albert2002, Lee2002]. The degree distribution of scale-free networks follows a power law: $P(k) \approx Ak^{-\gamma}$ where A is a normalisation constant and γ is a degree exponent. For example, the degree distribution of metabolic networks and protein interaction networks typically obey power laws with $2\langle\gamma\langle 3$ [Jeong2000, Guelzim2002]. The important distinction of a network being a scale-free network is that these generally show several “highly connected” nodes, the so called Hubs.

b) Connectivity, path length, efficiency and paths

A path in a metabolic network represents a sequence of chemical reactions that transform the compound represented by the source node of the path into that represented by the sink node of the path. The *distance* between two nodes in networks is the minimum number (or sum of the weights in edge-weighted graphs) of the edges in any path connecting these nodes. Two nodes are *connected* if a sequence of adjacent nodes, a path, links these two nodes [Bolloba´s1979]. The *average distance* d_{ij} is the average number of edges in the shortest path between any two nodes i and j, in the given network [Dijkstra1959]. The *global graph efficiency* is defined as $\langle 1/d_{ij} \rangle$ [Latora2001, Latora2003]. Directed graphs in which every pair of nodes is connected by a directed path are called strongly connected graphs. We know that cellular networks are not strongly connected, in general, but it is advantageous to identify the maximal subgraphs inside these networks which are strongly connected, the so-called strong components. The strong components of a graph are connected with each other in an acyclic fashion. The analysis of the connectivity structure of a metabolic network can give useful hints to its functional organization [Ma'ayan2004].

c) Clustering coefficient

The clustering coefficient C_i, is a measure of the extent to which a node's first neighbourhood is connected [Watts1998].

$$C_i \equiv \frac{2E_i}{k_i(k_i - 1)}$$

Here k_i is the degree of node *i*, and E_i is the number of edges connecting the immediate neighbours of node i. The average clustering coefficient of a network, calculated by averaging the clustering coefficients of all of its nodes, is a useful measure of the strength of connectivity inside a network. A large average clustering coefficient suggests a high level of cohesiveness and redundancy [Wagner2001, Ravasz2002].

d) Betweenness centrality

A node is termed a *source* if it has only outgoing edges, and a *sink* when it had only incoming edges. The *Betweenness centrality* $C_B(x)$ of a node x which is neither sink nor source is defined as the number of shortest paths from node i to node j passing through node x, divided by the total number of shortest ij – paths, for a graph with X nodes and E edges(Equation 9). The *Betweenness centrality* of a

$$C_B(x) = \sum_{\substack{i \neq x \neq j \in X \\ i \neq j}} \frac{\sigma_{ij}(x)}{\sigma_{ij}}$$ **Equation 9**

node indicates the importance of that node for the propagation of flow within the network [Anthonisse1971, Freeman1977]. Holme *et al.* have shown that ubiquitous substrates in the biochemical networks may not have the highest degrees in the network, but they often have the highest betweenness centralities [Holme2003].

4.1.3. KEGG pathway maps

KEGG refers to the "Kyoto Encyclopedia of Genes and Genomes" [Kanehisa1997a]. KEGG is a knowledge base for systematic analysis of gene functions in terms of the networks of genes and molecules [Ogata1999]. It comprises three databases namely PATHWAY (repository of knowledge of molecular pathways and complexes), GENES (repository of gene catalogs of completely sequenced genomes and partial genomes) and LIGAND (repository of chemical compounds and chemical reactions) [Goto1998]. The PATHWAY database uses a graph-theoretic form for data representation. In this graph representation, a *node* is a gene product or complex and an *edge* is a protein-protein interaction. This protein-protein interaction could be direct physical interaction, iso-enzyme relation or gene expression relation among given gene products. GENES database has a collection of genes for all the organisms in KEGG. A typical entry in GENES database contains following information: organism name, gene name, functional description, functional hierarchy, chromosomal position, codon usage, nucleotide sequence and amino acid sequence. The LIGAND database contains information about chemical compounds, enzyme molecules, enzymatic and non-enzymatic reactions. PATHWAY contains protein-protein interaction networks for various cellular processes. In the KEGG pathway maps, DNA and chemical compounds are not considered as nodes but rather form the edges in the network. A protein-protein interaction can be 1) a direct physical interaction such as protein modification, protein binding, or protein cleavage; 2) an indirect interaction representing association of two enzymes that catalyze two successive reaction steps; 3) an indirect interaction involving gene expression, namely, the relation between the genes encoding a transcription factor and a target gene product. The KEGG databases and computational tools are used in a semi-automated fashion to find *cores (i.e. basic wiring diagram of molecules in biological systems)* of known pathways using the knowledge present in KEGG reference pathways. The organism-specific KEGG pathways are generated by extension of these *cores* by integrating additional partners that are associated at the genome level (for example genes in the same operon), the transcriptome level (for example co-expressed genes), and the proteome level (for example binding partners). Figure 16 depicts the KEGG pathway map for lysine biosynthesis in the *Saccharomyces cerevisiae*. Both boxes and circles are clickable objects for retrieving detailed molecular information. A circle represents a metabolic compound. Each box represents an enzyme with the corresponding EC number inside it. The shading of the box indicates whether that gene product is present in the genome under study or not [Ogata1999]. The boxes (enzymes) whose genes are present in the genome under study are colored green.

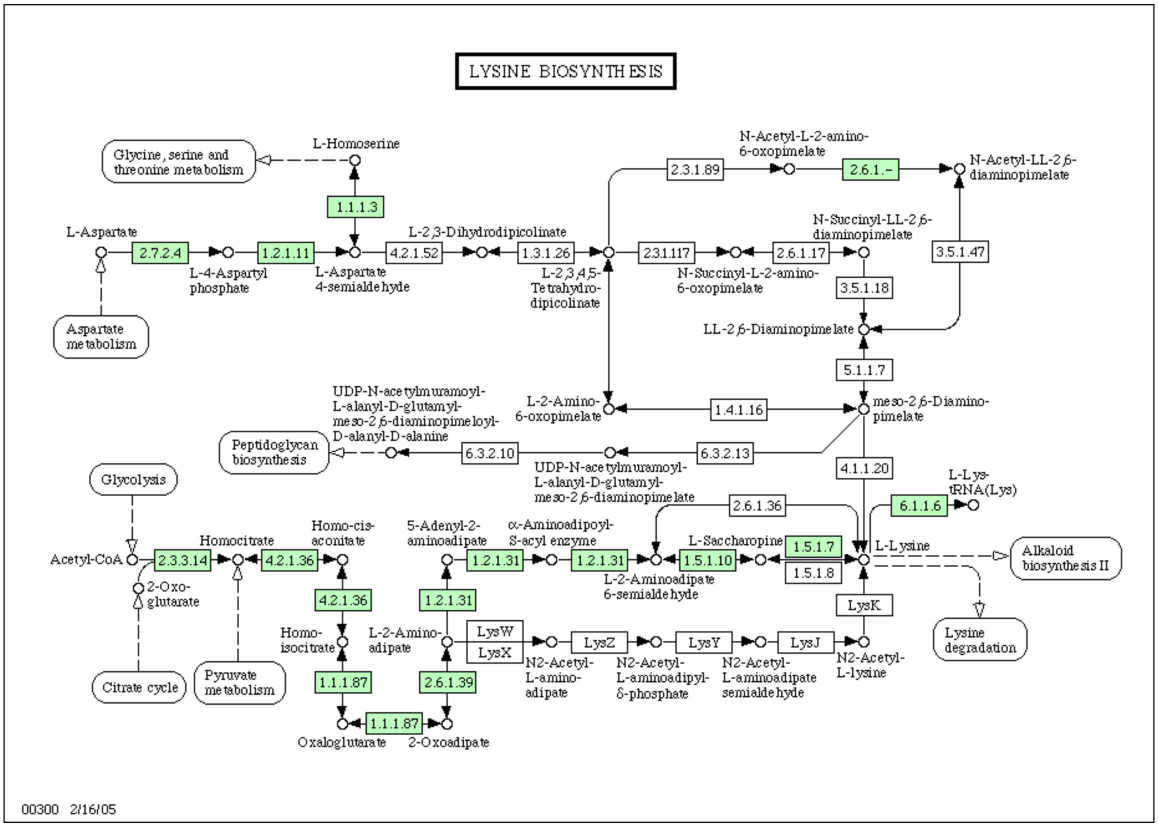


Figure 16 *KEGG pathway map of lysine biosynthesis in Saccharomyces cerevisiae.*

4.2. Mathematical approaches for pathway analysis

The idea of these mathematical approaches has been to exploit as much system level information as available, for pathway analysis. Both schools of thought namely the one following bottom-up construction and the one following top-down construction have paid increased attention to the development of rigorous mathematics approaches for pathway analysis.

Mathematical approaches for pathway analysis have been fundamental in the area of systems biology and metabolic engineering. These approaches have found direct application in modelling, simulation and optimization of metabolic pathways. These approaches can be divided into (i) Structural approaches, (ii) Stoichiometric approaches, (iii) Carbon flux approaches, (iv) Stationary and non-stationary mechanistic approaches and (v) Approaches based on gene regulation modeling [Wiechert2002]. (i) Structural model building starts with assimilation of the known knowledge on the mechanisms and components of reactions, published work and public data repositories like KEGG [Kanehisa1999], (ii) Stoichiometric models are a step ahead of structural models as they incorporate quantitative data on the cellular concentration of various reaction components. The stoichiometric modelling concept uses various abstraction levels like pooling of intracellular metabolites and lumping the intermittent steps in various subpathways, for which no quantitative data are available. Stoichiometric modeling approaches use a quasi-steady state assumption for modeling pathways, (iii) The carbon flux approaches are similar in essence to the stoichiometric approaches except that they use additional quantitative data to further dissect and balance the degree of freedom of stoichiometric balances. These quantitative data are generated from various labeling experiments in which a tracer like ^{13}C is used as a label. This label is used for the growth experiments. The distribution of the label as part of various cellular intermediates, at steady state is used as additional information to dissect various metabolite conversion steps in pathway models, (iv) Mechanistic modeling approaches incrementally model individual reactions steps for the underlying pathway model. These individual steps are then subjected to the theory coming from standard enzyme kinetics. (v) Modeling approaches involving gene regulation are still in their infancy. The basic idea of the modeling with gene regulation is to use the constraints coming from gene regulatory mechanisms, for modeling reaction steps in the pathway model.

Formal models of metabolic networks along with “metabolite snapshots” methods (the comprehensive measurement of metabolite concentrations) are powerful approaches for understanding the perturbation in cellular metabolism due to genomic perturbations, for example knocking out genes, in mutants. Metabolite snapshot comparison amounts to the comparison of metabolite concentrations of mutants deleted for genes of unknown function, with metabolite concentrations of mutants deleted for genes of known function. It is not a new strategy but has existed in the area of population genetics for decades. The fitness defect can be thought of as the global effect of the all phenotypic changes that occur as a result of a genetic perturbation namely, knocking out single or multiple genes. These studies compare the knockout mutant strain to the wild type strain. Any change in the knockout mutant is hypothesized to be the result of verifiable/non-verifiable effects of the absence of particular genes. For yeast and other organisms, it was generally found that these knockout mutations usually show little or no fitness defect compared to the wild type strain [Drake1998, Keightley1999, and Lynch1999].

4.3. Stoichiometric analysis of metabolic networks

Metabolic networks have been extensively used for understanding principles of metabolic organization. The phenotype of a strain could be regarded as the experimentally observable behaviour of the underlying metabolic networks and the interactions of several components of these networks. These interactions could not be intuitively studied which led to an ever expanding area of research called mathematical modeling of cellular networks. Stoichiometric analysis is one branch of mathematical modeling and analysis. Stoichiometric analysis exploits the structural nature of metabolic networks. It is a useful method to identify the constraints on existing paths between two components and the biochemical capabilities of a metabolic network [Varma1994, Schuster2000, Edwards2000, Stelling 2002, Famili2003 and Price2003].

4.3.1. *S. cerevisiae* genome-scale metabolic models and their construction

Genome-scale metabolic models of micro-organisms are important tools for model-driven data analysis and can be used for calculating experimentally verifiable phenotypic predictions. Genome-scale metabolic model construction involves the assimilation of published biochemical, physiological and genomic information for a given organism. In addition to the information available in the books and journal publications, public data repositories like MIPS, SGD, Yeast Proteome Database, KEGG Database, ExPASy Biochemical Pathways, ExPASy Enzyme Database, ERGO and Swiss-Prot provide a basis for metabolic model reconstruction. In 2003, Price et al. proposed a naming convention of these *in silico* genome-scale metabolic models in the following manner. For example genome-scale model iAA#ORF is an abbreviation in which “i” stands for an *in silico* model, AA are the initials of the first scientist who reconstructed that model, and #ORF is the total number of genes accounted for in the model. iFF708 and iND750 are two major genome-scale metabolic models of *Saccharomyces cerevisiae* [Price2003, Forster2003, Famili2003].

4.3.2. Stoichiometric matrix

A stoichiometric matrix provides a detailed description of a biochemical network and is a useful mathematical formalism for representing the chemical interactions in a metabolic network. A stoichiometric reconstruction is performed by careful integration of data on the chemical transformations in a system with defined boundaries and in accordance with the principle of conservation of mass. The result is a matrix representation of data on network components and the interactions between these network components. The rows of the matrix correspond to the network components and the columns represent the chemical transformations (reactions) between the components. The elements of the matrix correspond to the stoichiometric coefficients of the associated chemical transformations. These elements are assigned a sign. Usually, a negative sign signifies that the node represented b the row of the matrix element is an “input (reactant)” and a positive sign represent an “output (product)”.

4.3.3. Sparsity of the stoichiometric matrix

Any set of biochemical transformations can be described by system of ordinary differential equations as follows [equation 10]:

$dX / dt = Nv(X)$ **Equation 10**

Where N v and X denote the stoichiometric matrix, the vector of reaction rates and the vector of concentrations of "internal (metabolites with the variable concentrations)" metabolites respectively. Similarly those metabolites which are buffered are named as "external" metabolites. At stationary state, the system can be represented as in equation 11:

$Nv = 0$ **Equation 11**

Equation 11 is in essence defines the energy, mass and redox potential contraints in the metabolic network. This in turn defines the constraints as well as capabilities of a given metabolic genotype.

Also flux vector of the irreversible reactions, v_{irr}, must follow equation 12.

$v_{irr} \geq 0$ **Equation 12**

To decide whether a given enzyme set is actually a functionally coherent set in metabolism, it must be determined whether the corresponding flux vectors can fulfill equations 11 and 12 [Nuno 1997, Pfeiffer1999]. The region encompassed by these flux vectors is known as region of admissible (attainable) flux vectors (i.e the metabolic flux distributions that did not violate the energy, mass or redox balance constraints) [Rockafellar1970 and Nozicka1974].

The stoichiometric matrix provides concise information about the metabolic network that it represents. Stoichiometric matrices are generally sparse, i.e. they contain few nonzero elements, because only few metabolites are connected by a chemical reactions and reactions involve few metabolites (no more than 3, generally). The complete set of vectors v satisfying the equation 11 defines a region called the "null space" of N. The stoichometric matrix represents a set of linear equations representing components of metabolic machinery of the organism [Lay1997]. In the past, a large number of linear algebra techniques have been applied to studying fundamental system properties [Clarke1988, Reder1988]. The "null space" defines all the possible and impossible capabilities of a given metabolic genotype [Schilling1999]. The null space, K, can be mathematically represented as a matrix whose columns are linearly independent vectors spanning this subspace.

$$NK = 0 \qquad \textbf{Equation 13}$$

The “null space” can be used to identify metabolite production capabilities for a given metabolic network, ease of conversion of carbohydrates into other biomolecules for a given network, as well as to find the critical links(bottlenecks) in the metabolic network [Edwards1998, Varma1994].

4.3.4. Related work

Various computational approaches use the concept of stoichiometric analysis as the basis for the further method development. One such approach is called *Flux Balance Analysis* (FBA) [Varma1994, Schilling1999, and Palsson2000]. The general idea comes from concept of reduction of admissible flux space [see section 4.3.3]. As described in the previous section, stoichiometric matrices are sparse and the linear systems resulting from them are underdetermined. The feasible flux distributions (distribution which satisfies equation 11) of a network having r reactions are restricted to the null-space of the stoichiometric matrix, and can be described by the only r - rank (stoichiometric matrix) free parameters instead of full r unknown reaction rates [Heinrich1996, Klamt2002]. The work by [Palsson2002, Papin2003, Holzhutter2004, Stephanopoulos2004 and Covert2001] present some of the applications of FBA approach [Bonarius1997, Edwards2002, Kauffman2003]. The FBA approach is a useful technique for quantification of metabolic capabilities (~production) of cellular systems. The system is assumed to be optimised with respect to functions such as maximisation of biomass production or minimisation of nutrient utilisation. This is followed by solving the system to obtain a steady-state flux distribution. This flux distribution is then used to interpret the metabolic capabilities of the system.

$$\frac{dx}{dt} = Sv \qquad \textbf{Equation 14}$$

$$v = \{v_1 v_2 .....v_n b_1 b_2 ....b_{next}\}^T$$

The dynamic mass balance of the metabolic system is described using the stoichiometric matrix, relating the flux rates of enzymatic reactions, $\mathbf{v}_{n \times 1}$ to time derivatives of metabolite concentrations, $\mathbf{x}_{m \times 1}$ as equation 14. Here v_i represents the internal fluxes (i.e system of fluxes that affect a particular intracellular metabolite), b_i represents the exchange fluxes (i.e. fluxes which bring the metabolites into and out of the system boundries) in the system and $next$ is the number of external metabolites in the system. External metabolites are the sources and sinks of the network. Also concentrations of external metabolites are assumed to be buffered. Internal metabolites (intermediates) have to be balanced with respect to production and consumption at steady state. Also since $m < n$, the system is under-determined and could be solved using optimisation criterion [Raman2005]

Similar in essence to FBA is another approach known as Elementary Flux Modes (EFM) [Schuster2000]. A mode of a system is a relative flux distribution that fulfils the steady state condition for the intermediates and the sign constraints for irreversible reactions. EFM is based on the exhaustive enumeration of all feasible flux vectors (v) for the equation 11. An EFM describes the minimal number of reactions capable of working together in a steady state and thereby indicating various modes of behaviour of a given system. EFM actually acts as a generating basis for all possible flux distributions and, thus, are minimal (constructive) description of the solution space. The algorithms for computing EFMs are generally from computational geometry; more specifically the algorithms for enumeration of extreme rays of polyhedral cones which and are combinatorially complex.

Stoichiometric analysis of metabolic networks has been increasingly successful in terms of its predictive power compared to the topological approaches which are based on simple graph-theoretic methods. Another advantage of stoichiometric analysis is its scalability and feasibility even in the absence of the knowledge about kinetic parameters and rate equations, as compared to kinetic modeling approaches [Steuer2007].

4.3.5. Data

Section 4.3.5.1 and 4.3.5.2 describes the data coming from the yeast genome-scale models.

4.3.5.1. iFF708 genome-scale metabolic model of yeast

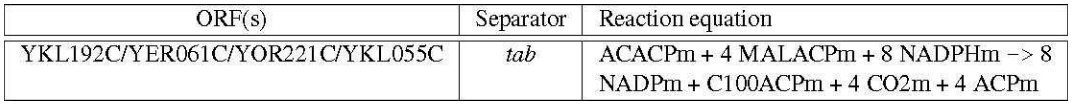
Foerster et al. build the first genome scale metabolic model of yeast. This model is called iFF708 [Forster2003]. iFF708 stands for "*in silico*" yeast model proposed by "Foerster and Famili", accounting for 708 genes. This reconstructed metabolic model was the first comprehensive network for a eukaryotic organism. The initial model accounted for a total of 708 open reading frames (ORFs) corresponding to 1035 metabolic reactions [Foerster2003]. In this model, all metabolic reactions are assigned to three cellular localizations namely mitochondria, cytosol and extracellular space. All *in vivo* reactions belonging to other compartments as well as the reactions, for which no cellular localization information is available, are assumed to be cytosolic. iFF708 also provides information on whether a given reaction is reversible or irreversible. A reaction for which there is no directionality information available is assumed to be reversible. Two-thirds of the reactions in the iFF708 are assumed to be irreversible.

iFF708 reaction format

Each comment in the reaction text file has to be marked by a leading # mark (hash-mark). A reaction line consists of one or more ORF names participating in the reaction and a reaction equation. The list of ORF(s) and the reaction equation have to be separated by a tab. If more than one ORFs influence a reaction, they have to be separated by a slash (/). The reaction equation is denoted in the common chemical notation for reactions. The names of the reactants have to be abbreviated as given in the metabolite text file [5.2.3.1]. There has to be a blank between coefficients, names, plus signs and the reaction arrow. Possible reaction arrows are:

- -> for a irreversible reaction and
- <-> for a reversible reaction.

The following example contains all relevant cases.



ORF(s)	Separator	Reaction equation
YKL192C/YER061C/YOR221C/YKL055C	*tab*	ACACPm + 4 MALACPm + 8 NADPHm -> 8 NADPm + C100ACPm + 4 CO2m + 4 ACPm

Table 8 *Example reaction for iFF708* coding style

iFF708 Metabolite format

A metabolite line consists of the following columns separated by tabs: a) Abbreviation and b) Metabolite name. Both the columns are mandatory. Both the columns are mandatory. The following examples represents metabolite format for few cases in iFF708 format.

Abbreviation	Name
HIS	L-Histidine
ATP	ATP
ASP	L-Aspartate

Table 9 *Examples for Metabolites in iFF708 coding style*

4.3.5.2. iND750 genome scale metabolic model of yeast

In the year 2004, Duarte et al. proposed a fully compartmentalized genome-scale model of *Saccharomyces cerevisiae*. The iND750 stands for "*in silico*" yeast model proposed by "Natalie C. Duarte", accounting for 750 genes. The iND750 metabolic model is much more elaborate than the earlier iFF708 model. The iND750 summarizes the currently available information on ORFs, transcripts and proteins of yeast. Essentially the iND750 model differs from the iFF708 model in the following manner: (a) Localization: five additional compartments were included namely peroxisome, nucleus, golgi apparatus, vacuole and endoplasmic reticulum, (b) Revision of functional assignments of the gene products based on newly published results and description of the model in terms of elementally(mass conservation)and charge balanced reactions(charge conservation). (c) cell-wide proton balance. The iND750 file format is more detailed in terms of encoding reactions occurring in the metabolism of *S. cerevisiae*. While keeping the information on the reaction and the corresponding ORFs, it also includes information, like corresponding EC numbers, protein names as well as the biological processes to which it belongs and gives every reaction a unique reaction abbreviation. Furthermore it comprises higher number of compartments than the earlier iFF708 model.

iND750 reaction format

A reaction line consists of the following columns separated by tabs:

- Abbreviation of the reaction
- Name of the reaction
- Reaction equation
- EC Number
- Biological Process
- ORF(s)
- Protein(s) encoded by the corresponding ORF(s)

The reaction equation is denoted in the common chemical notation for reactions. If the reaction takes place in only one compartment, the equation itself is preceded by a short tag representing the compartment. This tag is separated from the equation itself by a ":" mark. Encoded compartments and their tags are:

Compartment	Tag
Extracellular	[e]
Peroxisome	[x]
Cytosol	[c]
Mitochondrion	[m]
Vacuole	[v]
Endoplasmic reticulum	[r]
Golgi apparatus	[g]
Nucleus	[n]

Table 10 *Table of the compartments used in the iND750 model and their corresponding tags*

In case of reactions connecting multiple compartments, the compartment tag at the beginning of the equation is omitted and the tag corresponding to its location is appended to every metabolite name.

The names of the reactants have to be abbreviations given in the metabolite file. All coefficients other than 1 (default when no coefficient is given explicitly) have to be given in brackets, see Table 12 for an example. There has to be a blank between coefficients, names, plus signs and the reaction arrow. Possible reaction arrows are:

- --> for a irreversible reaction
- <==> for a reversible reaction.

