Modelling Approaches for the Integration of Different Omics and Database-Information for *Clostridium acetobutylicum*

vorgelegt von
Dipl.-Ing.
Sebastian Curth
aus Berlin

von der Fakultät III – Prozesswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Leif-Alexander Garbe
1. Gutachter: Prof. Dr. Peter Neubauer
2. Gutachter: Prof. Dr. Peter Götz
3. Gutachter: Prof. Dr. Reinhard Guthke

Tag der wissenschaftlichen Aussprache: 21. März 2014

Berlin 2014
D 83

To family

Acknowledgements

First of all, I want to thank Professor Peter Götz for giving me the great opportunity to work on the COSMIC2 project. I owe him the start of this work by his innovative dynamic model from which all approaches in this thesis could germ. I am very grateful for the moral support when I started working with principal components to develop new ideas. Many thanks go as well to Peter Neubauer, to whom I owe the possibility to develop this thesis in the Biotechnology department of TU Berlin. I am also thankful to Reinhard Guthke who confidently acknowledged to become referee to this thesis.

Discussion with Professor Große-Wiesmann on data analysis and engineering-approaches always showed me different aspects and facets of the problem. They did not reduce the number of questions, on the contrary, but they helped finding pragmatic approaches a lot.

I want to thank Katy Wolstencroft, who introduced me to Taverna. Your enthusiastic support for workflow-creation allowed the development of core aspects of this thesis.

Many thanks go to my colleagues from all the different projects at the Biotechnology department of Beuth-Hochschule. Many many hours were spent in front of confusing graphs and diagrams and I was trying to make sense of them with your support: Susanne Wickert, Julia Rosenlöcher, Katharina Tomschek, Susanne Fischer, Sabrina Fischer, Kunigunde Stephani-Kosin, Tanja Westphalen.

COSMIC2 also consisted of a lot of experimental work. Many very fruitful discussions were held with Dr. Stefan Junne to start experimental work with Clostridia. Thank you Mirja Rothe, your excellent technical skills were vital for the project and inspiring for my own work. Thank you as well my two bachelor students, Marcel Mehl, who established fermentations, and Andriy Grygorenko who standardised experimental proceedings and measurements, data acquisition would not have been possible without you. Further thanks for technical support in many practical questions go to Thorsten Jamrath, Barbara Ritsche, Nabeel Fattohi and Harald Gerullis.

Thanks to the entire COSMIC2 consortium, who provided spores, protocols and

support in data acquisition.
This work was made possible through financial support by the BMBF project 0315782B (Verbundprojekt COSMIC2 - Systembiologie des Butanolproduzenten *Clostridium acetobutylicum*: Eine neue Quelle für Treibstoffe und Grundchemikalien; Teilprojekt B.)

Abstract

This work establishes methods to investigate the butanol formation of *Clostridium acetobutylicum*. In particular, the generation of two types of models will be extensively discussed. Therefor, the required formal basis for the model evaluation and the information technological standards will be introduced, e.g. the construction of a local database of clostridial annotation in KEGG.
The first model is a static pathway-model that provides the integration of transcriptome data into a metabolic-network model for visualisation and analysis purposes. It is proposed to use a novel rule from boolean logic for data integration to facilitate visual access to characteristics of the metabolic network. As consequence, the postulation of experimental hypotheses is facilitated: The possibility of a 3-hydroxybutyrate dehydrogenase activity in *C. acetobutylicum* is illuminated. The resulting priority list from annotation transfer contains functional and regulatory aspects of the data and the databases and it hereby offers an optimal starting point to initiate experimental work.
The second model is a dynamic model that is used to map metabolome and transcriptome data from fermentation experiments together. Its unique structure allows a number of new analyses - and shows new problems. Its simulation results suggest that the pH-shift in *C. acetobutylicum* can be solely related to transcript dynamics. Optimisation strategies on the transcript level and the parameter level of the model will be implemented and their results discussed.
Finally, the principal component analysis will be used to optimise computation times of such a model and from this, two novel methods will be derived: dynamic aspects of transcriptome data will be alternated to construct regulatory similar expression profiles with different amplitudes, and genes will be classified according to their regulatory similarity in a novel clustering approach.

Abstract

Diese Arbeit umfasst die Methoden-Erstellung zur Erforschung der Butanol-Bildung von *Clostridium acetobutylicum* über *in silico* Modelle. Zwei dieser Modelle werden hier ausführlich besprochen und die notwendige Basis aus mathematischen Formalismen für die Evaluierung sowie informationstechnischen Herangehensweisen eingeführt, wie z.B. die Etablierung einer lokalen Datenbank der Clostridien-Annotation aus KEGG.

Das erste zu besprechende Modell ist ein statisches Pathway-Model, das es ermöglicht Transkriptom-Daten unter Zuhilfenahme eines metabolischen Netzwerkes darzustellen und zu analysieren. Insbesondere wird eine boolesche Logik diskutiert, die die Daten-Integration vollzieht. Charakteristische Eigenschaften des Netzwerkes werden so hervorgehoben und metabolische Zustände visuell zugänglich gemacht. Die Bildung experimenteller Hypothesen wird erleichtert: Hier wird die Möglichkeit einer 3-Hydroxybutyrate dehydrogenase Aktivität in *C. acetobutylicum* näher beleuchtet. Die resultierende Prioritätenliste des Annotations-Transfers beinhaltet sowohl funktionale als auch regulatorische Informationen aus Datenbanken und Experimenten und bietet somit einen optimalen Startpunkt, experimentelle Forschung zu initiieren.

Das zweite zu besprechende Modell ist ein dynamisches Modell, das benutzt wird, um Metabolom- und Transkriptom-Daten aus Fermentations-Experimenten zu vereinigen. Seine besondere Struktur ermöglicht eine Vielzahl neuartiger Analysen - bringt aber ebenso neuartige Probleme mit sich. Die Resultate deuten darauf hin, dass der pH-Shift in *C. acetobutylicm* allein von der Dynamik der Transkriptom-Daten abhängt. Optimierungsstrategien auf Transkript-Ebene und auf Parameter-Ebene des Modells werden implementiert und ihre Resultate diskutiert.

Schließlich wird über Hauptkomponenten-Analyse sowohl eine Methode zur Optimierung der Laufzeiten eines solchen Modells gegeben als auch zwei neue Methoden geschlussfolgert: Eine, um die zeitliche Dynamik der Transkriptom-Daten geeignet zu variieren ohne regulatorische Profile zu verändern, die andere um Gene anhand ihrer regulatorischen Ähnlichkeit zu klassifizieren.

Contents

List of Abbreviations and Symbols

Δ_G	network diameter
γ	parameter indicating community structure from the distribution of node neighbours
ϱ_X	cell density
b	boundary for up-regulation or down-regulation
C	counter of genes that occur in subsets from annotation methods
c_i, c_o	concentration of intracellular or extracellular compound
c_X	biomass concentration
$C_{ecc/clo/rad/str/bet}$	node centralities
$e \in \mathcal{E}$	edges in the MMM, reactions
F	pump rate
G, G^a, H	Graphs of the MMM, standard, augmented, standard & augmented
$J(a, b)$	Jaccard index between objects a and b
$j = 1..N_J$	gene numbers
$k_{11\|2}$	maximal conversion rate of acetate kinase
$k_{1\|3}$	maximal conversion rate of thiolase
$k_{1\|9}$	maximal conversion rate of ethanol dehydrogenase

$k_{1\mid 11}$	maximal conversion rate of phosphotransacetylase
$k_{23\mid 17}$	maximal conversion rate of the CoA-transferase that uses acetic acid
$k_{3\mid 4}$	maximal conversion rate of lumped reaction of 3-hydroxybutyry-CoA dehyrogenase, butyryl-CoA dehydrogenase and 3-hydroxybutyryl-CoA dehydrogenase
$k_{4\mid 10}$	maximal conversion rate of butanol dehydrogenase
$k_{4\mid 5}$	maximal conversion rate of phosphotransbutyrylase
$k_{5\mid 6}$	maximal conversion rate of butyrate kinase
$k_{63\mid 47}$	maximal conversion rate of the CoA-transferase that uses butyric acid
$k_{x\mid y}$	maximal conversion rate that uses substrate x and transforms it into product y
m_X	biomass weight
$n_{\mathcal{P}}((a,b))$	similarity of two genes a,b according to their Pfam-motifs
n_i, n_o	amount of intracellular or extracellular compound
$Pfa, \mathcal{P}$	function that maps genes onto pfam-motifs
R^2	coefficient of determination
R_{sink}	loss of metabolites by outflow
r_i, r_o	reaction rate of intracellular/ extracellular compound
r_{sxp}	reaction rate from substrate s to product p
$Rct, \mathcal{R}$	function that maps genes onto reactions
s	integrated similarity score from multiple methods
$v \in \mathcal{V}$	nodes in the MMM, compounds
V_C, V_R, V_L	cell volume, reactor volume, liquid volume
$x \in \mathcal{X}, \mathcal{X}_0, \mathcal{X}_{M_i}$	all genes, genes without reaction annotation, genes as candidates retrieved by Method M_i
$x \in \mathcal{X}^b(s)$	genes being up-regulated
$x \in \mathcal{X}_b(s)$	genes being down-regulated

3-HBDH	3-hydroxybutyrate dehydrogenase
B. subtilis/BS	*Bacillus subtilis*
ABE	Acetone-Butanol-Ethanol
Ac	acetate
Ac-CoA	acetyl-CoA
Ac-P	acetyl-phosphate
AcAc	acetoacetate
AcAc-CoA	acetoacetyl-CoA
ack	acetate kinase
AcOn	acetone
bcd	butyryl-CoA dehydrogenase
BdhAB,AdhE	alcohol dehydrogenases
BHBu-CoA	3-hydroxybutyryl-CoA
BL3D	BioLayout Express 3D
BLAST	Basic Local Alignment Search Tool
bn	billion
BRENDA	BRaunschweig ENzyme DAtabase
BSU	KEGG gene identifiers for *B. subtilis*
Bu	butyrate
Bu-CoA	butyryl-CoA
Bu-P	butyryl-phosphate
buk	butyrate kinase
BuOH	butanol
CA	*C. acetobutylicum*
CA_C/CA_P	KEGG gene identifiers for *C. acetobutylicum* genes on chromsome or plasmid
CNA	CellNetAnalyser

CPD:	KEGG compound identifiers
crt	crotonase
Crt-CoA	crotonyl-CoA
ctfAB	CoA-transferase A/B
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
E/N	edges to nodes fraction
EtOH	ethanol
FBA	Flux Balance Analysis
G(s)	graph at state s derived from up-regulation
H(s)	graph at state s derived from up-regulation and augmentation
hbd	hydroxybutyrate dehydrogenase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD_{50}	lethal dose
MCL	Markov Clustering
MetaCyc	Encyclopedia of Metabolic Pathways
MJ	Megajoule
MMM	Metabolite-Metabolite-Mapping
OD	optical density
OECD	Organisation for Economic Co-operation and Development
ORF	open reading frame
p.a.	per annum
PCA	Principal Component Analysis
PEP	phosphoenolpyruvic acid
pSOL1	Megaplasmid of C. acetobutylicum

PTS	phosphotransferase system
RN:	KEGG reaction identifiers
RVP	Reid Vapour Pressur
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SED-ML	Simulation Experiment Description Markup Language
SOAP	Simple Object Access Protocol
thl	thiolase
yEd	Graph Editor by yWorks

List of Figures

List of Tables

Chapter 1

Preliminaries

They both savoured the strange warm glow
of being much more ignorant than ordinary people,
who were ignorant of only ordinary things.
Terry Pratchett

This chapter introduces the two aspects in which this thesis in embedded, a problem-centred project, SysMO-COSMIC2, and a systems biological approach that integrates different types of data.

On Usefulness

Systems biological strategies for several organisms were funded by the transnational initiative SysMO to enhance European wide collaborations, This initiative is divided into several sub-projects, one of them is SysMO-COSMIC2. This project is the starting point of this dissertation because it follows an engineering approach on a well known, widely treated problem: renewable energy production and chemical key compound generation. In the scope of dwindling crude oil reserves this problem requires recapitulation with modern biological and information technological techniques. Such a sustainable technology is under development by investigating the fermentation of *Clostridium acetobutylicum* and optimizing its productivity.

On Work Distribution in SysMO-COSMIC

COSMIC funding was sustained over two periods of three years each. While in the first period, development of suitable standard operating procedures (SOPs), development of cloning techniques and fermentations of the wildtype of *Cl. acetobutylicum* were in focus of research, the second period was used to emphasise on mutant generation and mutant fermentation.

The development of a SOP for fermentations was necessary to allow comparison of different experiments carried out at different sites. This procedure describes the set up of a continuous chemostat culture that is shifted between two distinct metabolic states by fermenting at two different pH values. The recorded responses are thought to give major insights into the regulatory mechanisms of the organism. This work was distributed along three modelling groups and five experimental groups and included the generation of suitable mutation technology, the culturing of mutant strains, the establishment of downstream protocols and the generation of models to describe and design experiments.

On Data Deposit

Parallel to the proposition of experiments, a computer scientific issue is challenged within all SysMO projects: Quantity and size of data require meaningful ways of organization in a database, of annotation and of standardisation. Identically, modelling approaches require description of the inherent model structure, graphical representation of the model details, and simulation annotation. To tackle this problem a sustainable platform called SysMO-SEEK was established to allow exchange and future use of data and models. This is achieved by implementing several standard formats.

On Standards

The System's Biology Markup Language (SBML) [Hucka et al., 2003] aims at an unified description of biological models. It is a XML-based format, distributed as level 3, and it is widely acknowledged. Usually, model deposition into online resources as e.g. Biomodels [Li et al., 2010] requires SBML format. Rapid browsing through published models, downloading them and checking the published results is one feature of this standard format. Reproducibility of published results can be ensured using SED-ML, the Simulation Experiment Description Markup Language [Waltemath et al., 2011]. Standardisation of graphical representations recently started by the use of SBGN, the Systems Biology Graphical Notation [Klipp et al., 2007]. Finally, integration of software tools into web-pages, like JWS [Snoep and Olivier, 2002] in SysMO-SEEK allow also the experimentalist to access and use models. Also the use of Taverna, which will be used later (3.1) is only possible through this standardisation. The ongoing research on suitable SBML features provides also a framework to standardise experiments. It will be shown that the transferability of XML formats between different tools is an issue however [Alves et al., 2006] and also SBML is not yet flexible enough to consider all models of a certain type. Lately, efforts were successful in unifying standards for pathway models in with a plug-in for Cytoscape [Shannon et al., 2003] which is named BiNoM. It is a promising tool for integration of data and interoperability of standards [Bonnet et al., 2013].

On Automation

The different approaches for data-analysis, data-curation and data-integration easily outnumbers the data-creation effort [Palsson and Zengler, 2010]. Automation represents a possible tool to increase reproducibility of results and reduce this effort. It will be shown that even knowledge discovery can be undertaken by an automation approach [Aksenov et al., 2005]. Therefor, the scope of this work is broaded. It is focussed on *Cl. acetobutylicum* research but it may be equally used for any other organism.

On Storage of Results

Data and models presented in this thesis will be linked to my personal profile[*] on SEEK.

Reading of The Thesis

I encourage the electronic reading of this thesis, because the representation of large networks is only poor in printed style. High zooms are supported by most

[*] https://seek.sysmo-db.org/people/319

of the images. A digital copy[†] is deposited in SEEK.
The three main chapters are not strongly dependent on each other, cross-references are given when necessary. Technical details and equations are given in the appendices.

[†] https://seek.sysmo-db.org/presentations/88

Chapter 2

Introduction

Here's what I think the truth is:
We are all addicts of fossil fuels in a state of denial,
about to face cold turkey.
Kurt Vonnegut

This introduction is dedicated to give the background of this thesis, microbial butanol production. First, it will give a historical motivation for the research of the biological process that uses *Clostridium acetobutylicum* (2.1). Past and current studies focussed on the biological and biochemical factors involved in butanol synthesis (2.2). Also, a bouquet of process variants were researched to increase productivity (2.3). Combining these information will help in focussing and understanding several key experiments from literature. These acquired data will play a dominant role in the following three chapters (2.4).
Finally the thesis proposal (2.5) and the thesis outline (2.6) will be given.

2.1 Motivation of Research

This thesis' core is the production of one chemical compound, butanol. From the overview of its economical relevance (2.1.1), the necessity of alternative production routes becomes obvious. Historically, a similar scenario was present when oil refinery industry had not yet been developed. In these times, a biological process had been used to generate acetone. It was named after its main products, acetone, butanol and ethanol, the ABE-fermentation. This historical process was already subjected to several optimisation approaches (2.1.2).

2.1.1 Butanol Resource Management

The Uses of Butanol

Butanol (CAS:71-36-3) serves as gasoline additive [Duerre, 2007] and as intermediate for the chemical production of acrylates, glycol ethers, resins, and various esters. It further serves as solvent for various products, e.g. paints, gums, fats, waxes, rubber, as a swelling agent and colour carrier in textile industry and as an extraction agent for various drugs, antibiotics and hormones. It is also an additive for the cosmetic and the cleaning industry [Company, 2006, SE, 2008]. In 2011, the market volume of butanol is accounted to be 3 million tonnes, an increase by 2.1 % compared to the previous year [Tanya Rezler, 2012]. The total market price is estimated to rise from \$5.9 *bn* from 2011 to \$9.2 *bn* in 2015.

The Chemical Production of Butanol Relies on Oil Resources

The chemical process of butanol synthesis comprises the hydroformylation of propylene with carbon monoxide to butyraldehyde in the presence of a rhodium-based catalyst, which is followed by hydrogenation of the aldehyde to the alcohol [Siegel and Himmele, 1980]. In contrast to this oxosynthesis, the Reppe synthesis directly produces butanol from propylene, it is however more expensive [Lee et al., 2008b]. Since propylene is a product of the oil cracking process [Li et al., 2007], butanol stands in direct relation with the availability of oil resources.

New Opportunities Through Rising Oil Prices

The annual report of British Petroleum (BP) from 06/2012 [Dudley, 2012b] summarises that the price for a barrel oil rose by 40 % from 2010 to 2011, which makes the actual price \$111.26 per barrel. This dramatic increase is not only the result of dwindling crude oil resources, it can be also understood as a question for a new process design of fuel production, that is covering both, world wide demand and economical feasibility. Demand is increasing slower than supply (0.6 million barrels per day vs 1.3 million barrels per day) which corresponds to growth by

0.3 % and 1.3 %, respectively. The largest increase in consumption was registered in China, 505.000 barrels/day. These numbers may sound enthusiastic, yet, projections from 01/2012 into the year 2030 estimate the increase of the global energy consumption by 1.6 % p.a, leading to an increase by 39 % compared to nowadays [Dudley, 2012a]. The major increase is due to non OECD-countries. Any estimate that the consumption of raw materials may be regressive would be a fatal error and new supply methods must be found. A second major issue is the import dependency of crude oil for any country. Europe's market is constantly and almost entirely relying on the import of oil (94 % in 2030) and this dependency is even increasing for gas (80 %). Industrial research consequently increases investments into renewables, which are the fastest growing fuels by 8.2 % p.a. Butanol as biofuel has a market value of $2.5 per gallon [Pfromm et al., 2010], while for the chemical industry this price increases to $5-6 per gallon [Doris de Guzman, 2011].

Is Bio-Butanol a Competitor of Bio-Ethanol?

Bio-butanol production is considered an antagonistic product to bio-ethanol: As fuel additive it has superior properties compared to ethanol. The Reid vapour pressure (RVP) which is a measure of evaporative emissions is decreased by a factor of 6 which makes butanol safer to handle. More importantly, its improved hydrophobicity helps blending at higher concentrations with gasoline. Engine modifications are not required. Ethanol can only be blended up to 85 %. Hydrophobicity and decreased corrosiveness to metallic compounds of the pipelines make butanol an ecologically safer compound regarding ground water. Last but not least, the caloric value of butanol is just 10 % less than gasoline, and 50 % higher than ethanol [Brekke, 2007, Duerre, 2007]. Environmental risk estimations consider it as readily biodegradable under aerobic conditions. Acute toxicity in water is reached at 0.5 $\frac{g}{L}$. Importantly, it has low potential to accumulate in a biological system. Studies showed that in animal systems the LD_{50} ranged between $0.8 - 4\ g/kg$ body mass. *In vivo* hydrolysis of butanol occurs fast, $20min$ after application of radioactive butyl acetate (30.2mg/kg of body weight) to rats, the hydrolysis product butanol was not detectable anymore [Hernandez, 2004].
Nevertheless, it is argued that the current production process can not compete with bio-ethanol production as long as feed-stock costs represent a major factor of the production costs. A megajoule energy costs $0.07 for butanol and $0.03 for ethanol. The authors further argue that the current process robustness is not sufficient for an industrial scale production [Pfromm et al., 2010].

2.1.2 The Historical ABE Production Process

This section summarises the very exhaustive review by Jones and Wood [Jones and Woods, 1986] if not stated otherwise.

Research Aimed at Acetone as Primary Product

Originating from a research project for rubber synthesis in Great Britain, in the period from 1912-1914, Chaim Weizmann isolated a butanol and acetone producing strain that was able to grow on potato starch and a broad range of other polysaccharide substrates, as root crops, nitrogen-fixing legumes, cereal crops and, more generally agricultural soil. The resulting patent became the initial point of research for the following generations. Acetone as a base chemical for colloidal production had economical priority. The historical process of acetone production took calcium acetate and disrupted it into calcium oxide and acetone with the help of heat. Since during the First World War imports of calcium acetate were stopped, the British economy had to find alternative production routes, and the several clostridial species were promising candidates (table 2.1), one of them is nowadays in focus of research, *Clostridium acetobutylicum*.

Table 2.1: Clostridium species of interest to chemical industry in the DSMZ database

Species	DSMZ ID
C. acetobutylicum	792
C. saccharobutylicum	13864
C. butyricum	10702
C. pasteurianum	525
C. beijerinckii	791
C. tyrobutyricum	2637

Competition Inspired Optimization of Substrate Utilisation

The research focus has been the growth of *Clostridium acetobutylicium* on a multitude of substrates: monosaccharides like lactose, polysaccharides like cellulose, and complex substrates like maize, waste sulfite liquor from paper industry and molasses [Beesch, 1952]. It is not surprising that consequently a huge effort was spent on optimal media research.

A first production plant for acetone production was erected in 1915, in the later years similar plants followed in Canada, India, France and the United States. The rising of automobile industry around 1920 led to an increased demand of butanol which was hitherto an unwanted by-product of acetone formation. In parallel, petrol-industry developed, and competitiveness of the process became an issue. This inspired research on cultures using starch more efficiently as energy source. Little success was granted to this approach and research was diverted to the investigation of fermentations on a multitude of monosaccharides from hydrolysates of complex carbon sources. The found strain CSC no.8 was able to ferment up to 6.5 % of the sugars, leading to a butanol yield of 2 %. Increasing this yield became a major task in research. In particular for Great Britain that was heavily depend on import, the different carbon sources became the limiting factor. 60 % of the butanol production costs were caused exclusively by the substrate.

Competitiveness of Bio-Butanol Lasted Until the End of World War II

The improvement of strains remained an issue until the 2nd World War where the demand of acetone again drastically increased. As a result, semi-continuous fermentations were operated with a multi-column continuous distillation downstream to extract alcohols. Until 1960 the use of fermentation as production route virtually ceased in all Western Countries, a plant in South Africa remained operational until 1983.

2.2 Biological Facts

This section will deal with the general introduction to the biology of *C. acetobutylicum* (2.2.1), which is then followed by a summary of biochemical production pathways (2.2.2) and genetic optimisation of strains (2.2.3).

2.2.1 What is Clostridium acetobutylicum?

Clostridium acetobutylicum is a member of the firmicutes genus. It is an obligate-anaerobe, Gram-positive and spore forming organism that is able to ferment a variety of different sugars and convert them to acetic acid, butyric acid and solvents as acetone, butanol and ethanol in the typical ABE-fermentation [Duerre, 2005]. Its genome sequence was recorded and annotated in 2001 [Noelling et al., 2001]. It consists of one main chromosome (3.94 Mb) and a mega plasmid pSOL1 (192 kb), which contain 3740 protein-coding open reading frames and 107 RNA genes.
The life cycle consists of three distinct phases [Luetke-Eversloh and Bahl, 2011]:

- acidogenesis: In this phase the cells are exponentially growing and the products acetic acid and butyric acid prevail.

- solventogenesis: In this phase the cells take up the excreted acids and metabolise them to the corresponding alcohols, ethanol and butanol, as well as acetone, in a fixed ratio depending on the substrate. For glucose as substrate the ratio of ABE products is 3:6:1.

- sporulation: In this phase, productivity ceases and cells transform into a durable state until environmental conditions ameliorate.

The major part of solventogenic genes is located on the pSOL1 plasmid [Grimmler et al., 2011]. Losing the plasmid thereby results into a solvent negative strain [Rogers, 2002]. In order to prevent this loss, it is required to apply several stress parameters on the cultures e.g. addition of acids, decreases in pH, changes in dilution rate or temperature [Barbeau et al., 1988].

2.2.2 Biochemical Pathways

Carbohydrate Uptake

There are two distinct uptake systems in bacteria, either ion channel mediated uptake along an ion gradient, mainly H^+ and Na^+ ions, or active import by cleavage of high-energy bonds, mainly ATP or phosphoenolpyruvate (PEP). The latter of both mechanism is the predominant in clostridial species. A multitude of sugars can be imported by different phosphotransferase systems (PTS), in total 13 systems are known in *C. acetobutylicum*, including one on the pSOL1 plasmid [Duerre, 2005, p.155ff]. It was shown, that glucose uptake is pH dependent [Yerushalmi et al., 1986b].

Glycolysis and Acid and Solvent Pathways

Biochemical studies on glycolysis are rare in *C. acetobutylicum*, the pathway is mainly inferred from the genome sequence. Only the glyceraldehyde-3-phosphate dehydrogenase has been analysed [Duerre, 2005, p.675]. However, the main pathway for acid and solvent production are extensively studied. This section summarises [Duerre, 2005, p.671ff]. As shown in figure 2.1, activation of pyruvate (Pyr) to acetyl-CoA (Ac-CoA) is achieved via the pyruvate ferredoxin-oxidoreductase (pfor). This system produces hydrogen and carbondioxide. Lactate (La) is produced via a lactate deyhdrogenase (ldh). Acetyl-CoA is the key compound in acid and solvent production. Condensation of two molecules acetyl-CoA via thiolase (thlAB) leads to one molecule acetoacetyl-CoA (AcAc-CoA). Acetoacetyl-CoA

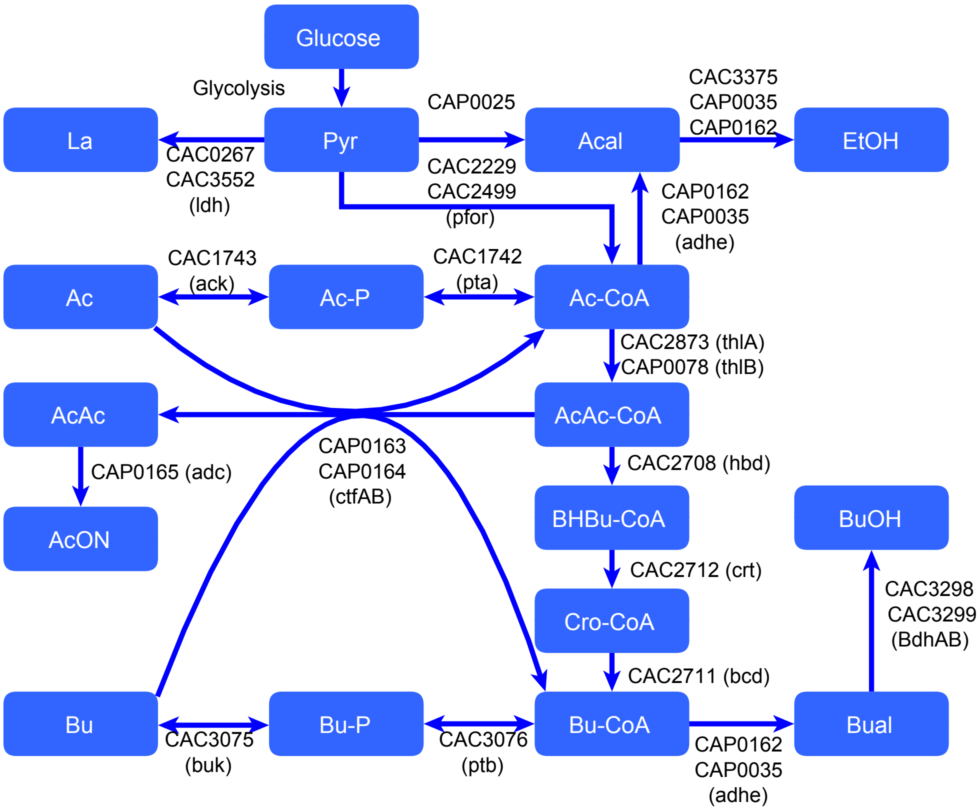

Figure 2.1: Production pathways of butyric and acetic acid and the solvents acetone, ethanol and butanol. Adopted from [Duerre, 2005, p.674] and [Lee et al., 2008b]. Abbreviations are explained in the text.

is converted to β-hydroxy-butyryl-CoA (BHBu-CoA) that is subsequently dehydrated to crotonyl-CoA (Crt-CoA) which then is converted to butyryl-CoA (Bu-CoA), via the three enzymes 3-hydroxybutyryl-CoA dehydrogenase, crotonase, butyryl-CoA dehydrogenase (hbd, crt, bcd) respectively. These three genes form an operon.