Columns not used in a reaction (e.g. if the EC number is unknown) are to be left blank. However the tabs separating the columns must not be omitted.

The following examples contain all relevant cases.

Abbreviation	Name	Equation	EC-Number	Process	ORF(s)		Protein(s)
ASNS1	Asparagine syn	[c] : asp-L + atp + gln-L + h2o	EC-6.3.5.4	Alanine and aspartate	(YGR124W	or	(Asn2) or (Asn1) or
	thase (glutamine	--> amp + asn-L +glu-L +h+		metabolism	YPR145W	or	(Asn3)
	hydrolysing)	ppi			YML096W)		
TREH	alpha, alpha -trehalase	[*c*] : *h*2o + *tre*-->(2)glc-D	EC-3.2.1.28	Alternate Carbon Metabolism	(YDR001C YBR001C)	or	(Nth1) or (Nth2)
O2ter	O2 endoplasmic reticulum transport	o2[c] <==> o2[r]		Transport, Endoplasmic Reticular			

Table 11 *Reaction examples for iND750 coding style*

iND750 Metabolite format

A metabolite line consists of the following columns separated by tabs:

- Abbreviation of the metabolite
- Name
- Compartment
- Formula
- Charge

Of these items, the abbreviation and the compartment are mandatory. The name and formula can be left empty as we do not use these for mapping the metabolites to their participating reactions. For the compartment the full names as stated in the reaction section above have to be used. The following examples contain all relevant cases of metabolites notation of iND750 format:

Abbreviation	Name	Compartment	Formula	Charge
his-L	L-Histidine	cytosol	C6H9N3O2	0

atp	ATP	cytosol	C10H12N5O13P3	-4
atp	ATP	mitochondrion	C10H12N5O13P3	-4

Table 12 *Examples for Metabolites in iND750 coding style*

4.4. Isotopomer analysis

Abelson and Hoering et al. discovered the relative enrichment of C-13 in the carboxyl group of amino acids in nature [Abelson1961]. This was the first indication of the existence of isotope distributions in biological compounds. Galimov et al. proposed that both in chemically equilibrated and non-equilibrated reaction systems, a microscopic reversibility of the enzymatic reactions is the cause of thermodynamically ordered isotope distributions [Galimov1985]. Schmidt et al. proposed that the kinetic isotope effects on the enzymatic reactions are the primary cause of isotope discriminations. In the last decade, various NMR and MS techniques have been developed to capture the stable isotope distributions (labeling pattern) of metabolites (see section 2.1.2).

Isotopomer is an abbreviation for "Isotope Isomer". Since a given carbon atom can either be labeled (C-13) or non-labeled (C-12) and also due to rule of numbering the carbon atom positions in a molecule, it is possible to code labeling patterns of metabolites as a sequence of ones and zeros. Similar to the construction of the stoichiometric matrix, the nonzero elements of an isotopomer correspond to the isotope paths from source isotopomers to target (product) isotopomers. Isotopomer distributions provide the maximum amount of information which can be derived from the C-13 tracer studies. In this section, we will give the definitions and describe the theory of computational approaches for studying isotopomer distributions [Schmidt1997]. A molecule which has n C atoms can have a theoretic maximum of 2^n isotopomers. The vector of all isotopomers of a given metabolite is called IDV (Isotopomer Distribution Vector). IDV is the vector containing mole fractions of metabolite molecules that are labeled in a specific pattern. The labeling of the metabolites can be represented as 0 (C-12) or 1(C-13). These sequences of zeros and ones can then be interpreted as binary numbers, the conversion of which to decimal numbers provides a unique way of ordering labeling patterns and thereby indexing them as elements of the $IDVs$. For example, in glucose IDV there are $2^6 = 64$ elements [Equation 15]. The first element of this vector is indexed as 0 and is depicted as $I_{glc}(0)$. The element at index 1 contains the mole fractions of the glucose molecules labeled by the binary number 000001_{bin}, i.e. a single labeled carbon at the sixth position. The mole fraction of the glucose labeled at first carbon position will likewise be the element at index 32 because this fraction will be represented as 100000_{bin}. The complete labeling state of glucose is given by.

$$IDV_{glu} = \begin{pmatrix} I_{glc}(0) \\ I_{glc}(1) \\ \ldots \\ \ldots \\ I_{glc}(63 \end{pmatrix} = \begin{pmatrix} I_{glc}(000000_{bin}) \\ I_{glc}(000001_{bin}) \\ \ldots \\ \ldots \\ I_{glc}(111111_{bin}) \end{pmatrix}$$

$$where \sum_{i-0}^{63} I_{glc}(i) = 1$$ **Equation 15**

4.4.1. Isotopomer mapping matrices

Generally, individual reactions in the metabolic network have one or more reactants and one and more product molecules. Additionally, there is a large number of isotopomers in each metabolite pool; hence it becomes cumbersome to write a single equation for each individual isotopomer. To solve this issue, another concept called *Isotopomer Mapping Matrices* ($IMMs$) was introduced in the literature. $IMMs$ are constructed to sum up all pairs of reactant isotopomers, which produce the respective product isotopomer in all positions of the product IDV [Schmidt1997]. For a single biochemical carbon exchange reaction, there will be single IMM defined for each pair of reactant and product molecules. The number of columns of an IMM equal the number of vector elements of the reactant IDV. The number of rows of IMM equals the number of vector elements in the product IDV. See equation below for the complete $IMM_{pyr>oaa}$ of conversion of pyruvate (**pyr**) to oxaloacetate (**oaa**)

$$IMM_{pyr>oaa} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Equation 16

4.4.2. Atom mapping matrices

Atom mapping matrices (AMM) describe the conversion of atoms of a substrate metabolite to the atoms of the product by a given reaction. The AMM elements are constants and are defined *a priori* for every reaction [Zupke1995]. In the web server (MetaModel), the AMM format is simple. The first line is the description line and the following several lines store the actual AMM. Different $AMMs$ have to be separated by a blank line. Every description line begins with a # sign, followed by the reaction which is described by the AMM in the form: Substrate Name → Product Name; followed by the number of carbons of the two compounds taking part in the reaction is given as: $Ns = x; Np = y$

#PYR -> Ac-CoA; Ns =3; Np =2

0 1 0
0 0 1

#Cit -> Ac-CoA; Ns =6; Np =2

1 0 0 0 0 0
0 1 0 0 0 0

Table 13 *Example for two AMMs*

Every *AMM* line consists of the coefficients of one row of the *AMM* separated by spaces. No leading spaces in front of the first coefficient and after the last one are allowed.

#G6P -> Ru5P;Ns=6;Np=5 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1	#Cit -> a-KG;Ns=6;Np=5 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0	#Ru5P -> F6P;Ns=5;Np=6 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1
#G6P -> CO2;Ns=6;Np=1 1 0 0 0 0 0	#a-KG -> CO2;Ns=5;Np=1 1 0 0 0 0	#Mal -> CO2;Ns=4;Np=1 0 0 0 1
#F1,6biP -> DHAP;Ns=6;Np=3 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0	#F1,6biP -> GAP;Ns=6;Np=3 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1	#Mal -> Pyr;Ns=4;Np=3 1 0 0 0 0 1 0 0 0 0 1 0
#DHAP -> GAP;Ns=3;Np=3 0 0 1 0 1 0 1 0 0	#Ru5P -> GAP;Ns=5;Np=3 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1	#Pyr -> OAA;Ns=3;Np=4 1 0 0 0 1 0 0 0 1 0 0 0
#Ac-CoA -> Cit;Ns=2;Np=6 1 0 0 1 0 0 0 0 0 0 0 0	#OAA -> Cit;Ns=4;Np=6 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0	#CO2 -> OAA;Ns=1;Np=4 0 0 0 1
#PYR -> Ac-CoA;Ns=3;Np=2 0 1 0 0 0 1	#Cit -> Ac-CoA;Ns=6;Np=2 1 0 0 0 0 0 0 1 0 0 0 0	#Cit -> CO2;Ns=6;Np=1 0 0 0 0 0 1
#a-KG -> OAA;Ns=5;Np=4;f17 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0	#a-KG -> OAA;Ns=5;Np=4;f18 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1	#Cit -> OAA;Ns=6;Np=4 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
#PYR -> CO2;Ns=3;Np=1 1 0 0		

Table 14 *AMM for metamodel*

4.4.3.Bottlenecks

In various studies, determination of a large number of isotopomer mapping matrices is a complex task. In our webserver, we have automated the conversion of atom mapping matrices into IMM. The estimation of the initial AMM set is an unsolved problem and there is no general method or repository available.

4.4.4. Data

The primary input for isotopomer analysis comprises the atom mapping matrices which present a comprehensive view of the atom flow in a set of metabolite pool. Basically, all chemical reactions need to be represented in terms of *AMM* among individual metabolites involved in these reactions. Table 14 comprises the *AMMs* that we used for isotopomer analysis.

AMM list: Here each box represents a AMM conversion table which captures a given cellular reaction. The first line the actually reaction, for example “#G6P ->Ru5P; Ns=6; Np=5” represents a reaction in which G6P gets converted to Ru5P, also Ns is the number of carbons in the substrate(G6P) and Np is the number of carbons in the product(Ru5P)

4.5. Synthetic accessibility of metabolites

4.5.1. Synthetic accessibility: Definition

The *synthetic accessibility* S_i of a metabolite i is the minimal number of metabolic reactions needed to produce i from the network inputs [Wunderlich2006]. The *total synthetic accessibility* of biomass S is the summation of the synthetic accessibility over all components of the biomass.

$$S = \sum_i S_i$$

Equation 17

The algorithm for computing the synthetic accessibility is based on an iterative breadth-first search. The algorithm initially examines all the reactions that require a specific metabolite as a reactant. It then labels the reactions for which all the reactants are available, as “accessible” and subsequently all the output metabolites of these reactions as “accessible”. The algorithm iteratively examines all the reactions which require one of the newly “accessible” labeled metabolites as a reactant, determines whether all reactions are accessible or not based on the “accessibility” of the reactants until no new metabolite is found to be “accessible”.

4.5.2. Measures of production capability

In the past, many topological measures of network production capability have been reported. These include enzyme usage, node degree, graph diameter. Node degree refers to the number of incoming/outgoing edges linked to a node. The nodes with the higher degrees tend to be more important for the network functionality as they generally act as hubs and critical members for a large number of chemical transformations [Jeong2001, Albert2000]. Enzyme usage is another measure which is somewhat similar to the synthetic accessibility measure. It is defined as the number of times the reactions catalyzed by each enzyme are used to produce the biomass components in the wild type strain [Newman2001]. Also, increase in the graph diameter compared to the graph diameter of the wild type, can be a rough measure for inviability of a knockout.

4.5.3. The scope of a metabolite

The concept of *scope* of a metabolite for studying network expansion was first introduced by Handorf et al. [Handorf2005]. The concept of scope of a metabolite exploits the inherent hierarchical ordering of the metabolic reactions in metabolic pathways and networks. Scopes are defined as sets of metabolites that can be synthesised by a metabolic network when it is provided with given seeds (Sets of initial metabolic compounds). Thus, scopes represent synthesizing capacities of the seeds in the network [Handorf2006].

4.5.4. Growth medium

With reference to the synthetic accessibility measure, the *growth medium* comprises all initial input metabolites. The input metabolites are chosen to cover the real composition of a minimal medium as much as possible. For example, as our wild type strain is amino acid auxotrophic in nature, meaning thereby it requires certain metabolites namely the amino acids histidine, leucine, and uracil for normal growth. Hence these metabolites are included as the input metabolite in order to compensate for amino-acid auxotrophy. Other metabolites like oxidized form of thioredoxin, H1 (in the endoplasmic reticulum), NADPH (in the endoplasmic reticulum), and dolichol should also be included as inputs.

This is due to the fact that in the absence of these components, the wild type strain would also be rendered “inviable” (see section 4.5.6.1).

4.5.5. Standard biomass components

Similar to the concept of growth medium, the synthetic accessibility algorithm has a basic assumption that all standard biomass components should be synthesizable, hence should be part of the “accessible” output metabolites (see section 4.5.6.2).

4.5.6. Data

The following subsections list the input and output metabolites that are used to mimic growth under minimal media for yeast.

4.5.6.1. Input medium components

Amino Acids		
L-Alanine	L-Arginine	L-Asparagine
L-Aspartate	L-Cysteine	L-Glutamine
L-Glutamate	Glycine	L-Histidine
L-Isoleucine	L-Leucine	L-Lysine
L-Methionine	L-Phenylalanine	L-Proline
L-Serine	L-Threonine	L-Tryptophan
L-Tyrosine	L-Valine	
Nucleotides		
Adenine	Cytosine	Guanine
Thymine		
Other Metabolites		
O_2	K+	Na+
SO_4	Thioredoxin ox.	Trehalose
Uracil	H_2O	Ammonium
CO_2	H	Inorganic phosphate
Dolichol		
Carbon Source		
Glucose		

Table 15 *Standard medium composition assumed by the Synthetic Accessibility method*

Table 15 depicts the standard input medium component assumed in the synthetic accessibility method in MetaModel.

4.5.6.2. Output biomass components

Table 16 depicts the standard input biomass component assumed in the synthetic accessibility method in MetaModel.

1,3-beta-D-Glucan	L-Leucine
AMP	L-Lysine
L-Arginine	Mannan
L-Asparagine	L-Methionine
L-Aspartate	Phosphatidate, yeast-specific
ATP	Phosphatidylcholine, yeast-specific
CMP	Phosphatidylethanolamine, yeast-specific
L-Cysteine	L-Phenylalanine
dAMP	L-Proline
dCMP	Phosphatidylserine, yeast-specific
dGMP	Phosphatidyl-1D-myo-inositol, yeast-specific
dTMP	L-Serine
Ergosterol	Sulfate
L-Glutamine	L-Threonine
L-Glutamate	Trehalose
Glycine	Triglyceride, yeast-specific
Glycogen	L-Tryptophan
GMP	L-Tyrosine
H2O	UMP
L-Histidine	L-Valine
L-Isoleucine	Zymosterol

Table 16 *Standard biomass composition assumed by the Synthetic Accessibility method*

Chapter 5

5. *Implementation of the Web server for metabolic network analysis*

In this chapter, we describe the implementation, functionality and design of various analysis modules of our webserver, MetaModel.

5.1. Introduction

In the current implementation, MetaModel facilitates the analysis of the *Saccharomyces cerevisiae* metabolic models iFF708 and iND750, and of user-defined custom models. The web server has three modules namely, ***Stoichiometry***, ***Isotopomer Path Tracing*** and ***Optimization.***

5.2. Implementation

Metamodel has been entirely coded in the Python language. It is currently available at
http://mpiat3502.ag3.mpi-sb.mpg.de/metamodel/index.php.
For further details, refer to the programmer's guide in appendix 2.

5.2.1. Schematic view

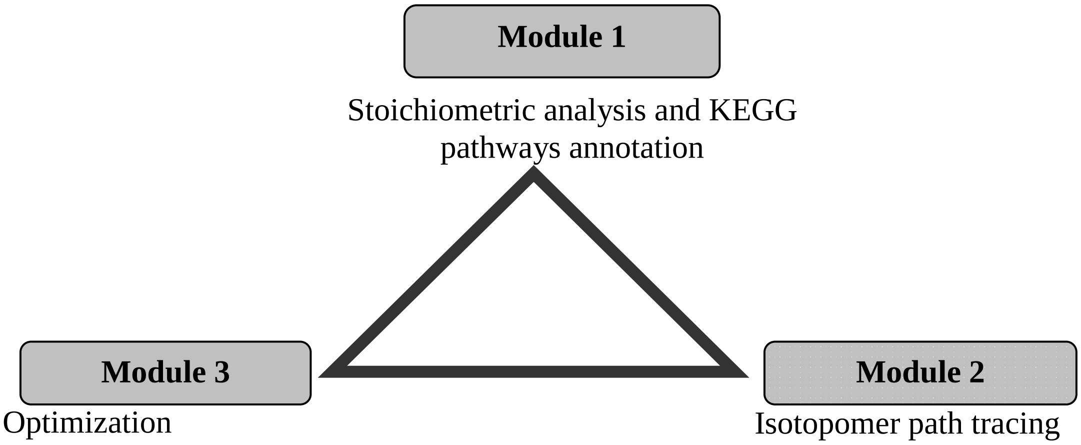


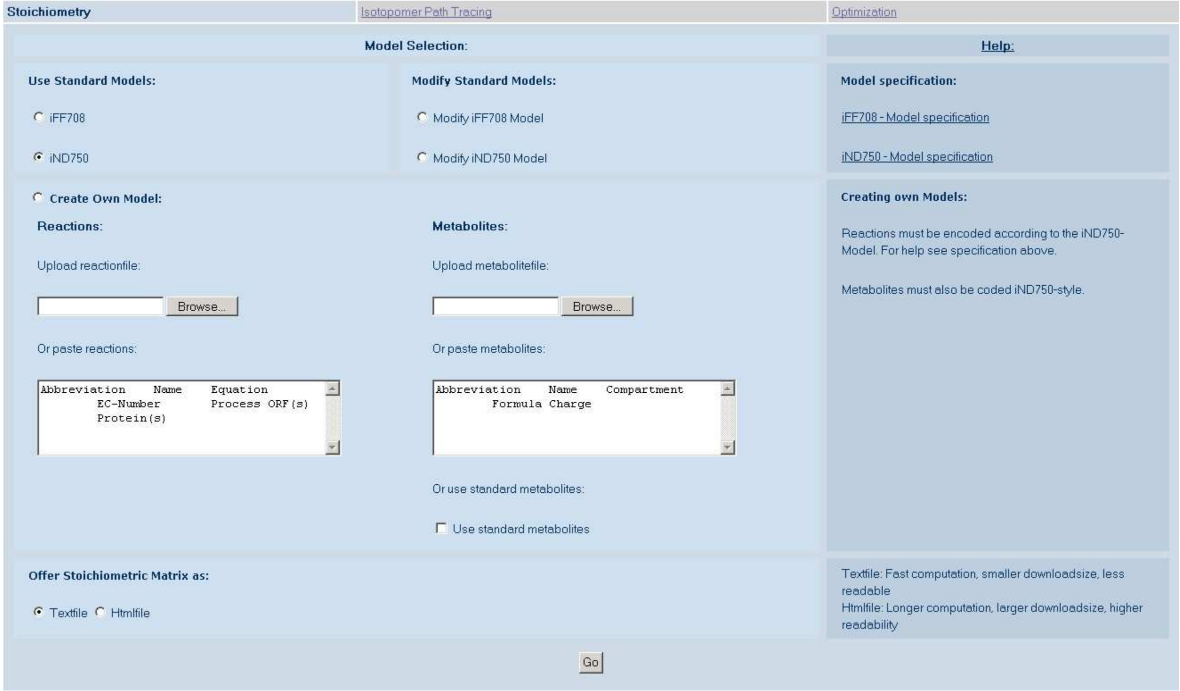
Figure 17 *Metamodel webserver:* basic modules

Figure 17 represents the basic modules coded in Metamodel webserver. Module 1 deals with construction of stoichiometric matrices for literature defined or user defined genome scale model of yeast and other organisms. It also provides a functionality to edit the literature models and encodes search functions for metabolite-reactions and vice versa associations. Module 1 also has a script for finding out the KO pathway annotation for already annotated or unannotated ORF present in the genome scale model understudy. Module 2 deals with the calculation of isotopomer mapping matrices (IMM) from atom mapping matrices (AMM) for a given reaction set. Module 3 works on finding out the node type for all the metabolites present in the genome scale model understudy, as well as calculation of synthetic accessibility score for single and multiple gene knockout mutants. Sections 5.2.2-5.2.4 presents these three modules in further details.

5.2.2. Module I: *Stoichiometry*

5.2.2.1. Design and Implementation

This module performs stoichiometric analysis on a given metabolic network. Figure 18 depicts the user interface of the module I Stoichiometry



Model Selection:

Use Standard Models:

○ iFF708

● iND750

Modify Standard Models:

○ Modify iFF708 Model

○ Modify iND750 Model

Model specification:

iFF708 - Model specification

iND750 - Model specification

○ Create Own Model:

Reactions:

Upload reactionfile:

Or paste reactions:

```
Abbreviation    Name    Equation
        EC-Number       Process  ORF(s)
        Protein(s)
```

Metabolites:

Upload metabolitefile:

Or paste metabolites:

```
Abbreviation    Name    Compartment
        Formula Charge
```

Or use standard metabolites:

□ Use standard metabolites

Creating own Models:

Reactions must be encoded according to the iND750-Model. For help see specification above.

Metabolites must also be coded iND750-style.

Offer Stoichiometric Matrix as:

● Textfile ○ Htmlfile

Textfile: Fast computation, smaller downloadsize, less readable
Htmlfile: Longer computation, larger downloadsize, higher readability

Figure 18 *Stoichiometry module's user interface*

- Uploading a Metabolic Model

There are three ways to input a metabolic model in MetaModel:

- Use Standard Models:

There are two models to choose from: iFF708 and iND750. For a detailed description of model format and the differences between the two models please use the online help on the web-server.

- Modify Standard Models:

When using this option you will see a text area containing the selected model's reactions in raw format. You can modify the reaction set at this stage. This includes the possibility of deleting reactions, adding new reactions and modifying reactions. After pressing the Save Modifications button, you will be forwarded to the result page.

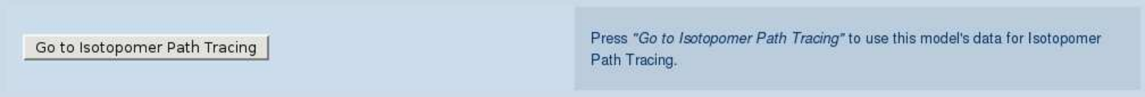
- Create Own Model:

You can input custom-defined models by uploading the reactions and metabolites of the network, in iND750 like format (see section 4.3.5.2.). You can input the reactions and metabolites in the following ways: a) Uploading a file, b) Paste the text in the textbox, offered by the user interface. Alternatively the user can select to use the standard metabolites if not specified explicitly.

Finally the user can specify the output format for the stoichiometric matrix. The text file is a tab-separated presentation of the matrix whereas the html file presents the data in the form of an html table. If the user selects one of the above models as well as the output format and clicks on the "Go" button, the next page will be showing the reactions comprising the selected model and the user gets forwarded to the result page.

5.2.2.2. Functionality

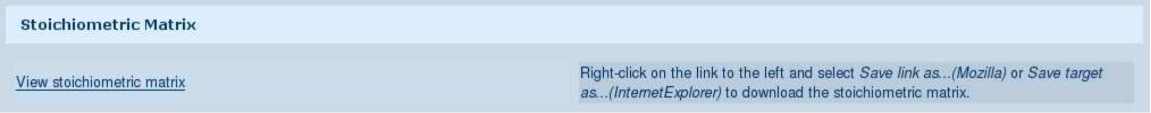
The Stoichiometry module calculates the stoichiometric matrix as well as statistics about the defined network and offers functions for searching for reactions and/or pathways containing given metabolites (Figure 22). The figure 22 depicts the two metabolite centric search functions encoded in the Stoichiometry module namely i) searching all the reactions in which a given metabolites participates and ii) searching all the processes in which a given metabolite participates. At this stage the user can go directly to module 2 "Isotopomer Path Tracing" or proceed to the KO-based annotation method (Figure 23).



Press *"Go to Isotopomer Path Tracing"* to use this model's data for Isotopomer Path Tracing.