Depending whether acidogenic or solventogenic conditions prevail, the fluxes from acetyl-CoA and butyryl-CoA are diverted into different directions: During acidogenesis, each CoA-derivative is phosphorylated with inorganic phosphate by their respective phosphotransferase, phosphotransacetylase or phosphotransbutyrylase (pta and ptb), into acetyl-phosphate (Ac-P) and butyryl-phosphate (Bu-P). These phosphates act as donor for the reaction of two distinct kinases, acetate kinase (ack) or butyrate kinase (buk) to generate ATP and the acids, acetate and butyrate (Ac and Bu).

During solventogenesis, acid re-uptake acts either by the reverse reaction of the kinases, or via an acetoacetyl-CoA: acetate/butyrate-coenzyme A transferase (ctfAB) consisting of two subunits. ctfAB accepts acetoacetyl-CoA as CoA-donor and transfers the CoA to the acids, the products are the respectve CoA-acid derivative and acetoacetate. Acetoacetate decarboxylase (adc) acts on acetoacetate and produces acetone (AcON). The two CoA-derivatives are reduced to their respective aldehydes (adhE) and alcohols via unspecific alcohol dehydrogenases (BdhAB). A complete view on reaction mechanisms of these enzymes is given elsewhere [Gheshlaghi, 2009].

Uptake of Acids

The investigation of the effect of propionic and acetid acid uptake on the metabolic spectrum during batch fermentation showed that two pathways are active, the CoA-transferase pathway and the kinase-phosphotransbutyrylase pathway. The latter was dominant and led to an increase of solvent yields [Huesemann and Papoutsakis, 1990]. Phosphotransbutyrylase acts on its substrates in a ping-pong like mechanism. Its activity is highly sensitive to pH changes in the physiological range, mainly in the butyryl-phosphate direction but not so much in the butyryl-CoA direction. Further it is inhibited by ATP, suggesting a role of ATP for the determination of reaction direction [Wiesenborn et al., 1989]. Both re-uptake paths of acids are confirmed by several pulse experiments in batch culture. Flux balance analysis showed that in batch culture the short-term response to an acetic acid pulse in acidogenesis, the flux from acetyl-CoA to acetoacetyl-CoA is increased, but not the CoA-transferase activity. The long-term response showed increased acetone and reduced butanol concentrations. A butyrate pulse in chemostat culture showed that butyrate is mainly taken up via the CoA-transferase pathway, however the butyrate synthesis still prevailed [Junne, 2010].

The Metabolic Shift

The metabolic shift is the transitional phase *C. acetobutylicum* undergoes when acidogenesis halts and solventogenesis starts. It is reversible [Haus et al., 2011] and acidogenic conditions are maintained above pH 5.7 in continuous culture, solventogenic conditions are established below a pH of 4.3 [Janssen et al., 2010]. The conversion of acids into solvents is seen as a de-acidification mechanism [Monot et al., 1984, Huang et al., 1985]. Since the internal pH of *C. acetobutylicum* cannot be maintained at a constant level [Gottwald and Gottschalk, 1985], this suggests that the switch is linked to the pH. Indeed, in pH-uncontrolled culture a surplus in acid production or addition of high amounts of acetic acid (up to 200 mM) cause an acid-crash which blocks solventogenesis [Maddox et al., 2000, Cho et al., 2012]. These results were confirmed by addition of less then 2 mM formic acid in two separate studies [Wang et al., 2011, Cho et al., 2012]. The importance of internal pH and the pH-gradient between cell and the reactor are extensively discussed elsewhere [Papoutsakis et al., 1987].

As undissociated butyrate concentrations are linked to internal pH, its precursors like butyryl-phosphate may also have signalling function [Desai et al., 1999, Paredes et al., 2005]. Indeed a butyrate kinase deletion mutant showed a significant earlier start of solventogenesis compared to the wildtype. Additionally, the concentration of butyryl-phosphate was bimodal with one peak corresponding to solvent production and one corresponding to carboxylic acid reutilization [Zhao et al., 2005]. The onset of solventogenesis showed correlations with butyryl-CoA spikes in batch culture [Boynton et al., 1994] and to undissociated butyric acid levels around $6-13$ mM [Monot et al., 1984, Huang et al., 1985, Terracciano and Kashket, 1986, Huesemann and Papoutsakis, 1988]. A current study of a deletion mutant of the butyric acid production pathway re-evaluates the hypothetical role of butyrylphosphate and butyryl-CoA and it comes to the conclusion that neither is necessary for the onset of solventogenesis [Lehmann and Luetke-Eversloh, 2011]. Similarly, the role of acetic acid on the onset of the pH-shift is also controversial. Studies suggest both, that internal acetic acid has no effect [Bahl et al., 1982a, Zhao et al., 2005], while acetate stimulus experiments suggest it has an effect [Junne, 2010].

The alteration of electron flow is suspected to induce the shift [Meyer et al., 1986, Rao and Mutharasan, 1987]. Also a characteristic increase of the redox-potential is observed during the shift [Grupe and Gottschalk, 1992, Peguin et al., 1994]. The deletion mutant of one of the two acid kinases and the CoA-transferase is unable to produce solvents, the authors consider this a result of the inability to control electron flow of *C. acetobutylicum M5* [Sillers et al., 2008]

The DNA-topology is influenced by different environmental conditions and it was shown that relaxation of the coiling increased acetoacetate decarboxylase transcription [Duerre, 2005, p.680].

Finally, sigma factors are suspected to play a significant role, in particular Spo0A phosphorylation has a primary role in the onset of solventogenesis. Spo0A deletion mutants neither show sporulation nor butanol production [Ravagnani et al., 2000]. However, over-expression of this gene did not alter the onset time of solventogenesis [Alsaker et al., 2004]. The general organization and regulation of solventogenic genes in operons is reviewed by [Duerre et al., 2002].

In *C. acetobutylicum P262*, acidogenesis and solventogenesis seem to operate in parallel - the cells undergo cyclic changes in productivity during a chemostat at different dilution rates [Clarke et al., 1988]. The authors suggest, that it is unlikely that both pathways are operating in parallel in one cell, a mixed population hypothesis may be reasonable. Current research re-investigates this hypothesis and models indicate such a possibility [Millat et al., 2013b].

2.2.3 Engineering a Butanol Production Strain

A summary of pathway remodelling approaches is given by [Lee et al., 2008b, Luetke-Eversloh and Bahl, 2011]. Butanol increase is reported by non-replicative plasmid inactivation of butyrate kinase [Green et al., 1996]. Combination of the butyrate-kinase mutant with over-expression of an alcohol dehydrogenase yielded a strain that produced 16.7 $\frac{g}{L}$ of butanol [Harris et al., 2000]. A transcriptional repressor for solvent synthesis was recognised and studied, its inactivation resulted in a deregulated solvent production strain with better yields of solvents [Nair et al., 1999]. Overexpression of the alcohol dehydrogenase and downregulation of the CoA transferase using antisense RNA (asRNA) yielded an increase of ethanol concentrations to 9 $\frac{g}{L}$ and butanol levels comparable to the wildtype [Tummala et al., 2003a].

In a corresponding study, thiolase and alcohol dehydrogenase were overexpressed by using the phosphotransbutyrylase promotor and again asRNA for ctfAB silencing, a higher selectivity of butanol to acetone was achieved. Solvent titers reached 30 $\frac{g}{L}$ [Sillers et al., 2009].

Engineering of thiolase was performed using an *E. coli* library that was screened for optimal thiolase activity before it was inserted into *C. acetobutylicum*. The resulting strain showed less growth and optimised alcohol titers for both, ethanol and butanol [Mann and Luetke-Eversloh, 2013].

Deletion and overexpression analysis of Spo0A revealed a fundamental role in both, sporulation and initiation of solvent production as transcriptional regulator, which also regulates other sporulation factors, similar to *B. subtilis* [Harris et al., 2002, Thormann et al., 2002].

The newly developed ClosTron insertion mutation II technology [Heap et al., 2007] can be used to achieve highly reproducible knock-out mutants. Use of this system is reported for transforming *C. acetobutylicum* into an ethanol producer by inactivation of 3-hydroxybutyryl-CoA-dehydrogenase. Inactivation of this

whole pathway did neither alternate sporulation nor the onset of solventogenesis [Lehmann and Luetke-Eversloh, 2011]. Deletion of the phosphotransacetylase alone does not change acetate production and the deletion of the acetoacetate decarboxylase resulted in a drastically reduced acetone production. Combined deletion of both genes increased the flux to butyryl-CoA leading to butyrate [Lehmann et al., 2012a]. Inactivation of the acetate kinase alone did not alternate the solvent production either, the double knock-out of butyrate and acetate kinases is currently under investigation [Kuit et al., 2012]. The deletion of phosphotransbutyrylase showed as well high ethanol and butanol yields and a disruption of butyrate production, pH control was necessary to allow the metabolic shift. Metabolic profiles of mutants are pH-dependent: With no pH-control there is no acetone, and accumulation of acetate. With pH-control there is acetone production and re-uptake of acetate. It was equally shown that butyrate may be re-assimilated in the ptb mutant [Lehmann et al., 2012b].
A different metabolic approach converts acetone to isopropanol: Insertion of dehydrogenases from other species allow the production of more than 20 $\frac{g}{L}$ of alcohol [Lee et al., 2012, Dusseaux et al., 2013]. Earlier, other authors reported an increase from 2 $\frac{g}{L}$ *to* 18.8 $\frac{g}{L}$ by simultaneous disruption of both kinases and the insertion of a mutated alcohol dehydrogenase [Jang et al., 2012], which is in contrast to another study, who reported that their double knock-out strain did not produce butanol [Sillers et al., 2008].

2.2.4 Concurrent Designs

Concurrent processes are established, the butanol production apparatus of *C. acetobutylicum* is shuttled into *E. coli* and other bacteria. Although the productivity is low, the authors suggest a high potential of this approach [Inui et al., 2008, Nielsen et al., 2009]. Still, a complete unadapted organism faces the same challenges of butanol stress on the cell membrane than the better adapted *Clostridia*. A completely different approach to butanol production is performed by employing monoxygenases and butane as substrate [Duerre, 2005, p.685].On this process a patent is pending (patent EP 0 987 348 A1).

2.3 Butanol Fermentation

This section reviews the necessary technological and environmental parameters that allow a sustained butanol production in *C. acetobutylicum*

2.3.1 Production of Butanol

Media Composition

A minimal medium with proteins and glucose supplemented with p-aminobenzoic acid and biotin was sufficient to promote growth [Beesch, 1952]. When the fermentation was operated in glucose or ammonia limited mode, no solvents but only acids were produced. It was equally shown that butyric acid at a pH less than 5.0 could shift the culture to solventogenesis with improved ratios of products [Bahl et al., 1982a].
The same authors showed that a phosphate limited medium could produce superior results to the hitherto existent fermentations. They were able to ferment 54 $\frac{g}{L}$ of glucose to 10 $\frac{g}{L}$ of butanol and 4 $\frac{g}{L}$ of acetone [Bahl et al., 1982b].
The role of ions in the fermentation broth was elucidated in the same year by another group, who showed that iron, magnesium and potassium ions can promote growth, only magnesium and manganese ions had a deleterious effect when applied in excess [Monot et al., 1982]. Conversely, it was shown that iron limitation and viologene addition as redox-agent enhance the butanol yield [Peguin and Soucaille, 1995].

Substrates, Product Yields and Product Spectra

The typical ratio of 6:3:1 of butanol, acetone, ethanol is reached on starch, saccharose, xylose and fructose, while a ratio of 5:4:1 is reached on arabinose [Beesch, 1952]. Using a mixture of glucose and xylose (1:1), it was found that uptake of xylose seems repressed by glucose uptake, since xylose but not glucose accumulated in the medium [Fond et al., 1986]. Solvent productivity on all the three, glucose, xylose and its mixture was at most 0.8 $\frac{g}{Lh}$, 0.58 $\frac{g}{Lh}$ and 0.94 $\frac{g}{Lh}$.
Another mixed substrate fermentation of glucose and low-grade glycerol (1:1 and 2:1) in a chemostat showed that butanol was the major endproduct (43 % and 63 %) and the culture could be maintained stable over 70 days. In two experiments, glucose was entirely consumed (15 $\frac{g}{L}$ and 30 $\frac{g}{L}$) and 43 % of added glycerol was used. Solvent productivity was assessed as 0.47 $\frac{g}{Lh}$ [Andrade and Vasconcelos, 2003].
Grown solely on glycerol, 1,3-propanediol is formed, when grown on rhamnose, also 1,3-propanediol, propionic acid and propanol are produced [Forsberg, 1987].
Complex substrates like corn fibre yield lower solvent productivity (0.2 $\frac{g}{Lh}$ to 0.4 $\frac{g}{Lh}$) and they require addition of xylanases to successfully start the batch fermentation. Uptake of glucose and arabinose was superior to uptake of galactose

and xylose [Qureshi et al., 2006]. Similarly, a complex substrate mixture from hydrolysates of distillers grain was studied on different clostridial species. *Clostridium acetobutylicum* could reach a solvent productivity in batch operation mode of about 0.25 $\frac{g}{Lh}$, fermenting the following sugars in decreasing order of preference: glucose, arabinose, galactose, cellobiose, xylose, mannose. Inhibitory effects of the hydrolysis products syringaldehyde, ferulic and p-coumaric acid, as well as the growth stimulating effect of furfural and hydroxymethyl furfural were demonstrated in this study [Ezeji and Blaschek, 2008].
Input of carbondioxide to the fermentation by gassing inhibits dehydrogenase activity and increases butanol yields [Kim et al., 1984]. In a glucose limited chemostat, carbon monoxide gassing leads to decreased growth rates but increased glucose uptake, with no acetone but sustained butanol production of 0.74 $\frac{g}{Lh}$ [Meyer et al., 1986].

2.3.2 Counteracting the Effects of Butanol

Butanol Effects

The accumulation of butyrate and acetate in the membrane does not create massive cell leakage [Huang et al., 1985]. Butanol however has a chaotropic effect on the membrane which results in early cell death and small productivity. It is one major focus in optimisation proceedings. It was shown that the internal pH could not be maintained and there was leakage of ATP [Bowles and Ellefson, 1985, Balodimos et al., 1988] or PEP and therefore a stop of the glycolytic flux [Gheshlaghi, 2009]. In contrast, a metabolome study showed that addition of 5 $\frac{g}{L}$ butanol did not drastically affect intracellular metabolite pools [Amador-Noguez et al., 2011]. This is confirmed by two butanol stress experiments in continuous culture, a pulse experiment [Janssen et al., 2012] and a stepwise forcing experiment [Schwarz et al., 2012]. Pulse experiments with butanol in batch culture during the acidogenic and solventogenic phase increased acetate uptake in the short term, but in the long term every reaction is reduced [Junne, 2010]. As a result of butanol stress, the fraction of saturated to unsaturated fatty acids showed a dose-dependent increase. Conversely, butanol challenges of 0.25 % vol/vol and 0.75 % vol/vol were used to study the tolerance of two strains, the pSOL1 deletion strain as control strain and the strain 824(pGROE1) that contains the chaperone system groESL under a thiolase promotor. Butanol addition increased the expression of the major stress responses and the solvent formation genes, while it decreased the expression of genes for fatty acid synthesis and glycolysis [Tomas et al., 2004].
A library enrichment study allows the identification of genes conveying butanol tolerance. It was undertaken by transferring stationary phase cultures into media with increased butanol concentrations. Strains containing a plasmid with the

CA_{C1869} gene showed a 81% increase in butanol tolerance compared to the wild type. It grew to higher cell densities and showed a prolonged metabolism. It remained an open question which regulations were changed due to over-expression of this gene [Borden and Papoutsakis, 2007].

Fermentation Operation Modes

A review on butanol toxicity and possibilities to overcome it, is given by [Ezeji et al., 2010].

First attempts to reduce the chaotropic effects of butanol were started by using liquid-liquid extraction with n-decanol saturated with butyric acid. A continuously operated membrane bioreactor was therefore connected to a mixer-settler cascade. A fourfold increase of butanol (0.51 $\frac{g}{Lh}$ to 1.96 $\frac{g}{Lh}$) could be noted. Direct contact to the decanol phase caused cell damage however. Additionally, butyric acid saturation of the extraction phase was necessary to prevent its removal from the fermentation process [Eckert and Schuegerl, 1987].

Pervaporation, that is evaporation of a liquid after diffusion through a membrane, was used to remove butanol in chemostat culture. It increased the product formation rate to 2.34 $\frac{g}{Lh}$ and higher [Izak et al., 2008].

Perstraction is a similar process where butanol is allowed to diffuse through a permeable membrane into an ionic liquid. An increase of butanol production from 0.057 $\frac{g}{Lh}$ to 0.21 $\frac{g}{Lh}$ was reached. An eightfold higher amount of lactose could be fermented during this batch fermentation compared to the usual batch operation mode [Qureshi and Maddox, 2005].

In all these methods the fermentation broth is in direct contact to the membrane and fouling of the membrane becomes an issue. A method like vacuum product recovery overcomes this problem. Although butanol has a higher boiling point, the azeotropic mixture with the other alcohols in the fermentation broth leads to a better vaporisation. This approach allows the cells to completely utilise glucose for higher growth and higher concentrated product streams [Mariano et al., 2011].

Biocatalysis by immobilised cells allows the conversion of substrate to the desired product, in a first attempt *C. acetobutylicum* was supplied with a medium that did not support cell growth and effects of different feeding stream metabolites were investigated. The packed-bed reactor was run in continuous mode, cells were immobilised in alginate beads. Productivity of this system was lowered by sporulation and cell death, still butyrate supply of 2 $\frac{g}{L}$ allowed production of 1.9 $\frac{g}{L}$ butanol after 10 h of cultivation. Activity regeneration in the same system can be reached by supply of ammonia and vitamins in the feed [Reardon and Bailey, 1989, Reardon and Bailey, 1992]. In a different setting, immobilised cells were examined for their extracellular α-amylase activity on starch in the feed flow. The total solvent yield was 1 $\frac{g}{L}$ [Badr et al., 2001]. Finally, biomass recycling was used to increase butanol yields. It was shown that under non-glucose limiting conditions recycling lowers the ATP demand and increases solvent yields, under limiting conditions only higher yields of acids were reached. In general, a range of total solvent productivity between 4.2 $\frac{g}{Lh}$ to 6.5 $\frac{g}{Lh}$ were reached and the authors proposed an experiment that would ultimately lead to a total solvent productivity of 12.4 $\frac{g}{Lh}$ [Meyer and Papoutsakis, 1989].

2.4 Published Data

This works aims at integration of transcriptome data and metabolome data and the evaluation of database information in view of transcriptome data. A short introduction into both omics will be given here (2.4.1). For completeness, some proteome studies are also mentioned. For a more in-depth introduction about opportunities and pitfalls, refer to this review on bacterial omics [Mashego et al., 2007].
In the second part of this section, three different experiments will be introduced in more detail (2.4.2). They will serve as data for the models that are going to be established throughout this thesis.

2.4.1 Analysis Techniques for Different Omics

A short review on omics is given by [Fiehn, 2001, Joyce and Palsson, 2006].

Metabolomics

The physico-chemical properties of the ABE-products allow an easy detection by gas chromatography [Fond et al., 1984, Green et al., 1996, Harris et al., 2000, Lehmann et al., 2012a]. However, since the evaporation of acids may impose some methodological problems, the use of a High-Performance Liquid Chromatography (HPLC) and a refractive index detector is also frequently encountered [Buday et al., 1990, Tomas et al., 2003, Tummala et al., 2003a, Kuit et al., 2012]. The coupling of a tandem mass spectrometer to the HPLC allows the determination of intracellular metabolites. This procedure requires several preparatory steps, e.g. rapid sampling and rapid quenching [Schaub et al., 2006, Schaedel and Franco-Lara, 2009]. Such approaches were used for determining intracellular metabolites of *E. coli* grown in C^{13}-glucose supplemented medium [Schaub, 2005, Bajad et al., 2006]. For *C. acetobutylicum*, one similar study of a batch culture is published. 121 metabolites were measured using a tandem mass spectrometry after addition of universally labelled C^{13}-glucose. Massive changes in all metabolites during the shift from acidogenesis to solventogenesis were observed. The carbon flux is redirected from biomass growth to solvent production [Amador-Noguez et al., 2011].
Online measurements procedures are published for metabolite analysis using a mid-infrared spectroscopy approach [Kansiz et al., 2001] and for redox balance determination using a fluorescent probe [Srivastava and Volesky, 1991].

Transcriptomics

The analysis of the complete transcriptome of *C. acetobutylicum* allows the temporal resolution nowadays. Numerous such data sets are available: Study of Spo0A overexpression [Alsaker et al., 2004], groESL overexpression [Tomas et al., 2003], ctfAB knockdown [Tummala et al., 2003b] and the transcriptional programme of sporulation [Alsaker and Papoutsakis, 2005, Jones et al., 2008] were performed. Responses to butanol addition [Alsaker et al., 2004, Alsaker et al., 2010, Janssen et al., 2012, Schwarz et al., 2012] and to several acids [Alsaker et al., 2010] were recorded. Reproduction of array results is usually undertaken by using a real-time PCR approach on few genes [Nolan et al., 2006, Lehmann and Luetke-Eversloh, 2011]. The quantities measured by both approaches are in general comparable [Dallas et al., 2005].

Proteomics

Stress response related proteins were detected using pulse-labelled proteins in a batch culture [Terracciano et al., 1988]. The proteome study of a phosphate-limited chemostat culture analysed 130 proteins and found 52 proteins being up-regulated two-fold during the onset of solventogenesis, and 34 proteins being downregulated by the same factor [Schaffer et al., 2002]. A more sensitive proteome protocol was developed and tested in a similar culture, yielding a resolution of over one thousand proteins on a 2D gel [Schwarz et al., 2007a]. In a phosphate-limited chemostat, 15 proteins could be specifically assigned to acidogenesis and 29 to solventogenesis [Janssen et al., 2010].

2.4.2 Data Treated Within This Thesis

Three different data sets will be shown here, the standard batch fermentation in complex medium, the acid stimulation experiments of that batch fermentation, and the continuous fermentation under phosphate limited conditions.

The Standard Batch Fermentation

The first fermentation in which metabolomic and transcriptomic data were both collected was done in the Papoutsakis laboratory [Alsaker and Papoutsakis, 2005, Jones et al., 2008]. The transcriptome data is large as it was collected over 25 samples along the whole fermentation time. A batch culture was grown in complex growth medium (CGM) and maintained at pH ≥ 5 until sporulation. Butyrate spiked after 16 h and butanol production started at the same time. The exponential growth became stationary and cells switched to solventogenesis. This fermentation yielded 11 $\frac{g}{L}$ of butanol after 45 h. Cells continued to morphologically change until 60h but productivity ceased. A K-means cluster algorithm split the transcriptional analysis data into five groups, corresponding to several phases, thus identifying the genes significantly up-regulated during these phases.

During the first phase, cell motility genes are up-regulated, as well as glucose transporter genes and nucleotide transporter genes.

The second phase is marked by an increase in energy production and the sporula-

tion relevant markers abrB and sinR are up-regulated.
The third cluster is overlapping with the second, however expression is sustained over a longer period and genes relevant for fatty acid biosynthesis and iron import are up-regulated. In this third cluster, the key solventogenic genes, Granulose formation genes and stress and heat shock proteins are up-regulated. Also branched amino acid synthesis seems up-regulated.
The fourth cluster contains numerous carbohydrate relevant uptake genes, and genes encoding for transport of inorganic ions and amino acids. Starch metabolism genes were also up-regulated and may be involved in granulose formation. Additionally, arginine biosynthesis genes were up-regulated, because arginine was probably depleted in the medium.

Acid Stimulus During Batch Fermentation

In the stimulation experiment of the Papoutsakis group, acetate, butyrate and butanol were added during the exponential phase [Alsaker et al., 2010]. Preparatory studies suggested that levels of 46, 78 and 107 mM of acetate, and levels of 17, 33 and 49 mM of butyrate negatively affect the metabolism. A summary of all metabolic effects is given in the paper.
While acetic acid stress (45 mM) down-regulates butyrate formation and vice versa, butyrate stress (50 mM) down-regulates acetate formation. Addition of acids up-regulated stress response and solventogenic genes but unexpectedly not sporulation relevant genes. Transporter proteins, post-translational modification proteins and energy metabolism genes are up-regulated after acetate and butyrate stress. The amino acid transport and metabolism show both, upregulation and downregulation. Spo0A is slightly up-regulated, but upregulation ceased for sporulation genes within 6 h. Glucokinase and the IIABC phosphotransferase-system were first highly expressed and later downregulated.

The Continuous Fermentation According to COSMIC-SOP

This section summarises the paper of. [Grimmler et al., 2011].
Acidogenic steady state conditions at pH 5.8 are reached after approximately 150 h, 25 mM of acetate are present and 51 mM of butyrate. After switching off pH-control, pH 4.5 is reached and then maintained actively by pH-control. Solventogenic conditions are thereby established, and 37 mM of butanol and 24 mM of acetone are produced. Ethanol concentrations remained unchanged in both conditions, 5 mM and 8 mM. Glucose showed a peak during the shift, from 51 mM under acidogenic conditions to 92 mM during the shift to 86 mM under solventogenic conditions. The peak of glucose corresponds to maximum concentrations of acids and optical density.
245 genes were differentially expressed and collected in 4 groups, up-regulated

genes under either acidogenesis (Group 1) and solventogenesis (Group 2), induced (Group 3) or repressed (Group 4) only during the shift:

Group 1 arginine biosynthesis, flagellin biosynthesis, acetyl-CoA conversion to crotonyl-CoA, alcohol dehydrogenase ,

Group 2 endoglucanases, glycerol-3-phosphate dehydrogenase, flavodoxin, cysteine and sulfur metabolism, fatty acid synthesis, glycosyltransferases, solventogenic genes, cellusomal-like genes

Group 3 pyruvate decarboxylase, stress response, predicted mannose uptake system

Group 4 carbon-monoxide dehydrogenase, glycerol-3-phosphate transport, tricarboxyacid cycle.

The increase of glucose during the shift is argued to reflect the increased carbon uptake of the organism through acid re-assimilation. Sporulation and solventogenesis are two separate events and also the transcription of stress related genes was already initiated when butanol was not present in the medium. Also, fatty acid synthesis may not be a consequence of butanol stress, as the relevant genes are up-regulated already during the transition.