Figure 19 *Link for module 2*

- Stoichiometric matrix:



Stoichiometric Matrix

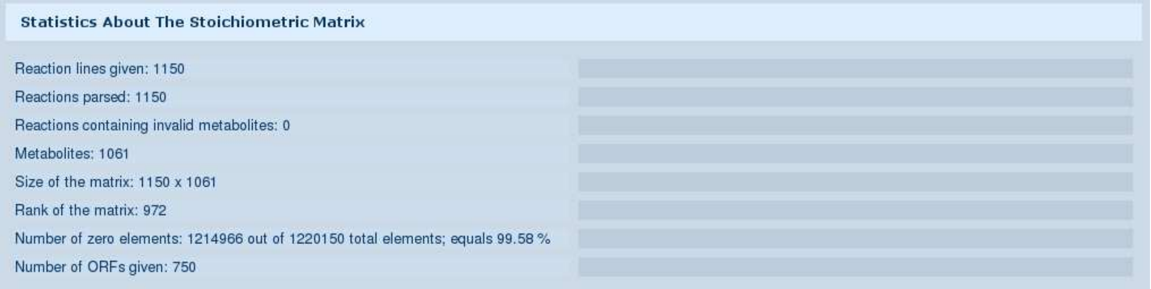
View stoichiometric matrix

Right-click on the link to the left and select *Save link as...(Mozilla)* or *Save target as...(InternetExplorer)* to download the stoichiometric matrix.

Figure 20 *Link for stoichiometric matrix download*

The stoichiometric matrix can be downloaded in the user-defined format.

- Statistics:



Statistics About The Stoichiometric Matrix

Reaction lines given: 1150

Reactions parsed: 1150

Reactions containing invalid metabolites: 0

Metabolites: 1061

Size of the matrix: 1150 x 1061

Rank of the matrix: 972

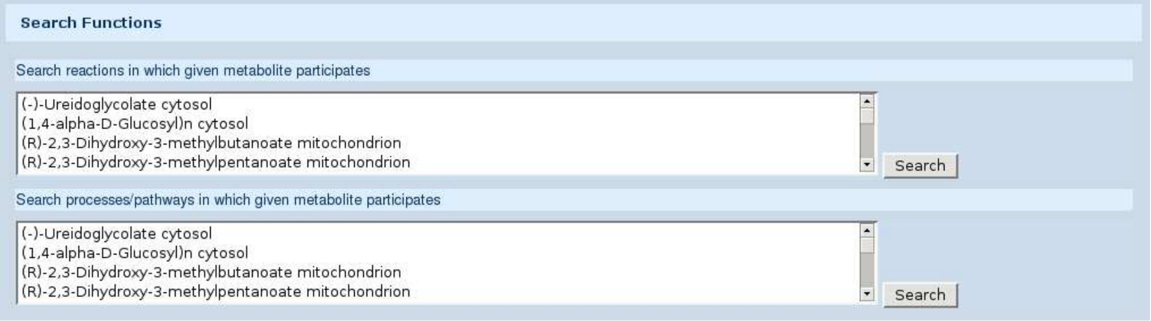
Number of zero elements: 1214966 out of 1220150 total elements; equals 99.58 %

Number of ORFs given: 750

Figure 21 *Statistics on the stoichiometry of the metabolic model*

Statistics include the following information namely, the number of reactions and metabolites which can be extracted from the input, a list of invalid reactions, the size of the matrix, the rank of the matrix, the sparseness of the matrix and the number of ORFs.

Search functions:



Search Functions

Search reactions in which given metabolite participates

(-)-Ureidoglycolate cytosol
(1,4-alpha-D-Glucosyl)n cytosol
(R)-2,3-Dihydroxy-3-methylbutanoate mitochondrion
(R)-2,3-Dihydroxy-3-methylpentanoate mitochondrion

Search processes/pathways in which given metabolite participates

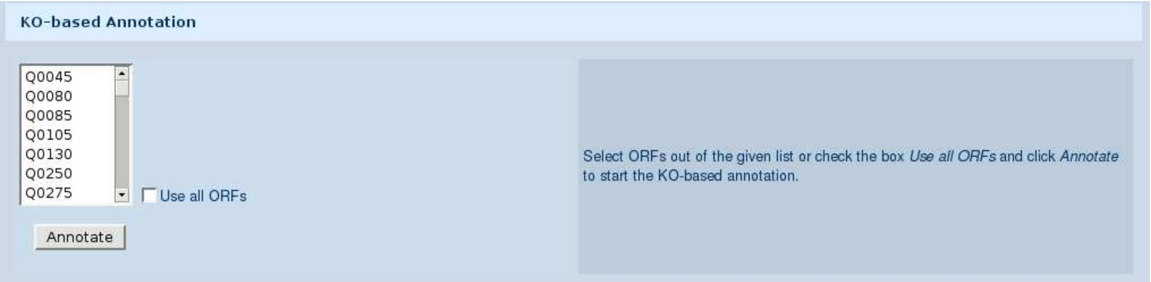
(-)-Ureidoglycolate cytosol
(1,4-alpha-D-Glucosyl)n cytosol
(R)-2,3-Dihydroxy-3-methylbutanoate mitochondrion
(R)-2,3-Dihydroxy-3-methylpentanoate mitochondrion

Figure 22 *Search function interface*

Depending on the format for encoding the reactions, which has been selected by the user, following search functions are offered:

- iND750 format:
    - list the reactions which include a given metabolite.
    - list the pathways/processes which include a given metabolite.
- iFF708 format:
    - Search reactions in which a given metabolites participates

▪ KO-based Annotation:



Select ORFs out of the given list or check the box *Use all ORFs* and click *Annotate* to start the KO-based annotation.

Figure 23 *Input interface for KO based annotation*

The KO-based annotation uses scripts from the KOBAS library. This library annotates given ORF identifiers with the corresponding KEGG Orthology (KO) terms and identifies pathways which are statistically enriched with these genes. KO was developed for integration of pathway and genomic information in KEGG. KO is an extension of ortholog identifiers and is composed of a DAG hierarchy of four flat levels. The top level consists of the following five categories: metabolism, genetic information processing, environmental information processing, cellular processes and human diseases. The second level divides the five functional categories into finer sub-categories. The third level corresponds directly to the KEGG pathways, and the fourth level consists of the leaf nodes, which are the functional terms [Goto2000].

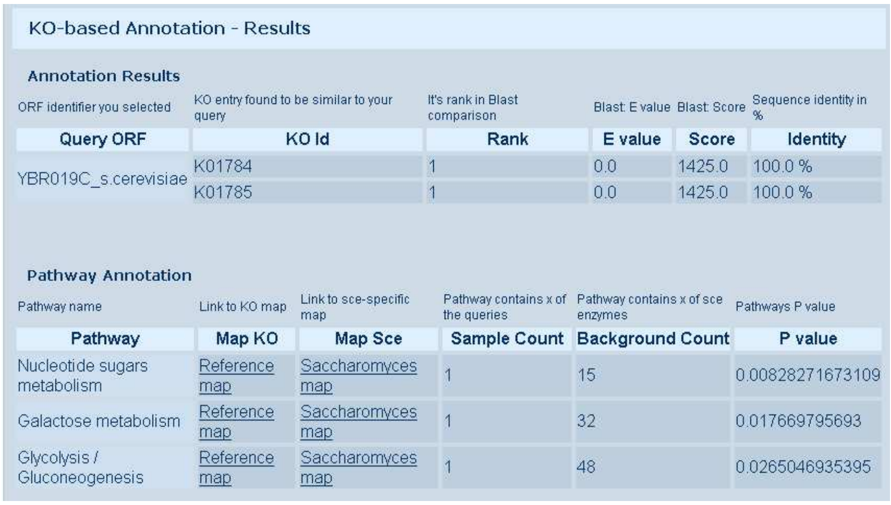
As input, the user can either select the relevant ORFs from the selection box or simply annotate all ORFs by checking the box ”Select all ORFs”. Another way of specifying the

ORFs which one wants to annotate is by entering them manually, separated by commas, in the text box supplied by the user interface. The output includes already annotated KO terms and a list of enriched pathways with the link to their corresponding KEGG-map. We used the KOBAS package to carry out the KO-based annotation [Wu2006].

5.2.2.3.Validation

Exploration of YBR019C (Gal10) gene knockout

GAL10(YBR019C) encodes a bifunctional enzyme with mutarotase and UDP galactose 4-epimerase activities. Both of these functions are key in the process of galactose catabolism; mutarotase converts beta-D-galactose into its alpha form and galactose 4-epimerase catalyzes the reversible conversion between UDP-galactose and UDP-glucose (Majumdar2004, Fukasawa1980, DE1958, Cherry1998). Loss of Gal10p activity renders cells unable to grow when galactose is the sole carbon source (Douglas1964). In this work, we selected the "YBR019C" orf in the *KO-based annotation* script of the "Stoichiometry" module. Each annotation is accompanied by the p-values calculated using KO terms.



KO-based Annotation - Results

Annotation Results

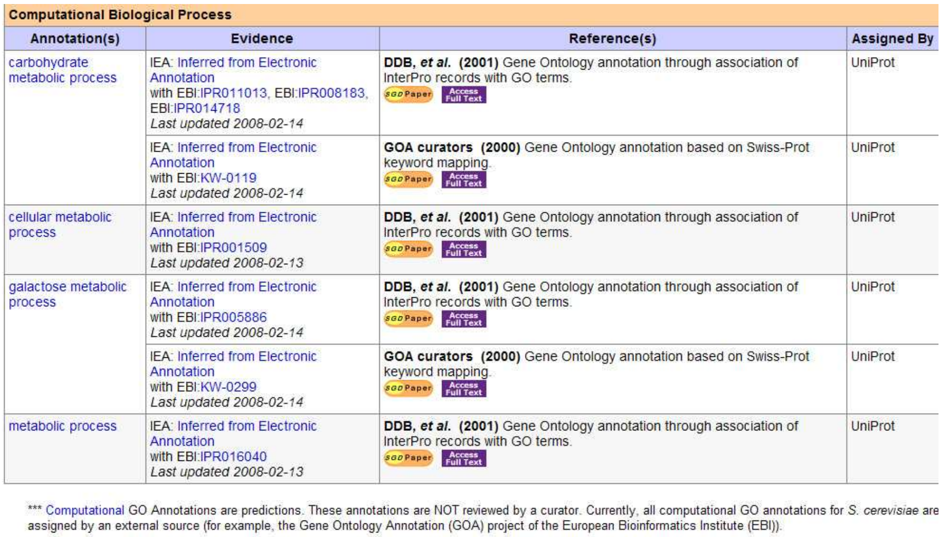
ORF identifier you selected	KO entry found to be similar to your query	It's rank in Blast comparison	Blast E value	Blast Score	Sequence identity in %
Query ORF	**KO Id**	**Rank**	**E value**	**Score**	**Identity**
YBR019C_s.cerevisiae	K01784	1	0.0	1425.0	100.0 %
	K01785	1	0.0	1425.0	100.0 %

Pathway Annotation

Pathway name	Link to KO map	Link to sce-specific map	Pathway contains x of the queries	Pathway contains x of sce enzymes	Pathways P value
Pathway	**Map KO**	**Map Sce**	**Sample Count**	**Background Count**	**P value**
Nucleotide sugars metabolism	Reference map	Saccharomyces map	1	15	0.00828271673109
Galactose metabolism	Reference map	Saccharomyces map	1	32	0.017669795693
Glycolysis / Gluconeogenesis	Reference map	Saccharomyces map	1	48	0.0265046935395

Figure 24 *Pathway annotation for the YBR019C*

We found that the KO-based annotation also assigns the YBR019C to "Nucleotide sugars metabolism" (p-value 0.008), "Galactose metabolism" (p-value 0.018) and "Glycolysis/Gluconeogenesis" (p-value 0.026) (Figure 24). All these annotations agree with the literature knowledge for YBR019C. *KO-based annotation* is a straight forward tool which can be applied to gene annotation. In addition to the functional annotation of the ORF specified, MetaModel provides a platform for interfacing with the KEGG pathway annotation for the ORF understudy, on the fly. Figure 26 depicts the KEGG map, ”Nucleotide Sugars Metabolism”, with the corresponding enzyme classification number of the ORF under study.



Computational Biological Process

Annotation(s)	Evidence	Reference(s)	Assigned By
carbohydrate metabolic process	IEA: Inferred from Electronic Annotation with EBI:IPR011013, EBI:IPR008183, EBI:IPR014718 *Last updated 2008-02-14*	**DDB, *et al.* (2001)** Gene Ontology annotation through association of InterPro records with GO terms. SGD Paper Access Full Text	UniProt
	IEA: Inferred from Electronic Annotation with EBI:KW-0119 *Last updated 2008-02-14*	**GOA curators (2000)** Gene Ontology annotation based on Swiss-Prot keyword mapping. SGD Paper Access Full Text	UniProt
cellular metabolic process	IEA: Inferred from Electronic Annotation with EBI:IPR001509 *Last updated 2008-02-13*	**DDB, *et al.* (2001)** Gene Ontology annotation through association of InterPro records with GO terms. SGD Paper Access Full Text	UniProt
galactose metabolic process	IEA: Inferred from Electronic Annotation with EBI:IPR005886 *Last updated 2008-02-14*	**DDB, *et al.* (2001)** Gene Ontology annotation through association of InterPro records with GO terms. SGD Paper Access Full Text	UniProt
	IEA: Inferred from Electronic Annotation with EBI:KW-0299 *Last updated 2008-02-14*	**GOA curators (2000)** Gene Ontology annotation based on Swiss-Prot keyword mapping. SGD Paper Access Full Text	UniProt
metabolic process	IEA: Inferred from Electronic Annotation with EBI:IPR016040 *Last updated 2008-02-13*	**DDB, *et al.* (2001)** Gene Ontology annotation through association of InterPro records with GO terms. SGD Paper Access Full Text	UniProt

*** Computational GO Annotations are predictions. These annotations are NOT reviewed by a curator. Currently, all computational GO annotations for *S. cerevisiae* are assigned by an external source (for example, the Gene Ontology Annotation (GOA) project of the European Bioinformatics Institute (EBI)).

Figure 25 *Computational annotations for YBR019C from Saccharomyces Genome Database (SGD)*

Also, in case of uncharacterized ORFs, MetaModel provides most probable functional annotations along with the reference organism and the *S. cerevisiae* KEGG map.

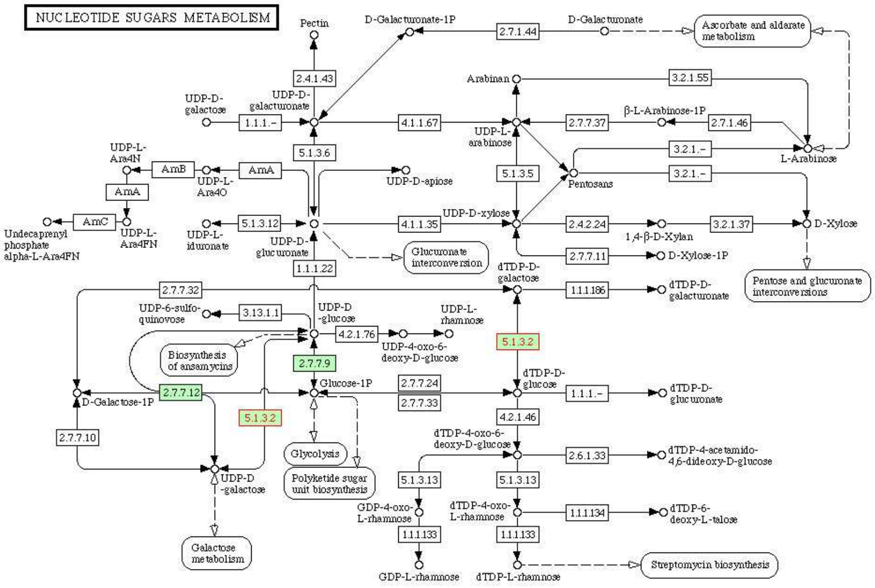
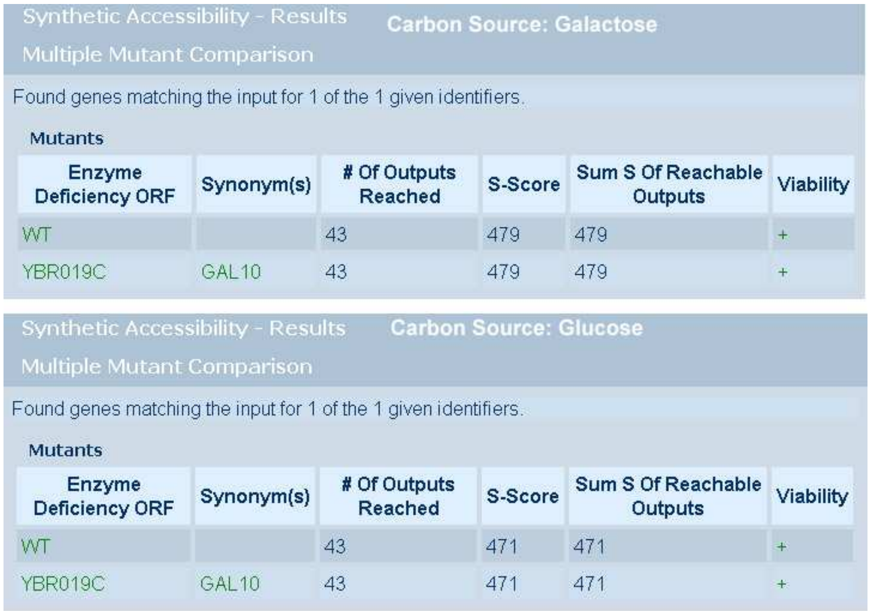


Figure 26 *KEGG pathway map of YBR019C (from MetaModel).* EC number 5.1.3.2 is the EC number of YBR019C.

Analysis of a set of genes

In this work, we worked on analyzing a set of genes which have known or putative association with a specific biological process, using *KO-based annotation* in "Stoichiometry" module. We searched SGD to find those ORFs which are associated to the term "glucose metabolic process".



Synthetic Accessibility - Results
Multiple Mutant Comparison
Carbon Source: Galactose

Found genes matching the input for 1 of the 1 given identifiers.

Mutants

Enzyme Deficiency ORF	Synonym(s)	# Of Outputs Reached	S-Score	Sum S Of Reachable Outputs	Viability
WT		43	479	479	+
YBR019C	GAL10	43	479	479	+

Synthetic Accessibility - Results
Multiple Mutant Comparison
Carbon Source: Glucose

Found genes matching the input for 1 of the 1 given identifiers.

Mutants

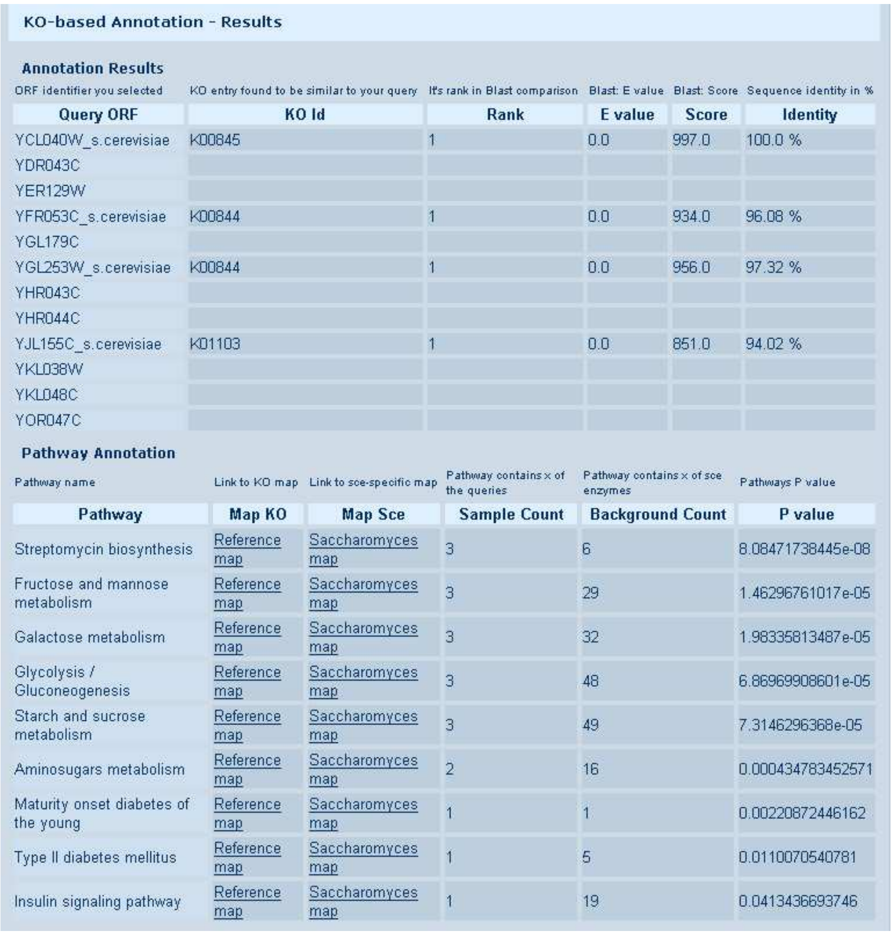
Enzyme Deficiency ORF	Synonym(s)	# Of Outputs Reached	S-Score	Sum S Of Reachable Outputs	Viability
WT		43	471	471	+
YBR019C	GAL10	43	471	471	+

Table 17 *Synthetic accessibility score of YBR019C calculated using "Optimization" module of MetaModel.*

There are 22 ORFs which are associated with the term "glucose metabolic process". Out of these 22 ORFs, 12 ORFs are manually annotated (K_{Manual} set; method used: ISS (Inferred from Sequence or Structural similarity)\IGI (Inferred from Genetic Interactions)\IMP(Inferred from Mutant Phenotype)\TAS(Traceable Author Statement)) to the above term and the rest 10 ORFs are computationally inferred($K_{Computational}$ set; method used: IEA i.e. Inferred from Electronic Annotation) to be associated with the above term. Table 18 lists the ORF set associated with the gene term "glucose metabolic process". We start with entering these two sets of ORFs individually in the input text-box in *KO-based annotation.* For the first ORF set, K_{Manual} , we find that out of 12 ORFs we could annotate 4 ORFs with the pathways which are part of more general "glucose metabolic process" term. In addition, we find certain pathways like "Type II diabetes mellitus” and “Insulin signalling pathways” which are close in terms of association to “glucose metabolic process” (Table 19).