2.5 Summary and Thesis Proposal

The necessity of researching alternative production routes for fuels in particular for butanol is widely accepted. An effective and competitive process would not only sustain the transition from fossil fuels to regenerative energy but it would equally offer independence of a national economy to imports of raw oil. Since from the historical perspective *C. acetobutylicum* was primarily an acetone producer, butanol production was not in focus. By considering the physico-chemical properties of butanol in combination to the lengthy optimisation approaches to overcome problems of process design, the design of an optimal production strain is a very demanding task and a serious disadvantage to this process. The different possibilities of genetic engineering propose a new approach to overcome the before mentioned problems.

In silico design of an organism helps in reducing the experimental workload associated with this engineering problem. It can suggest experiments on the one hand, and it procures the scientist with a tool to investigate and focus on key aspects of the system on the other hand. This thesis contributes to such a design by proposing approaches to the following tasks:

1. Investigation of data with a focus on the relationship of the reactome with the underlying proteomic or transcriptomic data profiles.

2. Hypothesis procurement for the filling of annotation gaps in the proteomic database of *C. acetobutlicum* based on database and data information.

3. Generation of a model that integrates transcriptome and metabolome data for the *in silico* description of different culture and process designs.

4. Hypothesis procurement for the directed generation of mutants from the estimated parameters of the dynamic model.

2.6 Outline

Chapter 3 provides answers for the first two tasks. It will present an automated modelling approach that integrates reaction annotations and temporal transcript level data from several experiments into a database. The evaluation of transcript data from a reactome point of view adds a further dimension for data evaluation besides gene annotation and may enhance readability of the data. From a graph-theoretic formalism approaches to investigate the transcriptome data are proposed. Reduction of complexity and several examples will be shown. Missing reaction annotations represent gaps in this database. The investigation how to find them and infer knowledge will occupy the second part. A comparative approach will be proposed that integrates again transcriptome data and database-information. This approach will be carried out for a specific enzyme, the 3-hydroxybutyrate dehydrogenase which is not annotated in *C. acetobutylicum* but in numerous other organisms.

Chapter 4 provides answers to the other two tasks by integrating time-series of transcript level data into a reaction network model of clostridial butanol synthesis. The model construction and model evaluation with respect to metabolomic data and mutation experiments occupies this chapter's first part. It will be shown that the metabolic shift can be solely explained through transcript dynamics without requiring considerations of pH. The second part then treats local and global sensitivity analysis and their use for finding an experimentally feasible optimal parameter sets that increases solvent productivity, which will be presented for several reactions.

Chapter 5 focusses on one implementation problem of this novel dynamic model type. High numerical effort for the calculation of the integrated transcript level profiles needs alleviation. Compression using the principal component analysis does not only allow to increase calculation speed, but it will also inspires a novel tool for model analysis in scope of optimisations by varying dynamic pattern in the data, and it will inspire a data analysis routine in scope of clustering of regulatory similar information.

Chapter 3

Automated Network Model Creation

The important thing in science
is not to so much to obtain new facts
as to discover new ways of thinking about them.
Sir William Bragg

Biological information tends to be very heterogeneous in its qualities: Transcriptome data stand beside the network structure of a biological pathway, protein sequences stand beside regulatory cascades. However, these information represent one same organism and belong together. Their integration represents one key challenge. This chapter proposes an automated procedure to answer some aspects of this challenge.

Starting from an overview of existing approaches of available pathway-databases for *C. acetobutylicum* (3.1), it is deduced that an independent and more flexible solution is required here. This database then is extended by integrating experimental transcript level time-series in a novel way (3.2). Some examples of derived models will be given in a graph format (3.3). Gaps in the present annotation represent a challenge that this model cannot cover. A strategy to overcome such gaps is annotation-transfer between species for which a methodology will be proposed (3.4). This will lead to the construction of hypotheses that can be experimentally verified and a second possibility to integrate a different type of data-base information, the enzyme annotation and Pfam-motifs, with transcriptome experiments. As case study the research of a 3-Hydroxybutyrate dehydrogenase in *C. acetobutylicum* will be undertaken (3.5). The conclusion is given in section 3.6.

3.1 Database-Harvesting

3.1.1 Existing Pathway-Models

A metabolic network, synonymously a pathway-model, is a mapping of compounds to compounds via reactions. Reactions are the result of an enzyme activity, which is the result of the protein product of a specific gene. The correct gene-reaction annotation is an ongoing process in the databases, e.g. not all reactions bear E.C. numbers, not all genes are identified as enzymes. Many tools are available to bridge some of such short-comes [Durot et al., 2009].

The online database MetRxn was only recently published. It harvests information from KEGG, BRENDA and MetaCyc. Extensive curation effort was put in this database by resolving naming inconsistencies, by balancing of stoichiometry and charges, and by reconciliation of missing information between the databases. The aim of MetRxn is the creation of pathway-models usable for flux balance analysis, the comparison of conserved reactions throughout species and the identification of all possible conversion routes from one substrate to one product. In total, 44 models of organisms are stored there [Kumar et al., 2012]. Notwithstanding the efforts, download of the pathway-model of *C. acetobutylicum* from MetRxn was not possible at writing time of this thesis and it could not be used. It consists of 430 metabolites connected by 363 reactions .

Earlier, a similar network was created and manually curated, containing 479 metabolites connected by 502 reactions [Lee et al., 2008a]. It aimed again at flux balance analysis and growth performance evaluation from fermentation data. Curation of the network was performed by several gap-filling procedures, as e.g. BLAST research of missing enzymes. Single gene deletion studies suggest that 194 reactions are essential, and 27 reactions are partially essential. Parallel to this work, other authors published their network [Senger and Papoutsakis, 2008]. It contains 422 metabolites in 552 reaction. Curation was performed using a maximal flux criterion for biomass production in conjunction with biomass constituting equations and a thermodynamical consideration of the free Gibbs energy. Flux balance analysis was again the aim and outcomes of single gene knock-outs were proposed.

All three authors made neither workflows for database creation nor the databases themselves publicly available. A reactome-knowledgebase has been established for various other organisms and allows integration of softwares, namely Cytoscape and R, as well as the interpretation of transcript expression data on the website [Matthews et al., 2009]. In the commercial software Insilico a published model for *C. acetobutylicum* exists that contains 507 reactions and 310 compounds.

3.1.2 Local Solutions

Published networks were not available and their construction tools were not accessible. Also the designation published softwares was very specific and did not suffice for several features that will be required later.

The first task for any local solution is to gain access to a database, e.g. KEGG. While there are several published softwares that access KEGG directly (e.g. YANAsquare [Schwarz et al., 2007b]), again none of them seems flexible enough to deal with the manifold of deposited information, nor was it possible to download entire genomes. A programme used for building pathway-models and integrate data is the very recently published RAVEN-toolbox [Agren et al., 2013]. Also the Vanted toolbox integrates data into SBGN models from KEGG [Junker et al., 2006]. However, the most flexible solution for this thesis was Taverna: A programme able to integrate a manifold of web-services, connect data-pipelines to one desired output [Hull et al., 2006]. In combination with MATLAB it is possible to pipeline the information for further use. Very recently, the creation of precise signalling models from KEGG has been treated and a KEGG translator for networks published [Wrzodek et al., 2013].

3.2 Network Analysis of the Clostridial Reactome

It is the scope of this section to create an on-place database that allows easy information retrieval and its manipulation, and human-readable maps of the stored annotations.

Through Taverna it was possible to automatically create a mapping from gene identifiers (*cac:*) to reaction identifiers (*rn:*) to reaction partners (*cpd:*) by harvesting KEGG with a workflow (B.1).

For the integration of transcriptome data to the downloaded reactions in a graph-based format, a graph-theoretical backbone is introduced: Necessary notations and the definitions of a graph are given (3.2.1) and a suitable software searched (3.2.2). Structural characterisation of a biochemical network will aid in visualisation as well as model analysis 3.2.3. Finally, a novel mathematical formalism will be presented to integrate data into the database (3.2.4).

3.2.1 Graph Definition and Notations

Several possibilities to describe a metabolic network exist. One could define reactions and metabolites as nodes and link both when the metabolite is part of the reaction. Using a line graph transformation, one could as well study the reactions linked by their respective metabolites [Nacher et al., 2005]. Here graphs with only one type of nodes are treated: metabolite nodes that are being connected with each other if and only if they are substrate and product of each other in the same reaction. This graph will be named the metabolite-metabolite mapping (MMM) or in mathematical notation, it is a graph $G = (\mathcal{V}, \mathcal{E})$ that contains two sets, a set $\mathcal{V}$ of nodes v connected by a set of edges $e \in \mathcal{E}$. For convenience, the genome of *C. acetobutylicum* is understood as set $\mathcal{X}$ of increasing numbers until the last gene with number N_J:

$$\mathcal{X} := \{j;\ j = 1...N_J\} \tag{3.1}$$

The mapping of the numbers in $\mathcal{X}$ to the corresponding reactions will be called *Rct*. *Rct* is not injective, since one reaction may be performed by several enzymes. The graph G is generated by applying *Rct* on $\mathcal{X}$. It is implicitly understood that no node without edge must exist and that the set of nodes is given at all times (e.g. all annotated metabolites in KEGG). *Rct* only acts on the edges:

$$G_{\mathcal{X}} = (\mathcal{V}_{\mathcal{X}}, \mathcal{E}_{\mathcal{X}}),\ \mathcal{E}_{\mathcal{X}} := Rct(\mathcal{X}). \tag{3.2}$$

Further, "$\cap$" and "$\cup$" denote the intersection and the union, respectively, and "$\backslash$" is the set difference.

A transcript level at time t_i of gene $j \in \mathcal{X}$ will be named $x_j(t_i)$.

3.2.2 Softwares For Visualisation of Graphs

Visualisation of large networks and their annotations is a persistent problem throughout systems biology. KEGG pathways are manually drawn and curated and thereby easy to overview, however the greater scope of one metabolite within the whole networks is lost. In order to gain a broader view on metabolite connectivity, the first ansatz was to find a software that makes annotations and pathways human-readable. Requisites to the softwares were: easy inter-operability with MATLAB and Excel, ease of layouting, and visualisation of multidimensional annotations. These softwares were tested:

- BioLayout Express 3D (BL3D)
- yED
- Cytoscape
- CellNetAnalyzer (CNA)

BL3D [Theocharidis et al., 2009] serves mainly for graphical layouting in two or three dimensional space, it further offers two handy applications. One is the possibility to search for a node in the internet by simple clicking it and accessing a user-defined web page. The dbget functionality of KEGG can be perfectly used for such purpose for all three types of identifiers (*cac:*, *rn:*, *cpd:*). Rapid annotation retrieval in KEGG, makes BL3D an excellent tool for storing and investigating graphs. A second important tool is the MCL-clustering algorithm that detects densely connected nodes [van Dongen, 2000]. This algorithm will be used in chapter 5. No interaction with Excel is possible.

yED has superior abilities compared to BL3D in aligning nodes and edges via a huge library of algorithms. Graphical manipulation of node properties allows a virtual designing of graphs in two dimensional space. Drawing of new nodes and connecting them to the network is made easy. Uploading of annotations is not possible. Again no interaction with Excel is possible.

Cytoscape's [Shannon et al., 2003] essential strength is graph theoretical analysis of network properties and still a large library of algorithms for graph formatting. The possibility to integrate various annotations makes this the most powerful working tool [Troyanskaya, 2005, Joyce and Palsson, 2006]. Interaction with Excel-files is implemented.

Finally, *CellNetAnalyzer* [Klamt et al., 2007] is a tool for the mathematical investigation of pathway models, its capabilities for visualisation are limited. Since it is programmed in MATLAB, interaction with other scripts is facilitated.

Interaction of these four programmes with each other is complicated, while the graphml-languages is supported by the first three softwares, CellNetAnalyzer accepts SBML, which the others do not. Still, the current version of graphml seems not be standardized at such point that several softwares can be automatically interlinked. This problem is reported for several other tools that use SBML as well [Joyce and Palsson, 2006]. For this reason, interaction is granted by using a simple node list. This however, makes re-layouting necessary. For this reason, the use of Cytoscape is preferred because it offers a vast library of algorithms.

3.2.3 Graph Characterisation

A human-readable graph is achieved by visualising graph parameters, e.g. distance measures. This step also helps in characterising the network further [Stelling et al., 2002, Klipp et al., 2004]. This approach focusses on certain network properties, e.g. centrality measures give a prioritisation of targets [Aittokallio and Schwikowski, 2006]. The following enumeration of network parameters relies on [Barabási and Oltvai, 2004].

Node Degree

The node degree is the number of neighbours of $v \in \mathcal{V}$. In a random network the node degree probability $P(k)$ of a node having exactly k connections is a gaussian function. In a biological network this probability has a sharp peak at the beginning and then it falls according to a power law: $P(k) \propto k^{\gamma}$ with $\gamma < 0$. Such a network is called scale-free. The parameter γ further characterises the networks robustness: For $-3 < \gamma < -2$ the emerging network properties are robust against failure of single nodes. For $\gamma \approx -2$, highly connected nodes are in contact with the major parts of all nodes, while for $\gamma \approx -3$ these highly connected nodes disappear and a random network emerges. Here, it is obtained by a non-linear regression routine of Cytoscape.

Network Diameter

The network diameter Δ_G is the maximal distance between any two nodes $v, w \in \mathcal{V}$. It serves as a measure for the compactness of the graph. However, only a small diameter is a reliable parameter since it truly shows that nodes are within close proximity, whereas a large diameter only shows that two nodes are distant, while the others may be compact. Compact networks suggest an easy and rapid communication of the interlinked nodes.

Centrality Measures

The eccentricity $C_{ecc}(v)$ is the reciprocal the shortest paths with maximal lengths from the node $v \in \mathcal{V}$ to other nodes $w \in \mathcal{V}$.

$$C_{ecc}(v) := \frac{1}{\max\{dist(v, w) : w \in \mathcal{V}\}} \tag{3.3}$$

Thus, a high eccentricity shows that all other nodes are in proximity, whereas a low eccentricity means that at least one node and its neighbours are very far.
The closeness C_{clo} is the reciprocal of the sum of all shortest paths that contain $v \in \mathcal{V}$.

$$C_{clo}(v) := \frac{1}{\sum_{w \in \mathcal{V}} dist(v,w)} \tag{3.4}$$

Likewise the eccentricity, high values are positive in the sense of proximity. The closeness gives a tendency how the node is embedded in the graph, if either isolated or central.
The radiality C_{rad} is the average of the difference of the graph diameter and the shortest paths from $v \in \mathcal{V}$ to all other nodes.

$$C_{rad}(v) := \frac{\left(\sum_{w \in \mathcal{V}} \Delta_G + 1 - dist(v,w)\right)}{n-1} \tag{3.5}$$

Hence, by consequently subtracting the shortest paths, the radiality becomes high if all the paths are short, the node is then in the centre. Conversely, if all the paths are long, then the node is in the periphery.
The stress centrality C_{str} stands for the number of shortest paths σ_{st} from any nodes $s \in \mathcal{V}, t \in \mathcal{V}$ different to $v \in \mathcal{V}$ passing through v.

$$C_{str}(v) := \sum_{s \neq v \in \mathcal{V}} \sum_{t \neq v \in \mathcal{V}} \sigma_{st}(v) \tag{3.6}$$

In biological terms, high stress shows how much a molecule is involved in the cellular processes, it may not however symbolise how much this node is necessary to hold together the different parts of the graph.
Likewise the stress centrality, the betweenness C_{bet} considers shortest paths from nodes s to t passing through a node v, however it weights this number with the total number of shortest paths connecting s and t, but not necessarily passing through v.

$$C_{bet}(v) := \sum_{s \neq v \in \mathcal{V}} \sum_{t \neq v \in \mathcal{V}} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.7}$$

Thereby, if $v \in \mathcal{V}$ is the only connection between $s \in \mathcal{V}$ and $t \in \mathcal{V}$, it gets a high betweenness value. As complementary information to the stress centrality, this value allows to assess the importance of a node to connect different parts of the network.
A visualisation of two such parameters is easy to fulfil in Cytoscape: The node colour and the node size are mapped to desired continuous graph parameters. Further mappings are possible, e.g.discrete graph parameters can be mapped to the node shape.

3.2.4 Data-Driven Network Generation - Methodology

The database that integrates transcriptome data and pathway information will be called a *data-driven pathway.* By this integration, analysis of data is possible by a perspective from the reaction network. Reduction of the whole network of achievable reactions to the achieved reactions will help in characterising the current status of the cell. It was suggested that the visualisation of networks during different metabolic states provides a beneficial analysis tool [Khatri et al., 2012]. Surprisingly, the integration of short time-series data into networks for this purpose is not encountered in literature, consider the review by [Dutta et al., 2009].

Boolean Rules for a Two-State System

The starting point is the entire graph as derived from the KEGG database

$$G_{\text{CAC}}^{\text{KEGG}} = (\mathcal{V}, \mathcal{E}),\ \mathcal{E} = \mathit{Rct}(\mathcal{X}). \tag{3.8}$$

$G_{\text{CAC}}^{\text{KEGG}}$ consists of 792 reactions that are connecting 852 metabolites.
For data-integration, a *filtering* approach will be used, so that pathway activity can be assessed from the sub-graphs of $G_{\text{CAC}}^{\text{KEGG}}$ [Aittokallio and Schwikowski, 2006, Reed et al., 2006]: Knowing transcriptome experiments from several culture states one can evaluate the bacterium's regulatory events and provide hints on the current necessity of enzyme synthesis and hereby activation of the conversion from substrates to products. More precisely, one analyses two non-overlapping culture states, e.g. acidogenesis vs solventogenesis or short term response vs long term response. For simplicity they are referred to as s_1 and s_2.

$\mathcal{X}^{b_u}(s)$ will denote the set of all transcript level expression values larger than some boundary b_u at state s:

$$\mathcal{X}^{b_u}(s) := \Big\{ j : x_j\big(t_s\big) > b_u,\ j \in \mathcal{X} \Big\} \tag{3.9}$$

and $\mathcal{X}_{b_l}(s)$ is the set of all transcript expression values smaller than some boundary b_l at state s:

$$\mathcal{X}_{b_l}(s) := \Big\{ j : x_j\big(t_s\big) < b_l,\ j \in \mathcal{X} \Big\}. \tag{3.10}$$

Having this partition, a third partition is immanent, the set transcripts of which is neither clearly repressed nor clearly induced, they are uncertain:

$$\mathcal{X}_{b_u}^{b_l}(s) := \{ j : b_l < x_j(t_s) < b_u,\ j \in \mathcal{X} \}. \tag{3.11}$$

Creation of Data-Driven Pathways From Logical Rules

Application of the three rules creates a boolean network from the initial graph $G_{\text{CAC}}^{\text{KEGG}}$. Two rules are shown here for state s_2:

1. genes up-regulated at s_2: $\mathcal{X}^{b_u}(s_2)$
2. genes down-regulated at s_1 and uncertain at s_2: $\mathcal{X}_{b_l}(s_1) \cap \mathcal{X}_{b_u}^{b_l}(s_2)$

For simplicity, symmetry of the boundary parameters is assumed, $\|b_l\| = \|b_u\| = b$. While the first logical rule is the intuitive approach when considering transcript data - up-regulation is considered as activation, the second rule transfers the available information from one state to the other. If indeed expression were first repressed during one state and it were relieved during the later state, it appears that if a repression were relieved. Vice versa, if it were repressed in the second state and uncertain during the first state, the organism appears to start repression. This is an *augmentation* of state s_2 with respect to s_1, because data from the uncertain regions is used that would have been otherwise neglected. By this *augmentation* is a positive statement about cell efficiency.

In order to distinguish the outcomes of these two rules, two types of graphs are generated, the augmentation graph (G^a-graph) and the induction graph (G-*graph*). For simplicity of evaluation a combination of both graphs, the H-*graph*, is suitable

$$H := G^a \cup G = (\mathcal{V}, \mathcal{E}^a \cup \mathcal{E}). \tag{3.12}$$

Note, that one must not augment when as reference an untreated culture is used, as e.g. in [Alsaker et al., 2010], since augmentation only makes sense when the reference state for microarray hybridisation is taken from the same culture, either at a separate time-point or as average over all time-points. An external reference is uninformative.

The following example illustrates the use of these rules to distinguish between two states, e.g. acidogenesis (s_1) and solventogenesis (s_2). The non-augmented graph G during s_2 reads:

$$G(s_2) := \big(\mathcal{V}, \mathcal{E} \cap \mathcal{E}(s_2)\big),\ \mathcal{E}(s_2) = Rct\Big(\mathcal{X}^b(s_2)\Big). \tag{3.13}$$

Consequently, the augmented graph H during state s_2 then reads as

$$H(s_2) := \big(\mathcal{V}, \mathcal{E} \cap \mathcal{E}(s_2)\big),\ \mathcal{E}(s_2) = Rct\Big(\mathcal{X}^b(s_2) \cup \big(\mathcal{X}_{-b}(s_1) \cap \mathcal{X}_b^{-b}(s_2)\big)\Big). \tag{3.14}$$

Similarly, the active reactions during s_1 after augmentation are defined as:

$$H(s_1) := \big(\mathcal{V}, \mathcal{E} \cap \mathcal{E}(s_1)\big),\ \mathcal{E}(s_1) = Rct\Big(\mathcal{X}^b(s_1) \cup \big(\mathcal{X}_{-b}(s_2) \cap \mathcal{X}_b^{-b}(s_1)\big)\Big). \tag{3.15}$$

Validation and Model Reduction

Considering the two sets of edges $\mathcal{E}(s_1)$ and $\mathcal{E}(s_2)$ closer, it becomes apparent that redundancy in biochemical pathways makes the intersection of both a non-empty set despite the underlying sets of genes being distinct. It will therefore be of interest to consider a third graph:

$$G_{c_X} := \big(\mathcal{V}, \mathcal{E} \cap \mathcal{E}(s_1) \cap \mathcal{E}(s_2)\big). \tag{3.16}$$

As two entirely different states are compared, this graph will help identifying sustained reactions. Finally, the graph of all inactive reactions will be necessary for validation:

$$G_{\text{inact}} := \Big(\mathcal{V}, \mathcal{E} \setminus \big(\mathcal{E}(s_1) \cup \mathcal{E}(s_2)\big)\Big) \tag{3.17}$$

Further reduction of graph size may become necessary to increase readability. One way of doing this without loosing connectivity information, is to assume again efficiency. A biological reaction from a substrate to a distant product takes a shortest path without deviating to far distant molecules. Betweenness or stress values larger than zero indicate that a node is used by at least one shortest path. Thereby, iterative elimination of nodes from G with no stress value reduces the network size. Ultimately, this means that solitary linear branches are cut one-by-one from the periphery of the network until a bifurcation is reached.

3.2.5 Summary

This section proposes two rules for the creation of a pathway-model that integrates transcriptome data with database knowledge to increase readability of data.

Use of the Data-driven Pathway

Literature shows that creation of meaningful flux balance network models is a laborious task because of the urgent need to create a realistic representation of the *in vivo* fluxes [Lee et al., 2008a]. There is no need to re-investigate such a model, because several flux balance models for *C. acetobutylicum* do already exist [Lee et al., 2008a, Senger and Papoutsakis, 2008].

The established data-driven pathway formalism serves for data visualisation in a graph-based format. It was shown that regulatory networks are better analysed in such a format [Freeman et al., 2007]. Therefor it was not urgently necessary to unify compound isomers and to determine reaction directions, or to fill gaps that are due to database errors. This explains the huge differences in metabolite and reaction number between the published models and the raw model downloaded from KEGG (900 in this study compared to 400 to 500 in other studies).

Boolean Rules for Integration of Omics

Some authors do report the application of boolean rules, but these rules are not given [Duarte et al., 2007]. Other authors specify their rules in a one-point measurement, however they do not include how they deal with continuous data [Covert et al., 2001]. Metabolic fluxes after deletion of unexpressed transcripts were used to relate transcript profiles and reaction profiles [Akesson et al., 2004]. Pathway models were generated from up-regulation assessments [Patil and Nielsen, 2005]. They proposed to track differentially expressed genes but made no use of repressed genes or the temporal structure of their data.

Projects like Reactome offer different clustering and data analysis algorithms, but not the type proposed here.
Augmentation as proposed here unveils information from a two-point comparison. Repression during one state and uncertain levels during a second state represent additional information to the usual events of up-regulation. This additional information is expected to reveal further information, as e.g. on constitutively expressed genes.

Boolean Rules Can be Expanded to Multiple States

Although a comparison of two different states already offers valuable information [Reed et al., 2006], this is a limitation that can be overcome easily. If there are more than two clearly distinguishable states or phenotypes, then splitting the time-series data into regions and using the correspondence of genes to these regions helps in robustification of the model. For standardisation of transcriptome data, a similar approach was suggested [Yang et al., 2003]. That is, a gene is counted as up-regulated or down-regulated if this occurs more than once in the corresponding region.

Rules and Data are Not Limiting

Boolean approaches has been shown to construct networks that contain a rich complexity to be studied [Dhaeseleer et al., 2000]. The here presented approach takes transcriptome data. In a similar fashion it could also make use of proteome data. Further, the use of transcriptome data is not limiting. *C. acetobutylicum* is not a fast growing organism, and it was shown that a mapping between transcripts and proteins is possible in a single cell study if samples are not taken directly after cell splitting [Golding et al., 2005].

Promising Results by Integrating Omics Data

Different Omics are already reported to be integrated into pathway models: Integration of metabolome measurements from stimulus-response experiments into pathways was treated by [Cakir et al., 2006]. They achieved the unification of metabolome and transcriptome-measurements which enabled them to assess whether genes are hierarchically or metabolically regulated. A successful approach of transcriptome integration is reported: A graph was generated from genes that were connected based upon Pearson correlation. Unfortunately, the biological meaning of strongly correlated transcript expression profiles to gene-gene correspondence is not discussed. The built model aided in mapping MCL-generated clusters to specific tissues [Freeman et al., 2007].
In contrast to hypothesis-driven research, work with huge data from transcriptome experiments allows knowledge discovery [Bassett Jr et al., 1999]. In this sense the

result of the integration of time-series data of transcript levels into a pathway model is a tool that should not be underestimated. Here, the visualisation of transcriptome data will eventually lead to knowledge discovery, as will be shown in the following sections.

Outlook: Further Integration Possibilities

This pathway can be further filtered by integrating more information: The enzyme's substrate specificity adds one filter criterion - substrates with a small specificity can be neglected, and the number of links between two metabolites effectively reduced. BRENDA offers such information and is readily accessible in Taverna through a SOAP-service [Chang et al., 2009]. This route was not undertaken here: Many enzymes have not yet been tested experimentally in *C. acetobutylicum*, hence also these information require comparison approaches, e.g. filling gaps by considering phylogenetically close relatives.
Reactions thermodynamics further discriminate the reaction directions. Further, implementation of this approach would include redox-potential considerations, intracellular pH measurements and energy balance determination [Kumar et al., 2012]. Yet, available data are insufficient for such purpose.

Outlook: Possibilities of Validation

Viability of the network is one type of network validation commonly suggested [Reed et al., 2006]. Building a data model from transcriptome data alone naturally is not sufficient to create a viable organism, as only regulatory events are detectable and constitutive genes are missing [Troyanskaya, 2005]. Validation can still be carried out on the experimental level, using knock-out mutants and phenotype comparison, enabling a check whether database entries are missing [Reed et al., 2006].