Manual annotated mutant set (K_{Manual})	Computationally annotated mutant set($K_{Computational}$)
YCL040W	YIL042C
YKL048C	YOR125C
YER129W	YNL241C
YGL179C	YGR192C
YHR044C	YKL127W
YJL155C	YMR278W
YKL038W	YJL052W
YHR043C	YMR105C
YFR053C	YJR009C
YOR047C	YJR090C
YDR043C	
YGL253W	

Table 18 *List of manually and computationally annotated genes which are associated with the term “glucose metabolic process”*



KO-based Annotation - Results

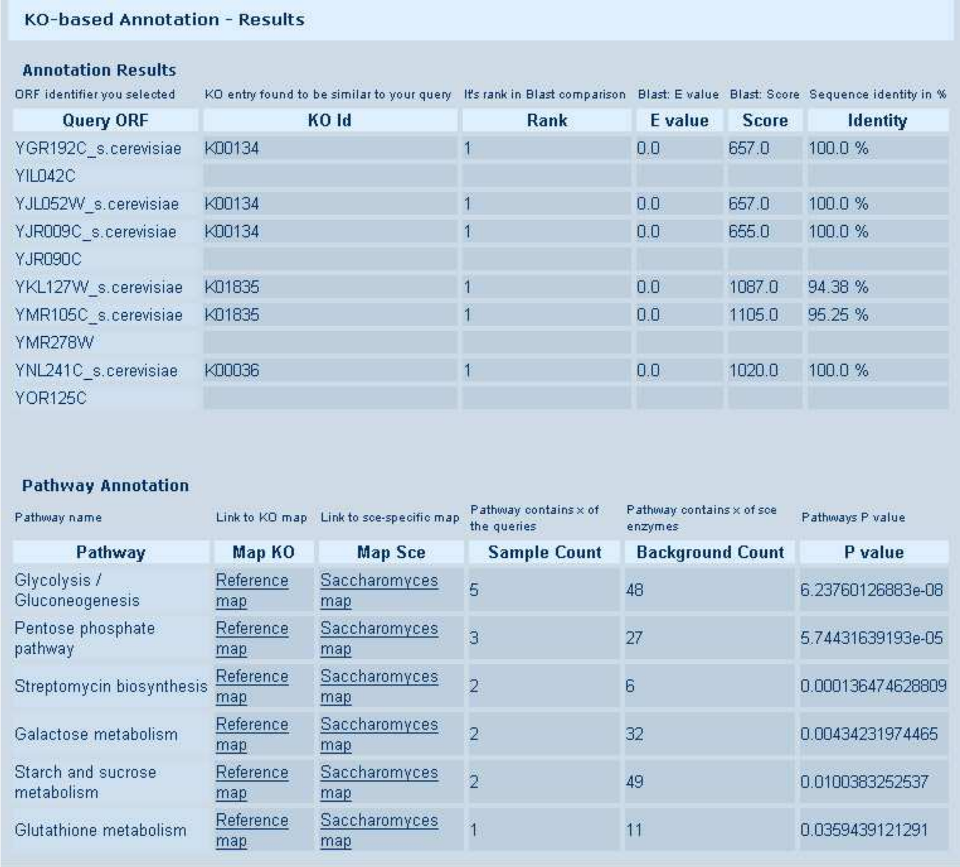
Annotation Results

ORF identifier you selected	KO entry found to be similar to your query	It's rank in Blast comparison	Blast: E value	Blast: Score	Sequence identity in %
Query ORF	**KO Id**	**Rank**	**E value**	**Score**	**Identity**
YCL040W_s.cerevisiae	K00845	1	0.0	997.0	100.0 %
YDR043C					
YER129W					
YFR053C_s.cerevisiae	K00844	1	0.0	934.0	96.08 %
YGL179C					
YGL253W_s.cerevisiae	K00844	1	0.0	956.0	97.32 %
YHR043C					
YHR044C					
YJL155C_s.cerevisiae	K01103	1	0.0	851.0	94.02 %
YKL038W					
YKL048C					
YOR047C					

Pathway Annotation

Pathway name	Link to KO map	Link to sce-specific map	Pathway contains x of the queries	Pathway contains x of sce enzymes	Pathways P value
Pathway	**Map KO**	**Map Sce**	**Sample Count**	**Background Count**	**P value**
Streptomycin biosynthesis	Reference map	Saccharomyces map	3	6	8.08471738445e-08
Fructose and mannose metabolism	Reference map	Saccharomyces map	3	29	1.46296761017e-05
Galactose metabolism	Reference map	Saccharomyces map	3	32	1.98335813487e-05
Glycolysis / Gluconeogenesis	Reference map	Saccharomyces map	3	48	6.86969908601e-05
Starch and sucrose metabolism	Reference map	Saccharomyces map	3	49	7.3146296368e-05
Aminosugars metabolism	Reference map	Saccharomyces map	2	16	0.000434783452571
Maturity onset diabetes of the young	Reference map	Saccharomyces map	1	1	0.00220872446162
Type II diabetes mellitus	Reference map	Saccharomyces map	1	5	0.0110070540781
Insulin signaling pathway	Reference map	Saccharomyces map	1	19	0.0413436693746

Table 19 K_{Manual} *mutant set’s pathway annotation using MetaModel.*



KO-based Annotation - Results

Annotation Results

ORF identifier you selected	KO entry found to be similar to your query	It's rank in Blast comparison	Blast: E value	Blast: Score	Sequence identity in %
Query ORF	KO Id	Rank	E value	Score	Identity
YGR192C_s.cerevisiae	K00134	1	0.0	657.0	100.0 %
YIL042C					
YJL052W_s.cerevisiae	K00134	1	0.0	657.0	100.0 %
YJR009C_s.cerevisiae	K00134	1	0.0	655.0	100.0 %
YJR090C					
YKL127W_s.cerevisiae	K01835	1	0.0	1087.0	94.38 %
YMR105C_s.cerevisiae	K01835	1	0.0	1105.0	95.25 %
YMR278W					
YNL241C_s.cerevisiae	K00036	1	0.0	1020.0	100.0 %
YOR125C					

Pathway Annotation

Pathway name	Link to KO map	Link to sce-specific map	Pathway contains x of the queries	Pathway contains x of sce enzymes	Pathways P value
Pathway	Map KO	Map Sce	Sample Count	Background Count	P value
Glycolysis / Gluconeogenesis	Reference map	Saccharomyces map	5	48	6.23760126883e-08
Pentose phosphate pathway	Reference map	Saccharomyces map	3	27	5.74431639193e-05
Streptomycin biosynthesis	Reference map	Saccharomyces map	2	6	0.000136474628809
Galactose metabolism	Reference map	Saccharomyces map	2	32	0.00434231974465
Starch and sucrose metabolism	Reference map	Saccharomyces map	2	49	0.0100383252537
Glutathione metabolism	Reference map	Saccharomyces map	1	11	0.0359439121291

Table 20 $K_{Computational}$ *mutant set's pathway annotation using MetaModel.*

For the second ORF set, $K_{Computational}$, we find that out of 10 ORFs we could annotate 6 ORFs with the pathways like "glycolysis", "pentose phosphate pathway", "streptomycin biosynthesis", "galactose metabolism", "starch and sucrose metabolism" and "glutathione metabolism" with pvalues less than 0.036(Table 20).

5.2.3. Module II: *Isotopomer Path Tracing*

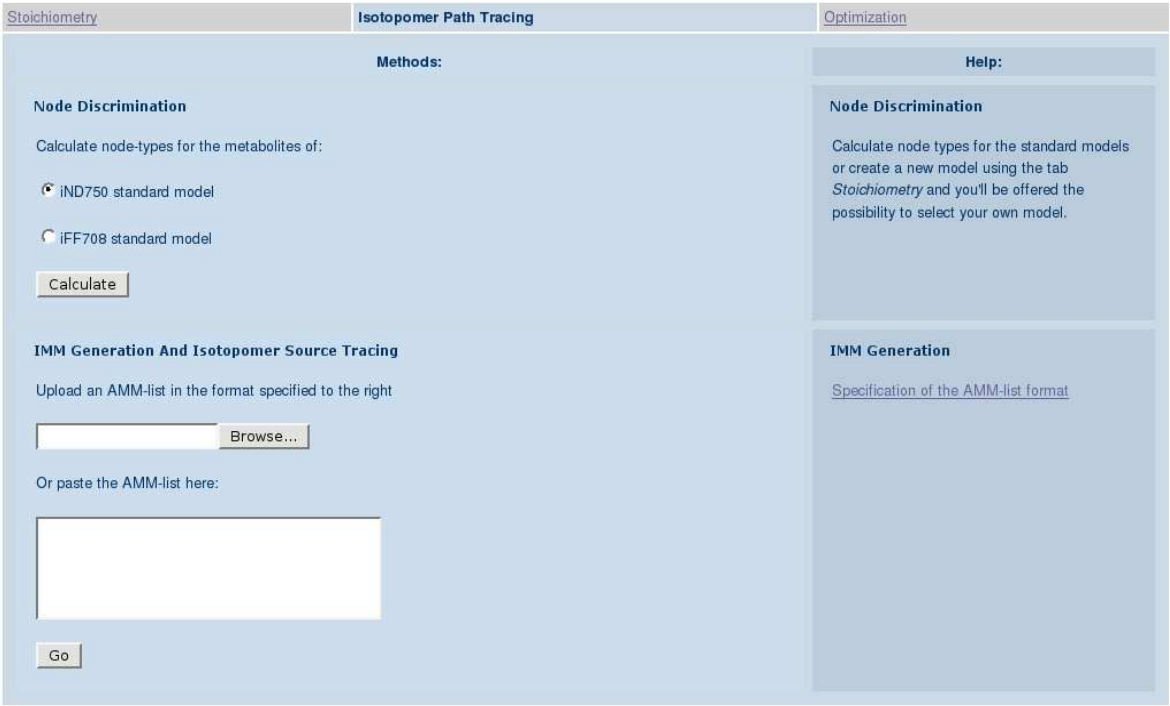
5.2.3.1. Design and implementation

Module II consists of two different methods namely,

- Node discrimination
- IMM generation and isotopomer path tracing

▪ Node discrimination

- The "Node discrimination" script calculates the types of metabolites in a given metabolic network. The type of a metabolite is assigned according to in-degree and out-degree of its corresponding node [Forbes2001].



Node Discrimination

Calculate node-types for the metabolites of:

● iND750 standard model

○ iFF708 standard model

Calculate

IMM Generation And Isotopomer Source Tracing

Upload an AMM-list in the format specified to the right

Or paste the AMM-list here:

Node Discrimination

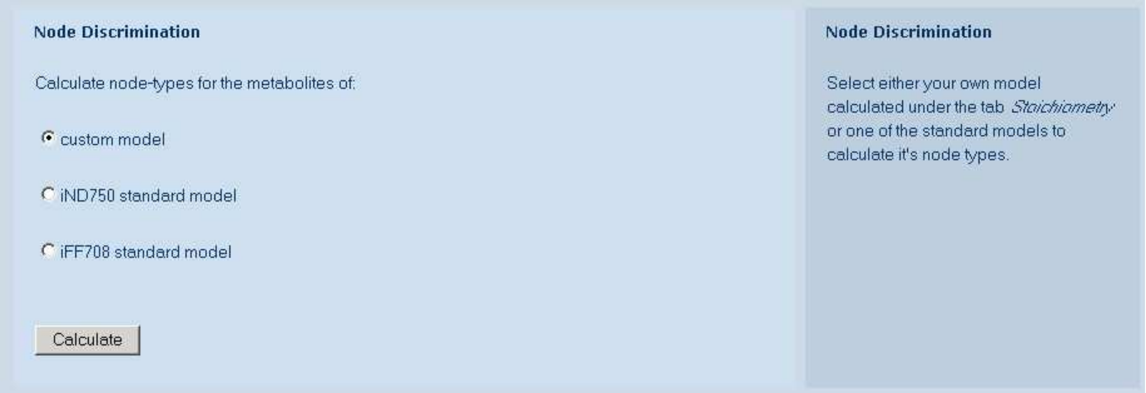
Calculate node types for the standard models or create a new model using the tab *Stoichiometry* and you'll be offered the possibility to select your own model.

IMM Generation

Specification of the AMM-list format

Figure 27 *Module 2: Input Interface*

▪ Input



Node Discrimination

Calculate node-types for the metabolites of:

● custom model

○ iND750 standard model

○ iFF708 standard model

Calculate

Node Discrimination

Select either your own model calculated under the tab *Stoichiometry* or one of the standard models to calculate it's node types.

Figure 28 *Input for the Node Discrimination method*

In this case the input is only the metabolic model. MetaModel performs node discrimination for three different metabolic models:

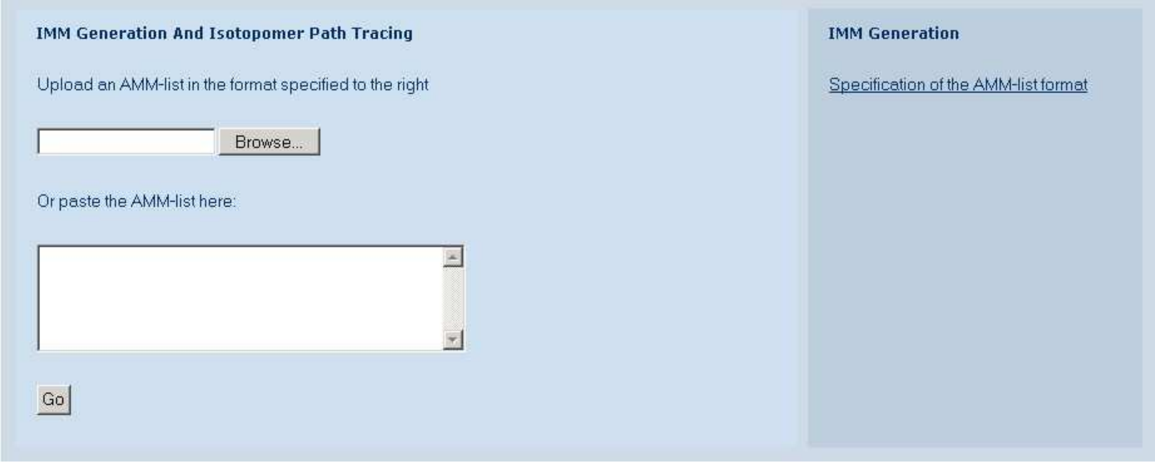
- iND750 model
- iFF708 model
- custom model

For being able to select your own model, you first have to generate it using Module 1. On the result page of the stoichiometry model, you find a button labeled “Go to Isotopomer Path Tracing”. After clicking on this button you will be redirected to the “Isotopomer Path Tracing module.”

- IMM Generation & Isotopomer Path Tracing

“IMM Generation and Isotopomer Path Tracing” is the second method offered by Module II. For more information about the principles underlying these conversions please refer to the paper by Forbes et al. [Forbes2001, section 4.4.1].

- Input



IMM Generation And Isotopomer Path Tracing

Upload an AMM-list in the format specified to the right

Or paste the AMM-list here:

IMM Generation

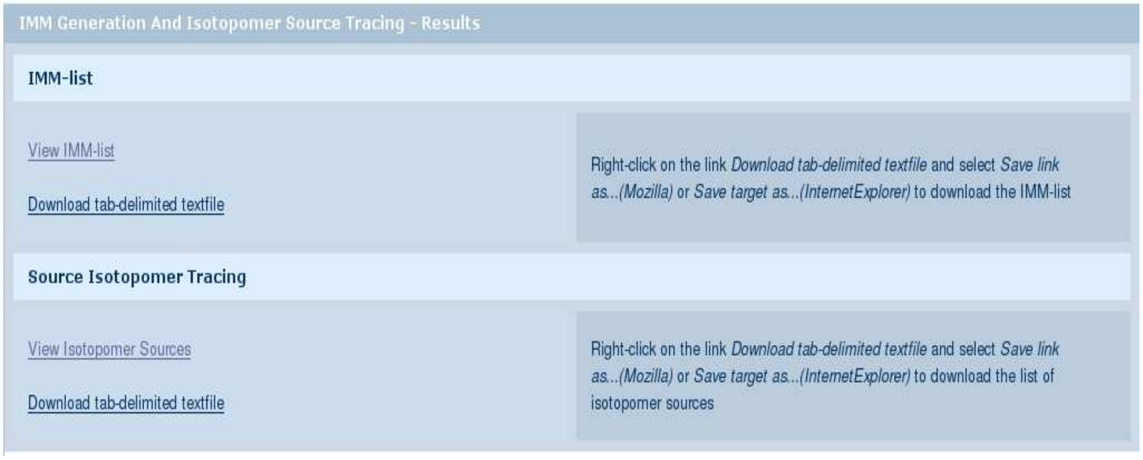
Specification of the AMM-list format

Figure 29 *User interface for the Isotopomer Path Tracing method*

The input consists of a list of Atom Mapping Matrices. The user can either upload a file containing this list or just paste the list in the textbox. For specification of the AMM format see either our specification page on the MetaModel online help, or section 4.4.2.

- Results

On the result page, the user can view the IMMs and Isotopomer Tracing results in html format and download these as tab-delimited text files.



IMM Generation And Isotopomer Source Tracing - Results

IMM-list

View IMM-list

Download tab-delimited textfile

Right-click on the link *Download tab-delimited textfile* and select *Save link as...(Mozilla)* or *Save target as...(InternetExplorer)* to download the IMM-list

Source Isotopomer Tracing

View Isotopomer Sources

Download tab-delimited textfile

Right-click on the link *Download tab-delimited textfile* and select *Save link as...(Mozilla)* or *Save target as...(InternetExplorer)* to download the list of isotopomer sources

Figure 30 *Result page of for the Isotopomer Path Tracing method*

5.2.3.2.Functionality

The user has the option of a) finding the number of nodes representing the four different node-types and all metabolites sorted by their node types, in a tabular format (*node discrimination*) and b) calculation of the source isotopomers from any traceable individual product isotopomer (*isotopomer path tracing*) [Figure 27-30]. The *node discrimination* function assigns all the model components under study into four types of nodes:

- **Merge**: Node having more than one incoming edge and at most one outgoing edge
- **Split**: Node having more than one outgoing edge and at most one incoming edge
- **Both**: Node having more than one incoming and more than one outgoing edge
- **None**: Node having maximum one incoming and one outgoing edge

The isotopomer path tracing function calculates a set of Isotopomer Mapping Matrices (IMM) out of given Atom Mapping Matrices (AMM) and based on these calculates a set of possible input isotopomers for a given product isotopomer in a given model [section 4.4.1]. The experimental isotopomer distribution is usually compared with the theoretical labeling patterns expected according to the pathways detected in any given genome. Our isotopomer path tracing helps in deciphering the flow of labeling pattern (^{13}C) between a set of metabolites.

5.2.3.3.Validation

In the present section, we validated whether what we could decipher in terms of isotopomer links between various atoms of amino acids in *Desulfovibrio vulgaris (D. vulgaris)*. *D. vulgaris* is a sulfate reducing bacterium. The *D. vulagris* genome is fully annotated and that is the precise reason why it has been used as a model organism for studying the sulfate reducing physiology and in various functional genomics studies [Chhabra2006, He2006, and Mukhopadhyay2006**]**. Isotopomer analysis is a powerful technique for understanding central metabolic pathways and fluxes under steady state in various organisms. The AMM set used in this study was obtained from Tang et.al. (Tang2006, Appendix 4). Table 21 enlists atom flow associations between a subset of metabolites, by finding the "Source Isotopomers" for the respective"Product Isotopomers", using Isotopomer tracing method in Module 2 of MetaModel. Also the user can download an individual Isotopomer Mapping Matrices (IMM) for any given AMM set (Appendix 4). Table 21 is a subset of the isotopomer mapping result obtained for the AMM under study.

We can clearly see that the similarity in isotopomer patterns in some amino acids is a result of an underlying association in terms of them being produced from the same precursor metabolites (shared biosynthetic pathway) [Weitzel2007, Jennings2008]. For further study, one could follow the biosynthetic pathways and validate the flow of labeled carbons from source isotopomers to respective isotopomer due to inherent biochemical associations. For example, we calculate that tyrosine and phenylalanine are derived from (i.e. source isotopomers) phosphoenolpyruvate and erythrose-4-phosphate. Furthermore, the actual IMMs and the source isotopomer results can be downloaded as simple text file which can be imported and used in further analysis using R or Matlab.

Product Isotopomer	Source Isotopomer(s)
PHE[000010000]	E4P[0 0 0 0] PEP[0 0 1]
PHE[000010001]	E4P[0 0 0 1] PEP[0 0 1]
PHE[000010010]	E4P[0 0 1 0] PEP[0 0 1]
PHE[000010011]	E4P[0 0 1 1] PEP[0 0 1]
PHE[000010100]	E4P[0 1 0 0] PEP[0 0 1]
PHE[000010101]	E4P[0 1 0 1] PEP[0 0 1]
PHE[000010110]	E4P[0 1 1 0] PEP[0 0 1]
PHE[000010111]	E4P[0 1 1 1] PEP[0 0 1]
PHE[000011000]	E4P[1 0 0 0] PEP[0 0 1]
PHE[000011001]	E4P[1 0 0 1] PEP[0 0 1]
PHE[000011010]	E4P[1 0 1 0] PEP[0 0 1]
PHE[000011011]	E4P[1 0 1 1] PEP[0 0 1]
PHE[000011100]	E4P[1 1 0 0] PEP[0 0 1]
PHE[000011101]	E4P[1 1 0 1] PEP[0 0 1]
PHE[000011110]	E4P[1 1 1 0] PEP[0 0 1]
TYR[000001000]	E4P[1 0 0 0] PEP[0 0 0]
TYR[000001001]	E4P[1 0 0 1] PEP[0 0 0]
TYR[000001010]	E4P[1 0 1 0] PEP[0 0 0]
TYR[000001011]	E4P[1 0 1 1] PEP[0 0 0]
TYR[000001100]	E4P[1 1 0 0] PEP[0 0 0]
TYR[000001101]	E4P[1 1 0 1] PEP[0 0 0]
TYR[000001110]	E4P[1 1 1 0] PEP[0 0 0]
TYR[000001111]	E4P[1 1 1 1] PEP[0 0 0]
TYR[000010000]	E4P[0 0 0 0] PEP[0 0 1]
TYR[000010001]	E4P[0 0 0 1] PEP[0 0 1]
TYR[000010010]	E4P[0 0 1 0] PEP[0 0 1]
TYR[000010011]	E4P[0 0 1 1] PEP[0 0 1]
TYR[000010100]	E4P[0 1 0 0] PEP[0 0 1]
TYR[000010101]	E4P[0 1 0 1] PEP[0 0 1]
TYR[000010110]	E4P[0 1 1 0] PEP[0 0 1]
TYR[000010111]	E4P[0 1 1 1] PEP[0 0 1]

Table 21 *Subset of source isotopomer tracing results*

5.2.4. Module III: *Optimization*

5.2.4.1. Design and implementation

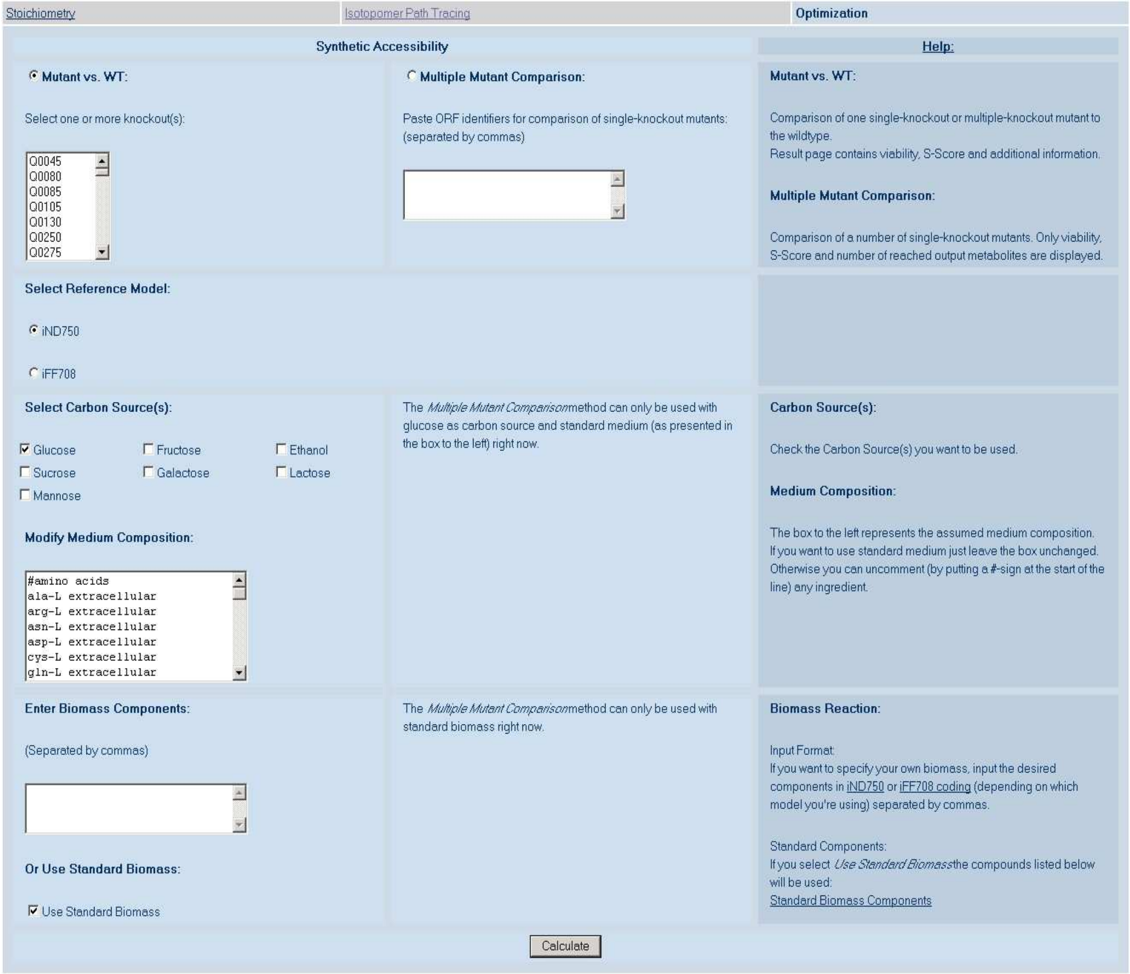
The third module is called “Optimization”. The optimization module currently encodes the notion of “Synthetic Accessibility”. The synthetic accessibility script has been motivated by the method introduced by Wunderlich & Mirny [Wunderlich2006]. In the present implementation we consider those substrates that are consumed in the given genome scale metabolic network, as input substrates. For example sugars, oxygen, nitrogen etc are regarded as part of the input substrates. The output substrates like amino acids, nucleotides and other components are regarded as the biomass. We calculate synthetic accessibility S_j of an output j as the minimal number of metabolite reactions needed to produce j from the network inputs. S_j equal to infinity, means that the metabolite j cannot be synthesized by the available input substrates. Total synthetic accessibility S is the simple summation of the synthetic accessibility of all components of the biomass [section 4.5.1]. For a more detailed description of how the method works and what are the underlying assumptions please refer to that paper. Here we will only state the required inputs and explain the results (prediction of viability/growth speed of the yeast strains) generated by our webserver. Our method can perform viability calculations in the following two settings:

- comparison of single mutant vs. wild type (multiple gene knockouts)
- comparison of multiple mutant vs. wild type (single gene knockouts)

Comparison of single mutant vs. wild type:

In this case, multiple mutants with given enzyme deficiencies are compared to the wild type, one at a time. To specify the mutant's deficiencies, the user should select the appropriate ORF identifiers given in the selection box. Furthermore, the user needs to select the medium composition on which the computation should be based (for standard medium composition see section 4.5.6.1). The user can perform the procedure on multiple carbon sources. The remaining medium ingredients are displayed in a textbox and the user can modify the medium by commenting them there. Next, the user can specify the biomass components (for the standard biomass composition, see section 4.5.6.2). The biomass represents the metabolites the organism should be able to produce, in order to be viable. The user can either input them manually or select a set of metabolites considered important for viability by the respective authors of the underlying metabolism models (by checking the box *'Use Standard Metabolites'*.