3.3 Data-Driven Pathways

With the available integration scheme it is now possible to re-consider published data. First, a possibility to fix a suitable boundary parameter b must be found (3.3.1), then three experiments are introduced: the standard batch fermentation in complex medium (3.3.3, G^{batch}, [Jones et al., 2008]), the acetic acid addition experiment in batch culture (3.3.4, $G^{\text{batchpulse}}$, [Alsaker et al., 2010]), the pH-shift experiment in continuous culture under phosphate limitation according to COSMIC-specifications (3.3.5, G^{conti}, [Grimmler et al., 2011]).
For visualisation purposes the node for H_2O with all its connections was deleted. An overview of the considered states is given in table 3.1. Further, a summary of some graph properties is given in table 3.2. The γ-parameter shows that the initial network and its data-driven derivatives are organised in communities. Although these networks are not scale-free by common definition, the different filtering approaches, both rules, and different boundary parameters do not drastically change this parameter. Only for the inactive reaction graph it can be noticed that no organisational structure is preserved, which is expected.

Table 3.1: Considered states of data-driven networks in three published experimental settings.

	G^{batch}	$G^{\text{batchpulse}}$	G^{conti}
s_1	10h	15min post pulse	pH 5.8 (acidogenesis)
s_2	40h	20h post pulse	pH 4.5 (solventogenesis)

3.3.1 Derivation of the Boundary Parameter

For the determination of the boundary parameter, this section proposes to access a structural property, the fraction of edges to nodes (E/N). This number allows to track whether constitutive edges or peripheral edges are being eliminated when a stricter, increasing, boundary parameter is used. Deletion in the periphery will more likely create solitary nodes that are not considered further. In the converse, a decreasing boundary parameter allows to assess whether new nodes are added to the graph or existing nodes interlinked. This entirely refers to the boundary parameter determination shown in figure 3.2.

Different Metabolic States are Distinguished by the Edges to Nodes Fraction

To deduce which states should be considered as s_1 a scan over ranges of b is necessary. An example of such a scan over is shown for the stimulated batch experiment in figure 3.1: Stress induced through the acetate pulse becomes apparent here, only few genes are up-regulated directly after the pulse (t_1). The long-term response shows a steady increases of metabolic active genes until this

Table 3.2: Summary of generated graphs
none: no regression of γ was possible because $R^2 \leq 0.6$.

Graph	b	Nodes	Edges	γ (R^2)	Δ_G
$G_{\text{CAC}}^{\text{KEGG}}$	-	902	2176	-1.29 (0.81)	12
$G^{\text{conti}}(s_1)$	1.1	280	384	-1.33 (0.79)	15
$H^{\text{conti}}(s_1)$	1.1	346	475	-1.35 (0.75)	16
$G^{\text{conti}}(s_2)$	1.1	362	523	-1.51 (0.86)	14
$H^{\text{conti}}(s_2)$	1.1	502	910	-1.40 (0.83)	14
$G^{\text{conti,cX}}$	1.1	144	181	none	11
$H^{\text{conti,cX}}$	1.1	165	204	none	13
$G^{\text{conti,inact}}$	1.1	400	355	-2.24 (0.97)	19
$H^{\text{conti,inact}}$	1.1	231	164	-2.65 (0.96)	7
$H^{\text{batch}}(s_1)$	0.8	306	447	-1.48 (0.78)	17
$H^{\text{batch}}(s_1)$	1.3	211	292	-1.42 (0.71)	18
$H^{\text{batch}}(s_2)$	0.8	394	585	-1.38 (0.77)	16
$H^{\text{batch}}(s_2)$	1.3	197	263	-1.41 (0.76)	16
$H^{\text{batch,cX}}$	0.8	68	82	none	4
$H^{\text{batch,cX}}$	1.3	46	61	none	4
$H^{\text{batch,inact}}$	0.8	320	257	-2.57 (0.97)	11
$H^{\text{batch,inact}}$	1.3	574	504	-2.70 (0.92)	22
$G^{\text{batchpulse}}(s_1)$	1.1	70	76	-1.59 (0.79)	6
$G^{\text{batchpulse}}(s_2)$	1.1	354	519	-1.40 (0.82)	15
$G^{\text{batchpulse,inact}}$	1.1	512	465	-2.50 (0.92)	15

number attains its maximum at t_8. Accordingly, reaction and metabolite numbers follow this course. Major differences only become visible when the edges to nodes fraction is calculated. The network which is most invariant to changes of b is found at t_4. Choosing t_1 and t_7 as representative time-points for two different metablic states, the difference between both networks is maximal, and they are clearly distinguishable from the network at t_4.

Choice of b - Discrimination by Defects of E/N

The fraction of edges is expected to monotonically decrease, otherwise a *defect* has occurred: The centrality measures of the initial graph $G_{\text{CAC}}^{\text{KEGG}}$ (figure 3.3) indicate that the major part of the nodes is densely connected - betweenness values are small - and there is not a large set of nodes lying in the graphs periphery - closeness values are small. Additionally, by application of the boolean rules the graph diameter decreases for each experiment but γ stays unaltered. These observations show that the graphs integrity is loosened by deletion of edges not nodes. The behaviour of $H^{\text{batch}}(s_2)$ for $b > 1.4$ indicates therefore an undesired defect in the graph's topology. Here the graph is split into numerous smaller

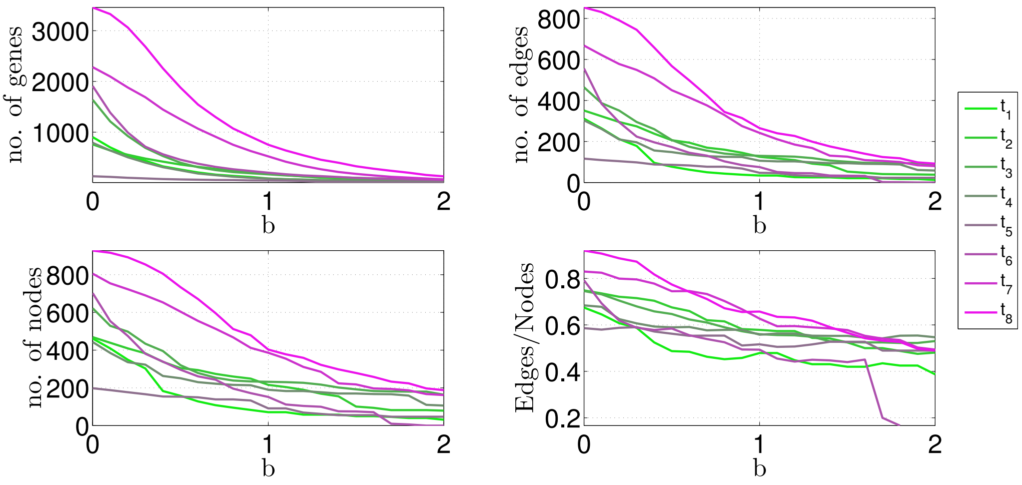

Figure 3.1: Complete scan of b through all times ($t_1, ..., t_8$) of the stimulated batch experiment, $G^{\mathrm{batchpulse}}$.
Upper left: up-regulated genes
Upper right: activated reactions (edges)
Lower left: activated metabolites (nodes)
Lower right: edges to nodes fraction

sub-graphs that are not connected.
With the same reasoning a sudden steep descent as present in figure 3.1 for t_7 shows a second defect, here the number of nodes and edges is approaching zero because at $b > 1.5$ most nodes are being deleted.

Choice of b - Discrimination by Levels and Descent

The steepness of descent of the E/N can be regarded as uncertainty of the boundary parameter b - in the close proximity of a chosen value the graphs topology should not undergo strong alternations. For this reason, an analysis of the graph should be carried out at some flat point of the curve.
In G^{batch} E/N has a similar descent during both states until $b = 1.5$, then the defect occurs. The same behaviour without defect for large values is visible for G^{conti}, here the descent is weaker for the state s_1. Both graphs start at the same level of E/N. It can be expected that the topology properties of the graphs are similar too. Considering the stimulated batch culture, here both states are immanently different, the short-term response is 0.2 smaller than the long-term response, also their descents differ between each other: While it is strong for low b and then decreases close to zero for s_1, it is constant for s_2. The steeper descent accounts for randomly spread edges. A constant level accounts for the deletion of the same number of edges and nodes, hence nodes from the networks periphery are being deleted for increasing b.

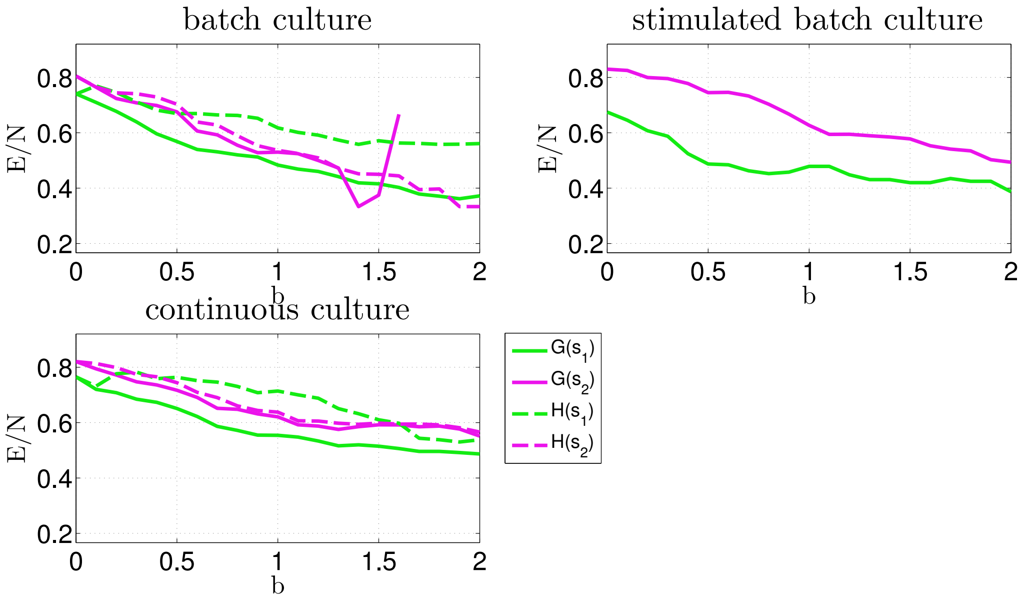

Figure 3.2: Scan of b for different transcriptome experiments. E/N is the fraction of edges to nodes.

In order to study the effects of the choice of b, two different values, $b_1 = 0.8,\ b_2 = 1.3$, will be compared for the batch culture. Since the stimulated batch culture is referenced to an unstimulated batch culture, it is reasonable to choose $b = 1.1$ as intermediate value of this interval.

3.3.2 Augmentation Characterises Solventogenesis

Acidogenesis and solventogenesis are two distinct states that are expected to be visible in these networks in either experiment, batch culture or continuous culture. For both it does not matter whether the graph is augmented during state s_2, since not much additional information is gained with respect to E/N. However, there is a difference up to 0.2 between $G(s_1)$ and $H(s_1)$ in both experiments. This eventually shows that a number of unconsidered reactions in s_1 are of no need in s_2 anymore and are therefore shut down. This indicates that solventogenic phase is an adaptation to more hostile conditions. While the main metabolism from s_1 is largely preserved, additional reactions ensure the survival of the organism. This becomes obvious in the continuous culture, $H^{\text{conti}}(s_2)$ is much larger than $H^{\text{conti}}(s_1)$ (table 3.2) and their intersection $H^{\text{conti,cX}}$ is large, it contains one fourth to one third of both graphs. In the converse $H^{\text{batch}}(s_2)$ is smaller by 0.2 than $H^{\text{conti}}(s_2)$, here metabolic activity ceased because of starting sporulation. Reactions that are relevant for sporulation are not covered by the KEGG database.

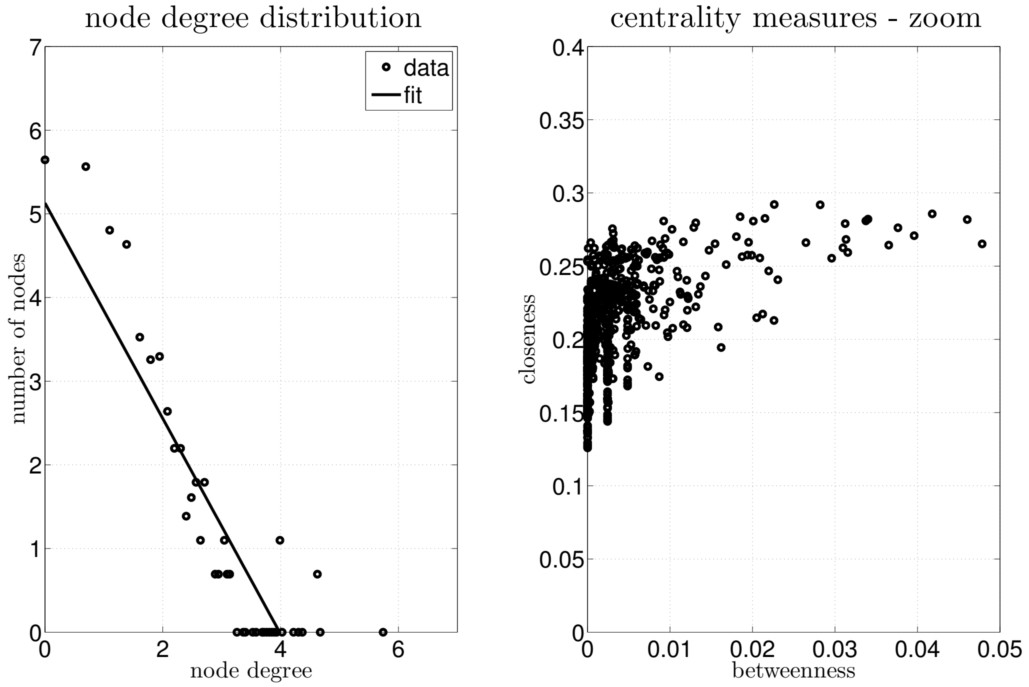

Figure 3.3: Left: Node degree distribution in $G_{\mathrm{CAC}}^{\mathrm{KEGG}}$ with the linear fit for determination of γ.
Right: Centrality statistics in $G_{\mathrm{CAC}}^{\mathrm{KEGG}}$. Further centrality measures are not shown as betweenness and stress are strongly correlated, as well as closeness, eccentricity and radiality are strongly correlated.

3.3.3 Visualisation of the Standard Batch Fermentation

In the standard batch fermentation [Jones et al., 2008] in complex medium s_1 corresponds to $t = 10\ h$ (figure 3.4) when solvent production starts, and s_2 corresponds to $t = 40\ h$ (figure 3.5) when the culture enters the sporulation-state. The reference state is the average over all transcripts and measured time-points. The authors distinguish in their paper, six different clostridial stages occurring in the temporal transcript expression data. The here chosen states are well distinguishable according to these stages. Visualisation is focussed on revealing the outcomes of two different boundary parameters.

Early Phase

The reactions ($b = 1.3$) during s_1 are sulfur aminoacid and serine metabolism, co-factor synthesis, parts of sugar metabolism yielding butanoic acid and parts of cell wall synthesis.
This view is complemented with the more uncertain reactions ($b = 0.8$), in particular for leucine- and gluthatione-synthesis. The pathway for mureine synthesis

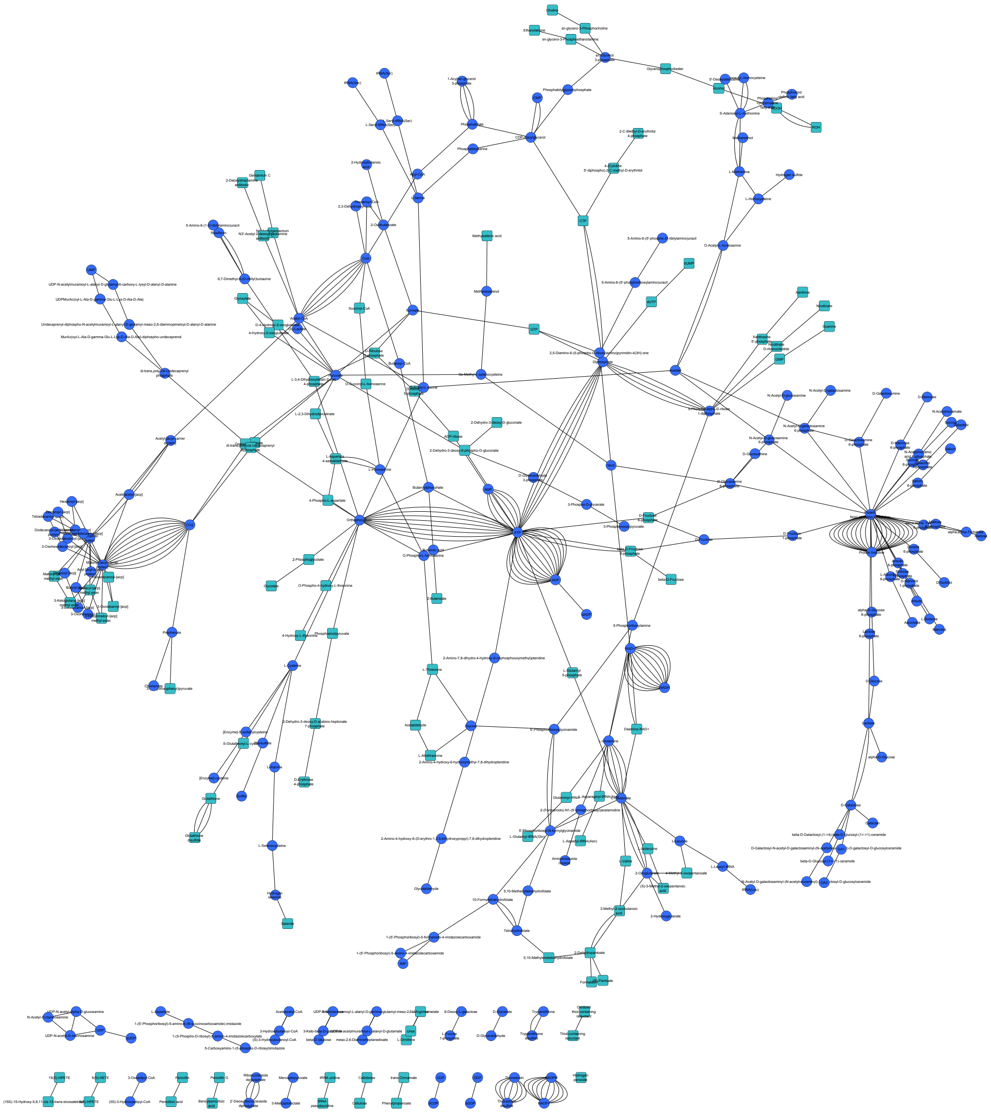

Figure 3.4: Early Batch Experiment ($H^{\text{batch}}(s_1)$), darkblue and lightblue: $b > 0.8$; darkblue: $b > 1.3$

is added. A delta-2-oxidreductase for crotonic acid becomes apparent. Also a threonine lyase is active.

From the graph it is easily visible, that lowering b does not interlink existing compounds but instead, it adds new branches to the existing network.

Late Phase

Compared to the early phase, the late phase contains less reactions and metabolites for the stronger certitude than for the weaker - here, one obtain the same number of metabolites which are more densely connected. For $b = 1.3$ the energy metabolism,

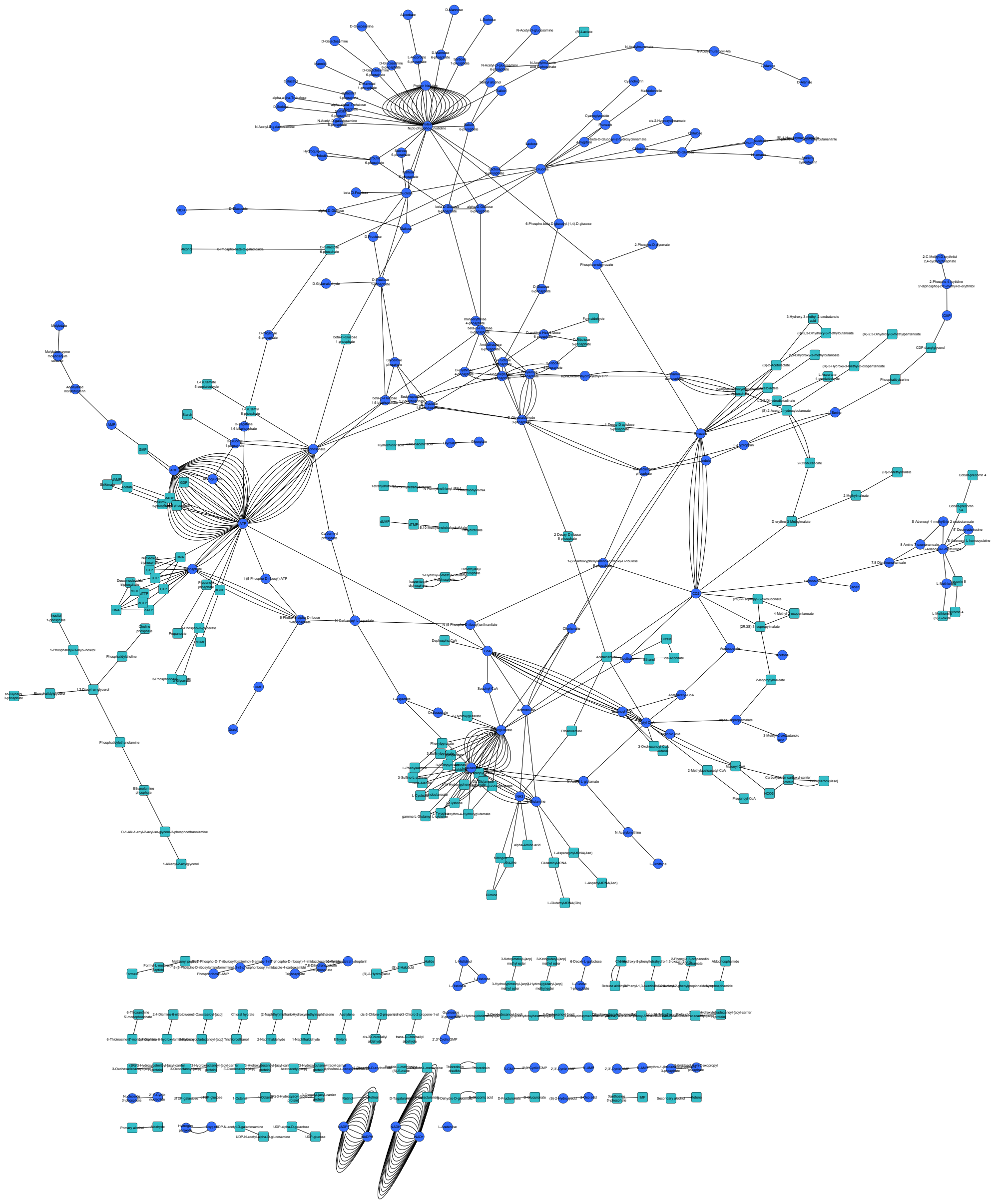

Figure 3.5: Late Batch Experiment ($H^{\text{batch}}(s_2)$), darkblue and lightblue $b > 0.8$, darkblue: $b > 1.3$

glycolytic paths, pentose pathways and secondary metabolism become apparent, aspartate metabolism and glutamate are central. On a less certain level nucleotide synthesis, butanal production and membrane biosynthesis are seen. The reactions for $b = 0.8$ are connections of the more certain reactions for $b = 1.3$: Acetate, O-acetyl-serine and acetyl-CoA are the most stressed metabolites (not shown) for $b = 0.8$. As expected, they play a central role during s_1. For $b = 1.3$, these metabolites move back in ranking to places 8, 5 and 6 respectively.

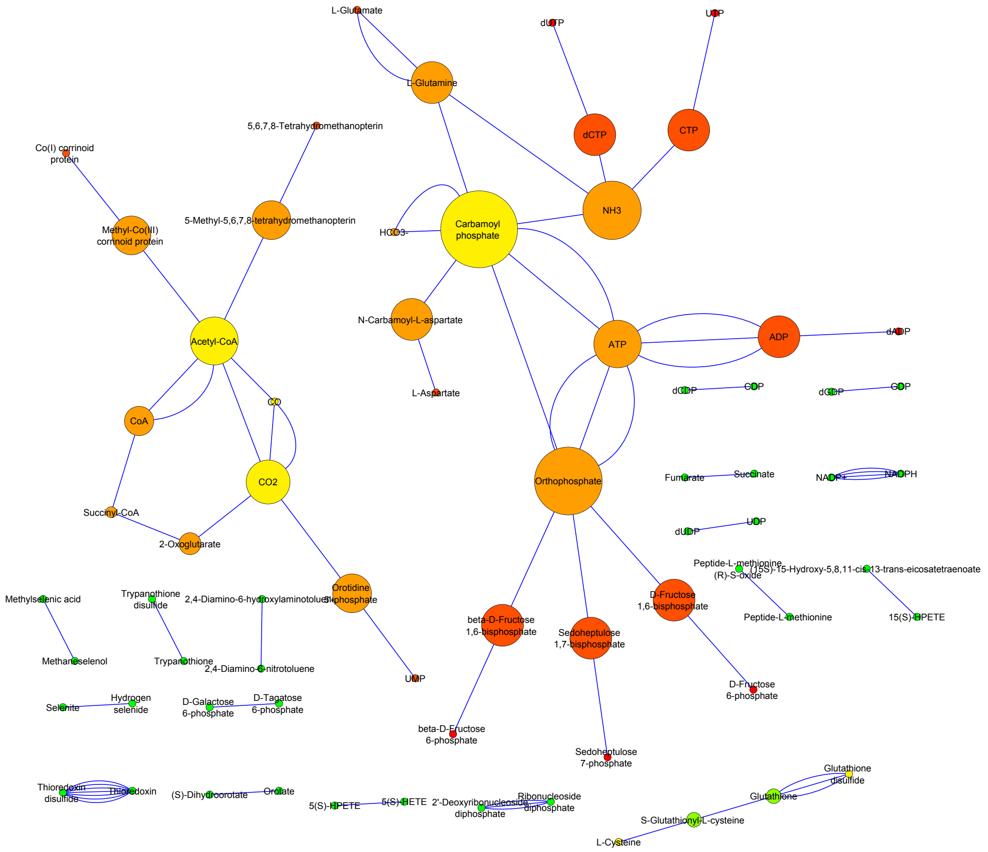

Figure 3.6: Early Batch Pulse Experiment ($G^{\text{batchpulse}}(s_1)$), colour:green to red for decreasing eccentricity, size: small to large for increasing stress

3.3.4 Visualisation of the Acetic Acid Pulse Experiment in Batch

In the acetic acid pulse experiment [Alsaker et al., 2010], s_1 corresponds to the short-term response after 0.25 h (figure 3.6) and s_2 to the long-term response after 20 h (figure 3.7). Here, the reference state is an independent batch culture and for each measured time-point, an untreated reference is used. Therefor, one must not augment the G-graph. Visualisation is focussed on the different graph topological parameters stress and eccentricity and their role to increase human-readability of the network.

Short-term Response

For the short-term response, the network is small - only few paths are activated after the pulse. Notably, reactions around ammonia and phosphate become apparent, conversion of fructose-bis-phosphate and seduheptulose phosphates, glutamine and aspartate conversion. Further, few reactions are involving acetyl-CoA. Thioredoxin utilization increases by multiple reactions.

Long-term Response

The long-term response is more complex than the short-term response. It involves a variety of CoA-reactions, a multitude of carbon dioxide involving reactions. here are butanoyl-CoA originating reactions, synthesis of different amino acids and acetyl-CoA driven synthesis of branched small fatty acids. Further, the ABC-transporters for sugars are activated. One also finds upgregulated sugar import in the batch culture, this suggests that acetic acid has a stimulating effect on glucose uptake. This was seen also in continuous culture (data not shown).
From a comparison with $H^{\text{conti}}(s_2)$, one recognises that more than two third of the involved reactions are identical to the late stimulated network. Acetic acid addition seems indeed be an inducer for a state comparable to solventogenesis, this also was seen in continuous culture, acetone and butanol were produced after short and sustained stimuli with acetic acid (data not shown).