- Input



Stoichiometry | Isotopomer Path Tracing | Optimization

Synthetic Accessibility

● Mutant vs. WT:

Select one or more knockout(s):

Q0045
Q0080
Q0085
Q0105
Q0130
Q0250
Q0275

○ Multiple Mutant Comparison:

Paste ORF identifiers for comparison of single-knockout mutants: (separated by commas)

Select Reference Model:

● iND750

○ iFF708

Select Carbon Source(s):

☑ Glucose ☐ Fructose ☐ Ethanol
☐ Sucrose ☐ Galactose ☐ Lactose
☐ Mannose

Modify Medium Composition:

```
#amino acids
ala-L extracellular
arg-L extracellular
asn-L extracellular
asp-L extracellular
cys-L extracellular
gln-L extracellular
```

The *Multiple Mutant Comparison*method can only be used with glucose as carbon source and standard medium (as presented in the box to the left) right now.

Enter Biomass Components:

(Separated by commas)

Or Use Standard Biomass:

☑ Use Standard Biomass

The *Multiple Mutant Comparison*method can only be used with standard biomass right now.

Help:

Mutant vs. WT:

Comparison of one single-knockout or multiple-knockout mutant to the wildtype.
Result page contains viability, S-Score and additional information.

Multiple Mutant Comparison:

Comparison of a number of single-knockout mutants. Only viability, S-Score and number of reached output metabolites are displayed.

Carbon Source(s):

Check the Carbon Source(s) you want to be used.

Medium Composition:

The box to the left represents the assumed medium composition. If you want to use standard medium just leave the box unchanged. Otherwise you can uncomment (by putting a #-sign at the start of the line) any ingredient.

Biomass Reaction:

Input Format:
If you want to specify your own biomass, input the desired components in iND750 or iFF708 coding (depending on which model you're using) separated by commas.

Standard Components:
If you select *Use Standard Biomass*the compounds listed below will be used:
Standard Biomass Components

Figure 31 *User interface for the synthetic accessibility method*

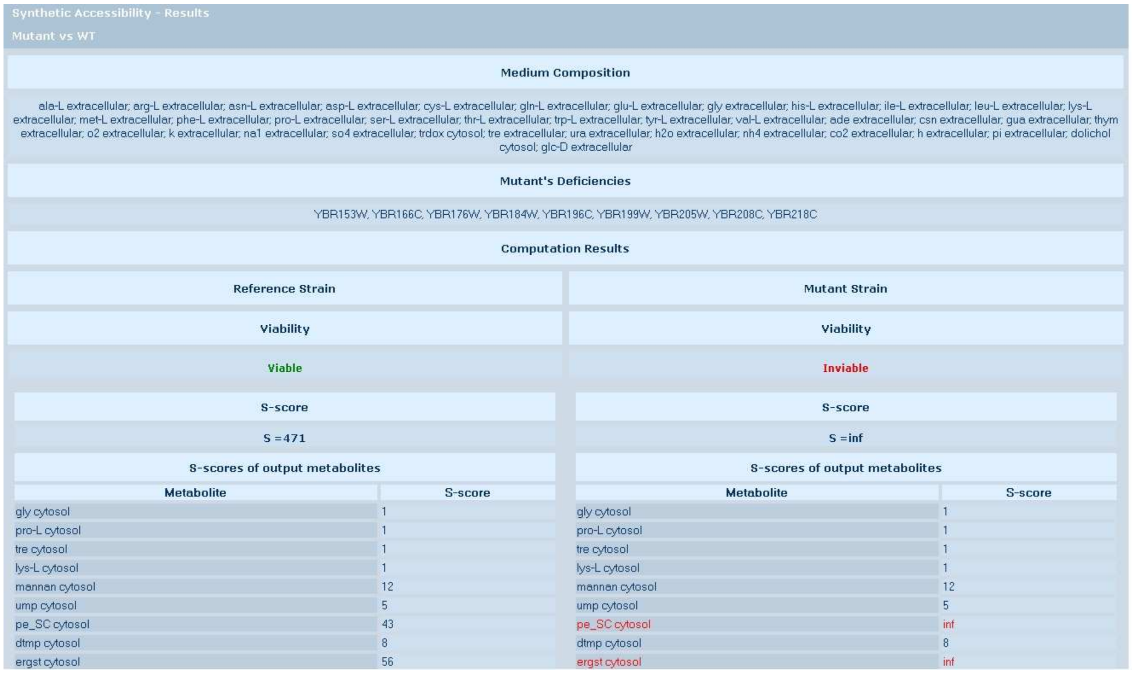
Comparison of multiple mutants vs. wild type: Using this version the user can generate a comparison table for multiple mutants (with single gene knockouts). The user can paste a list of ORF identifiers or gene names separated by commas in the textbox. This method assumes standard medium and biomass composition and is not affected by changes to the medium or the biomass via the respective input fields of the "*Single mutant Vs WT*" method.

Additionally for both of the methods, the user has to select the model to be used for calculations. The current models are iND750 and iFF708.

5.2.4.2. Functionality

The result page differs between the two alternatives of the “Synthetic Accessibility” method implementation:

- Single mutant vs. wild type comparison: First the invocation parameters, namely selected deficiencies and medium composition are stated.



Synthetic Accessibility - Results

Mutant vs WT

Medium Composition

ala-L extracellular; arg-L extracellular; asn-L extracellular; asp-L extracellular; cys-L extracellular; gln-L extracellular; glu-L extracellular; gly extracellular; his-L extracellular; ile-L extracellular; leu-L extracellular; lys-L extracellular; met-L extracellular; phe-L extracellular; pro-L extracellular; ser-L extracellular; thr-L extracellular; trp-L extracellular; tyr-L extracellular; val-L extracellular; ade extracellular; csn extracellular; gua extracellular; thym extracellular; o2 extracellular; k extracellular; na1 extracellular; so4 extracellular; trdox cytosol; tre extracellular; ura extracellular; h2o extracellular; nh4 extracellular; co2 extracellular; h extracellular; pi extracellular; dolichol cytosol; glc-D extracellular

Mutant's Deficiencies

YBR153W, YBR166C, YBR176W, YBR184W, YBR196C, YBR199W, YBR205W, YBR208C, YBR218C

Computation Results

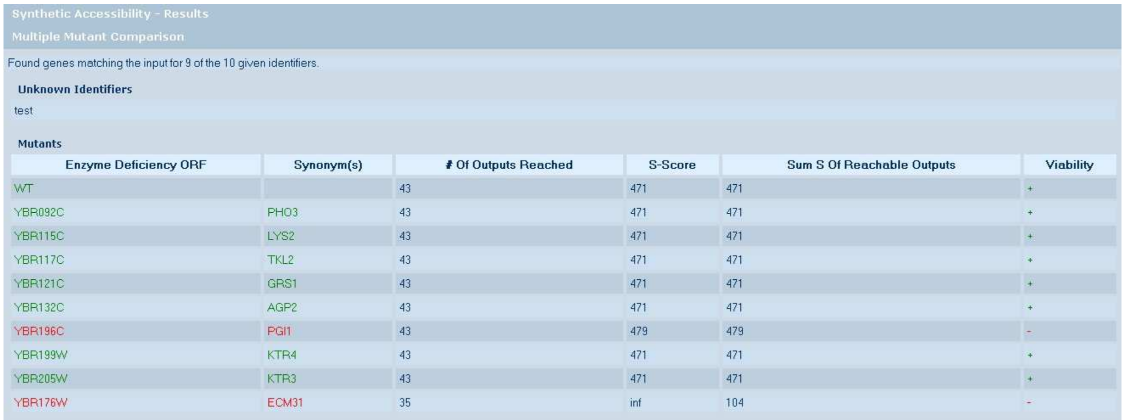
Reference Strain		Mutant Strain	
Viability		Viability	
Viable		Inviable	
S-score		S-score	
S =471		S =inf	
S-scores of output metabolites		S-scores of output metabolites	
Metabolite	S-score	Metabolite	S-score
gly cytosol	1	gly cytosol	1
pro-L cytosol	1	pro-L cytosol	1
tre cytosol	1	tre cytosol	1
lys-L cytosol	1	lys-L cytosol	1
mannan cytosol	12	mannan cytosol	12
ump cytosol	5	ump cytosol	5
pe_SC cytosol	43	pe_SC cytosol	inf
dtmp cytosol	8	dtmp cytosol	8
ergst cytosol	56	ergst cytosol	inf

Figure 32 *Typical result page for the single mutant vs. wild type method*

Next, the total synthetic accessibility score, S, of the wildtype and the mutant as well as the S scores of the different output metabolites (biomass components) are calculated.

- Multiple mutants vs. wild type comparison:

Additionally, if the calculation was based on the iND750 model a tabular representation of the usage percentages of the different pathways of the wildtype and the mutant is given. The output in this case consists of a list of erroneous ORF names, if any, followed by a comparative table for the different mutants, with respect to viability and overall S scores.



Synthetic Accessibility - Results

Multiple Mutant Comparison

Found genes matching the input for 9 of the 10 given identifiers.

Unknown Identifiers

test

Mutants

Enzyme Deficiency ORF	Synonym(s)	# Of Outputs Reached	S-Score	Sum S Of Reachable Outputs	Viability
WT		43	471	471	+
YBR092C	PHO3	43	471	471	+
YBR115C	LYS2	43	471	471	+
YBR117C	TKL2	43	471	471	+
YBR121C	GRS1	43	471	471	+
YBR132C	AGP2	43	471	471	+
YBR196C	PGI1	43	479	479	-
YBR199W	KTR4	43	471	471	+
YBR205W	KTR3	43	471	471	+
YBR176W	ECM31	35	inf	104	-

Figure 33 *Typical result page for multiple mutants vs. wild type comparison method*

5.2.4.3.Validation

In this work, we compared the performance of our optimization module with the literature dataset (Duarte2004, Giaever2002). This literature dataset comprises of 562 single gene knockouts of yeast. For the wild type strain, the total synthetic accessibility score, S, was equal to 471 and the number of reachable outputs was equal to 43[Appendix 5] The iND750 model by Palsson et al. has been earlier studied and validated by the use of Flux Balance Analysis (Giaever2002, Steinmetz2002) technique. In all these studies, a wide variety of growth conditions are considered.

Experimental	# ORF
Viable	486
Non Viable	76
MetaModel	
Viable	535
Non Viable	27
True Positive(TP)	462
True Negative(TN)	3

Table 22 *Comparison of MetaModel result and the FBA result from Palsson2004.* Total number of ORF used= 562. Growth condition used for calculation is YPD (glucose).

In our MetaModel server, we can replicate these conditions namely growth under defined carbon source like glucose(YPD), galactose(YPGal), glycerol(YPG), ethanol(YPE) and lactate(YPL) etc. For comparison, we run the “optimization” module on YPD data and compared with the literature (“experimental”) results from Duarte2004 and Giaever2002. The growth condition is YPD using glucose as a carbon source. For the ORF list and the detailed result, refer appendix 5.We calculated the sensitivity as $((TP \times 100)/(TP + FP))$ (%) and specificity as $((TN \times 100)/(TN + FN))$ (%) values using the result in Table 22. The sensitivity equals 86.5 % and specificity equals to 11.1% for the dataset understudy. The high overall sensitivity and low specificity can be attributed to the underlying assumption that since most genes are non essential, in our method we assign all those metabolic genes which are not included in the metabolic model, as viable (Duarte2004). Hence, the synthetic accessibility method works better in prediction of viable mutants than non-viable mutants. The performace of method is limited by the incomplete information on the underlying metabolic model. For further discussion on the performance of the method, refer the work by Mirny etal. [Mirny2006].

Conclusions

We have presented various computational techniques for metabolic network analysis based on metabolomics data. We have approached a niche area of development of web server for metabolic network analysis for genome scale metabolic models. Following are the problems that we attempted to solve in the present work:

Automation of GC-MS spectrometric data analysis: The developed software tool, CalSpec, is useful for efficient processing of ^{13}C labeling data from MS measurements in ^{13}C flux analysis. These MS data sets are generated in huge numbers due to (i) replicate measurements of one sample to assess the confidence in the measured values and estimation of error; (ii) replicate measurements of one experiment to check for isotopic steady-state; or (iii) different measurements of one sample with different protocols to obtain additional labeling information via alternative fragments. Our software tool, CalSpec, can process these large MS datasets in fast (running time few seconds) and automated fashion.

Mutant classification based on metabolomics data: We also explored various statistical techniques for the analysis of metabolic profiles and for understanding the effects of gene knockouts on the metabolic network functionality of wild type yeast. The methods is coded in R-language and uses the in-house developed methods for automation of GC-MS spectra analysis, quantification of summed fractional labeling of proteogenic amino-acid fragments, estimation of the extent of mutant association based on the global features growth rate μ ,biomass yield Y_{xs}, ethanol yield Y_p rate of biomass production Q_s and rate of ethanol production Q_p, followed by integration of transcript co-response profiles for mutant differentiation. In this framework, we have introduced a scheme for estimation of cluster quality in analysis of metabolic profiling data.

Method for mutant differentiation in the case of high similarity in metabolomics data: In the general case of when the genotypic perturbations (knockouts) are not sufficient for discrimination of mutant knockout metabolic profiles, we were able to find highly significant feature combinations for each individual mutant present in the original dataset. This method is coded in R –language and is a useful method for fast characterization for the metabolic profiling datasets for large scale knockout analysis. We show that in the absence of strong phenotypic perturbations , for example in our case where the metabolic profiles prove not be sufficient in finding any underlying functional associations among majority of the mutant set, outlier detection method can be used for a more granular analysis of each individual knockout mutant.

Web server for pathway analysis using genome scale metabolic models: In the recent years of computational systems biology research, a large number of theoretical methods have been developed for studying chemical transformations of substances. Also, in the last 5 years, several genome scale metabolic models have been reconstructed. We have approached the problem of analysis of various experimental data in the background of these genome scale models as well as prediction of viability of *in-silico* gene knockout mutants. In the same direction, we developed a new web server called MetaModel, for the analysis of genome-scale metabolic networks of eukaryotic organisms.

Appendix 1

Table 23 *List of mutants studied by Outlier detection method in section 3.3.1.3*

Source of Gene ontology data: Amigo 29 May2007. (http://amigo.geneontology.org)

Mutant_ORF	Glu	Fru	Gal	Molecular function	Biological process	Cellular component
ACE2			√	transcriptional activator activity	G1-specific transcription in mitotic cell cycle	Cytosol, nucleus
ADR1			√	trnscription factor activity	transcription, regulation of carbohydrate metabolic process , peroxisome organization and biogenesis, negative regulation of transcription from RNA polymerase II promoter by glucose	Nucleus
CAT8	√		√	specific RNA polymerase II transcription factor activity	positive regulation of gluconeogenesis , positive regulation of transcription from RNA polymerase II promoter	Nucleus
CYB2	√		√	L-lactate dehydrogenase (cytochrome) activity	electron transport	mitochondrial intermembrane space, mitochondrion
DLD2	√	√	√	lactate metabolic process, actin binding	lactate metabolic process	mitochondrial matrix, mitochondrion
FBP1	√	√	√	fructose-bisphosphatase activity	gluconeogenesis	Cytosol
FBP26	√		√	fructose-2,6-bisphosphate 2-phosphatase activity, 6-phosphofructo-2-kinase activity	glucose metabolic process	Cytosol
GAD1	√		√	glutamate decarboxylase activity	response to oxidative stress, glutamate catabolic process	cytoplasm
GAL10	√	√		UDP-glucose 4-epimerase activity, aldose 1-epimerase activity	galactose catabolic process	soluble fraction
GAL4	√			transcriptional activator & factor activity	positive regulation of transcription by galactose, galactose metabolic process	nucleus

GAL7	√	√		UTP:galactose-1-phosphate uridylyltransferase activity	galactose catabolic process	cytoplasm
GAL80	√			specific transcriptional repressor activity	galactose metabolic process, positive regulation of transcription by galactose	nucleus, cytoplasm
GLK1	√		√	glucokinase activity	glucose import, glucose metabolic process, glycolysis, mannose metabolic process	cytosol
GLO1	√			lactoylglutathione lyase activity	methylglyoxal catabolic process to D-lactate, glutathione metabolic process	nucleus, cytoplasm
HXK2			√	hexokinase activity	fructose import, fructose metabolic process, glucose import, glucose metabolic process, glycolysis, mannose metabolic process, regulation of cell size, regulation of transcription by glucose, replicative cell aging	cytosol, mitochondrion, nucleus
IMP2	√	√	√	peptidase activity, mitochondrial inner membrane peptidase activity, transcription coactivator activity	carbohydrate metabolic process, mitochondrial protein processing, DNA repair	mitochondrial inner membrane peptidase complex, cytoplasm
LAT1	√	√		dihydrolipoyllysine-residue acetyltransferase activity	pyruvate metabolic process	mitochondrion, mitochondrial pyruvate dehydrogenase complex
LEU4	√	√	√	2-isopropylmalate synthase activity,	leucine biosynthetic process	mitochondrion, cytoplasm
MAE1	√		√	malic enzyme activity	pyruvate metabolic process, amino acid metabolic process	mitochondrion
MAL33	√	√	√	transcription factor activity	carbohydrate metabolic process , regulation of transcription, DNA-dependent	nucleus
MIG1	√			specific transcriptional repressor activity, sequence-specific DNA binding	negative regulation of transcription from RNA polymerase II promoter by glucose	nucleus, nuclear envelope lumen
MIG2	√			specific transcriptional repressor activity, sequence-specific DNA	negative regulation of transcription from RNA polymerase II promoter by	nucleus

				binding	glucose	
MSN4	√		√	transcription factor activity	cellular response to glucose starvation, heat acclimation, regulation of transcription from RNA polymerase II promoter in response to stress, replicative cell aging, response to freezing, hydrostatic pressure, response to osmotic stress, oxidative stress and stress	nucleus, cytoplasm
NRG1	√	√		transcriptional repressor activity, DNA binding	response to pH, regulation of transcription from RNA polymerase II promoter, pseudohyphal growth, invasive growth (sensu Saccharomyces), glucose metabolic process, biofilm formation	nucleus
NRG2	√	√		transcriptional repressor activity	pseudohyphal growth, invasive growth (sensu Saccharomyces), biofilm formation,	nucleus
PCK1	√	√	√	phosphoenolpyruvate carboxykinase (ATP) activity	gluconeogenesis	cytosol
PFK26	√	√	√	6-phosphofructo-2-kinase activity	regulation of glycolysis, fructose 2,6-bisphosphate metabolic process	cytoplasm
PFK27	√		√	6-phosphofructo-2-kinase activity	regulation of glycolysis, fructose 2,6-bisphosphate metabolic process	cytoplasm
PGU1	√	√		polygalacturonase activity	pseudohyphal growth, pectin catabolic process	extracellular region
RBK1	√	√		ribokinase activity, ATP binding,	D-ribose metabolic process	nucleus, cytoplasm
RGT1	√			transcriptional repressor activity, transcriptional activator activity, transcription corepressor activity, RNA polymerase II transcription factor activity, DNA binding	regulation of glucose import, negative regulation of transcription, glucose metabolic process,	nucleus
SFA1	√	√	√	formaldehyde dehydrogenase (glutathione) activity,	formaldehyde catabolic process	mitochondrion, cytoplasm

				alcohol dehydrogenase activity		
SIP3	√		√	transcription cofactor activity	transcription initiation from RNA polymerase II promoter	nucleus
SNF11	√		√	general RNA polymerase II transcription factor activity,	chromatin remodeling	SWI/SNF complex, chromatin remodeling complex
SNF2	√	√		general RNA polymerase II transcription factor activity, DNA-dependent ATPase activity,	double-strand break repair, chromatin remodeling	SWI/SNF complex, chromatin remodeling complex
SUC2	√	√		beta-fructofuranosidase activity,	sucrose catabolic process,	mitochondrion
TYE7	√		√	transcription factor activity	transcription, positive regulation of glycolysis, G1/S-specific transcription in mitotic cell cycle	nucleus
UGA1			√	4-aminobutyrate transaminase activity	nitrogen utilization	intracellular
UGA2	√	√	√	succinate-semialdehyde dehydrogenase [NAD(P)+] activity	response to oxidative stress, glutamate decarboxylation to succinate, gamma-aminobutyric acid catabolic process	cytoplasm
XKS1	√			xylulokinase activity	xylulose catabolic process	cytoplasm
YBR184W	√		√	Unknown	Unknown	Unknown
YDR248C		√		Unknown	Unknown	Unknown

Appendix 2

MetaModel web server scripts

a) Module 1: Stoichiometry

Python (CGI)-scripts

- decision.cgi:
  - Calls: sce.php or stoi.cgi
  - Is called by: sce.php
  - Function:
    decision.cgi is used to handle the processing of parameters such as file format, output format and wether the user has given any reaction or metabolite files. It either calls sce.php, which gives the user the possibility to change the standard model or calls stoi.cgi to calculate a new model using the user-supplied reaction file.
- kegg_annot.cgi:
  - Calls: stoi.php
  - Is called by: stoi.php
  - Function:
    kegg_annot.cgi uses the KOBAS script blast2ko.py to create a KO based annotation of a user-selected set of ORFs. stoi.php is used for displaying ORF list from which the user can select the ORFs to be annotated and for displaying the KOBAS output as well.
  - Annotation: Isn't fully implemented yet, because of problems with the KOBAS-package.
- search1.cgi:
  - Calls: stoi.php
  - Is called by: stoi.php
  - Function:
    search1.cgi implements the search for reactions in which a user-selected metabolite takes part. As in the case of kegg_annot.cgi, stoi.php gives the user the possibility to select an input for the search and afterwards displays the results.
- search2.cgi:
  - Calls: stoi.php
  - Is called by: stoi.php
  - Function:
    search1.cgi implements the search for processes in which a user-selected metabolite takes part. As in the case of kegg_annot.cgi, stoi.php gives the user the possibility to select an input for the search and afterwards displays the results.
- stoi.cgi:

- Calls: stoi.php
- Is called by: decision.cgi or sce.php
- Function:
  stoi.cgi represents the actual calculation of the stoichiometric matrix. The parameters given either by sce.php or decision.php are used as the input for the calculation. After the model and several statistics have been calculated, the results are displayed by the script stoi.php.

b) Module 2: Isotopomer path tracing

- ammparse.cgi:
  - Calls: mutantmap.php
  - Is called by: sce.php
  - Function:ammparse.cgi parses the AMMs supplied either by uploading a file or pasted in the textbox, handles the AMM to IMM conversion, uses the created IMMs for Isotopomer Source Tracing and creates the output, which will be shown on the page mutantmap.php.
- nodediscrimination.cgi:
  - Calls:
  - Is called by: sce.php
  - Function: nodediscrimination.cgi calculates the node types for the metabolites supplied in the selected model. It displays the results in tabular format.

c) Module 3: Optimization

- syn acc.cgi:
  - Calls: mopt.php
  - Is called by: sce.php
  - Function: syn_acc.cgi calculates the synthetic accessbility scores of mutants and either displays them directly (in case of invocation of ’Mutant vs WT’ functionality), or formats them for output in mopt.php (if the ’Mutant Comparison’ functionality is used).

Appendix 3

Mutant name	mue	Qs	Qp	QO2	Mutant name	mue	Qs	Qp	QO2
CYB2_glc	0.279	15.6	24.4	1.3	FBP26_fru	0.348	21.002	9.778	0.73
GAL80_glc	0.307	19.8	19.6	1.3	ADR1_fru	0.412	26.639	31.017	1.98
GLO1_glc	0.342	21.3	26.7	2.2	MIG2_fru	0.354	22.524	20.949	1.4
GAD1_glc	0.312	20.8	26.1	1.4	HXK2_fru	0.273	12.842	16.739	0.92
CAT8_glc	0.35	23.4	27.9	1.8	UGA1_fru	0.332	19.089	25.604	2.47
SIP3_glc	0.329	19.2	30.1	1.6	YDR248C_fru	0.364	22.972	19.654	2.53
TYE7_glc	0.333	21.8	30.4	1.3	PCK1_fru	0.412	26.499	19.856	1.98
GAL4_glc	0.28	16.9	27.1	2.2	PGU1_fru	0.379	24.752	25.996	1.97
YBR184W_glc	0.292	16	27.9	1	NRG1_fru	0.403	26.853	22.487	1.75
SNF11_glc	0.4	22.9	39.2	2.5	MAL33_fru	0.332	17.906	17.49	1.18
XKS1_glc	0.396	22.6	33.6	1.8	RBK1_fru	0.4	28.782	31.606	1.6
GLK1_glc	0.32	21.3	23	1.2	SFA1_fru	0.394	24.535	24.1	1.64
MAE1_glc	0.282	16	27.9	1	DLD2_fru	0.341	23.02	13.341	1.21
RGT1_glc	0.294	23.3	29.5	1.3	UGA2_fru	0.377	22.747	21.087	1.76
MSN4_glc	0.294	16	23.9	0.9	GAL7_fru	0.326	21.54	17.252	0.78
FBP26_glc	0.288	15.8	26.2	1.1	GAL10_fru	0.317	18.275	22.355	1.08
MIG2_glc	0.288	13.5	22.2	0.9	NRG2_fru	0.401	25.382	20.771	0.98
MIG1_glc	0.289	21	32	0.6	PFK26_fru	0.392	24.208	26.092	1.7
PFK27_glc	0.286	19.3	27	0.9	SUC2_fru	0.371	20.016	18.986	0.31
PCK1_glc	0.338	22.9	29.9	2.1	LAT1_fru	0.381	27.467	21.51	0.6
PGU1_glc	0.311	20.4	30.7	1.1	LEU4_fru	0.378	20.132	9.045	1.58
NRG1_glc	0.232	13.1	15.2	0.8	FBP1_fru	0.354	20.852	26.611	0.7
MAL33_glc	0.306	20.6	34.1	1.2	IMP2_fru	0.385	24.558	40.216	0.71
RBK1_glc	0.341	24.9	36.5	1.5	SNF2_fru	0.287	16.8	19.93	0.71
SFA1_glc	0.351	21.6	28.1	3.3	Reference_fru	0.236	17.977	22.749	0.31
DLD2_glc	0.303	18.1	30.5	1	CYB2_gal	0.307	14.341	11.711	5.61038
UGA2_glc	0.24	14	17	0.6	GLO1_gal	0.316	13.387	12.434	4.166667
GAL7_glc	0.306	20	19.6	1.1	PFK2_gal	0.27	7.413	9.806	3.835227
GAL10_glc	0.338	22.1	35.7	2.9	GAD1_gal	0.316	10.606	16.682	5.642857
NRG2_glc	0.237	15.1	17.5	0.7	CAT8_gal	0.276	10.368	15.432	0.345
PFK26_glc	0.243	15.8	22.2	0.8	SIP3_gal	0.272	11.313	14.527	4.473684
SUC2_glc	0.287	23.7	22.1	3.2	TYE7_gal	0.216	8.849	9.898	3.970588
LAT1_glc	0.293	16.6	28	1.3	YBR184W_gal	0.259	12.075	12.985	2.475153
LEU4_glc	0.314	15.8	22.3	1.3	SNF11_gal	0.207	7.469	11.879	0.294034
FBP1_glc	0.225	12.4	20.9	0.7	XKS1_gal	0.297	19.623	17.299	5.557635
IMP2_glc	0.296	16.4	19.6	0.9	GLK1_gal	0.243	11.033	10.171	4.963235
SNF2_glc	0.236	16.2	20.1	1.8	ACE2_gal	0.225	19.288	17.69	4.210329
Reference_glc	0.32	13.9	18.2	2.3	MAE1_gal	0.26	7.98	11.461	5.762411
CYB2_fru	0.401	22.22	19.653	2.49	MSN4_gal	0.261	11.755	16.381	4.338431
GAL80_fru	0.373	19.523	16.576	2.49	FBP26_gal	0.236	9.914	9.742	3.782051
PFK2_fru	0.331	16.05	19.913	1.34	ADR1_gal	0.17	4.312	4.506	3.964552
GAD1_fru	0.317	18.517	20.548	1.46	MIG2_gal	0.34	13.462	19.267	2.361111
CAT8_fru	0.406	24.973	23.52	2.39	HXK2_gal	0.212	6.469	9.65	4.416667
SIP3_fru	0.4	23.628	38.489	2.98	UGA1_gal	0.277	13.755	12.522	3.170788
TYE7_fru	0.293	16.219	13.932	2.13	PFK27_gal	0.268	10.3	6.366	1.495536
GAL4_fru	0.369	22.935	18.489	1.28	PCK1_gal	0.181	6.376	3.261	3.269509
YBR184W_fru	0.436	24.465	10.041	2.48	MAL33_gal	0.196	5.938	3.7	3.223684
SNF11_fru	0.273	13.465	11.529	2.08	SFA1_gal	0.253	9.58	15.344	6.588542
XKS1_fru	0.336	18.281	16.383	2.28	DLD2_gal	0.199	5.835	4.259	4.783654
GLK1_fru	0.297	18.364	15.186	0.93	UGA2_gal	0.221	7.885	4.409	5.525
ACE2_fru	0.244	14.018	9.677	1.09	PFK26_gal	0.234	5.53	6.757	5.668605
MAE1_fru	0.418	26.156	10.853	3.53	LEU4_gal	0.22	5.108	5.874	5.926724
RGT1_fru	0.416	25.151	40.493	1.65	FBP1_gal	0.335	8.684	6.306	9.264381
MSN4_fru	0.412	23.461	32.52	1.84	IMP2_gal	0.293	9.931	8.201	8.102876
					Reference_gal	0.223	6.504	8.262	5.402132

Appendix 4

AMM list for used in section 5.2.3.3.

```
# Pyr -> PEP;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#PEP -> PGA;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#PGA -> T3P;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#CoA -> Acetate;Ns=2;Np=2
1 0
0 1

#C1 -> GLY;Ns=1;Np=2
0
1

#ICT -> CO2;Ns= 6;Np=1
0 0 0 0 0 1

#ACoA -> MAL;Ns= 2;Np=4
0 0
0 0
0 1
1 0

#OXO -> CO2;Ns= 5;Np=1
1 0 0 0 0

#OXO -> GLU;Ns=5;Np=5
1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1

#OXO -> SUCC;Ns=5;Np=4
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1

#ASP -> MET;Ns= 4;Np= 5
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1
0 0 0 0

#PYR -> MAL;Ns=3;Np=4
1 0 0
0 1 0
0 0 1
0 0 1
```

```
#ACoA -> LEU;Ns=2;Np=6
1 0
0 1
0 0
0 0
0 0
0 0

#PYR -> LEU;Ns= 3;Np=6
0 0 0
0 0 0
0 1 0
0 1 0
0 0 1
0 0 1

#C1 -> MET;Ns=1;Np=5
0
0
0
0
1

#C1 -> HIS;Ns=1;Np= 6
0
0
0
0
0
1

#E4P -> TYR;Ns= 4;Np=9
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#E4P -> F6P;Ns=4;Np=6
0 0 0 0
0 0 0 0
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#E4P -> PHE;Ns=4;Np=9
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1
```

```
#CO2 -> OAA;Ns= 1;Np=4
0
0
0
1

# F6P -> E4P;Ns=6;Np=4
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1

#F6P -> G6P;Ns=6;Np=6
1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1

#F6P -> S7P;Ns=6;Np=7
1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0

#T3P -> F6P;Ns=3;Np=6
0 0 1
0 1 0
1 0 0
1 0 0
0 1 0
0 0 1

#F6P -> C5P;Ns=6;Np=5
1 0 0 0 0 0
0 1 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0

#ICT -> OXO;Ns=6;Np=5
1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0

#MAL -> OAA;Ns=4;Np=4
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#SER -> C1;Ns=3;Np=1
0 0 1
```

```
#F6P -> G6P;Ns=6;Np=6
1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1

#G6P -> C5P;Ns= 6;Np=5
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1

#GLX -> MAL;Ns=2;Np=4
1 0
0 1
0 0
0 0

#GLY -> C1;Ns=2;Np=1
0 1

#GLY -> CO2;Ns=2;Np=1
1 0

#GLY -> SER;Ns=2;Np=3
1 0
0 1
0 0

#PYR -> CO2;Ns=3;Np=1
1 0 0

#PYR -> CoA;Ns=3;Np=2
0 1 0
0 0 1

#C1 -> SER;Ns=1;Np=3
0
0
1

#CO2 -> MAL;Ns=1;Np=4
0
0
0
1

#PYR -> OAA;Ns=3;Np=4
1 0 0
0 1 0
0 0 1
0 0 0

#ICT -> GLX;Ns=6;Np=2
1 0 0 0 0 0
0 1 0 0 0 0

#ICT -> SUC;Ns=6;Np=4
0 0 0 0 1 0
0 0 0 1 0 0
0 0 1 0 0 0
0 0 0 0 0 1

#MAL_OAA -> CO2;Ns=4;Np=1
0 0 0 1

#AcoA_DvH -> ICT;Ns=2;Np=6
1 0
0 1
0 0
0 0
0 0
0 0

#ACoA_Ecoli -> ICT;Ns=2;Np=6
0 0
0 0
0 0
0 1
1 0
0 0

#OAA_DvH -> ICT;Ns=4;Np= 6
0 0 0 0
0 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1
1 0 0 0

#OAA -> ICT_Ecoli;Ns=4;Np=6
0 0 0 1
0 0 1 0
0 1 0 0
0 0 0 0
0 0 0 0
1 0 0 0

#PYR -> ALA;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#OAA -> ASP;Ns=4;Np=4
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#PYR -> VAL;Ns=3;Np=5
1 0 0
0 1 0
0 1 0
0 0 1
0 0 1

#PYR -> LYS;Ns=3;Np=6
0 0 0
0 0 0
0 0 0
0 0 0
0 1 0
0 0 1

#ACoA -> C1;Ns=2;Np=1
 0 1


#PYR -> LEU;Ns=3;Np=6
0 0 0
0 0 0
0 1 0
0 1 0
0 0 1
0 0 1

#OAA -> CO2;Ns=4;Np=1
0 0 0 1

#PEP -> OAA;Ns=3;Np=4
1 0 0
0 1 0
0 0 1
0 0 0

#OAA -> PEP;Ns=4;Np=3
1 0 0 0
0 1 0 0
0 0 1 0

#PEP -> PYR;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#C5P -> F6P;Ns=5;Np=6
1 0 0 0 0
0 1 0 0 0
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0

#C5P -> S7P;Ns=5;Np=7
1 0 0 0 0
0 1 0 0 0
1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1

#G6P -> CO2;Ns=6;Np=1
1 0 0 0 0 0

#PEP -> TYR;Ns=3;Np=9
1 0 0
0 1 0
0 0 1
0 1 0
0 0 1
0 0 0
0 0 0
0 0 0
0 0 0

#PEP -> PHE;Ns=3;Np=9
1 0 0
0 1 0
0 0 1
0 1 0
0 0 1
0 0 0
0 0 0
0 0 0
0 0 0

#T3P -> C5P;Ns=3;Np=5
0 0 0
0 0 0
1 0 0
0 1 0
0 0 1
```

```
#C5P -> HIS;Ns=5;Np=6
1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1
0 0 0 0 0

#C5P -> S7P;Ns=5;Np=7
1 0 0 0 0
0 1 0 0 0
1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1

#C5P -> T3P;Ns=5;Np=3
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1

#PGA -> SER;Ns=3;Np=3
1 0 0
0 1 0
0 0 1

#ACoA -> CO2;Ns=2;Np=1
 1 0
```

```
#SUC -> MAL;Ns=4;Np=4
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#T3P -> E4P;Ns=3;Np=4
0 0 0
1 0 0
0 1 0
0 0 1

#S7P -> E4P;Ns=7;Np=4
0 0 0 1 0 0 0
0 0 0 0 1 0 0
0 0 0 0 0 1 0
0 0 0 0 0 0 1


#S7P -> C5P;Ns=7;Np=5
1 0 1 0 0 0 0
0 1 0 1 0 0 0
0 0 0 0 1 0 0
0 0 0 0 0 1 0
0 0 0 0 0 0 1
```

```
#S7P -> F6P;Ns=7;Np=6
1 0 0 0 0 0 0
0 1 0 0 0 0 0
0 0 1 0 0 0 0
0 0 0 0 0 0 0
0 0 0 0 0 0 0
0 0 0 0 0 0 0

#SER -> GLY;Ns=3;Np=2
1 0 0
0 1 0

#E4P -> S7P;Ns=4;Np=7
0 0 0 0
0 0 0 0
0 0 0 0
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1

#E4P -> T3P;Ns=4;Np=3
0 1 0 0
0 0 1 0
0 0 0 1
```

Appendix 5

Synthetic accessibility results from section 5.2.4.3.

MetaModel applied to Palsson2004		**MetaModel Prediction**	**Experimental (+ (Viable); - (NonViable)**	**MetaModel applied to Palsson2004**		**MetaModel Prediction**	**Experimental (+ (Viable); - (NonViable)**
ORF	*Gene product*			ORF	*Gene product*		
WT(Number of reachable output=43, S score=471)				WT(Number of reachable output=43, S score=471)			
YMR056C	AAC1	+	+	YBR205W	KTR3	+	+
YNL141W	AAH1	+	+	YBR199W	KTR4	+	+
YKL106W	AAT1	+	+	YPL053C	KTR6	+	+
YLR027C	AAT2	+	+	YNL071W	LAT1	+	+
YNR033W	ABZ1	-	+	YJL134W	LCB3	+	+
YGR037C	ACB1	-	+	YOR171C	LCB4	+	+
YBL015W	ACH1	+	+	YLR260W	LCB5	+	-
YLR304C	ACO1	+	-	YGL009C	LEU1	+	+
YAL054C	ACS1	+	+	YNL104C	LEU4	+	+
YAR015W	ADE1	+	+	YOR108W	LEU9	+	+
YLR028C	ADE16	+	+	YFL018C	LPD1	+	-
YMR120C	ADE17	+	+	YDR503C	LPP1	+	+
YGR204W	ADE3	+	+	YOR142W	LSC1	+	+
YMR300C	ADE4	+	+	YGR244C	LSC2	+	+
YGL234W	ADE57	+	+	YNL268W	LYP1	+	+
YGR061C	ADE6	+	+	YIR034C	LYS1	+	+
YDR408C	ADE8	+	+	YIL094C	LYS12	+	+
YOL086C	ADH1	+	-	YDL182W	LYS20	+	+
YMR303C	ADH2	+	+	YDL131W	LYS21	+	+
YMR083W	ADH3	+	+	YDR234W	LYS4	+	+
YGL256W	ADH4	+	+	YGL154C	LYS5	+	+
YBR145W	ADH5	+	+	YNR050C	LYS9	+	+
YDR226W	ADK1	+	-	YKL029C	MAE1	+	+
YER170W	ADK2	+	+	YGR289C	MAL11	+	+
YJR105W	ADO1	+	-	YGR292W	MAL12	+	+
YCL025C	AGP1	+	+	YBR298C	MAL31	+	+
YFL055W	AGP3	+	+	YBR299W	MAL32	+	+
YMR170C	ALD2	+	+	YOR221C	MCT1	+	-
YMR169C	ALD3	+	+	YKL085W	MDH1	+	+
YOR374W	ALD4	+	+	YOL126C	MDH2	+	+
YER073W	ALD5	+	+	YDL078C	MDH3	+	+
YPL061W	ALD6	-	-	YGR121C	MEP1	+	+
YNL270C	ALP1	+	+	YNL142W	MEP2	+	+
YML035C	AMD1	+	+	YPR138C	MEP3	+	+
YDR242W	AMD2	+	+	YKR069W	MET1	+	+
YPR128C	ANT1	+	+	YFR030W	MET10	+	+
YCL050C	APA1	+	+	YPL023C	MET12	+	+
YDR530C	APA2	+	+	YGL125W	MET13	+	+
YML022W	APT1	+	+	YKL001C	MET14	+	+
YDR441C	APT2	+	+	YPR167C	MET16	+	+

YBR149W	ARA1	+	+
YOL058W	ARG1	+	+
YJL071W	ARG2	+	+
YJL088W	ARG3	+	+
YHR018C	ARG4	+	+
YER069W	ARG5,6	+	+
YDR127W	ARO1	-	+
YDR380W	ARO10	+	+
YGL148W	ARO2	-	+
YDR035W	ARO3	+	+
YBR249C	ARO4	+	+
YPR060C	ARO7	+	+
YGL202W	ARO8	+	+
YHR137W	ARO9	+	+
YDL100C	GET3, ARR4	+	+
YPR145W	ASN1	+	+
YGR124W	ASN2	+	+
YDR321W	ASP1	+	+
YPR026W	ATH1	+	+
YBL099W	ATP1	+	+
YLR295C	ATP14	+	-
YPL271W	ATP15	+	-
YDR377W	ATP17	+	-
YML081C-A	ATP18	+	+
YJR121W	ATP2	+	-
YPR020W	ATP20	+	+
YPL078C	ATP4	+	-
YDR298C	ATP5	+	-
YKL016C	ATP7	+	-
YOR011W	AUS1	+	+
YBR068C	BAP2	+	+
YDR046C	BAP3	+	+
YJR148W	BAT2	-	+
YGR282C	BGL2	+	+
YGR286C	BIO2	+	+
YNR058W	BIO3	+	+
YNR057C	BIO4	+	+
YNR056C	BIO5	+	+
YJR025C	BNA1	+	+
YJR078W	BNA2	+	+
YBL098W	BNA4	+	+
YLR231C	BNA5	+	+
YFR047C	BNA6	+	+
YCR032W	BPH1	+	+
YEL063C	CAN1	+	+
YPL111W	CAR1	+	+
YLR438W	CAR2	+	+
YML042W	CAT2	+	+
YLR307W	CDA1	+	+
YLR308W	CDA2	+	+
YLR303W	MET17	+	+
YOL064C	MET22	+	-
YJR010W	MET3	+	+
YER091C	MET6	+	+
YOR241W	MET7	+	-
YLL062C	MHT1	+	+
YJR077C	MIR1	+	+
YBR084W	MIS1	+	+
YNL117W	MLS1	+	+
YLL061W	MMP1	+	+
YPL104W	MSD1	+	-
YOL033W	MSE1	+	-
YPR047W	MSF1	+	-
YNL073W	MSK1	+	-
YGR171C	MSM1	+	-
YHR091C	MSR1	+	-
YDR268W	MSW1	+	-
YPL097W	MSY1	+	-
YKR080W	MTD1	+	+
YGR055W	MUP1	+	+
YHL036W	MUP3	+	+
YGR007W	MUQ1	-	+
YLR382C	NAM2	+	-
YDL040C	NAT1	+	+
YGR147C	NAT2	+	+
YMR145C	NDE1	+	+
YDL085W	NDE2	+	+
YML120C	NDI1	+	+
YLR138W	NHA1	+	+
YJL126W	NIT2	+	+
YLR351C	NIT3	+	+
YLR328W	NMA1	+	+
YGR010W	NMA2	+	+
YOR209C	NPT1	+	+
YGL067W	NPY1	+	+
YDR001C	NTH1	+	+
YBR001C	NTH2	+	+
YKL120W	OAC1	+	+
YKL055C	OAR1	+	-
YPL134C	ODC1	+	+
YOR222W	ODC2	+	+
YJR073C	OPI3	-	+
YOR130C	ORT1	+	+
YJR051W	OSM1	+	+
YDR538W	PAD1	+	+
YIL145C	PAN6	-	+
YKR097W	PCK1	-	+
YGR202C	PCT1	+	+
YER178W	PDA1	+	+
YBR221C	PDB1	+	+