3.3.5 Visualisation of the pH-Shift Experiment in Continuous Culture

The pH-shift experiment [Grimmler et al., 2011] is more useful with respect to the batch fermentation experiment because of two aspects: The two phases are clearly separated before and after the shift and no sporulation occurs in the continuous culture. Acidogenic conditions correspond to s_1, early solventogenic condition, when the pH is stabilised to s_2. Visualisation is focussed on the comparison of augmented and non-augmented graph and centralities of the augmented graph.

The Graph of Acidogenesis

Graph centrality measures are shown in figure 3.8 and complemented with a third dimension, the G- and H-graph (figure 3.9). Glutamate and joint amino acids like glutamine and aspartate take a central role of this pathway. In a more distant region different CoA and phosphate derivatives are present. The most outside regions are occupied by several fatty acid metabolites and vitamines. Most interestingly for this network in contrast to solventogenesis, the carbon dioxide node is only loosely connected to the overall network.
There is the carbon monoxide fixation pathway, known to be active for several organisms and Clostridia, but not in *C. acetobutylicum* [Koepke et al., 2011], the decarboxylation of isocitrate to oxoglutarate and the oxidative decarboxylation of oxoglutarate to succinyl-CoA. All these reactions belong to the H-graph, hence are strongly down-regulated in solventogenesis. There is no other decarboxylase up-regulated in acidogenesis.
In addition, the whole conversion path from acetyl-CoA to butanal is already

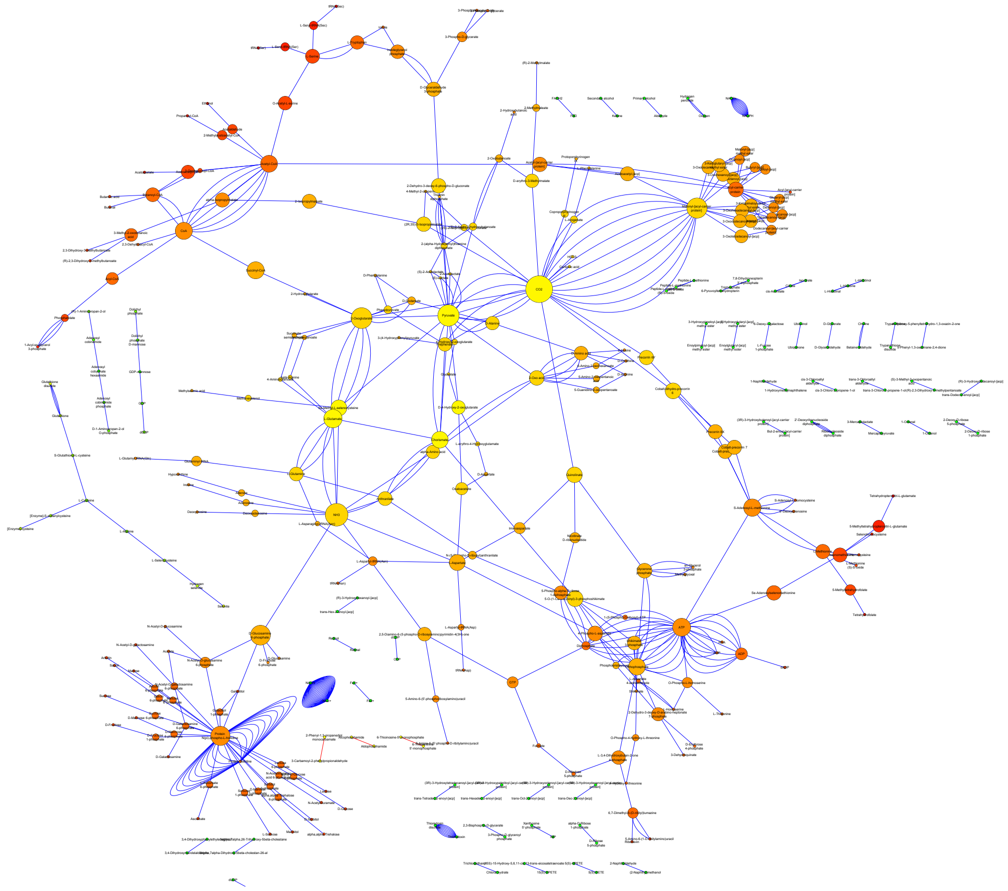

Figure 3.7: Late Batch Pulse Experiment ($G^{\text{batchpulse}}(s_2)$), colour:green to red for decreasing eccentricity, size: small to large for increasing stress

present. Membrane lipid synthesis/degradation and hydrofolate synthesis uniquely belong to the H-graph.

Graph of the Solventogenesis

The solventogenic graph is larger than the acidogenic graph. Graph centralities are shown in figure 3.10 which is again complemented with the G-graph, and H-graph (figure 3.11). The most central metabolites are glutamate, ATP, further pyruvate and phosphoenolpyruvate. In the medium and long range, a multitude of sugar and nucleotide involving reactions are found. In contrast to acidogenic conditions, carbon dioxide plays a central role together with pyruvate, ATP, ammonia and glutamate. Surprisingly, also the position of butanoyl-CoA has changed, it is shifted to the periphery. The production of carbondioxide is related to membrane lipid conversion,, to pyruvate decarboxylation. Further reactions are

1. rn:R06895: coproporphyrinogen-III:S-adenosyl-L-methionine oxidoreductase

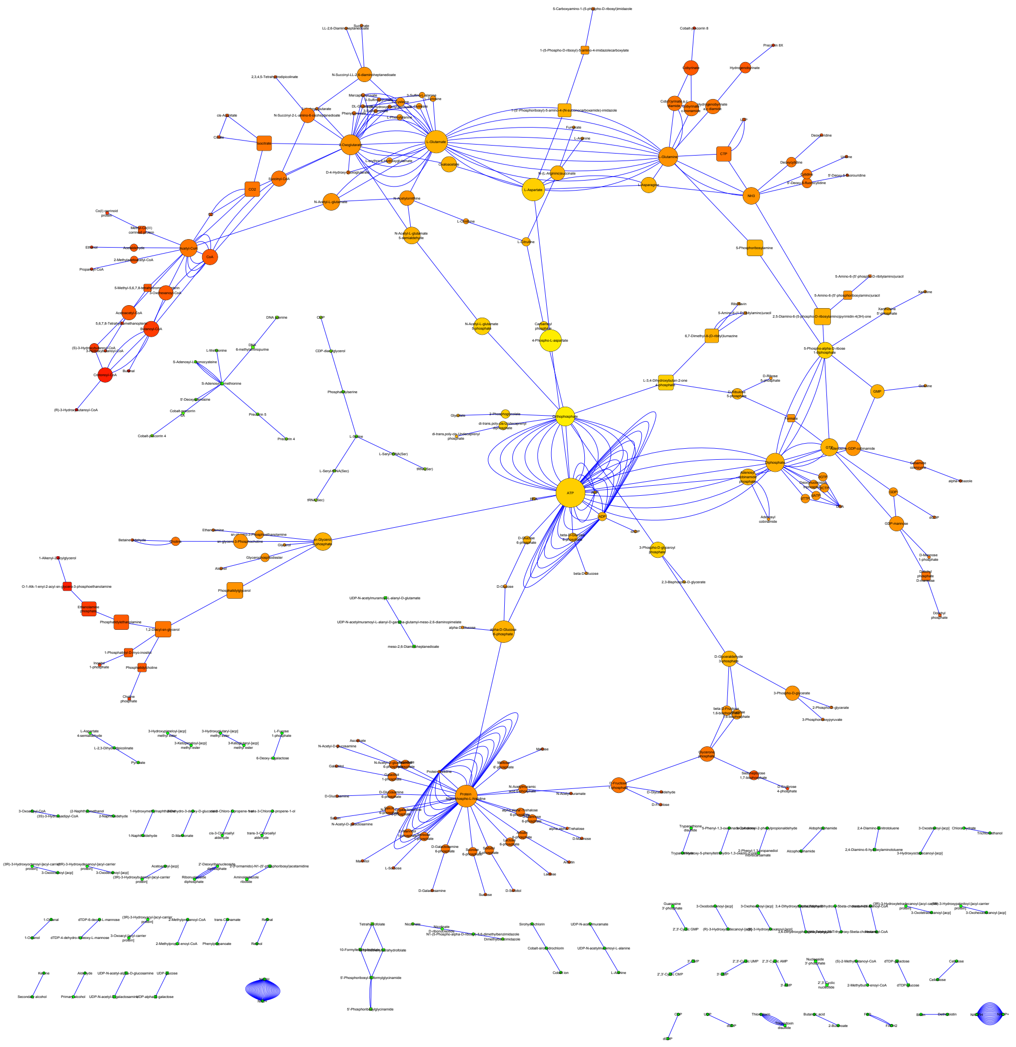

Figure 3.8: Continuous Experiment, Acidogenesis ($H^{\mathrm{conti}}(s_1)$), colour: green to red for decreasing eccentricity, size: small to large for increasing stress

2. rn:R03508: 1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose-5-phosphate carboxy-lyase

3. rn:R03348: Nicotinate-nucleotide:pyrophosphate phosphoribosyltransferase

4. rn:R01728: Prephenate:NAD+ oxidoreductase

5. rn:R01366: Acetoacetate carboxy-lyase

6. rn:R00965: orotidine-5'-phosphate carboxy-lyase

7. rn:R00451: meso-2,6-diaminoheptanedioate carboxy-lyase

8. rn:R00178: S-adenosyl-L-methionine carboxy-lyase

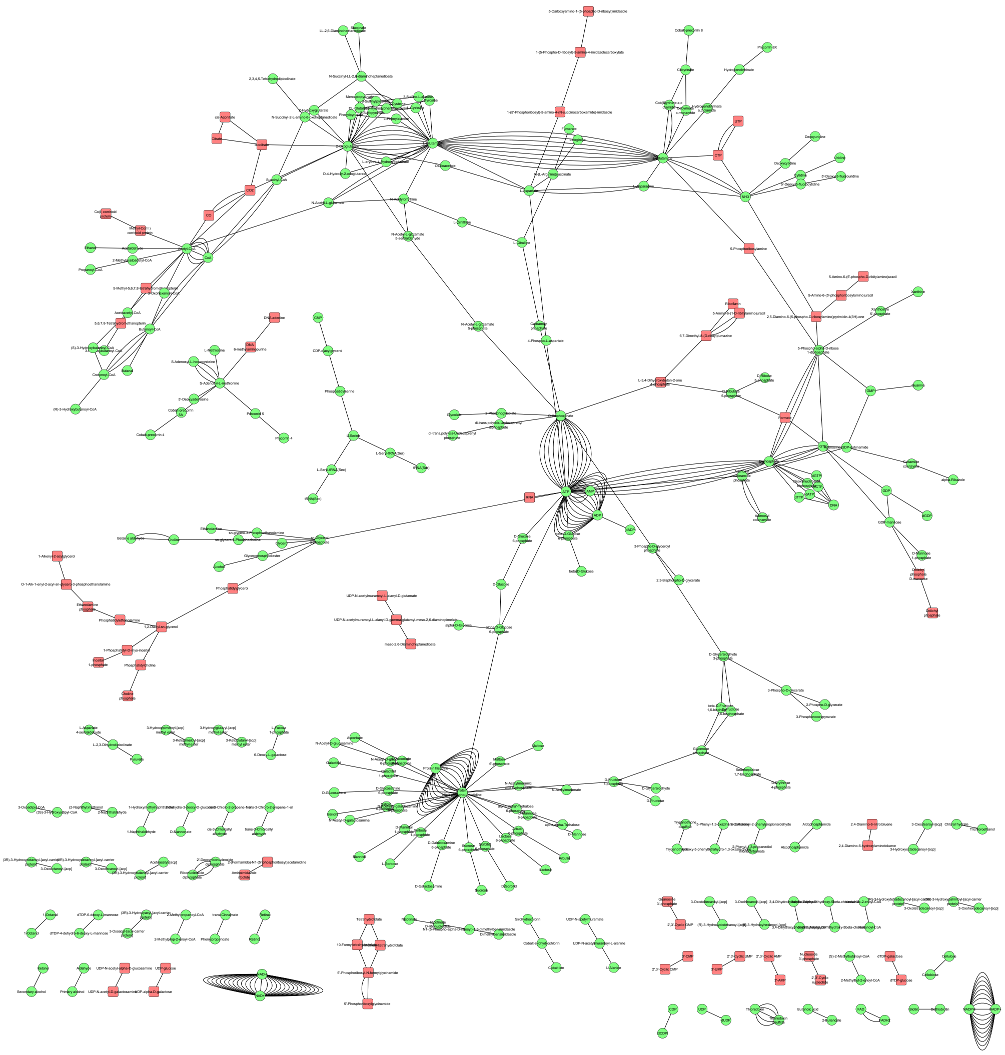

Figure 3.9: Continuous Experiment, Acidogenesis, with Augmentation, green and red: $H^{\mathrm{conti}}(s_1)$, red: $G^{\mathrm{conti}}(s_1)$

One finally observes several sulfuric aminoacid relevant reactions in the H-graph.

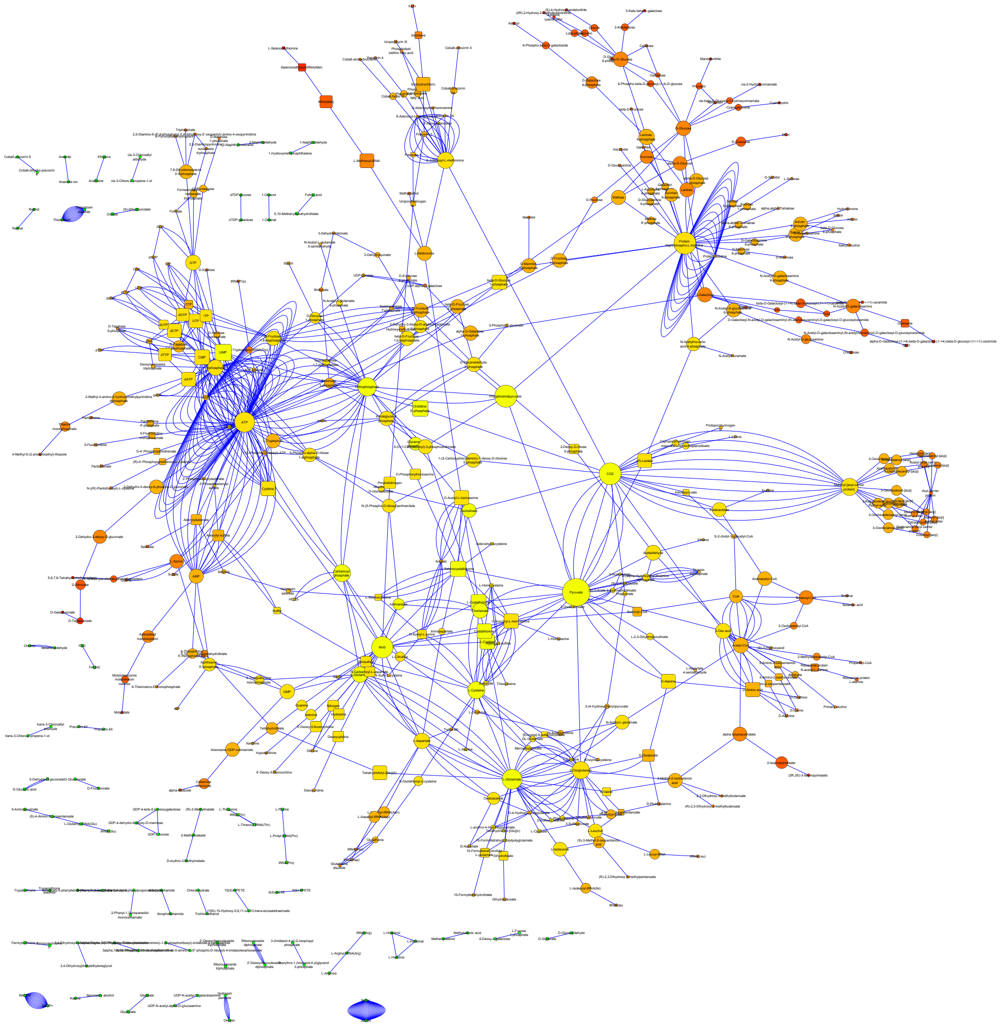

Figure 3.10: Continous Experiment, Solventogenesis ($H^{\mathrm{conti}}(s_1)$), colour:green to red for decreasing eccentricity, size: small to large for increasing stress

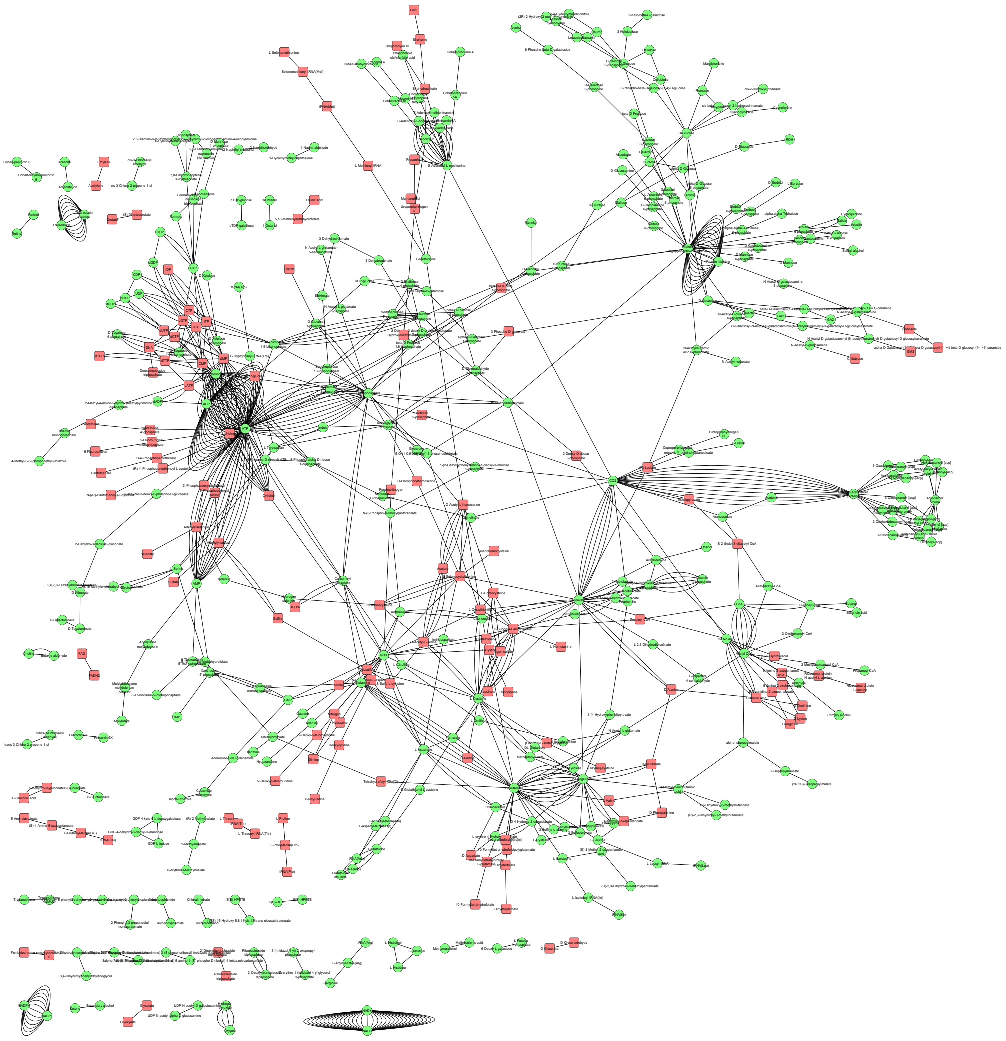

Figure 3.11: Continous Experiment, Solventogenesis, with Augmentation, green and red: $H^{\text{conti}}(s_1)$, red: $G^{\text{conti}}(s_1)$

3.3.6 Conclusions

In this section several data were presented with the previously derived formalism for data-driven pathway generation. As this model was used for visualisation only purposes, it was not curated. As soon as the download from published curated databases is possible, visualisation of these pathways can be re-done easily because of script automation (appendix B.2).

Biological Activity is Critical for the Choice of the Boundary Parameter

As first challenge, a scheme was researched at which transcript expression level a reaction is considered active in the network. The boundary parameter b was derived using a general graph topological trait, the edges to nodes fraction in combination to general graph statistics. Similar proceeding are known to be fruitful in pathway recognition [Khatri et al., 2012]. The precise network and its outcome still heavily depend on the choice of b. While its definition is intuitive, it includes the assumption that it is single-valued for each experiment. Transcriptional activity, defined as the activity of RNA polymerase, is known to differ along the bacterial life cycle [Golding et al., 2005]. Consequently, each state could also obtain its individual b accounting for different transcriptional activities. On the one hand, the design of the reference state in the microarray experiment copes with that problem: By taking the average over all time-points as reference, the variations in transcriptional activity can be accurately covered. On the other hand, other data reference to a state at the beginning or the end of cultivation, and thereby they distort the data. Time-dependent choice of b is comparable to a segmented normalisation of data, proven superior before [Yang et al., 2003].

Improved Evaluation of the Boundary Parameter by Qualitative Biological Knowledge?

Statistical tools to assess significant changes in expression pattern were used [Cakir et al., 2006], this is comparable to the determination of a significant b. It is equally reported that statistical tools work on the edge of their intended capability, non-statistical tools are proven more worthy [Huang et al., 2009]. An optimal parameter for the transcriptome evaluation can be found by an alternative route - complementation with pathway information. Expectations on compound connectivity are biologically testable, e.g. the number of reactions converting ATP or NADH is one feasible criterion.

Alternatively, changes in product concentrations can be mapped to changes in respective transcripts. Accumulation of metabolite pools suggest that there are more influx than outflux reactions. This naturally requires a better curated network.

From a comparative perspective, one can also note that the transcriptome data can be split into two functionally different sets, a set of enzyme-coding transcripts and a set of non-enzyme coding transcripts as will be shown in the next section for a different purpose (equation 3.21). Both subsets are correlated because they represent the vital organism. If it were possible to access all transcriptional regulators in *C. acetobutylicum* as it is for *E.coli* [Gama-Castro et al., 2008] with their respective open reading frames (ORFs), deduction of b requires that these sets correlate for a given b.

Augmentation Requires More Extensive Studies

After the proposition of the augmentation rule (section 3.2.4) its application showed that it acts very differently on the solventogenic state than on the acidogenic state - while in acidogenesis an influence could be noted, solventogenesis did not show alternated graph topologies.

This result relies on the sizes of the three different regions that are constructed by choosing the boundary parameter b. This choice is critical at the point where genes vary close to the boundary δ_b

$$\delta_b = |x(s_1) - x(s_2)| = (-b - \epsilon - (-b + \epsilon)) = 2\epsilon \approx 0$$

Such genes are falsely considered as augmented: In continuous culture 24% of all augmented genes from acidogenesis differ by less than one order of magnitude, during solventogenesis that is 20%.

It is suggested to further investigate other augmentation rules, e.g. $\mathcal{X}_b^0(s)$ that circumvents this problem.

Multidimensional Visualisation is a Challenge

Multi-dimensionality easily arises in the biological context. Here, the visualisation of KEGG database was undertaken by a metabolite-metabolite mapping with integrated data. The underlying structure is however more complicated because reactions usually require more than one substrate to produce more than one product. Such graphs are hypergraphs; sets of nodes are connected by sets of edges. The visualisation of such problems is only at its beginnings [Junghans, 2008].

Already for the simple graphs shown here visualisation is a challenge in Cytoscape. It was the aim to facilitate hypothesis generation. In particular, visualisation of ontologies for genes, enzymes, reactions, metabolites and pathways require side-by-side visibility. Several graphs and their different attributes were shown, the two boundary parameters in the batch experiment, the two types of logical rules in the continuous experiment, the network centrality measures in the stimulated batch experiment. More annotation needs to be fed to the graph in order to allow easier interpretation, up to five dimensions are possible in one graph, still Cytoscape is not equipped for such a purpose. Enrichment tools are frequently

encountered [Huang et al., 2009] and efforts are spent in visualising ontologies in web-interfaces [Dennis et al., 2003], in trees [Chevenet et al., 2006], a tool for multidimensional annotation visualisation in graphs seems missing so far.
Dynamic visualisation by movies shows the emergence of paths and their disappearance. Usability assessments of this approach indicate that static networks are easier to treat [Farrugia and Quigley, 2011]. Dynamic and static possibilities to visualise different timescales are reviewed by [Secrier and Schneider, 2013].

Use of this Model

From the networks of batch culture and stimulated batch culture it was shown that a state comparable to solventogenesis was induced by acetate addition. This view is supported by experiments in continuous culture according to COSMIC specifications in which acetate stimuli ($50\, mM$) as pulse or as step-function were applied during acidogenesis. The acetate was used for the production of acetone and butanol. It was also found in these cultures that glucose uptake is stimulated and growth starts.
It was further shown that positions of metabolites within the data-driven network can be monitored across different metabolic states and pathways. The appearance of crotonate indicates a pathway that contains this metabolite. It will be shown in the next section that the observance of this metabolite leads to the formulation of a new pathway and the annotation of yet unannotated proteins.

3.4 Identification of Missing Reactions

Missing annotations for reactions are frequently encountered. One possibility to deal with these gaps is to transfer annotation from other species [Forslund, 2011]. This section lays the fundamentals to compare organisms based on a KEGG-database query and Pfam-motifs (3.4.1). Comparing a close relative to *C. acetobutylicum* which is *B. subtilis* (3.4.2) is carried out and hypotheses can be constructed side to side to the MMM.

3.4.1 Comparative Approach

Motivation

The original annotation of *C. acetobutylicum* is incomplete, a large set of genes has no annotation. Recently reported experiments suggest that there may be branches in the acid and solventogenic pathways missing, because a knock-out of the transacetylase and transbutyrylase still yielded an acetate and butyrate positive mutant [Lehmann et al., 2012a, Lehmann et al., 2012b]. Also the hitherto unannotated tricarboxylic acid-cycle was only recently discovered by a metabolome study [Crown et al., 2011]. Determination of function is known to be possible through homologies in close relatives [Durot et al., 2009]. The use of Pfam-motifs [Punta et al., 2012] and Pfam-motif architecture has therefor gained increasing interest during the last years [Ofran et al., 2005, Lin et al., 2006, Koestler et al., 2010].