YER061C	CEM1	+	-
YCL064C	CHA1	+	+
YGR157W	CHO2	-	+
YBR023C	CHS3	+	+
YNR001C	CIT1	+	+
YCR005C	CIT2	+	+
YPR001W	CIT3	+	+
YLR133W	CKI1	+	+
YBR003W	COQ1	+	+
YNR041C	COQ2	+	+
YOL096C	COQ3	+	-
YML110C	COQ5	+	-
YGR255C	COQ6	+	-
YBL045C	COR1	+	-
YPL172C	COX10	+	+
YLR038C	COX12	+	+
YNL052W	COX5A	+	+
YIL111W	COX5B	+	+
YHR051W	COX6	+	-
YMR256C	COX7	+	-
YLR395C	COX8	+	+
YDL067C	COX9	+	-
YOR303W	CPA1	+	+
YJR109C	CPA2	+	+
YNL130C	CPT1	+	+
YOR100C	CRC1	+	+
YDL142C	CRD1	+	+
YBR036C	CSG2	+	+
YDR256C	CTA1	+	+
YBR291C	CTP1	+	+
YGR088W	CTT1	+	+
YEL027W	CUP5	+	-
YML054C	CYB2	+	+
YAL012W	CYS3	+	+
YGR155W	CYS4	+	+
YOR065W	CYT1	+	+
YML070W	DAK1	+	+
YFL053W	DAK2	+	+
YIR027C	DAL1	+	+
YIR029W	DAL2	+	+
YIR032C	DAL3	+	+
YIR028W	DAL4	+	+
YJR152W	DAL5	+	+
YIR031C	DAL7	+	+
YFL001W	DEG1	+	-
YHR011W	DIA4	+	-
YLR348C	DIC1	+	+
YPL265W	DIP5	+	+
YDL174C	DLD1	+	+
YLR172C	DPH5	+	+
YLR044C	PDC1	+	+
YLR134W	PDC5	+	+
YGR087C	PDC6	+	+
YGL248W	PDE1	+	+
YOR360C	PDE2	+	+
YBR035C	PDX3	+	-
YBL030C	PET9	+	+
YGR240C	PFK1	-	+
YMR205C	PFK2	-	-
YIL107C	PFK26	+	+
YOL136C	PFK27	+	+
YKL127W	PGM1	+	+
YMR105C	PGM2	+	+
YNL316C	PHA2	+	+
YBR092C	PHO3	+	+
YDR481C	PHO8	+	+
YML123C	PHO84	+	+
YCR037C	PHO87	+	+
YBR296C	PHO89	+	+
YJL198W	PHO90	+	+
YNR013C	PHO91	+	+
YPL268W	PLC1	+	-
YPL036W	PMA2	+	+
YCR024C-A	PMP1	+	+
YEL017C-A	PMP2	+	+
YDL095W	PMT1	+	+
YAL023C	PMT2	+	+
YOR321W	PMT3	+	+
YDL093W	PMT5	+	+
YGR199W	PMT6	+	+
YGL037C	PNC1	+	+
YLR209C	PNP1	-	+
YPL188W	POS5	+	-
YIL160C	POT1	+	+
YGL205W	POX1	+	+
YHR026W	PPA1	+	-
YMR267W	PPA2	+	-
YPL148C	PPT2	+	-
YDR300C	PRO1	+	-
YOR323C	PRO2	+	+
YER023W	PRO3	+	+
YHL011C	PRS3	+	-
YBL068W	PRS4	+	+
YOL061W	PRS5	+	+
YGR170W	PSD2	+	+
YPL212C	PUS1	+	+
YGL063W	PUS2	+	+
YNL292W	PUS4	+	+
YLR142W	PUT1	+	+
YHR037W	PUT2	+	+

YDR294C	DPL1	-	+
YDR284C	DPP1	+	+
YBR208C	DUR1,2	+	+
YHL016C	DUR3	+	+
YJR137C	ECM17	+	+
YBR176W	ECM31	-	+
YLR299W	ECM38	+	+
YMR062C	ECM40	+	+
YDR147W	EKI1	+	+
YGR254W	ENO1	+	+
YHR123W	EPT1	-	+
YML126C	ERG13	-	+
YMR202W	ERG2	+	+
YNL280C	ERG24	-	-
YLR056W	ERG3	+	+
YGL012W	ERG4	-	+
YMR015C	ERG5	+	+
YML008C	ERG6	+	-
YLR300W	EXG1	+	+
YDR261C	EXG2	+	+
YOR317W	FAA1	+	+
YIL009W	FAA3	+	+
YMR246W	FAA4	+	+
YFR019W	FAB1	+	-
YBR041W	FAT1	+	+
YER183C	FAU1	+	+
YLR377C	FBP1	+	+
YJL155C	FBP26	+	+
YPR062W	FCY1	+	+
YER056C	FCY2	+	+
YER060W	FCY21	+	+
YER060W-A	FCY22	+	+
YCR034W	FEN1	-	+
YCR028C	FEN2	+	-
YLR342W	FKS1	+	-
YMR306W	FKS3	+	+
YIL134W	FLX1	+	+
YBL013W	FMT1	+	+
YKR009C	FOX2	+	+
YLL043W	FPS1	+	+
YBL042C	FUI1	+	+
YPL262W	FUM1	+	+
YMR250W	GAD1	+	+
YBR020W	GAL1	+	+
YBR019C	GAL10	+	+
YLR081W	GAL2	+	+
YBR018C	GAL7	+	+
YKR039W	GAP1	+	+
YDR019C	GCV1	+	+
YMR189W	GCV2	+	+
YOR348C	PUT4	+	+
YPL147W	PXA1	+	+
YKL188C	PXA2	+	+
YGL062W	PYC1	+	+
YBR218C	PYC2	+	+
YOR347C	PYK2	+	+
YHR001W-A	QCR10	+	+
YFR033C	QCR6	+	+
YDR529C	QCR7	+	-
YJL166W	QCR8	+	+
YGR183C	QCR9	+	+
YDL090C	RAM1	+	-
YCR036W	RBK1	+	+
YIL053W	RHR2	+	+
YBL033C	RIB1	+	-
YEL024W	RIP1	+	+
YIL066C	RNR3	+	+
YGR180C	RNR4	+	-
YJL121C	RPE1	+	+
YLR180W	SAM1	+	+
YDR502C	SAM2	+	+
YPL274W	SAM3	+	+
YPL273W	SAM4	+	+
YMR272C	SCS7	+	+
YKL148C	SDH1	+	+
YLL041C	SDH2	+	+
YDR178W	SDH4	+	+
YOR184W	SER1	+	+
YGR208W	SER2	+	+
YER081W	SER3	+	+
YIL074C	SER33	+	+
YDL168W	SFA1	+	+
YJR095W	SFC1	+	+
YBR263W	SHM1	+	+
YLR058C	SHM2	-	+
YDL052C	SLC1	-	+
YNR034W	SOL1	+	+
YCR073W-A	SOL2	+	+
YHR163W	SOL3	+	+
YGR248W	SOL4	+	+
YKL184W	SPE1	+	+
YOL052C	SPE2	+	+
YPR069C	SPE3	+	+
YLR146C	SPE4	+	+
YOR190W	SPR1	+	+
YMR101C	SRT1	+	+
YKL218C	SRY1	+	+
YDR536W	STL1	+	+
YJR130C	STR2	+	+
YMR054W	STV1	+	+

YAL044C	GCV3	+	-
YEL042W	GDA1	+	+
YOR375C	GDH1	+	+
YDL215C	GDH2	+	+
YAL062W	GDH3	+	+
YCR098C	GIT1	+	+
YEL011W	GLC3	+	+
YCL040W	GLK1	+	+
YOR168W	GLN4	+	+
YML004C	GLO1	+	+
YDR272W	GLO2	+	+
YOR040W	GLO4	+	+
YPL091W	GLR1	+	+
YDL171C	GLT1	+	+
YEL046C	GLY1	+	-
YHR183W	GND1	+	-
YGR256W	GND2	+	+
YDR508C	GNP1	+	+
YDL022W	GPD1	-	+
YOL059W	GPD2	+	+
YPR160W	GPH1	+	+
YDL021W	GPM2	+	+
YOL056W	GPM3	+	+
YKL026C	GPX1	+	+
YBR244W	GPX2	+	+
YHR104W	GRE3	+	+
YGR032W	GSC2	+	+
YJL101C	GSH1	+	+
YOL049W	GSH2	+	+
YFR015C	GSY1	+	+
YLR258W	GSY2	+	+
YHL032C	GUT1	+	+
YIL155C	GUT2	+	+
YER014W	HEM14	+	-
YOR176W	HEM15	+	+
YDL205C	HEM3	+	+
YER055C	HIS1	+	+
YFR025C	HIS2	+	+
YOR202W	HIS3	+	+
YCL030C	HIS4	+	+
YIL116W	HIS5	+	+
YIL020C	HIS6	+	+
YBR248C	HIS7	+	+
YML075C	HMG1	+	+
YLR450W	HMG2	+	+
YBR034C	HMT1	+	+
YGL077C	HNM1	+	+
YDR305C	HNT2	+	+
YDR158W	HOM2	+	+
YER052C	HOM3	+	+
YIL162W	SUC2	+	+
YBR294W	SUL1	+	+
YLR092W	SUL2	+	+
YPL057C	SUR1	+	+
YDR297W	SUR2	+	+
YLR372W	SUR4	+	-
YLR354C	TAL1	+	+
YBR069C	TAT1	+	-
YOL020W	TAT2	+	+
YJL052W	TDH1	+	+
YJR009C	TDH2	+	+
YGR192C	TDH3	+	+
YDL185W	TFP1	+	-
YPL234C	TFP3	+	-
YOL055C	THI20	+	+
YPL258C	THI21	+	+
YPR121W	THI22	+	+
YPL214C	THI6	+	+
YLR237W	THI7	+	+
YHR025W	THR1	+	+
YCR053W	THR4	+	+
YIL078W	THS1	+	+
YPR074C	TKL1	+	+
YJR066W	TOR1	+	+
YDR050C	TPI1	+	+
YBR126C	TPS1	+	-
YDR074W	TPS2	+	+
YMR261C	TPS3	+	+
YDR007W	TRP1	+	+
YER090W	TRP2	+	+
YKL211C	TRP3	+	+
YDR354W	TRP4	+	+
YGL026C	TRP5	+	+
YHR106W	TRR2	+	+
YML100W	TSL1	+	+
YGR019W	UGA1	+	+
YBR006W	UGA2	+	+
YDL210W	UGA4	+	+
YKL216W	URA1	+	+
YMR271C	URA10	+	+
YJL130C	URA2	+	+
YLR420W	URA4	+	+
YML106W	URA5	+	+
YBL039C	URA7	+	-
YJR103W	URA8	+	+
YDR400W	URH1	+	+
YNR012W	URK1	+	+
YJR049C	UTR1	+	+
YPR036W	VMA13	+	-
YBR127C	VMA2	+	-

YJR139C	HOM6	+	+
YER062C	HOR2	+	+
YDR399W	HPT1	+	+
YFR053C	HXK1	+	+
YGL253W	HXK2	+	+
YHR094C	HXT1	+	+
YFL011W	HXT10	+	+
YNL318C	HXT14	+	+
YNR072W	HXT17	+	+
YMR011W	HXT2	+	+
YDR345C	HXT3	+	+
YHR092C	HXT4	+	+
YHR096C	HXT5	+	+
YJL214W	HXT8	+	+
YIR037W	HYR1	+	+
YER065C	ICL1	+	+
YPR006C	ICL2	+	+
YOR136W	IDH2	+	+
YDL066W	IDP1	+	+
YLR174W	IDP2	+	+
YNL009W	IDP3	+	+
YER086W	ILV1	+	+
YCL009C	ILV6	+	+
YLR432W	IMD3	+	+
YML056C	IMD4	+	+
YHR046C	INM1	-	+
YJL153C	INO1	-	+
YDR072C	IPT1	+	+
YPL040C	ISM1	+	-
YDR497C	ITR1	+	+
YOL103W	ITR2	+	+
YKL217W	JEN1	+	+
YIL125W	KGD1	+	+
YDR148C	KGD2	+	+
YDR483W	KRE2	+	+
YOR099W	KTR1	+	+
YKR061W	KTR2	+	+
YOR071C	YOR071C	+	+
YOR192C	YOR192C	+	+
YKR053C	YSR3	+	+
YJL139C	YUR1	+	+
YOR332W	VMA4	+	+
YKL080W	VMA5	+	-
YLR447C	VMA6	+	-
YGR020C	VMA7	+	-
YEL051W	VMA8	+	-
YOR270C	VPH1	+	+
YLR240W	VPS34	+	-
YGR194C	XKS1	+	+
YJR133W	XPT1	+	+
YLR070C	XYL2	+	+
YAR035W	YAT1	+	+
YBR184W	YBR184W	+	+
YBR284W	YBR284W	+	+
YCR024C	YCR024C	+	-
YDR111C	YDR111C	+	+
YEL041W	YEL041W	+	+
YEL047C	YEL047C	+	+
YER053C	YER053C	+	+
YER087W	YER087W	+	+
YFL030W	YFL030W	+	+
YFR055W	YFR055W	+	+
YGR012W	YGR012W	+	+
YGR043C	YGR043C	+	+
YGR125W	YGR125W	+	+
YGR287C	YGR287C	+	+
YHL012W	YHL012W	+	+
YJL045W	YJL045W	+	+
YJL068C	YJL068C	+	+
YJL070C	YJL070C	+	+
YJL200C	YJL200C	+	+
YJL216C	YJL216C	+	+
YKL132C	YKL132C	+	+
YLR089C	YLR089C	+	+
YLR164W	YLR164W	+	+
YML082W	YML082W	+	+
YML096W	YML096W	+	+
YMR084W	YMR084W	+	+
YMR085W	YMR085W	+	+
YMR118C	YMR118C	+	+
YMR293C	YMR293C	+	-
YKL067W	YNK1	-	+

Index of Figures

Index of Tables

Bibliography

Abelson, P. H. & T. C. Hoering. Proc. Natl. Acad. Sci. U. S. A., Volume 47, Pages 623-632, 1961.

Adams A. ***Metabolomics: Small-molecule omics.*** The Scientist, Volume 17, Pages 38-40, 2003.

Allen J. et al. ***High-throughput classification of yeast mutants for functional genomics using metabolic footprinting.*** Nature Biotech., Volume 21, Pages 692–696, 2003.

Chen P. Barbara, D. ***Using the fractal dimension to cluster datasets.*** ACM KDD 2000, 2000.

Barlow HB. ***Unsupervised learning.*** Neural Computation, Volume 1, Pages 295-311, 1989.

Pronk JT Boles E, Jong-Gubbels P de. ***Identification and charcterization of MAF1, the Saccharomyces cerevisiae structural gene encoding mitochondrial malic enzyme.*** J. Bacteriol., Volume 180, Pages 2875-2882, 1998.

Brindle JT et al. ***Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics.*** Nature Med., Volume 8, Pages 1439–1444, 2002.

Brown MP et al. ***Knowledge-based analysis of microarray gene expression data by using support vector machines.*** Proc. Natl Acad. Sci., Volume 97, Pages 262–267, 2000.

Carlson M. ***Glucose repression in yeast.*** Curr. Opin. Microbiol., Volume 2, Pages 202-207, 1999.

Hayes A. Mohammed S. Gaskell S. J. Castrillo, J. I. & S. G. Oliver. ***An optimised protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry.*** Phytochemistry, Volume 62, Pages 929-937, 2003.

Caussinus H & A. Roiz. ***Interesting projections of multidimensional data by means of generalized component analysis***. Compstat, Volume 90, Pages 121-126, 1990.

Ball C Chervitz SA Dwight SS Hester ET Jia Y Juvik G Roe T Schroeder M Weng S Botstein D Cherry JM, Adler C. ***SGD: Saccharomyces Genome Database.*** Nucleic Acids Res, Volume 26(1), Pages 73-80, 1998.

Bernhard O. Palsson Christophe H. Schilling, Stefan Schuster & Reinhart Heinrich. ***Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era***. Biotechnol. Prog., Volume 15(3), Pages 296 - 303, 1999.

Clarke BL. ***Stoichiometric network analysis.*** S Cell Biophys., Volume 12, Pages 237-253, 1988.

The Gene Ontology Consortium. ***Gene Ontology: tool for the unification of biology.*** Nature Genet., Volume 25, Pages 25-29, 2000.

Costanzo MC et.al. ***The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information.*** Nucleic Acids Res., Volume 28, Pages 73–76, 2000.

Vander Y, Heyden Daszykowsk iM., K. Kaczmarek & B. Walczak. ***Robust statistics in data analysis — A review.*** Chemometrics and Intelligent Laboratory Systems, Volume 85:2, Pages 203-219, 2007.

Dauner, M. & U. Sauer. ***GC–MS analysis of amino acids rapidly provides rich information for isotopomer balancing.*** Biotechnol. Prog., Volume 16, Pages 642–649, 2000.

Dave R. ***Validating fuzzy partitions obtained through c-shells clustering***. Pattern Recognition Letters, Volume 17, Pages 613-623, 1996.

Iyer I, DeRisi, J. & P. Brown. ***Exploring the metabolic and genetic control of gene expression on a genomic scale.*** Science, Volume 278, Pages 680-686, 1997.

Raffard G. Canioni P. Pradet A. Dieuaide-Noubhani, M. & P. Raymond. ***Quantification of compartmented metabolic fluxes in maize root tips using isotope distribution from 13C- or 14C-labeled glucose.*** J. Biol. Chem., Volume 270, Pages 13147–13159, 1995.

Rosiers C. D. Montgomery J. A. David F. Garneau M. Donato, L. D. & H. Brunengraber. ***Rates of gluconeogenesis and citric acid cycle in perfused livers, assessed from the mass spectrometric assay of the 13C labeling pattern of glutamate.*** J. Biol. Chem., Volume 268, Pages 4170–4180, 1993.

Palsson BO. Duarte NC, Herrgard MJ. ***Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.*** Genome Res., Volume 14(7), Pages 1298-309, 2004.

Dujon B. ***The yeast genome project: what did we learn?*** Trends Genet, Volume 12, Pages 263–270, 1996.

Dolinski K Ball CA Binkley G Christie KR Fisk DG Issel-Tarver L Schroeder M Sherlock G Sethuraman A Weng S Botstein D Cherry JM Dwight SS, Harris MA. ***Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).*** Nucleic Acids Res, Volume 30, Pages 69-72, 2002.

Kowald A,C. Wierling H. Lehrach E. Klipp, R. herwig. ***Systems Biology in Practice***. , Wiley-VCH, 2005.

Acuna E. & Rodriguez C. ***A Meta analysis study of outlier detection methods in classification***. , 2004.

Edwards BO, J. S.; Palsson. ***How will bioinformatics influence metabolic engineering.*** Biotechnol. Bioeng., Volume 58, Pages 162-169, 1998.

Palsson B.O. Edwards, J.S.. ***The Escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities.*** Proc. Natl. Acad. Sci., Volume 97 (10), Pages 5528–5533, 2000.

Hegemann JH Becher D Feldmann H Guldener U Gotz R Hansen M Hollenberg CP Jansen G Kramer W Klein S Kotter P Kricke J Launhardt H Mannhaupt G Maierl A Meyer P Mewes W Munder T Niedenthal RK Ramezani Rad M Rohmer A Romer A Hinnen A etal. Entian KD, Schuster T. ***Functional analysis of 150 deletion mutants in Saccharomyces cerevisiae by a systematic approach.*** Mol. Gen. Genet., Volume 262, Pages 683-702, 1999.

Rousseeuw PJ, F.R. Hampel, E.M. Ronchetti & W.A. Stahel. ***Robust Statistics: the Approach Based on Influence Functions*** , Wiley, New York, 1986.

Bruhat A. Fafournoux, P. & C. Jousse. ***Amino acid regulation of gene expression.*** Biochem. J., Volume 351, Pages 1-12, 2000.

Livak KJ Falco SC, Dumas KS. ***Nucleotide sequence of the yeast ILV2 gene which encodes acetolacetate synthase.*** Nucleic Acid Res., Volume 13, Pages 4011-4027, 1985.

Foerster J. Nielsen J. Palsson B.O. Famili, I.. ***Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.*** Proc. Natl. Acad. Sci., Volume 100 (23), Pages 13134–13139, 2003.

Fiehn O et.al. ***Metabolite profiling for plant functional genomics.*** Nature Biotech., Volume 18, Pages 1157–1161, 2000.

Garrett R.G. Filzmoser P. & C. Reimann. ***Multivariate outlier detection in exploration geochemistry.*** Computers & Geosciences, 2005.

Palsson BO Forster J, Famili I & Nielsen J. ***Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network***. Genome Res., Volume 13(2), Pages 244-53, 2003.

D.G Fraenkel. ***Carbohydrate metabolism. In Molecular Biology of the Yeast Saccharomyces. Metabolism and Gene Expression***, Cold Spring Harbor Laboratory Press, Plainview, New York, 1982.

Snoep J. L. Westerhoff H. V. de la Fuente, A. & P. Mendes. ***Metabolic control in integrated biochemical systems.*** Eur. J. Biochem., Volume 269, Pages 4399-4408, 2002.

Gancedo JM, Funayama, S. & C. Gancedo. ***Turnover of yeast fructose-1,6-bisphosphatase in different metabolic conditions.*** Eur. J. Biochem, Volume 109, Pages 61-66, 1980.

Kumm J, M. Proctor C. Nislow D. F. Jaramillo A. M. Chu M. I. Jordan A. P. Arkin G. Giaever, P. Flaherty & R. W. Davis. . Proc. Natl. Acad. Sci., Volume 101, Pages 793–798, 2004.

E. Galimov. ***The Biological Fractionation of Isotopes***, Academic Press, Orlando, FL, 1985.

Gancedo R, C. & Serrano. ***Energy-yielding metabolism. In The Yeasts, Vol. 3, Metabolism and Physiology of Yeasts.*** , Academic Press, New York, 1989.

Gancedo JM. ***Yeast Carbon Catabolite Repression***. Microbiol Mol Biol Rev, Volume 62(2), Pages 334-361, 1998.

Giaever G.; D. D. Shoemaker; T. W. Jones; H. Liang; E. A. Winzeler; A. Astromoff & R. W. Davis. ***Genomic profiling of drug sensitivities via induced haploinsufficiency.*** Nat Genet, Mar, Volume 21, Number 3, Pages 278-283, 1999.

Goffeau A.; B. G. Barrell; H. Bussey; R. W. Davis; B. Dujon; H. Feldmann; F. Galibert; J. D. Hoheisel; C. Jacq; M. Johnston; E. J. Louis; H. W. Mewes; Y. Murakami; P. Philippsen; H. Tettelin & S. G. Oliver. ***Life with 6000 genes.*** Science, Oct, Volume 274, Number 5287, Pages 546, 563-546, 567, 1996.

Gombert AK ; M. Moreira dos Santos; B. Christensen & J. Nielsen. ***Network identification and flux quantification in the central metabolism of Saccharomyces cerevisiae under different conditions of glucose repression.*** J Bacteriol, Feb, Volume 183, Number 4, Pages 1441-1451, 2001.

Kanehisa M. Goto S, Nishioka T. ***LIGAND: chemical database for enzyme reactions.*** Bioinformatics, Volume 14(7), Pages 591-9, 1998.

U Grenander. ***Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern***

Synthesis and Regular Structures. , Berlin: Springer-Verlag., Berlin, 1976-198, 1976-1981.

Williams H. J. Sang E. Clarke K. Rae C. Griffin, J. L. & J. K. Nicholson. ***Metabolic profiling of genetic disorders: a multitissue (1)H nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue.*** Anal. Biochem., Volume 293, Pages 16-21, 2001.

Taskinen L Haarasilta S. ***Location of three key enzymes of gluconeogenesis in baker's yeast.*** Arch. Microbiol, Volume 113, Pages 159-161, 1977.

A. S. Hadi. ***Identifying multiple outliers in multivariate data***. Journal of the Royal Statistical Society, Volume Series B, 54, Pages 761-771, 1992.

F.R. Hampel. ***A general qualitative definition of robustness***. Annals of Mathematical Statistics, Volume 42, Pages 1887-1896, 1971.

Ronchetti E.M. Rousseeuw P.J. Stahel W. Hampel, F.R. ***Robust Statistics. The Approach Based on Influence Functions.*** , Wiley, New York, 1986.

Kahn D Heinrich R. Handorf T, Ebenhöh O. ***Hierarchy of metabolic compounds based on their synthesizing capacity.*** IEE Proc Syst Biol. Sep, Volume 153(5), Pages 359-63, 2006.

Ebenhoh O, Handorf, T. & R. Heinrich.. ***Expanding metabolic networks: scopes of compounds, robustness, and evolution.*** J. Mol. Evol., Volume 61, Pages 498–512, 2005.