Database Query of Missing Reactions

Assume there is a second organism to which one can compare the clostridial reactome. Applying the following database query efficiently identifies missing reactions in one organism by comparing functional similar homologues in the other. First, the reactome and the occurring Pfam-motifs will be harvested. As before the set of reactions an enzyme can perform ($\mathcal{R}_{\mathcal{X}}$), is given by a map Rct from a specific set of genes $\mathcal{X} \subset X^{\text{KEGG}}$ which forms a subset of all genes in the database, here the genes in KEGG X^{KEGG}.

$$\mathcal{R}_{\mathcal{X}} := Rct(\mathcal{X}). \tag{3.18}$$

A similar map Pfa is given to determine the constituting Pfam-motifs for each protein from the gene sequence ($\mathcal{P}_{\mathcal{X}}$):

$$\mathcal{P}_{\mathcal{X}} := Pfa(\mathcal{X}). \tag{3.19}$$

It will be required to determine the inverse of the map Rct. For a specific reaction $r \in R_{\mathcal{X}}$ it is given by:

$$Rct^{-1}(r \in R_{\mathcal{X}}) := \Big\{ x \in \mathcal{X}^{\text{KEGG}} : r = Rct(x) \Big\}. \tag{3.20}$$

From this definition it becomes clear that the inverse takes values in the whole set of genes within KEGG. The inverse of a specific Pfam-motif is defined similarly. Not every protein is bearing catalytic functions, therefore the kernel $\mathcal{X}_0$ of *Rct* is useful to enhance computational efficiency:

$$Rct\big(\mathcal{X}_0\big) = \emptyset. \tag{3.21}$$

This kernel partitions the set of genes accordingly to their existing reaction-annotation $\mathcal{X}_{\text{React}}$.

$$\mathcal{X}_{\text{React}} := \mathcal{X} \setminus \mathcal{X}_0 \tag{3.22}$$

Pfam-Motif Comparison Between Two Organisms

Choosing a close relative of *C. acetobutylicum*, e.g. *B. subtilis*, one can assume that the reactome of *B. subtilis* is better curated than the one of Clostridium. In order to compare both reactomes, one first determines reactions specific to *B. subtilis* ($\mathcal{R}^{\text{BS}}_{\text{spec}}$) by subtracting common reactions from both reactomes:

$$\mathcal{R}^{\text{BS}}_{\text{spec}} := Rct\big(\mathcal{X}^{\text{BS}}_{\text{React}}\big) \setminus Rct\big(\mathcal{X}^{\text{CA}}_{\text{React}}\big) \tag{3.23}$$

The inverse of this map now shows all the genes specific to these unknown reactions in *Clostridium* but known in *Bacillus.* Intersection with the genes of Bacillus narrows the solution space to the genes of interest ($\mathcal{X}^{\text{BS}}_{\text{spec}}$) in this study:

$$\mathcal{X}^{\text{BS}}_{\text{spec}} := Rct^{-1}\big(\mathcal{R}^{\text{BS}}_{\text{spec}}\big) \cap \mathcal{X}^{\text{BS}} \tag{3.24}$$

The Pfam-motifs of interest $\mathcal{P}$ are now determined from this set of specific genes $\Big(Pfa\big(\mathcal{X}^{\text{BS}}_{\text{spec}}\big)\Big)$ and retrieved in the genes with no reaction annotation of *Clostridium* $\Big(P\big(\mathcal{X}^{\text{CA}}_0\big)\Big)$. The two set of genes ($\mathcal{X}_{\text{comp}}$) bearing at least one of these motifs then will serve as database for comparison of these two species:

$$\begin{aligned}
\mathcal{P} &:= Pfa\big(\mathcal{X}^{\text{BS}}_{\text{spec}}\big) \cap P\big(\mathcal{X}^{\text{CA}}_0\big) && (3.25)\\
\mathcal{X}^{\text{CA}}_{\text{comp}} &:= Pfa^{-1}\big(\mathcal{P}\big) \cap \mathcal{X}^{\text{CA}} && (3.26)\\
\mathcal{X}^{\text{BS}}_{\text{comp}} &:= \mathcal{X}^{\text{BS}}_{\text{spec}} && (3.27)
\end{aligned}$$

Similarity Measure for Functional Homologs

Connection of the two gene sets $\mathcal{X}^{\text{CA}}_{\text{comp}}$ and $\mathcal{X}^{\text{BS}}_{\text{comp}}$ according to their Pfam-motifs $\mathcal{P}$ yields a map.

This map however requires reduction to be useful for manual inspection: Similarity in one motif only is not sufficient to hypothesise on the same function. Consequently, it is necessary to restrain the connections by assuming, that two genes of *Bacillus* and *Clostridium* are connected only when at least a percentage of their motifs is similar. Established similarity measures for Pfam-motif comparison are given in [Lin et al., 2006]. The Jaccard-index $J(x,\hat{x})$, $x \in \mathcal{X}^{\mathrm{CA}}, \hat{x} \in \mathcal{X}^{\mathrm{BS}}$ [Levandovski and D, 1971] is one of them:

$$J(x,\hat{x}) = \frac{\mathcal{P}^{\mathrm{CA}}(x) \cap \mathcal{P}^{\mathrm{BS}}(\hat{x})}{\mathcal{P}^{\mathrm{CA}}(x) \cup \mathcal{P}^{\mathrm{BS}}(\hat{x})} \tag{3.28}$$

This study proposes to use a different measure $n_{\mathcal{P}}(x,\hat{x})$ for such purpose as will become clear instantly:

$$n_{\mathcal{P}}^{\mathrm{CA}}(x,\hat{x}) \quad := \quad \frac{\left|\mathcal{P}^{\mathrm{CA}}(x) \cap \mathcal{P}^{\mathrm{BS}}(\hat{x})\right|}{|\mathcal{P}^{\mathrm{CA}}(x)|} \tag{3.29}$$

$$n_{\mathcal{P}}^{\mathrm{BS}}(x,\hat{x}) \quad := \quad \frac{\left|\mathcal{P}^{\mathrm{BS}}(\hat{x}) \cap \mathcal{P}^{\mathrm{CA}}(x)\right|}{|\mathcal{P}^{\mathrm{BS}}(\hat{x})|} \tag{3.30}$$

$$n_{\mathcal{P}}(x,\hat{x}) \quad := \quad \frac{1}{2}\big(n_{\mathcal{P}}^{\mathrm{CA}}(x,\hat{x}) + n_{\mathcal{P}}^{\mathrm{BS}}(x,\hat{x})\big). \tag{3.31}$$

Comparison of Similarity Measures

Consider the three following general cases of two sets $P(g_1)$ and $P(g_2)$:

1. case: $P(g_1) = \{P_1, P_2\}$ versus $P(g_2) = \{P_1, \cdots, P_{10}\}$
2. case: $P(g_1) = \{P_1, \cdots, P_4\}$ versus $P(g_2) = \{P_3, \cdots, P_{12}\}$
3. case: $P(g_1) = \{P_1, \cdots, P_4\}$ versus $P(g_2) = \{P_3, \cdots, P_6\}$

Obviously for case 1, $P(g_1)$ contains all items that are present in $P(g_2)$. It is possible that $P(g_2)$ functions as $P(g_1)$, only $P(g_1)$ has more motifs. The Jaccard-index $J\big(P(g_1), P(g_2)\big) = 0.2$ is less beneficial in this case then the here proposed similarity measure $n\big(P(g_1), P(g_2)\big) = 0.5(1 + 0.2) = 0.6$.
The overlap in case 2 is identical to case 1, now $P(g_1)$ contains two distinct motifs to $P(g_2)$, the sizes of proteins are as in case 1. Again, the Jaccard-index $J\big(P(g_1), P(g_2)\big) = \frac{1}{6}$ is less beneficial $n\big(P(g_1), P(g_2)\big) = 0.5(0.5 + 0.2) = 0.35$.
Case 3 studies the effect if the sizes of both proteins are equal, both indices are increased to $J\big(P(g_1), P(g_2)\big) = 0.33\bar{3}$ and $n\big(P(g_1), P(g_2)\big) = 0.5(\frac{2}{4} + \frac{2}{4}) = 0.5$. The Jaccard index indicates again a very low similarity, although half of the motifs in both proteins are equal. These three examples show that the Pfam-motif similarity $n_{\mathcal{P}}$ gives a bonus to small proteins. If their function is known annotation transfer to larger proteins is enhanced.

3.4.2 Comparison of B. subtilis and C. acetobutylicum

The calculated sets from the previous presented approach are listed in table 3.4.2. The number of genes with a reaction annotation is higher for *Bacillus* than for *Clostridium* (747 vs 600). Consequently, the number of reaction specific to Bacillus is more than twice as large as the number of reactions in *Clostridium* (388 vs 139).

Table 3.3: The download of information from KEGG concerning both organisms resulted in approximately equal sized genomes ($\mathcal{X}$). The *Bacillus* genome contains more genes with reaction-annotation ($X \setminus \mathcal{X}_0$) and more reactions ($\mathcal{R}_\mathcal{X}$) than the one of *Clostridium*. There are twice as much reactions that can be inferred from *B. subtilis* than from *C. acetobutylicum* ($\mathcal{R}_{\text{spec}}$) for the respective other organism. 267 genes are responsible for these reaction (X_{spec}). 512 motifs are found in this specific gene-set of *Bacillus*.

sets	***B. subtilis*** size	***C. acetobutylicum*** size
$\mathcal{X}$	4422	4021
$\mathcal{X} \setminus \mathcal{X}_0$	747	600
$\mathcal{R}_\mathcal{X}$	1041	792
$\mathcal{P}_\mathcal{X}$	5004	4494
$\mathcal{R}_{\text{spec}}$	388	139
$\mathcal{X}_{\text{spec}}$	267	145
$\mathcal{X}_{\text{final}}$	204	821

A map of $\mathcal{R}_{\text{spec}}^{\text{BS}}$ is given in figure 3.12. This map is one tool to track reactions that are not present in *C. acetobutylicum*. Connection of the genes according to their Pfam-motifs yields a second tool for rapid function suggestion (figure 3.13). Direct annotation transfer is possible for smaller connected components in this unfiltered map by considering high edge-weights because several protein functions are contained in few very specific Pfam-motifs, e.g. CoA-transferase activity. The correct reaction-annotation is given for several enzymes (table 3.4.2), however the protein does not contain an E.C. number in *C. acetobutylicum*. This proof of concept shows also that hypothetical proteins with unknown functions are mapped to proteins with functional annotation, suggesting hypotheses can be derived from such maps.

Possible model reduction is achieved by choosing a threshold for the edges weights filtering all edges with $n_\mathcal{P} < 0.5$, reduces the model size by 66% (figure 3.14).

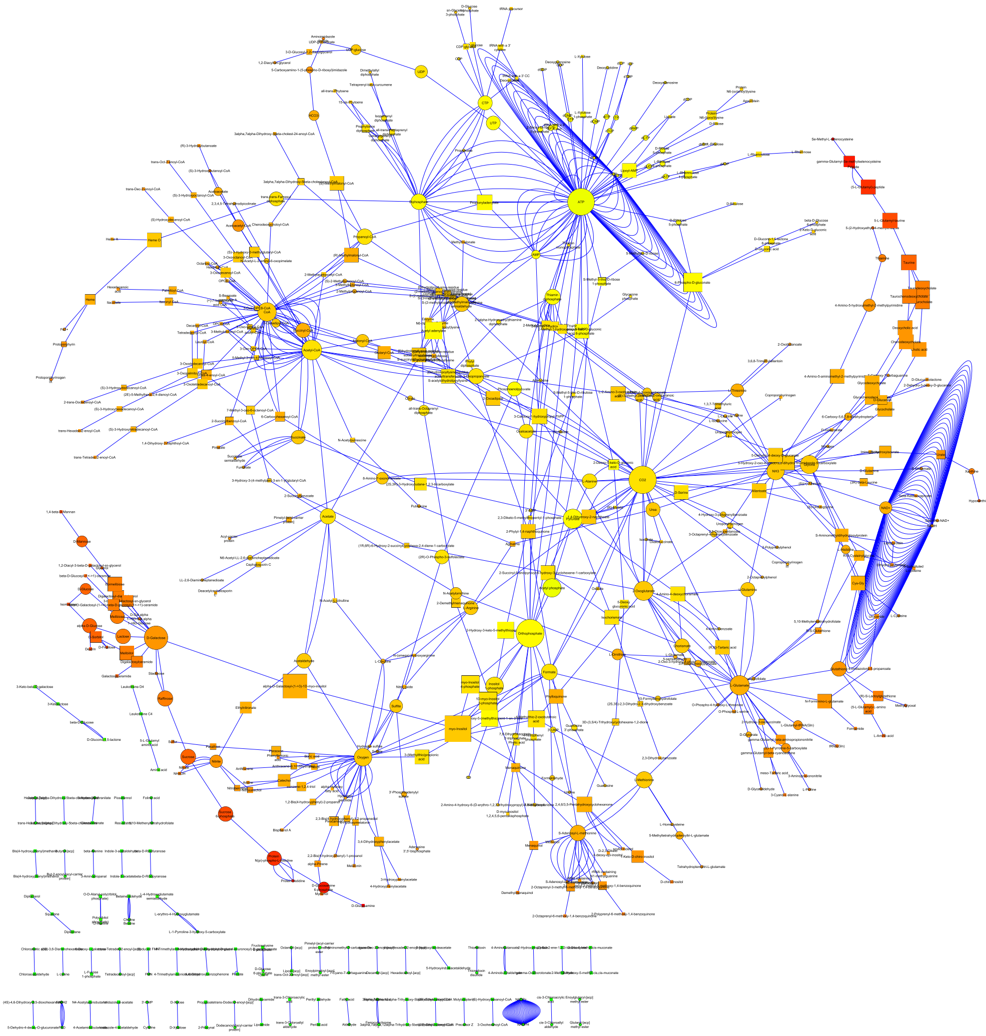

Figure 3.12: The reduced MMM of *B. subtilis* shows the reactions that are not annotated in *C. acetobutylicum*. As before, the size of the nodes correlates to stress and the colour to the eccentricity, red correspond to high values, and green to low values. This network consists of 577 metabolites, thereof 334 unique to *B. subtilis*.
rectangle: the compound is present only in *B. subtilis*.
ellipsoid: the compound when is present in both organisms.

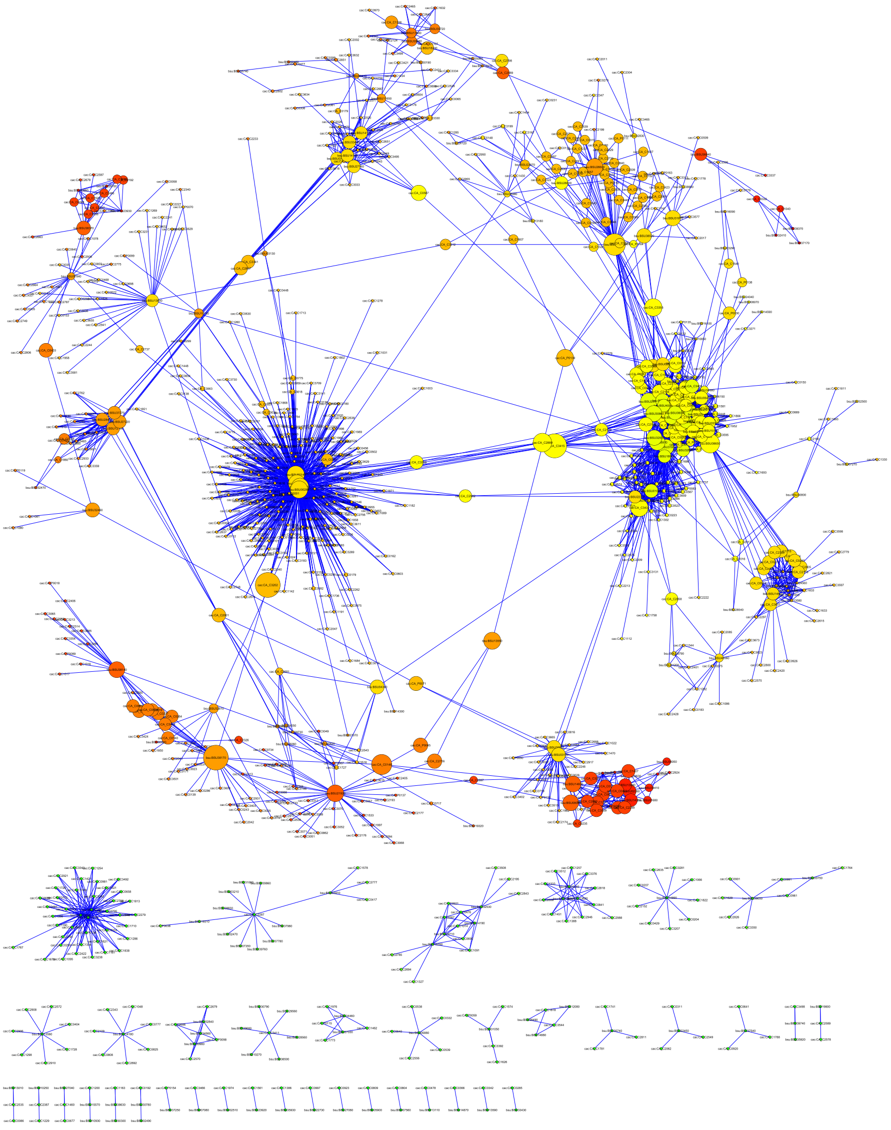

Figure 3.13: Genes of *C. acetobutylicum* are connected to the genes of *B. subtilis* which have an reaction annotation that is not found in *C. acetobutylicum*. The size of the nodes correlates with stress, and the colour corresponds to the eccentricity on an increasing green to red scale.

Table 3.4: Small connected components of this mapping either reproduce, expand or induce functions on proteins. The two columns of genes are connected to each other in the database. The first block of this table shows examples of enzymes that are correctly annotated and mapped together. These genes in *Clostridium* are not completely annotated as enzymes and represent gaps in the KEGG database. The second block represents the genes with poor or missing annotation, here the annotation can be considerably enriched with the *Bacillus* information.
DH = dehydrogenase, ST = sulfotransferase

x	**Annotation**	$\hat{x}$	**Annotation**
CA_{C0997}	nucleoside-diphosphate kinase	BSU_{22730}	same (EC:2.7.4.6)
CA_{C1200}	phospho-adenylylsulfate ST	BSU_{10930}	same (EC:1.8.4.8)
CA_{C1200}	phospho-adenylylsulfate ST	BSU_{15570}	same (EC:1.8.4.8)
CA_{C1462}	levanase/invertase	BSU_{34460}	same (EC:3.2.1.65)
CA_{C1462}	levanase/invertase	BSU_{27030}	same (EC:3.2.1.65)
CA_{C3498}	ribokinase sugar kinase	BSU_{35920}	same (EC:2.7.1.15)
CA_{C1574}	4-hydroxybutyrate DH	BSU_{31050}	choline DH (EC:1.1.1.-)
CA_{C3392}	butanol DH	BSU_{31050}	choline DH (EC:1.1.1.-)
CA_{P0059}	alcohol DH	BSU_{31050}	choline DH (EC:1.1.1.-)
CA_{C0804}	Pectate lyase related protein	BSU_{07560}	pectate lyase (EC:4.2.2.2)
CA_{C1190}	Fe-S-cluster redox protein	BSU_{32330}	lipoyl synthase [EC:2.8.1.8]
CA_{C1229}	hypothetical protein	BSU_{10250}	lipoate-protein ligase (EC:6.3.2.-)
CA_{C3238}	hypothetical protein	BSU_{32330}	lipoyl synthase [EC:2.8.1.8]

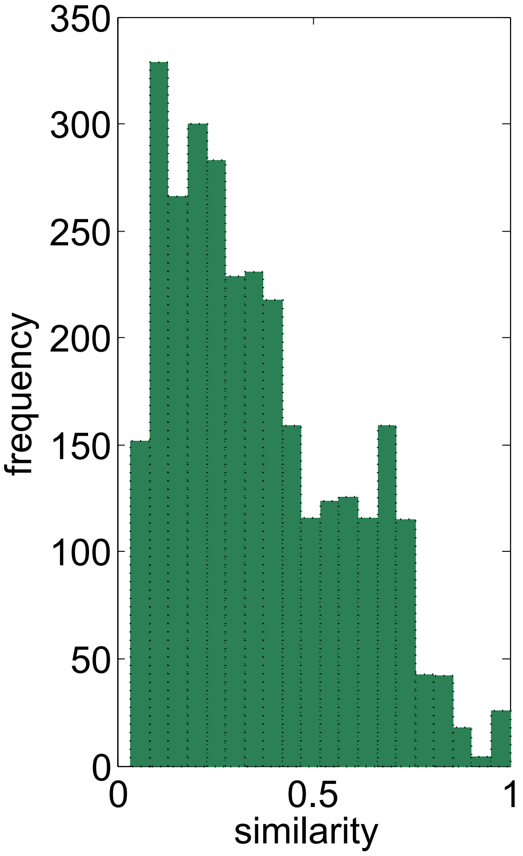

Figure 3.14: The distribution of edge weights of the comparative approach.

3.5 Annotation Transfer a Case-Study: 3-HBDH Activity in Clostridium

Here the example research of an enzyme activity in *C. acetobutylicum* will be carried out. After introduction of the general approach (3.5.1), experimental indications and database information are collected (3.5.2 and 3.5.3), then hypotheses will be given (3.5.4) and methods results will be shown (3.5.5). Finally some conclusions will be drawn (3.5.6).

3.5.1 Annotation Transfer Methods

Available Annotation Transfer Methods

Three different methods (M_1, M_2, M_3) will be considered for annotation transfer.

M_1: BLASTP mapping [Altschul et al., 1997] of two protein sequences, e.g. a protein of *C. acetobutylicum* to a close relative like *B. subtilius* was reported to detect enzymes with a comparable function [Rost et al., 2003] and architecture [Lee et al., 2008a].

M_2: Phylogenetic approaches have been used with success for annotation transfer [Pellegrini et al., 1999]. Comparing the Pfam-motifs of a chosen protein throughout different species gives a second approach. Indeed, combinations of domains are enriched in some functional classes [Forslund, 2011]. Pfam-domain architecture is accessible by several tools, e.g. the Weighted Domain Architecture Comparison Tool (WDAC) [Lee and Lee, 2009] or the Feature Architecture Comparison Tool (FACT) [Koestler et al., 2010]. Here, the Pfam-motifs responsible for an enzymatic activity in other organisms will be determined from a frequentist point of view and then these motifs will be retrieved in the clostridial annotation. This can be considered a preselection step before the more exhaustive Pfam-motif architecture approach is calculated.

M_3: Enzymes of one pathway are known to build stoichiometric complexes that channel the substrate [Srere, 1987]. Clustering of gene expression data is frequently used to reveal open reading frames and co-regulated genes [Tavazoie et al., 1999, Dhaeseleer et al., 2000]. By fixing a regulatory assumption, e.g. by choosing a gene of a target pathway, all possibly co-regulated genes are identified.

Integration of Annotation Transfer Methods

Any annotation transfer method M should be able to create subsets from the whole set of genes, they are named *candidates* $\mathcal{X}_{M_i}$

$$\mathcal{X}_{M_i} := M_i(\mathcal{X}). \tag{3.32}$$

In order to evaluate the retrieval of a candidate by a method, a score-matrix (C_M) is defined with the N_M methods as columns and the N_J genes as rows:

$$\mathbf{C}_M := \begin{cases} 1; x \in \mathcal{X}_{M_i} \\ 0; x \notin \mathcal{X}_{M_i} \end{cases} \tag{3.33}$$

Combining these counters in a ranking $\mathbf{s}_M$ with a column-vector of weights $\mathbf{w}$ enables the integration of all methods to one numeric value.

$$\mathbf{s}_M := \mathbf{C}_M \frac{\mathbf{w}}{||\mathbf{w}||} \tag{3.34}$$

3.5.2 Collection of Experimental Indications

Indications for the Presence of Crotonate

It was mentioned earlier that during the batch fermentation data a delta-2 oxidoreductase is up-regulated during acidogenesis and down-regulated during solventogenesis (3.3.3). The same pattern is obvious in the two stationary states of the continuous culture (figure 3.9 and figure 3.11). Measurements confirm that indeed crotonate is present in small amounts during continuous culture [*]. This suggests that there may be a pathway that uses crotonate.

Indications for a Unreckoned Butyrate Production Pathway

Mutants of acetate kinase or butyrate kinase did always produce minor amounts of both acids [Green et al., 1996] and butyrate was taken up by an unknown pathway - the established reverse kinase pathway and CoA-transferase activity were knocked-out [Lehmann et al., 2012b]. An acetoacetate decarboxylase knock-out mutant produces increased amounts of butyrate when supplemented with calcium carbonate and methyl viologen [Jiang et al., 2009]. In a different culture, knock-out mutants of acetoacetate decarboxylase have increased butyrate concentration and acetoacetate does not accumulate [Lehmann et al., 2012a]. Acetoacetate addition in pH-uncontrolled culture leads to increased butanol and butyrate production. In pH-controlled culture this effect is less pronounced [Papoutsakis et al., 1987]. This suggests that butyrate may be processed back to acetoacetate and vice versa under particular circumstances.

[*] personal communication, Kengen Laboratory, Wageningen

3.5.3 Collection of Database Information

Simplistic Approach to Alternative Production Pathways

Investigation of the clostridial reactome shows that no other reaction than the delta-2 oxidoreductase (E.C. 1.3.1.31, rn:R01689) is able to use crotonate as substrate or product. It seems unreasonable that an enzyme exists without any further integration of its substrates or products in other pathways. A simplistic approach would acknowledge that the enzymes (CA_{C2708}, CA_{C2710}, CA_{C2711}) using the CoA-derivates acetoacetyl-CoA, crotonyl-CoA, 3-hydroxybutyryl-CoA have a broad specificity towards non-CoA derivates. However, this unspecific reaction is not annotated in KEGG for these enzymes, also the crotonase is not annotated to process CoA-derivatives. It is further known that *C. acetobutylicum* possesses several distinct enzymes that do the same reactions, e.g. butanol dehydrogenases [Grimmler et al., 2011], [Duerre, 2005, p.678].

Complex Approach to Alternative Production Pathways

Plausible other possibilities of this alternative pathway are presented in figure 3.15. Variant A assumes that other CoA-transferases than ctfAB are producing the corresponding acids and CoA-derivates. Variant B assumes an unannotated dehydrogenase in *C. acetobutylicum* that uses acetoacetate as substrate and produces 3-hydroxybutyrate (B1) which is then the substrate for an unannotated dehydratase to produce crotonic acid (B2). Variant C suggest that this pathway can be inversed in direction.

There are No Unknown CoA-Transferases

The determination of variant A is straightforward through Pfam-motifs. The motif for a general CoA-transferases is the CoA_trans-motif. Which is available in *C. acetobutylicum* only in the already mentioned proteins ctfAB (CA_{P0163} and CA_{P0164}). This finding is complemented by a transacetylase inactivation strain that was not able to re-assimilated acids [Green et al., 1996]. Substrate specificity of the CoA-tranferases is broad, crotonate may be used as substrate, with an activity loss of 39% [Hartmanis et al., 1984]. Other authors report various other CoA-thioester substrates and cofactors for this enzyme [Barker et al., 1978].

No 3-Hydroxybutyrate Dehydrogenase is Known in C. acetobutylicum

Variant B1 is performed by the 3-hydroxybutyrate dehydrogenase (3-HBDH, EC 1.1.1.30, rn:R01361). Interestingly, no reaction for variant B1 is found in *C. acetobutylicum*, the only way metabolising acetoacetate is via decarboxylation [Papoutsakis et al., 1987]. However, in 750 other organisms there is this activity,

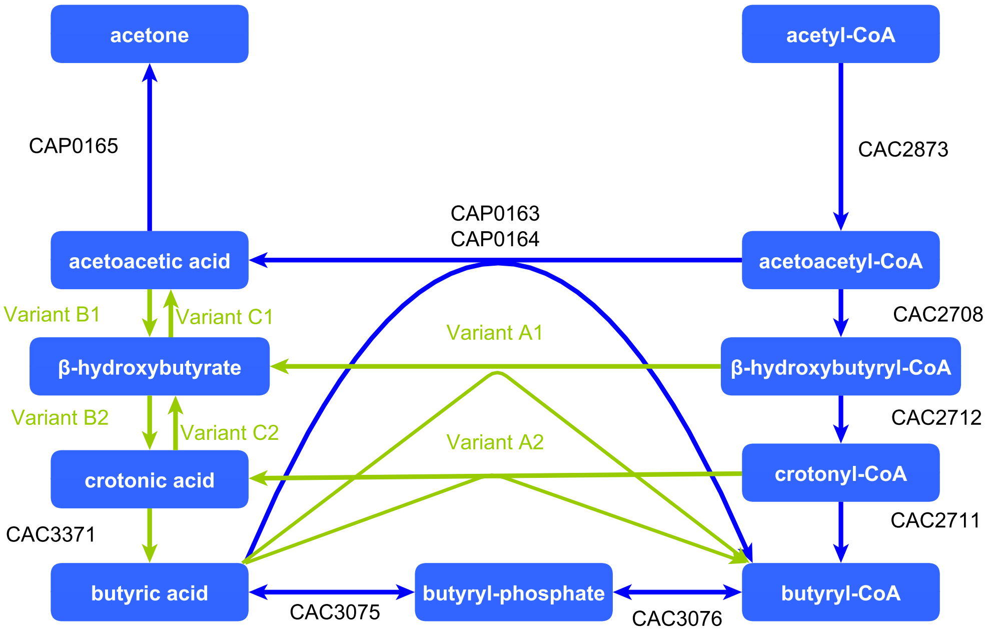

Figure 3.15: Alternative models of butyric ccid production suggest three different variants of crotonate production via CoA-transferases (variant A), acetoacetate consumption (variant B) or the inversed pathway (variant C)

one of them is *B. subtilis* (BSU_{38970}, yxjF). By considering the two established tools for comparison of *C. acetobutylicum* to *B. subtilis* the 3-HBDH reaction is visible in the metabolic network (figure 3.12) and several proteins similar to the *Bacillus* 3-HBDH are found (figure 3.13).

No 3-Hydroxybutyrate Dehydratase is Known

Finally, variant B2 is a dehydratase activity. While a motif is reported for CoA-dependent dehydratases, there is no known reaction catalysing the conversion from 3-hydroxybutyrate to crotonate. This reaction can only be an unspecific byproduct of another enzyme, or a product of a multi-step reaction within one enzyme. Fortunately, Pfam-motifs have further annotations, e.g. the Epimerase-motif which has a hydratase annotation.

3.5.4 Hypotheses for Annotation Transfer

M_1: BLASTP of the clostridial proteome against yxjF (BSU_{38970}) using KEGG.

M_2: A phylogenetic comparison of Pfam-motifs from all 750 annotated 3-HBDH to the clostridial genome using Taverna and MATLAB (B.1 and B.2).

M_3**:** Clustering of batch (ba) [Jones et al., 2008] and continuous (cu) culture data [Grimmler et al., 2011] according to three different regulatory scenarios:

A_1**:** Expression of the 3-HBDH is co-regulated to genes of the pathway that converts acetoacetyl-CoA to butyryl-CoA under acidogenic conditions, the responsible enzymes are encoded in the transcripts CA_{C2709}, CA_{C2711}, CA_{C2712} [Jones et al., 2008, Grimmler et al., 2011] (ac).

A_2**:** Expression of the 3-HBDH is co-regulated to genes of the pathway that converts acetoacetyl-CoA to butyryl-CoA under solventogenic conditions, the responsible enzymes are encoded in the transcripts CA_{C2009}, CA_{C2012}, CA_{C2016} [Jones et al., 2008, Grimmler et al., 2011] (so)

A_3**:** Expression of the 3-HBDH is co-regulated to the gene coding for the enoate-reductase CA_{C3371} (er).

Clustering is performed by Genesis, [Sturn et al., 2002] using a kmeans algorithm and the euclidean distance metric.

3.5.5 Results

Results of the BLASTP Approach

The gene BSU_{38970} from *B. subtilis* was used as example for an annotated 3-HBDH in a close relative. The BLASTP matches are noted in table 3.5. A small E-value shows significant matches to the target structure. However, the size of the protein may be significantly smaller than the target structure.

Results of the Phylogenetic Approach

750 organisms contain annotated 3-HBDH, harvest of their Pfam-motif showed that the average value of motif per enzyme is nine. Within the frequent motifs (table 3.6) the two adh-motifs and the KR-motif are predominant. They could serve as first criterion for research of the 3-HBDH. Following the list, several NAD-motifs are preserved throughout most of the species. One further observes the 3HCDH_N-motif and the Epimerase motif as being characteristic.
With the adh_short and the adh_short_C2 motif, 749 organisms are already covered. The only organism not covered is *Brucella melitensis*, its 3-HBDH contains only low abundant motifs: the TrkA_N, the 3HCDH_N and the 2-Hacid_dh_C. Eleven genes contain only the adh_short_C2-motif, they are found in the genus *Rickkettsia*. Another eleven organisms contain only the adh_short-motif. These are more heterogeneously spread than for the adh_short_C2-motif, containing higher vertrebrates (*Pongo abelii* or *Sus scrofa*) and bacteria, e.g. *Desulfobacterium autotrophicum*.

Table 3.5: BLASTP of the aminoacid sequence of BSU_{38970} to the proteome of *C. acetobutylicum*

Gene-ID	E-value
CA_{C2607}	5.00E-38
CA_{C3574}	5.00E-37
CA_{C3462}	9.00E-28
CA_{C0361}	3.00E-27
CA_{C2626}	3.00E-23
CA_{C1423}	7.00E-17
CA_{C3335}	9.00E-17
CA_{C1576}	5.00E-15
CA_{C2992}	8.00E-15
CA_{C1331}	1.00E-13
CA_{C0536}	4.00E-10
CA_{P0051}	6.00E-09
CA_{C3484}	5.00E-08
CA_{P0001}	7.00E-07
CA_{C3355}	8.00E-04

The number of considered motifs determines the number of candidates. From the pure frequentist point of view, the first four motifs seem promising as model for the 3-HBDH activity if not the *Brucella* gene would be annotated with the same function but without these motifs. In order to increase the list of candidates also low abundant motifs (at least 10% matching) will be considered as relevant. The corresponding solution set contains 79 proteins.

Results of the Regulatory Approach

The four genes CA_{C2708}, CA_{C2710}, CA_{C2711}, CA_{C3371} from the regulatory assumptions, were located in three different clusters during batch and continuous culture (refer to figure 3.16). For the operon that is up-regulated under solventogenic conditions (CA_{C2009}, CA_{C2012}, CA_{C2016}) the same cluster partition was used. Naturally within the results, all genes from the assumptions occur in these clusters. Three other genes are co-regulated in both experiments (*ba*, *co*) and bearing similar motifs: CA_{C2713}, the redox-sensing transcription repressor Rex, which is in the same open reading frame, a ketopantoate reductase named PanE/ApbA, CA_{C2937} and a nucleoside-diphosphate-sugar epimerase, CA_{C2166}.
The ketopantoate reductase, shares several motifs with CA_{C2708}: a NAD_binding_2-motif, three different dehydrogenases, the NAD_Gly3P_dh_N-motif, the 3HCDH_N-motif, responsible for the reduction of 3-hydroxyacyl-CoA and NAD-binding and the UDPG_MGDP_dh_N-motif. Finally, the ApbA-motif responsible for ketopantoate reductase activity is found in both proteins. Domains unique to the keto-

Table 3.6: List of 3-HBDH motifs most frequently occurring in 750 annotated KEGG species.

Motif	Frequency
adh_short	0.98
adh_short_C2	0.98
KR	0.97
Epimerase	0.70
Eno-Rase_NADH_b	0.42
3HCDH_N	0.40
NAD_binding_10	0.35
Polysacc_synt_2	0.23
Saccharop_dh	0.21
DUF1776	0.19
TrkA_N	0.16
THF_DHG_CYH_C	0.15
3Beta_HSD	0.14
RmlD_sub_bind	0.13
AdoHcyase_NAD	0.13
Shikimate_DH	0.11
2-Hacid_dh_C	0.10

pantoate reductase are a synthase for heptaprenyl diphosphate HEPPP_synt_1, DUF1879, a domain of unknown function, and a C-terminal ApbA-motif.
It should be noted, that CA_{C3371} shares many similar functions to CA_{C2708} as well, suggesting that this gene may also have more activities than already annotated.

Integration

The precedent approaches, the BLASTP-search, the phylogenetic comparison and the clustering all show different candidates (3.7). These weights are chosen to add to 9 for the experimental and for the database assisted methods. Since a gene can only be contained within one cluster for either culture method, the maximal score achievable is 6. The only exception to this occurs when two assumptions are contained within the same cluster, e.g. the CAC_{3335} is contained in the clusters of CA_{C2708}, CA_{C2709}, CA_{C2711} in batch culture, and the A1 and A3 condition partially overlap for three candidates. The score further prefers the database methods to the experimental methods by three points (9 versus 6 for batch and continuous culture). BLASTP is considered much inferior to Pfam-motif search (2 versus 7). The final ranking is shown side-to-side with the Pfam-motif similarity $n_{\mathcal{P}}$ in table 3.8. It is obvious from the table, that integration of different methods drastically changes the ranking compared to the Pfam-motif frequency alone. The sugar epimerase CA_{C2166}, the ketopantoate

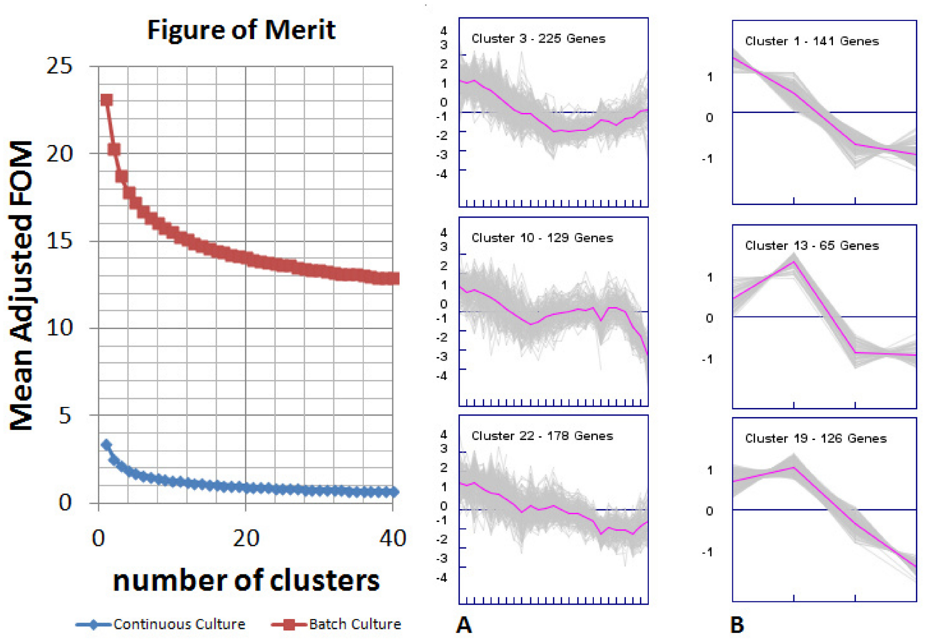

Figure 3.16: left: Figure of Merit [Yeung et al., 2001] of batch and continuous culture transcript expression levels as a function of number of clusters. right: (A) batch culture experiment. (B) continuous culture experiment. CA$_{C2708}$ belongs to A: cluster 22 and B: cluster 1, CA$_{C2710}$ belongs to A: cluster 10 and B: cluster 13, CA$_{C2711}$ belongs to A: cluster 10 and B: cluster 19, CA$_{C3371}$ belongs to A: cluster 3 and B: cluster 19.
The pink line is the average over the cluster. Grey lines are the candidate transcript expression level profiles.

reductase CA$_{C2937}$ and the short chain alcohol dehydrogenase CA$_{C3335}$ all seem very probable from both points of view. A Pfam-motif map of the three candidates is shown in figure 3.17. The ketopantoate reductase is apart from the two other candidates. The sugar epimerase only contains one adh_short-motif. Which is untypical for most of the 3-HBDH found in the phylogenetic research. Despite the SEFIR-motif, the unannotated dehydrogenase contains only motifs from the phyologenetic comparison. The ultimate step of identification can now be easily done.

Terminal Evaluation

From this short list, the Weighted Domain Architecture Comparison Tool (WDAC) [Lee and Lee, 2009] is able to rapidly calculate the domain-architectures. Somewhat surprisingly, however not completely unexpected from the BLASTP similarity, the domain architecture of the *B. subtilis* 3-HBDH and CA$_{3335}$ are identical to

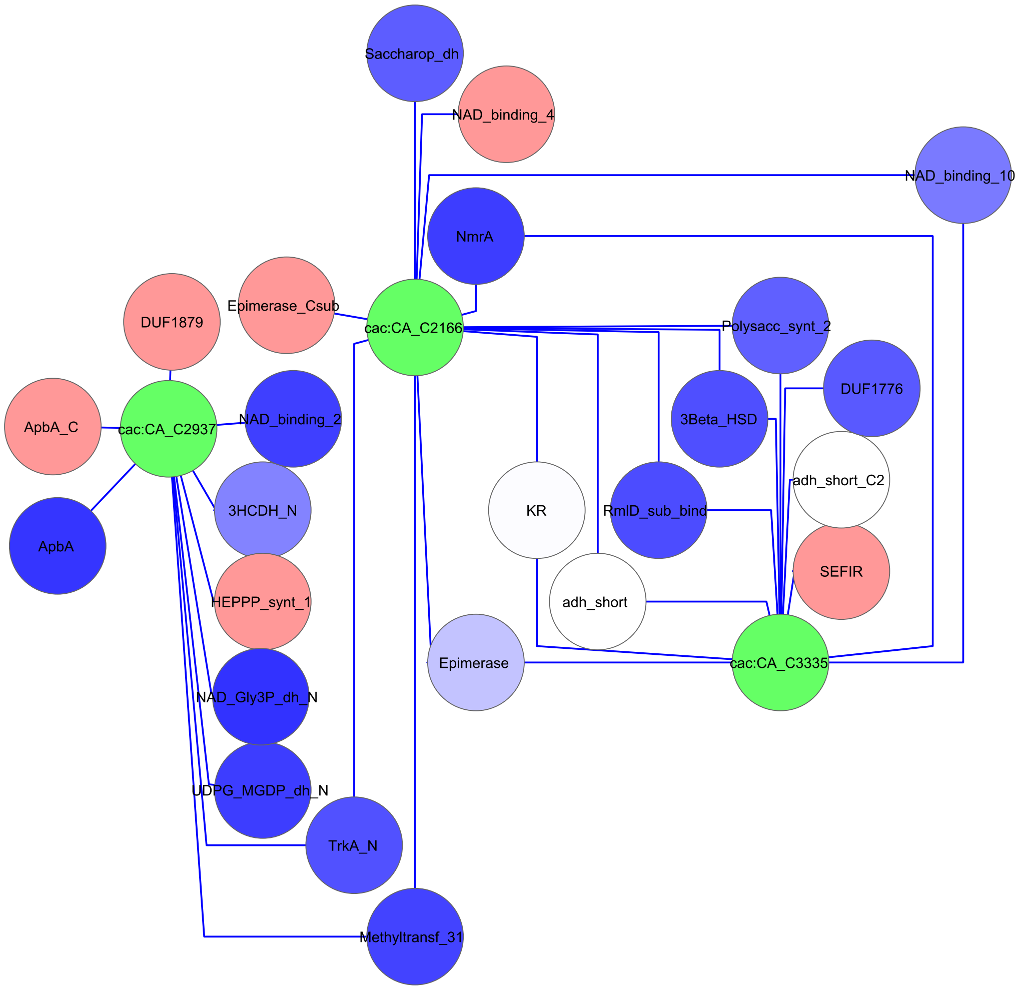

Figure 3.17: Map of 3-HBDH candidate genes and their Pfam-motifs.
green: gene identifier, red: motifs not retrieved by the phylogenetic Pfam-research, darkblue to white: increasing frequency of occurrence in the phylogenetic Pfam-research

Table 3.7: Integration of the three approaches with the assumptions and different datasets results in a data matrix **C** that is truncated to high score genes.

weights $\mathbf{w}$	**2**	**7**	**3**	**3**	**3**	**3**	**3**	**3**
Genes	M_1	M_2	$M_{3ba,ac}$	$M_{3ba,so}$	$M_{3ba,er}$	$M_{3co,ac}$	$M_{3co,so}$	$M_{3co,er}$
CA$_{C2166}$	0	1	1	0	1	1	0	0
CA$_{C3371}$	0	1	1	0	1	0	0	1
CA$_{C0267}$	0	1	1	0	1	0	0	0
CA$_{C2009}$	0	1	0	1	0	0	1	0
CA$_{C2708}$	0	1	1	0	0	1	0	0
CA$_{C2713}$	0	1	1	0	0	1	0	0
CA$_{C2937}$	0	1	1	0	0	0	0	1
CA$_{C3335}$	1	1	0	0	0	0	0	1
CA$_{C3355}$	1	1	0	1	0	0	0	0

Table 3.8: Application of the defined numerical weights produced priorities the candidates, the concomitant mapping of similarity to the *B. subtilis* Pfam-annotation represents a second dimension of information.

Genes g	Annotation	s_M	$n_{\mathcal{P}}(g, BSU_{38970})$
CA$_{C2166}$	nucleoside-diphosphate-sugar epimerase	0.59	0.65
CA$_{C3371}$	2-enoate reductase	0.59	0.06
CA$_{C0267}$	L-lactate dehydrogenase	0.48	0.23
CA$_{C2009}$	3-hydroxyacyl-CoA dehydrogenase	0.48	0.28
CA$_{C2708}$	3-hydroxybutyryl-CoA dehydrogenase	0.48	0.25
CA$_{C2713}$	redox-sensing transcriptional repressor Rex	0.48	0.07
CA$_{C2937}$	ketopantoate reductase PanE/ApbA	0.48	0.34
CA$_{C3335}$	Short-chain alcohol dehydrogenase enzyme	0.44	0.63
CA$_{C3355}$	polyketide synthase	0.44	0.21

each other. If there is a 3-HBDH in *C. acetobutylicum* experiments should aim at this gene first and then the two others.

3.5.6 Critical Evaluation

Starting from the research of unannotated reactions, a tool was constructed that compared two organisms based on their statistical Pfam-motif similarity. Differences of annotated reactions are readily visualised by a second tool, the mapping of R_{spec} in a MMM. From these two tools and experimental data, an alternative production pathway of butyrate was proposed. The researched 3-HBDH activity was conducted by integration of three different methods for annotation transfer, thereof one established, the BLASTP, one current approach in its simpler version, the Pfam-motif search, and regulatory investigation to integrate experimental data by clustering. As a result, candidates for the 3-HBDH in *Clostridium acetobutylicum* were proposed.

Remarks on Data

There is not yet strong evidence that a crotonate utilisation pathway is missing. The deletion of the 3-hydroxybutyryl-CoA dehydrogenase results in a butyrate and butanol deficient strain [Lehmann and Luetke-Eversloh, 2011]. This suggests that at least one product of this enzyme has a fundamental role for this pathway, e.g. crotonyl-CoA as CoA donor for hydroxybutyrate. Experiments with supplementation of crotonic acid or acetoacetate in minimial medium in pH-uncontrolled batch culture were inconclusive regarding the role of crotonate in the wildtype (results not shown).

Remarks on Tools

The comparison of *C. acetobutylicum* and *B. subtilis* via their Pfam-motif similarity $n_{\mathcal{P}}$ created a similarity map. This tool for the identification of annotation gaps can be used for any other organism and also by employing different similarity measures, as suggested by [Lin et al., 2006]. It is extendable by further organisms, however visualisation then gets a bigger challenge. Also, the retrieval of a suitable cut-off value for the similarity measure is not evident. Here an external criterion must be found.

Remarks on BLASTP

BLASTP was earlier encouraged [Rost et al., 2003]. It assumes the two types of enzyme are very similar, so it is required that yxjF is not a unique type enzyme as e.g. the 3-HBDH of *Brucella melitensis*. The Pfam-motif map of candidates from BLASTP is shown in figure 3.18. The 15 candidate proteins are centred around mainly six motifs: adh_short, adh_short_C2, KR, epimerase, NAD_binding_10, and 3HCDH_N. As was shown in the phylogenetic approach, these motifs are characteristic for most 3-HBDH and yxjF is indeed an adequate role-model for a 3-HBDH activity. However, the use of BLASTP is now discouraged [Forslund, 2011] because it is reckoned it is better suited to calculate a phylogenetic distance than a functional distance. It is a comparative approach which can only identify homologies, not analogies.

Remarks on Pfam-Motif Comparison

The here presented statistical approach for Pfam-motif comparison along different phylogenies is only possible if the motif is unspecific. Targets like a CoA-transferase, a fumarase, or a crotonyl-hydratase are readily identifiable by their respective Pfam-motif. Still, the information content of two different Pfam-motifs is not

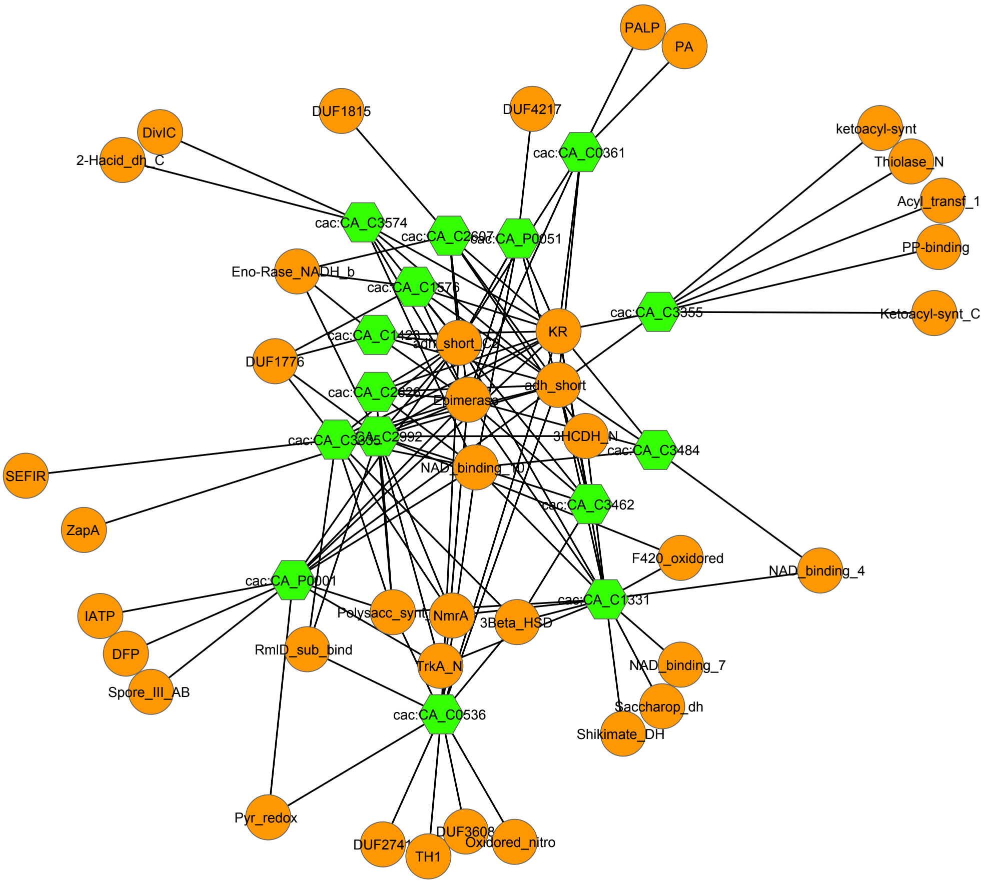

Figure 3.18: Pfam-motif map of 15 BLASTP candidates with low E-values. KEGG gene identifiers are marked in green hexagons, Pfam-motifs in orange circles.

identical and requires weighing according to their promiscuity along the kingdoms [Lee and Lee, 2009]. For the phylogenetic comparison of 750 different species and their 3-HBDH this was not urgently necessary because these enzymes are spread through all kingdoms and only few domains were highly promiscuous, e.g. the adh_motif.

A sole statistic remains erroneous, for this reason online tools like WDAC and FACT were used for the ultimate candidate selection. These tools require longer calculation times and a full genomic research is a computationally demanding task.

Pfam-motifs do not represent the sole possibility for annotation transfer, active-site profiling has been proven successful for the identification of highly conserved three dimensional structures of kinases from sequence data [Cammer et al., 2003].

Remarks on Clustering

Clustering has been proven worthy for functional annotation [Dhaeseleer et al., 2000]. Two plausibility criteria were used to restrain the candidate space. The first assumes that the known pathway for acetoacetyl-CoA and acetoacetate reduction runs in parallel to the unknown pathway. A similar approach was assumed by [Brown et al., 2000] who related tumor proteins to ribosomal proteins.
Nevertheless, false negative candidates are a risk with clustering as with any other data criterion since abstract assumptions are imposed: by choice of data curation and standardisation, by choice of clustering algorithm, by choice of distance metric [Brown et al., 2000, Brohee and van Helden, 2006, Freeman et al., 2007]. Despite this drawback, clustering induces a beneficial partition of the candidate space and partition of the candidate space through clustering allows the step-by-step elimination of invalid assumptions. If there is strong evidence that a gene cannot be co-regulated to a member of a cluster, all other members are also eliminated.

3.6 Final Conclusions

Overview

The manifold uses of pathway models were this chapter's main topic: the contextualisation of data, the guidance for metabolic engineering, the hypothesis-driven discovery and the network property discovery [Oberhardt et al., 2009]. This chapter started from the KEGG database and proposed Taverna as suitable tool for the harvest of metabolite reaction networks, multidimensional annotation and Pfam-motifs. For the analysis of the clostridial reactome, a formalism was introduced that allowed the integration of transcriptome data and the formal reaction-database to a data-supplemented database or a data-driven database. Several visualisation softwares and visualisation methods were tested to allow manual investigation of this database. From this, research of the 3-hydroxybutyrate dehydrogenase activity within existing annotations was motivated. A scheme was proposed to reveal candidates from integration of three different methods, database related and experiment related.

Contextualisation of Data and Network Property Discovery

Huge amounts of data are produced and deposited, metabolomic data and transcriptomic data stand side by side, the integration of omics is a necessary step in research [Joyce and Palsson, 2006]. In 2011 the HITS-Institute published an article that data-driven science represents a challenge for computer sciences and a re-thinking of the roles of hypothesis driven approaches to organisation of data into meaningful sets [Reuter, 2011]. Information retrieval and evaluation is facilitated when data is partitioned into smaller sets [Khatri et al., 2012]. In this thinking, the here presented pathway model approach represents one possibility to coherently, self-consistently organise different data from different experiments: Evaluation of transcriptome data on the reactome level reduces the number of considered transcripts and it enables the evaluation the data in terms of graph analysis.

The first step in this organisation is to make sense of the data and try then to infer structures from the data. This is defined as a top-down approach by [van Riel, 2006]. In several aspects, the here presented model contains a top-down approach, because it integrated the data by using two logical rules, the result is a subgraph that serves for several evaluations: Graph centralities help in manual ranking metabolites according to their position in the network, the edges to nodes fraction was able distinguish solventogenesis and acidogenesis, qualitative assessments like the connectivity of a single metabolite through out the different states provides first directions for its importance.

Known approaches of pathway analysis remain valid for this type of database in order to create further metadata: Dynamic properties of the network can still be derived [Klipp et al., 2004], network-motifs can be enriched [Joyce and Palsson, 2006], elementary modes and cut sets can be calculated [Klamt and Gilles, 2004].

Hypothesis-Driven Discovery and Guidance For Metabolic Engineering

The observation of crotonate synthesis is just one example how hypotheses are built from such a model. Gene-Reaction Networks are frequently encountered in literature, still models concentrate on flux balance calculation [Durot et al., 2009], the alternative evaluation of transcriptome data in the here proposed graph-based format seems underrepresented for hypothesis finding and metabolic engineering strategies. One reason for this is the focus on statistical evaluation of differentially expressed genes [Patil and Nielsen, 2005]. Differential expression can be understood as a majority criterion for data reduction [Yang et al., 2005]. Within the spirit of personalised medicine and individual treatments [W and BM, 2007, Katsnelson, 2013], the focus on single genes and their position within the whole network needs to be restrengthened. This type of data-model should be understood as a complement to statistical science. Crotonate synthesis is unreported in *C. acetobutylicum*, what other annotations are missing in the published models?

The annotations of genomes is a long lasting process [Khatri et al., 2012]. This work proposes therefor a comparison tool and integration schemes of data and the Pfam-database for annotation discovery. Within this scope the here introduced metabolic networks, comparison tools and the ranking score construct several what-if scenarios that aid in metabolic engineering [Aittokallio and Schwikowski, 2006, Durot et al., 2009]. For several well studied organisms a reactome knowledgebase is established [Matthews et al., 2009], this work is a methodological contribution to it.

The experimental investigation of a hypothesis is the ultimate step. It seems to be carried out with increasing ease: Multiple-site mutants of *C. acetobutylicum* become more and more frequent, e.g. [Jiang et al., 2009, Sillers et al., 2009, Lehmann et al., 2012a, Lehmann et al., 2012b].