Hansen J. & P. F. Johannesen. ***Cysteine is essential for transcriptional regulation of the sulfur assimilation genes in Saccharomyces cerevisiae.*** Mol. Gen. Genet., Volume 263, Pages 535-542, 2000.

Clark J et.al Harris MA. ***The Gene Ontology (GO) database and informatics resource.*** Nucleic Acids Res, Volume 32, Pages D258-61, 2004.

Schuster S. Heinrich, R. ***The Regulation of Cellular Systems.*** , Chapman & Hall, New York, 1996.

Holzer H. ***Catabolite inactivation in yeast.*** Trends Biochem. Sci., Volume 1, Pages 178-181, 1976.

Huber PJ. ***Robust Statistics***. , Wiley, New York, 1981.

Huhman LW, D. V. & Sumner. ***Metabolic profiling of saponins in Medicago sativa and Medicago truncculata using HPLC coupled to an electrospray ion-trap mass spectrometer.*** Phytochemistry, Volume 59, Pages 347–360, 2002.

Geva. B., I. Gath. ***Unsupervised Optimal Fuzzy Clustering***. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 11, 1989.

Penny K. I. & Jolliffe I. T. ***A comparison of multivariate outlier detection methods for clinical laboratory safety data***. The Statistician, Volume 50(3), Pages 295-308, 2001.

Murthy MN JAIN A.K. & PJ Flynn. ACM Computing Surveys, Volume 31, Pages 3, 1999.

Quackenbush J. ***Computational genetics: Computational analysis of microarray data***. Nature Reviews Genetics, Volume 2, Pages 418-427, 2001.

Johnston M. ***Feasting, fasting and fermenting. Glucose sensing in yeast and other cells.*** Trends Genet., Volume 15, Pages 29-33, 1999.

Johnston, M. & M. Carlson. ***Regulation of carbon and phosphate utilization. In E. W. Jones, J. R. Pringle, and J. R. Broach (ed.), The molecular and cellular biology of the yeast Saccharomyces, vol. 2. Gene expression.*** , Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1992.

Nielsen J, John Villadsen Karsten Schmidt, Morten Carlsen. ***Modeling Isotopomer Distributions in Biochemical Networks Using Isotopomer Mapping Matrices***. Biotech and Bioengg., Volume 55(6), Pages 831-840, 1997.

L. Kaufman & P.J. Rousseeuw. ***Finding Groups in Data: An Introduction to Cluster Analysis.*** , Wiley, New York, 1990.

Kell DB & R. D. King. ***On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning.*** Trends Biotechnol., Volume 18, Pages 93-98, 2000.

Zelder O. Kiefer P., Heinzle E. & Wittmann C.. ***Comparative metabolic flux analysis of lysine producing***

Corynebacterium glutamicum cultured on glucose or fructose. Appl. Environ. Microbiol., Volume 70(1), Pages 229-239, 2004.

Sumegi B. Dietmeier K. Bock I. Gajdos G. Tomcsanyi T. Sandor A. Kispal, G. ***Cloning and sequencing of a cDNA encoding Saccharomyces cerevisiae carnitine acetyltransferase.*** J. Biol. Chem., Volume 268, Pages 1824–1829, 1993.

Schuster S. Gilles E.D. Klamt, S.. ***Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria.*** Biotechnol. Bioeng., Volume 77 (7), Pages 734–751, 2002.

E. Knorr & R. Ng. ***A unified approach for mining outliers***. Proceedings Knowledge Discovery KDD, 1997.

Knorr, E. & Ng. R. ***Algorithms for mining distance-based outliers in large datasets.*** 24th Int. Conf. Very Large Data Bases (VLDB), 1998.

Weckwerth W. Linke T. & Fiehn O. Kose, F. ***Visualizing plant metabolomic correlation networks using cliquemetabolite matrices.*** Bioinformatics, Volume 17, Pages 1198–1208, 2001.

Ken Gable Tamsin Tarling Delphine Rebérioux Jenny Bryan Raymond J. Andersen Teresa Dunn Phil Hieter Kristin Baetz, Lianne McHardy & Michel Roberge. ***Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action.*** PNAS, Volume 101(13), Pages 4525-4530, 2004.

Slattery M G, L. L. Newcomb, J. A. Diderich & W. Heideman. ***Glucose Regulation of Saccharomyces cerevisiae Cell Cycle Genes***. Eukaryot. Cell, Volume 2(1), Pages 143 - 149, 2003.

D. C. Lay. ***Linear algebra and its applications***. , Addison-Wesley Longman Inc.: Reading, MA, 1997.

Hendrik Luesch. ***Towards high-throughput characterization of small molecule mechanisms of action.*** Mol. BioSyst., Volume 2, Pages 609-620, 2006.

Kanehisa M. ***A database for post-genome analysis.*** Trends Genet., Volume 13(9), Pages 375-6, 1997.

Olyarchuk JG Wei L. Mao X, Cai T. ***Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary***. Bioinformatics, Volume 21(19), Pages 3787-93, 2005.

Matt Kaeberlein, Christopher R Burtner & Brian K Kennedy. ***Recent Developments in Yeast Aging***. PLoS Genet., Volume 3(5), Pages e84, 2007.

Michal G. ***Biochemical Pathways: an Atlas of Biochemistry and Molecular Biology.*** , Wiley, New York, 1998.

Myatt GJ. ***Computer aided estimation of synthetic accessibility.*** *School*: University of Leeds, Leeds, UK, 1994.

Müller, D. & H. Holzer. ***Regulation of fructose-1,6-bisphosphatase in yeast by phosphorylation/dephosphorylation.*** Biochem. Biophys. Res. Commun., Volume 103, Pages 926-933, 1981.

R.T. Ng & Han J.. ***Efficient and Effective Clustering Methods for Spatial Data Mining.*** Very Large Data Bases Conference, 1994.

Yeh H,. J. C. Spande T. F. Noronha, S. B. & J. Shiloach. ***Investigation of the TCA cycle and the glyoxylate shunt in Escherichia coli BL21 and JM109 using 13C-NMR/MS.*** Biotechnol. Bioeng., Volume 68, Pages 316–327, 2000.

Nowlan SJ. ***Maximum likelihood competitive learning. In Advances in Neural Information Processing Systems***., Morgan Kaufmann, 1990.

Hollatz H Nozicka Fm Guddat J & Bank B. ***Theorie der linearen parametrischen optimierung***. , Akademie-Verlag, Berlin, 1974.

Perez-Iratxeta C et al Nuno J. C, Sanchez-Valdenebro I. ***Network organization of cell metabolism : mono saccharide interconversion.*** Biochem. J. h, Volume 324, Pages 103-111, 1997.

Fiehn O. ***Metabolomics- The link between genotypes and phenotypes.*** Plant Mol. Biol., Volume 48, Pages 155-171, 2002.

Sato K Fujibuchi W Bono H Kanehisa M. Ogata H, Goto S. ***KEGG: Kyoto Encyclopedia of Genes and Genomes.*** Nucleic Acids Res., Volume 27(1), Pages 29-3, 1999.

Oliver SG. ***Functional genomics: lessons from yeast.*** Philos. Trans. Roy. Soc. B., Volume 357, Pages 17-23, 2002.

Stepaniants SB,G. Cavet M. K. Wolf J. S. Butler J. C. Hinshaw P. Garnier G. D. Prestwich A. Leonardson P. Garrett-Engele C. M. Rush M. Bard G. Schimmack J. W. Phillips C. J. Roberts P. Y. Lum, C. D. Armour & D. D. Shoemaker. . Cell, Volume 116, Pages 121–137, 2004.

Vozza A. Agrimi G. De Marco V. Runswick M.J. Palmieri F. Walkers J.E. Palmieri, L. ***Identification of the yeast mitochondrial transporter for oxaloacetate and sulfate.*** J. Biol. Chem., Volume 274, Pages 22184–22190, 1999.

B.O. Palsson. ***The challenges of in silico biology.*** Nat. Biotechnol., Volume 18, Pages 1147–1150, 2000.

Kitawaga H. Gibbons P.G. Papadimitriou, S. & C. Faloutsos. ***LOCI: Fast Outlier Detection Using the Local Correlation Integral***. , 2002.

Shaw R. C. Sinskey A. J. Park, S. M. & G. Stephanopoulos. ***Elucidation of anaplerotic pathways in Corynebacterium glutamicum via 13C-NMR spectroscopy and GC–MS.*** Appl. Microbiol. Biotechnol., Volume 47, Pages 430–440, 1997.

Nuño JC Montero F Schuster S. Humboldt Pfeiffer T, Sánchez-Valdenebro I. ***METATOOL: for studying metabolic networks.*** Bioinformatics, Volume 15(3), Pages 251-7, 1999.

Reed J.L. Papin J.A. Wiback S.J. Palsson B.O. Price, N.D. ***Network-based analysis of metabolic regulation in the human red blood cell.*** J. Theor. Biol., Volume 225, Pages 185–194, 2003.

Steensma H.Y. van Dijken J.P. Pronk, J.T. ***Pyruvate metabolism in Saccharomyces cerevisiae.*** Yeast, Volume 12, Pages 1607–1633, 1996.

J. Reiber. R. Rezaee, B. Lelieveldt. ***A new cluster validity index for the fuzzy c-mean***. Pattern Recognition Letters, Volume 19, Pages 237-246, 1998.

Rockafellar R. ***Convex analysis***. , Princeton University Press, Princeton NJ, 1970.

Teusink B. Broadhurst D. Zhang N. Hayes A. Walsh M. C. Berden J. A. Brindle K. M. Kell D. B. Rowland J. J. Westerhoff H. V. van Dam K. Raamsdonk, L. M. & S. G. Oliver. ***A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.*** Nat Biotechnol., Volume 19, Pages 45-50, 2001.

Lengauer T, Rahnenfuhrer J. ***Classification of genes. In Bioinformatics: From Genomes to Therapies;*** Wiley-VCH Verlag, 2006.

Rajkó R. ***Treatment of model error in calibration by robust and fuzzy procedures***. Analytical Letters, Volume 27, Pages 215–228., 1994.

Reder C. ***Metabolic control theory: A structural approach.*** J. Theor. Biol., Volume 135, Pages 175-201, 1988.

Hettema E.H. van der Berg M. Tabak H.F. Wanders R.J.A. van Roermund, C.W.T.. ***Molecular characterization of carnitine-dependent transport of acetyl-CoA from peroxisomes to mitochondria in Saccharomyces cerevisiae and identification of a plasma membrane carnitine transporter, Agp2p.*** EMBO J., Volume 18, Pages 5843–5852, 1999.

Roessner U et. al. ***Metabolite profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems.*** Plant Cell, Volume 13, Pages 11–29, 2001.

Roessner-Tunali U et.al. ***Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development.*** Plant Physiol., Volume 133, Pages 84–99, 2003.

Lim F. Wallace J.C. Rohde, M. ***Electron microscopic localization of pyruvate carboxylase in rat liver and Saccharomyces cerevisiae by immunogold procedures.***. Arch. Biochem. Biophys., Volume 290, Pages 197–201, 1991.

Leory A. Rousseeuw P. ***Robust Regression and Outlier Detection***, Wiley Series in Probability and Statistics, 1987.

Rousseeuw P. ***Multivariate estimation with high breakdown point.*** , Mathematical Statistics and Applications Vol. B, Akademiai Kiado, 1985.

Rousseeuw PJ. & A.M. Leroy. ***Robust Regression and Outlier Detection***. , John Wiley & Sons Inc., New York, 1987.

Warwick B. Dunn George G. Harrigan Royston Goodacre, Seetharaman Vaidyanathan & Douglas B. Kell. ***Metabolomics by numbers: acquiring and understanding global metabolite data,*** Trends in Biotechnology, Volume 22(5), Pages 245-251, 2004.

Koutroubas K, S. Theodoridis. ***Pattern recognition***. , Academic Press, 1999.

Hatzimanikatis V. Bailey J. E. Hochuli M. Szyperski T. Sauer, U. & K. Wutrich. ***Metabolic fluxes in riboflavin-producing Bacillus subtilis.*** Nature Biotechnol., Volume 15, Pages 448–452, 1997.

Lasko D. R. Fiaux J. Hochuli M. Glaser R. Szyperski T. Wuthrich K. Sauer, U. & J. E. Bailey. ***Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism.*** J. Bacteriol., Volume 181, Pages 6679–6688, 1999.

Lauer M. & Fritsch H. Sauter, H.. ***Metabolite profiling of plants — a new diagnostic technique.*** Pap. Am. Chem. Soc., Volume 195, Pages 129, 1988.

Schuster S. Palsson B.O. Heinrich R. Schilling, C.H.. ***Metabolic pathway analysis: basic concepts and scientific applications in the postgenomic era.*** Biotechnol. Prog., Volume 15, Pages 296–303, 1999.

Nielsen J. Schmidt, K. & J. Villadsen. ***Quantitative analysis of metabolic fluxes in Escherichia coli using two-dimensional NMR spectroscopy and complete isotopomer models.*** J. Biotechnol., Volume 71, Pages 175–190, 1999.

Norregaard L. C. Pedersen B. Meissner A. Duus J. O. Nielsen J. Schmidt, K. & J. Villadsen. ***Quantification of intracellular fluxes from fractional enrichment and 13C–13C coupling constraints on the isotopomer distribution in labeled biomass components.*** Metab.Eng., Volume 1, Pages 166–179, 1999.

Fell D.A. Dandekar T. Schuster, S.. ***A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic systems.*** Nat. Biotechnol., Volume 18, Pages 326–332, 2000.

Anuj Kumar & Michael Snyder. ***Emerging technologies in yeast genomics.*** Nature Reviews Genetics, Volume 2, Pages 302-312, Luedemann A Selbig J Kopka J Steinhauser D., Junker BH. ***Hypothesis driven approach to predict transcriptional units from gene expression data.*** Bioinformatics, Volume 20, Pages 1928-1939, 2004.

Luedemann A. Thimm O. Steinhauser D., Usadel B. & Kopka J.. ***CSB.DB: a comprehensive systems-biology database.*** Bioinformatics, Volume 20(18), Pages 3647:51, 2004.

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

Steuer R. ***Computational approaches to the topology, stability and dynamics of metabolic networks***. Phytochemistry, 2007.

Szyperski, T. (1995). ***Biosynthetically directed 13C-fractional labeling of proteinogenic amino acids.*** Eur. J. Biochem., Volume 232, Pages 433–448, 1995.

Sârbu C & H.F. Pop. ***Fuzzy robust estimation of central location***. Talanta, Volume 54, Pages 128–130, 2001.

Hughes. TR . Cell, Volume 102, Pages 109–126, 2000.

Wittmann C Velagapudi V Heinzle E (2006) Accepted In: Talwar P, Lengauer T. ***Development of Computational Methods for Analysis of Metabolic Profiling Data***. International Conference on Systems Biology (ICSB 2006), Japan, 2006.

Wittmann C. Mangadu V. Talwar P., Lengauer T. & Heinzle E.. ***Towards cellular function through metabolite screening.*** In Proceedings of the 3rd Annual Conference on Metabolic Profiling: Pathways in Discovery, Princeton, New Jersey, Number 7, 2003.

Lengauer T. and Heinzle E. Talwar P., Wittmann C. ***Software tool for automated processing of 13C labeling data from mass spectrometry data.***. BioTechniques, Volume 35, Pages 1214-1215, 2003.

Baganz F. Westerhoff H. V. Teusink, B. & S. G. Oliver. ***Metabolic control analysis as a tool in the elucidation of the function of novel genes.*** Methods Microbiol., Volume 26, Pages 297-336, 1998.

R. N. Trethewey. ***Gene discovery via metabolic profiling.*** Curr. Opin. Biotechnol., Volume 12, Pages 135-138, 2001.

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

Friedman Trevor & Hastie, Tibshirani. ***Elements of Statistical Learning.*** , Springer, 2001.

Schipper D. Breedveld G.J. Mak P.R. Scheffers W.A. van Dijken J.P. van Urk, H. ***Localization and kinetics of pyruvate-metabolizing enzymes in relation to aerobic alcoholic fermentation in Saccharomyces cerevisiae CBS 8066 and Candida utilis CBS 621.***. Biochim. Biophys. Acta, Volume 992, Pages 78–86, 1989.

B. O. Varma, A.; Palsson. ***Metabolic flux balancing: basic concepts, scientific and practical use.*** Biotechnol. Bioeng., Volume 12, Pages 994-998, 1994.

Rohde M Devenish RJ Wallace JC Walker ME, Val DL. ***Yeast pyruvate carboxylase: identification of two genes encoding isoenzymes.*** Biochem.Biophys. Res. Commun, Volume 176, Pages 1210-1217, 1991.

Warner JR. ***The economics of ribosome biosynthesis in yeast.*** Trends Biochem. Sci., Volume 24, Pages 437-440, 1999.

Watkins SM & J. B. German. ***Metabolomics and biochemical profiling in drug discovery and development.*** Curr. Opin. Mol. Ther., Volume 4, Pages 224-228, 2002.

Weckwerth W & O. Fiehn. ***Can we discover novel pathways using metabolomic analysis?*** Curr. Opin. Biotechnol., Volume 13, Pages 156-160, 2002.

Heinzle E Wittmann C, Hans M. ***In vivo analysis of intracellular amino acid labelings by GC/MS.*** Anal Biochem, Volume 307(2), Pages 379:82, 2002.

Mirny LA. Wunderlich Z. ***Using the topology of metabolic networks to predict viability of mutant strains.*** Biophys J., Volume 91(6), Pages 2304-11, 2006.

Zimmermann KD, F.K.& Entian. ***Yeast Sugar Metabolism: Biochemistry, Genetics, Biotechnology and Applications.*** , Technomic Publishing, Lancaster, 1997.

Zupke, C. & G. Stephanopoulos. ***Intracellular flux analysis in hybridomas using mass balances and in vitro 13C NMR.*** Biotechnol. Bioeng., Volume 45, Pages 292–303, 1995.

Gene Ontology [http://www.geneontology.org/index.shtml]. ,

R Development Core Team ***R: A language and environment for statistical computing.*** *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2005 [http://www.R-project.org.]*. ,

Saccharomyces Genome Database [http://www.yeastgenome.org/].

The Yeasts: Metabolism and Physiology. , Academic Press, New York, 1989.

Molecular Biology of the Yeast Saccharomyces. Metabolism and Gene Expression. , Cold Spring Harbor Laboratory Press, Plainview, New York, 1982.

Bonarius HPJ, Schmid G, Tramper. ***Flux analysis of underdetermined metabolic networks: The quest for the missing constraints***. J. Trends Biotech, Volume 15, Pages 308–314, 1997.

Palsson BO. Edwards JS, Covert M. ***Metabolic modelling of microbes: The flux-balance approach.*** Environ Microbiol, Volume 4, 2002.

Julian L. Griffin, Catharine Goddard, Russell J., Mortishire-Smith, Brian, C. Sweatman, John N. Haselden, Kay Davies, Andrew A. Grace, Kieran Clarke, & Gareth L.A.H. Jones, Elizabeth Sang. ***A functional analysis of mouse models of cardiac disease through metabolic profiling***. JBC, 2004.

Griffin JL, Bollard ME. ***Metabonomics: its potential as a tool in toxicology for safety assessment and data integration.*** Curr Drug Metab. Volume 5(5), Pages 389-98., 2004.

Karthik Raman, Preethi Rajagopalan & Nagasuma Chandra.. ***Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs***. PLoS Comput Biol, Volume 1(5), Pages e46, 2005.

Kauffman KJ, Prakash P, Edwards JS. ***Advances in flux balance analysis.*** Curr Opin Biotech, Volume 14, Pages 491–496., 2003.

Duarte, N. C., M. J. Herrgard, and B. O. Palsson***. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model***. Genome Res, Volume 14, Pages 1298–1309, 2004.

Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, et al. ***Functional profiling of the Saccharomyces cerevisiae genome.*** Nature, Volume 418, Pages 387–391, 2002.

Steinmetz, L. M., C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman, T. Jones, A. M. Chu, G. Giaever, H. Prokisch, P. J. Oefner, and R. W. Davis. ***Systematic screen for human disease genes in yeast.*** Nat. Genet, Volume 31, Pages 400–404, 2002

De Robichon-Szulmajster H . ***Induction of enzymes of the galactose pathway in mutants of Saccharomyces cerevisiae.*** Science, Volume 127(3288), Pages 28-9, 1958.

Douglas HC and Hawthorne DC. ***ENZYMATIC EXPRESSION AND GENETIC LINKAGE OF GENES CONTROLLING GALACTOSE UTILIZATION IN SACCHAROMYCES.*** Genetics, Volume 49, Pages 837-44, 1964.

Fukasawa T, et al.***The enzymes of the galactose cluster in Saccharomyces cerevisiae. II. Purification and characterization of uridine diphosphoglucose 4-epimerase.*** J Biol Chem, Volume 255(7), Pages 2705-7, 1980.

Majumdar S, et al.***UDPgalactose 4-epimerase from Saccharomyces cerevisiae. A bifunctional enzyme with aldose 1-epimerase activity.*** Eur J Biochem, Volume 271(4), Pages 753-9, 2004.

Chhabra, S. R., Q. He, K. H. Huang, S. P. Gaucher, E. J. Alm, Z. He, M. Z. Hadi, T. C. Hazen, J. D. Wall, J. Zhou, A. P. Arkin, and A. K. Singh. ***Global analysis of heat shock response in*** **Desulfovibrio vulgaris** Hildenborough. J. Bacteriol., Volume 188, Pages 1817-1828, 2006.

He, Q., K. H. Huang, Z. He, E. J. Alm, M. W. Fields, T. C. Hazen, A. P. Arkin, J. D. Wall, and J. Zhou. 2006. ***Energetic consequences of nitrite stress in*** **Desulfovibrio vulgaris** ***Hildenborough, inferred from global transcriptional analysis.*** Appl. Environ. Microbiol., Volume 72, Pages 4370-4381, 2006.

Mukhopadhyay, A., Z. He, E. J. Alm, A. P. Arkin, E. E. Baidoo, S. C. Borglin, W. Chen, T. C. Hazen, Q. He, H. Y. Holman, K. Huang, R. Huang, D. C. Joyner, N. Katz, M. Keller, P. Oeller, A. Redding, J. Sun, J. Wall, J. Wei, Z. Yang, H. C. Yen, J. Zhou, and J. D. Keasling. 2006. ***Salt stress in*** **Desulfovibrio vulgaris** ***Hildenborough: an integrated genomics approach***. J. Bacteriol., Volume 188, Pages 4068-4078, 2006.

Tang Y, Pingitore F, Mukhopadhyay A,Phan R,Terry C. Hazen,and Jay D. Keasling ***Pathway Confirmation and Flux Analysis of Central Metabolic Pathways in Desulfovibrio vulgaris Hildenborough using Gas Chromatography-Mass Spectrometry and Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry.*** J. Bacteriol. Volume 189, Pages 940, 2006.

Mark E. Jennings and Dwight E. Mmatthews. ***Determination of Complex Isotopomer Patterns in Isotopically Labeled Compounds by Mass Spectrometry.*** Anal Chem. 2008.

Annik Nanchen, Tobias Fuhrer, Uwe Sauer**. *Determination of Metabolic Flux Ratios from 13C-Experiments and Gas Chromatography-Mass Spectrometry Data: Protocol and Principles*** in Metabolomics: Methods and Protocols. Series: Methods in Molecular Biology, 358,177-197, 2006**.**

Weitzel M, Wiechert W, Nöh K. ***The topology of metabolic isotope labeling networks.*** BMC Bioinformatics, Aug 29;8:315. 2007.

Wu J, Mao X, Cai T, Luo J, Wei L. ***KOBAS server: a web-based platfor for automated annotation and pathway identification.***Nucleic Acids Res.Jul 1;34(Web Server issue):W720-4,2006.