Chapter 4

Automated Dynamic Model Creation

Little by little, one travels far
John Ronald Reuel Tolkien

This chapter starts with the question which metabolic engineering strategies can be employed in order to increase butanol production. A dynamic model of butanol production that integrates time series of transcript data and metabolome data will be therefor used. Existing approaches are reviewed first (4.1) and the unique properties of this model introduced. From mass balance equations (4.2) a formalism for the implementation in the IT-architecture will be given (4.3). This model will be used for the parameter estimation of two different experiments (4.4). Hypotheses will be generated by employing global sensitivity analysis (4.5) which are followed by the conclusions (4.6).

4.1 Historical Perspective

Algebraic Rules For Flux Balance Analysis

Early studies on the stoichiometry of ABE-fermentations revealed simple algebraic rules to connect data with unknown information like the energy value. Their application was to check data inconsistencies [Yerushalmi et al., 1983, Papoutsakis, 1984]. These rules additionally lead to a program that enables the use of constraint flux balance analysis (FBA) that gives insights in the flux distributions of the organism. It was used by [Junne, 2010] to evaluate the outcome of stimulus response experiments in batch and continuous culture. Other applications are known [Desai et al., 1999, Lee et al., 2008a, Senger and Papoutsakis, 2008].

Integration of mRNA Yielded Superior Results

The early model by [Votruba et al., 1986] is a data-based model that is driven by curve fitting of different batch fermentation results to some function. Its kinetics are Michaelis-Menten type kinetics or directly proportional with butanol based inhibition terms. Direct relations between the different compounds through metabolic pathways are not considered. However, this model introduces a metabolic activity functional based on total RNA, which helps in describing the culture's history and consequently culture growth. In the same year, a model for glucose uptake was published by [Yerushalmi et al., 1986b], they were proposing an active site model that explains substrate internalisation and product externalisation. The same authors also extended their model to relate mRNA concentrations and butanol production from glucose. The mRNA is given the role to reflect culture states. They assume diffusion of compounds through the cell membrane and inhibition by butanol [Yerushalmi et al., 1986a, Yerushalmi et al., 1988].

Integration of Inhibitory Effects of Acids And Solvents

The model by [Jarzebski et al., 1992] aims at the understanding of a chemostat at different pH values and thereby dissociation states of butyric acid. This coupling influences growth and the onset of solventogenesis. A set of logical rules covers inhibitory effects of butyric acid. This model describes the data well: Two steady state values and a sustained oscillation are covered. Still, this type of logical rules is insensitive to major system changes, as e.g. fermentations of mutants or changes of medium composition. A product inhibition model was proposed by [Özilgen, 1988] to describe several fermentation experiments. It assumes logistic growth and inhibition of accumulated products.

Integration of Biochemical Pathway Information and pH

The Shinto model [Shinto et al., 2007] proposes the first mechanistic view on the ABE fermentation by considering the underlying biochemical pathways. Kinetics are Michaelis-Menten type, the CoA transferase reaction from butyrate or acetate to the corresponding CoAs is a random bi-bi mechanism. It is a batch fermentation model that does not distinguish between intracellular and extracellular metabolites. Similarly to the other models it considers biomass production that is inhibited by butanol. Here biomass production is proportional to acetyl-CoA levels. [Junne, 2010] also proposes a dynamic model for solvent formation. This model includes pH mediated dissociation of compounds. Enzyme concentrations are modelled as sigmoidal function accordingly to the available transcript expression levels during batch fermentation. Additional transport terms for acids are implemented. Finally, models were developed that assume the translation of enzymes is pH-dependent, these models explain the pH-shift in the continuous chemostat experiment as proposed by the COSMIC SOP [Haus et al., 2011, Millat et al., 2013a].

Integration of Time Series of Transcriptome Data in Kinetic Models

The integration of transcriptome data in a time series format and combine it with metabolome was introduced by [Götz and Reuss, 2009], this approach seems unique in literature. A stochastic model is proposed by another group that involves enzyme concentrations that are integrated as time profiles [Liebermeister and Klipp, 2006], integration of metabolome and transcriptome regulation in a flux balance model is described elsewhere [Covert et al., 2001]. A simpler approach of comparing two different concentrations is reported for a lactate dehydrogenase model, where ratios of different isozymes are incorporated [Downer et al., 2006]. In *E. coli* regulatory networks are inferred from mRNA data and integrated to the mass-balance as an ODE system [Carrera et al., 2009].

4.2 Derivation of the Dynamical Model

The creation of a model for an organism offers many tasks. Clearly, the manifold connections within a biological network, the interactions within one Omic and the interaction between Omics, they all represent a tremendous amount of data to be mapped into a model. Such model ideally covers the *in vitro* behaviour under the chosen environmental conditions. However, calculation and identification of all factors remains impossible, not only because functions of proteins remain obscure (3.5), the amount of available data does usually not suffice to predict a unique parameter set. Consequently, it is not the task to integrate all known interactions, but to find a minimal model to understand the function of the organism [Durot et al., 2009].

This chapter proposes such a model for butanol production in *C. acetobutylicum* and a formalism for the implementation of time series of transcript level data.

Model Structure

Within the cell, the biochemical network of butanol synthesis is considered as already presented (figure 2.1). Reduction of the pathways to the branching points yields the network shown in figure 4.1. Arrows in this pathway indicate the reaction direction as considered in the model. Touching arrows indicate the concomitant use of several substrates to several products. Stoichiometry is not shown in this model representation. The CAC and CAP numbers represent the transcripts considered as relevant for the reaction. A list of compounds and used abbreviations is given in table 4.1.

Reduction to branching points assumes that no intermediate molecule has regulatory functions. Since literature suggests a regulatory-role of acetyl-phosphate and butyryl-phosphate they remain included for monitoring purposes. Glucose uptake and glycolysis are combined to one reaction due to the fact that measurements of intermediates are difficult to access. Finally, kinetics of this model are oriented on Gheshlagi et al. [Gheshlaghi, 2009] and the PhD thesis of Stefan Junne [Junne, 2010].

A structured model approach is taken here, the cell and the reactor are two separate entities. Transport phenomena within each compartment will be furthermore neglected. Transport between compartment boundaries is simplified by assuming an intracellular substrate is converted into an extracellular product.

The goal of this model is to monitor the current status of the cell not simply by some pH-dependency but by following the expression of the relevant transcripts. Therefore, no pH description is necessary and the dissociation state of acids is unconsidered. A model of non-autonomous differential equations is achieved that combines three types of information: the biochemical pathways, the transcript level data for each enzyme involved in the pathways, and the enzyme kinetic information from literature.

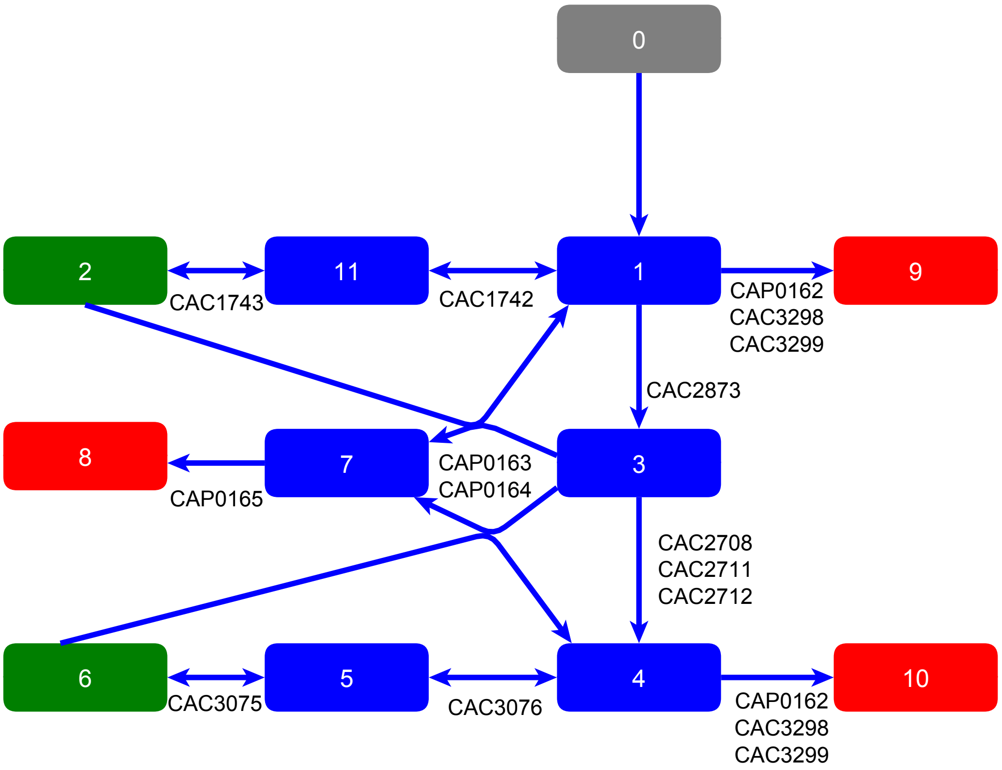

Figure 4.1: Minimal Model pathway structure for butanol synthesis by *C.acetobutylicum*. The grey box is the substrate *0:* glucose.
Red boxes show the solvents (*8:* acetone, *9:* ethanol, *10:* butanol).
Green boxes show the acids (*2:* acetate, *6:* butyrate). Blue boxes intracellular intermediates (*1:* acetyl-CoA, *3:* aceto-acetyl-CoA, *4:* butyryl-CoA,*5:* butyryl-phosphate, *7:* acetoacetate, *11:* acetyl-phosphate).

4.2.1 Derivation of the Model

Comparison of Compartment Volumes

The conversions take place in a cell, which is considered a compartment contained in the reactor. The cell density (ϱ_X) is defined as dry mass of cells (m_X) per cell volume (V_C). The mass of cells further corresponds to the measured biomass concentration (c_X) in the reactor volume (V_R):

$$\varrho_X = \frac{m_X}{V_C} = \frac{c_X V_R}{V_C} \qquad (4.1)$$

Compound number	compound	CHEBI-ID	Abbreviation
0	glucose	17234	Glc
1	acetyl-CoA	15351	ACoA
2	acetic acid	15366	ACE
3	acetoacetyl-CoA	15345	AACoA
4	butyryl-CoA	15517	BCoA
5	butyryl-phosphate	17260	BUP
6	butyric acid	30772	BU
7	acetoacetate	13705	AA
8	acetone	15347	ACN
9	ethanol	16236	ETOH
10	butanol	28885	BUOH
11	acetyl-phosphate	15350	ACP

Table 4.1: Compound Overview

The volume of the liquid phase (V_L) is the difference of reactor volume and cell volume:

$$V_L = V_R - V_C = V_R\Big(1 - \frac{V_C}{V_R}\Big) = V_R\big(1 - \frac{c_X}{\varrho_X}\big) \tag{4.2}$$

Biomass of *C. acetobutylicum* is not growing to high cell densities, hence we can assume $\frac{c_X}{\varrho_X} \ll 1$ and thereby we neglect the cell volume compared to the reactor volume:

$$V_L \approx V_R. \tag{4.3}$$

Relating Reaction Rates to Cell and Reactor

The total reaction rate (R) is the amount of one substance (n) being converted over time (t). Since the model is a two-compartment model, this reaction can be referenced to the cell (r^i) or to the reactor (r^o).

$$R \ = r^i V_C = r^o V_L \tag{4.4}$$

from equation 4.3 follows

$$r^o \ = r^i \frac{V_C}{V_R} \tag{4.5}$$

from equation 4.1 follows

$$r^o \ = r^i \frac{c_X}{\varrho_X} \tag{4.6}$$

The Time Law for Changes of Intracellular Concentrations

In the next step, the time law for the change of an intracellular component from the mole balance is derived. A substance n_i is produced and consumed by N_R reactions. The balance for an intracellular metabolite yields:

$$\frac{dn_{\mathrm{i}}}{dt} = \sum_{k=1}^{N_R} v_k R_k \tag{4.7}$$

$$v_k = \begin{cases} -1 & \text{if } n_i \text{ is a substrate} \\ +1 & \text{if } n_i \text{ is a product} \end{cases} \tag{4.8}$$

Expanding the left-hand side of equation 4.7:

$$\begin{aligned} \frac{dn_{\mathrm{i}}}{dt} &= \frac{d(c_i V_C)}{dt} && (4.9)\\ &= \frac{d(c_i V_R \frac{c_X}{\varrho_X})}{dt} && (4.10)\\ &= \frac{dc_i}{dt}\frac{c_X V_R}{\varrho_X} + \frac{d\frac{c_X}{\varrho_X}}{dt} c_i V_R + \frac{dV_R}{dt}\frac{c_X}{\varrho_X} c_i && (4.11) \end{aligned}$$

Now assume constant cell density and exponential growth at a rate μ and divide by the reactor volume:

$$\frac{dc_i}{dt}\frac{c_X}{\varrho_X} + \mu c_i \frac{c_X}{\varrho_X} + \frac{1}{V_R}\frac{dV_R}{dt}\frac{c_X}{\varrho_X} c_i = \sum_{k=1}^{N_R} v_k r_k^o \tag{4.12}$$

Factorise $\dfrac{c_X}{\varrho_X}$ and simplify the right-hand side according to equation 4.3

$$\frac{c_X}{\varrho_X}(\mu c_i + \frac{dc_i}{dt} + \frac{c_i}{V_R}\frac{dV_R}{dt}) = \sum_{k=1}^{N_R} v_k r_k^o \tag{4.13}$$

For batch and chemostat, the reactor volume remains constant $\left(\frac{dV_R}{dt} = 0\right)$ and after simplification and rearrangement one gets

$$\begin{aligned} \frac{c_X}{\varrho_X}(\mu c_i + \frac{dc_i}{dt}) &= \sum_{k=1}^{N_R} v_k r_k^o && (4.14)\\ \frac{c_X}{\varrho_X}(\mu c_i + \frac{dc_i}{dt}) &= \sum_{k=1}^{N_R} v_k \frac{R_k}{V_R} && (4.15)\\ \frac{dc_i}{dt} &= (\sum_{k=1}^{N_R} v_k r_k^i) - \mu c_i. && (4.16) \end{aligned}$$

The Time Law for Changes of Extracellular Concentrations

As before, a substance n_o is produced or consumed by N_R reactions. Additionally, it is transported out of the reactor by transport term T_{sink}.

$$\frac{dn_{\text{o}}}{dt} = \sum_{k=1}^{N_R} v_k R_k - T_{\text{sink}} \tag{4.17}$$

In continuous culture this sink is due to a pump and therefor dependent on the pump flow rate F:

$$T_{\text{sink}} = F c_o \tag{4.18}$$

Converting into concentrations and dividing by V_R, where $D = \dfrac{F}{V_R}$ is the dilution rate:

$$\begin{aligned}\frac{dc_o}{dt} &= \sum_{k=1}^{N_R} v_k \frac{R_k}{V_R} - \frac{F}{V_R} c_o && (4.19)\\ &= \sum_{k=1}^{N_R} v_k r_k^i - Dc_o && (4.20)\end{aligned}$$

4.2.2 Formalism

For the automated integration of the mathematical equations and biochemical network into the model, a formalism is required that is applicable to both compartments such that it is only necessary to specify whether a compound is located in one or the other.

In a first step, a compound is given an unique number either manually or by a reckoned online repository like from the database and ontology of Chemical Entities of Biological Interest (CHEBI). Using CHEBI-IDs is one step forward to sustainability of the model, they are unique and searchable through online services, and may be parsed in SysMO-SEEK.

The second step concerns the naming of reaction, as KEGG reactions are undirected and unrelated to compounds, the use of these identifiers is not recommended here. In order to give a reaction a direction, they are called $r_{s|p}$, where s and p are substrate-ID and product-ID delimited by the "|" character. Without loss of generality, multiple substrates and multiple products can be introduced $r_{s_1,s_2,\ldots,s_N|p_1,p_2,\ldots,p_M}$. Equations 4.16 and 4.20 can be generalised to equation 4.21:

$$\frac{dc_k}{dt} = \sum_j \tilde{V_j} r_{j|k} - \sum_i \tilde{V_i} r_{k|i} - D_k \cdot c_k \tag{4.21}$$

The parameter v_k is not necessary anymore, because the direction of the reaction is clear from the reaction identifiers, $r_{j|k}$ is the production of substance k,

reaction	kinetic	transcripts involved (cac:CA_...)
$r_{0\|1}$	glucose feed	
$r_{1\|11}$	MMT	$C1742$
$r_{11\|1}$	MMT	$C1742$
$r_{11\|2}$	MMT	$C1743$
$r_{2\|11}$	MMT	$C1743$
$r_{2,3\|1,7}$	bi-substrate MMT	$P0163$, $P0164$
$r_{1\|3}$	bi-substrate MMT	$C2873$
$r_{3\|4}$	substrate inhibition	$C2708$, $C2710$, $C2711$
$r_{4\|5}$	competitive product inhibition	$C3076$
$r_{5\|4}$	MMT	$C3076$
$r_{5\|6}$	MMT	$C3075$
$r_{6\|5}$	MMT	$C3075$
$r_{6,3\|4,7}$	bi-substrate MMT	$P0163$, $P0164$
$r_{7\|8}$	MMT	$P0165$
$r_{1\|9}$	MMT	$C3298$, $C3299$, $P0162$
$r_{4\|10}$	uncompetitive product inhibition	$C3298$, $C3299$, $P0162$
$r_{3\|7}$	$r_{2,3\|1,7} + r_{6,3\|4,7}$	$P0163$, $P0164$

Table 4.2: Model reaction kinetics and transcripts: The $r_{s|p}$ reaction-identifier shows the directed conversion from substrate s to product p using the transcript data from the respective transcript-identifiers. The kinetic model kin is specified in the second column. MMT: Michaelis-Menten type

$r_{k|i}$ is the consumption of substance k. However, the two process designs, batch and continuous culture, require a generalisation of the dilution D_k and the two compartment system requires a the factor $\tilde{V}$.

D takes three values:

$D_k = 0$: batch conditions, k is extracellular

$D_k = \mu$: batch conditions, k is an intracellular compound.

$D_k = \dfrac{F}{V_R}$: continuous conditions, k is either extracellular or intracellular

$\tilde{V}_j$ and $\tilde{V}_i$ also take three values:

$\tilde{V} = 0$: There is no reaction between the compounds k, i or k, j.

$\tilde{V} = \frac{\rho_X}{c_X}$: k is an intracellular compound.

$\tilde{V} = 1$: k is an extracellular compound.

Kinetic Laws and Integration of Transcript Levels

The amount of enzyme present in the cells is assumed to be time-dependent by explicitly integrating time-series data of the corresponding transcripts (table 4.2) into the rate equation. The type integration is explained in the next section, here it is an unspecified function f.

The reaction rate further depends on a specific kinetic (kin, table 4.2) that is a function of substrate concentrations (c_s), product concentrations (c_p) and in case of inhibitions also on any other species c_i.

The enzymes maximal rate $\bar{k}_{s|p}$ [mole/time unit] is related to this kinetic. This rate is the product of the maximal specific rate $k_{s|p}$ [mole/(time unit, g biomass and amount of enzyme)] the biomass c_X [g biomass] and the amount of enzyme which is a function of transcript levels f. This gives the generalised reaction rate equation from a substrate s to a product p:

$$\begin{aligned} r_{s|p} &= \bar{k}_{s|p}\mathrm{kin}_{s|p}(c_s, c_p, c_i) && (4.22) \\ &= k_{s|p} \cdot f(\text{transcript levels of } r_{s|p}) \cdot c_X \cdot \mathrm{kin}_{s|p}(c_s, c_p, c_i) && (4.23) \end{aligned}$$

By this definition, the established ODE-system of non-autonomous equations constitutes a descriptive model of the ABE-process.

4.2.3 Integration of Time-Dependent Data

Description of Growth And Glucose Consumption

Growth and glucose consumption are not modelled using mass balance, nor are they coupled to occurring reactions, e.g. acetyl-CoA was used for growth modelling [Shinto et al., 2007]. They are instead implemented as piecewise linear interpolations of the data. It is known that linear interpolations only poorly describe the data, however non-linear estimators or functional data analysis, require deep knowledge on data-structure and, more importantly, a sufficient large sampling set [Lehmann et al., 1999, Gustafsson et al., 2009]. Since replicates in fermentation experiments are rare, application of these advanced methods is difficult.

Therefore, it is assumed that the acetyl-CoA influx is directly proportional to the glucose uptake of the cell from the medium. The proportionality constant is the substrate yield Yp_{Glc}, it models the fraction of glucose used for growth compared to the glucose used for metabolite synthesis.

Growth is also considered directly proportional to the measured optical density in the medium.

Description of Transcript Levels

Transcript levels are also implemented as piecewise linear interpolations for the same reasoning as before. Data sparseness in the temporal dimension is usually more severe in that data than for growth or substrate profiles. Nevertheless, other studies were successful in establishing a transcription model that can be used for integration of transcript levels [Chen et al., 1999].

In order to use transcript data as a protein quantity it is necessary to assume that transcript levels map to protein levels. This assumption requires the study of two processes, protein translation and mRNA stability. Since *C. acetobutylicum* is not rapidly growing, the ribosome quantity remains constant during duplication of the cells and is not limiting [Golding et al., 2005]. Second it is necessary to scale the data to an upper bound. Since data are present in logarithmic format, the maximum should be scaled to zero to achieve maximal flux at least once during the time course of the experiment.

Third, *in vivo* transcript stability is an unknown parameter in this model. It is a function of the cell's status and may also differ for each transcript, a complete modelling of transcript translation and degradation was carried out earlier [Arnold, 2002]. It is clear from this modelling that the interplay of translation and degradation is not a linear function, and the transcriptome time series data needs to be shifted in a non-linear fashion in order to map to proteome data.

However, no proteome data in such temporal dimension is available and even if it were available the here applied methods would not change. Consequently, transcript expression data will be used as development standard until enough protein data is available.

Lumped reactions, e.g. the three reactions from acetoacetyl-CoA to butyryl-CoA are calculated as average of transcript levels, since usually the corresponding genes are organised in an operon and are expected to behave similarly.

Model Equations

The entire models equations are noted in appendix A.

4.3 Model Implementation

Various tools are available to implement ordinary differential equations for simulation and parameter estimation. The integration of temporal profiles from data for transcripts, biomass and glucose represent an additional requirement to the software.

Only two packages were found that fulfil this requirement, one is SBTOOLBOX2 [Schmidt and Jirstrand, 2006], a third party toolbox for MATLAB written by Henning Schmidt. Its computing power has been widely used in the biological community. It offers a graphical user interface that allows the execution of complex calculation tasks and it offers a library of scripts that can be freely accessed. The evaluation of C-script by MATLAB, called MEX-compilation, also greatly increases performance of the scripts.

The commercial SimBiology Toolbox by MATLAB can be used as well, however, its functionality is very limited when it comes to further analyses, e.g. sensitivity analysis.

Data Pre-Requisites

Transcript data must not have no missing values. Instead of linear interpolation, other imputation techniques are less error prone. The Metagenealyse-webpage uses principal component analysis (PCA) for imputation [Daub et al., 2003] of missing data in time-profiles of transcript data.

The Standard Format

A model standard-format is required to automatically integrate time-series data. It should be parsable by MATLAB and convertable to a readily calculable SBTOOLBOX2 model (B.3).

The standard-format is sketched in a toy-model in figure 4.2. First, the basic structure of the SBTOOLBOX2-model serves as core model and necessary parameters are predefined and therefor set to 0 (rGlcIn - the glucose influx into the organism, mue - the growth rate as determined from the change of optical density, cX - the biomass as determined from the optical density). The piecewise linear interpolations are calculated from matrix data. Second, the implementation of transcript data interpolation requires that transcript-identifiers $(T1, T2)$ can be separated and addressed. The format of a multidimensional function $f(T1, T2)$ makes this possible.

Initially, SBML models would be suitable as sustainable formats for model deposit and retrieval, as well as interactivity between several softwares for visualisation and calculation. However, this newly developed data-driven model type does not fit into a SBML-standard so far. SED-ML standards may cover this shortly. Until then this approach has to suffice.

```
********** MODEL NAME
Toy Model
********** MODEL NOTES
This model serves for illustration of the automated data implementation
procedure.
********** MODEL STATES
d/dt(x)=(rhoX/cX)*(rGlcIn - rxy)
d/dt(y)=(rhoX/cX)*rxy-mu*y
x(0)=0
y(0)=0
********** MODEL PARAMETERS
rhoX=300
kxy=0.1
K=0.1
********** MODEL VARIABLES
rGlcIn=0
cX=0
mu=0
********** MODEL REACTIONS
r0x=rGlcIn
rxy=f(T1,T2)*cX*kxy*x/(x+K)
********** MODEL FUNCTIONS
********** MODEL EVENTS
********** MODEL MATLAB FUNCTIONS
```

Figure 4.2: Structure of the standard model in SBToolbox2: The model is given a name that is used through out the model analysis. MODEL STATES are the compound under investigation given by the differential equation and its initial concentration. MODEL PARAMETERS govern the reaction kinetics. MODEL VARIABLES are an explicit function of time. MODEL REACTIONS are dependent on the MODEL STATES and the MODEL VARIABLES. MODEL FUNCTIONS allow the user to supply self-defined functions.

This toy model shows the automated data implementation method. The model consists of two differential equations for the compounds x and y, and two reactions $r_{0|x}$ and $r_{x|y}$, rGlcIn is the glucose influx into the organism and $r_{x|y}$ a conversion from x to y at a maximal rate of $k_{x|y}$ with Michaelis-Menten constant K. This conversion rate is happening inside a compartment of time-dependent volume c_X and constant density ϱ_X. μ is the associated compartment growth rate and governs the dilution of the compound x by growth. The conversion of x to y is governed by two different transcripts T_1 and T_2 that are combined by a function f, e.g. the average later on.

SimBiology Toolbox

As a commercially available software, the SimBiology-toolbox by MATLAB was tested. The implementation of the model can be undertaken by a GUI guided approach and by a scripting approach. Since the model is large, it is recommended to use batch processing of pre-defined equation files. A list of all involved reactions with relevant substrates and products was created and converted into an interpretable format for this toolbox. It turned out that data-dependent variation of transcript level profiles was not feasible *ab initio* in this toolbox. A programmed work-around in C that made it possible to implement it however for all three time courses, transcripts, biomass and glucose variation. However, it turned out that proprietary scripts as e.g. the sensitivity analysis in SimBiology were not supportive for varying compartment size so far. The investigation of this software was aborted at that point. Further softwares were tested, but the same effects as reported by Alves et al were observed: Interoperability of softwares, interfaces and documentation are poor for many of them and specific problems were either difficult or even impossible to be solved [Alves et al., 2006].

4.4 Evaluation of Experiments

4.4.1 Computer, Softwares, Data, Algorithms

Computer

All simulations were carried out on a AMD Phenom (TM) II X6 1090T processor with 3.20Ghz, 16GB of RAM.

Softwares

Simulations were run on Windows 7 SP1, 64bit. The list of all softwares is given in table 4.3.

MATLAB	7.13.0.564 (R2011b)
Bioinformatics Toolbox	4.0
Optimization Toolbox	6.1
Parallel Computing Toolbox	6.1
SimBiology	4.0
Statistics Toolbox	7.6
Symbolic Math Toolbox	5.2
SB Toolbox 2	Development
SBPD	Development

Table 4.3: Softwares

Data for Parameter Estimation and Cross-Validation

The first set of data is the batch fermentation in complex medium [Jones et al., 2008]. Its model is called the *batch model* (BM). Solvent production starts after the stop of growth in the transitional phase after 10 h.
The second set of data was collected during continuous fermentation in phosphate limited medium [Grimmler et al., 2011]. Its model is called the *continuous model* (CM). For the simulation the time frame of the data was shifted since data recording commenced only at 110 h after inoculation. The shift occurred after 160 h.

Algorithms for Model Simulation

BM and CM are constructed by the previously described scheme (section 4.2). Appendix A shows the model equations for the CM, for the BM, the same equations are valid but $D = 0$. Simulation of differential equations in SBTOOLBOX2 and SBPD follows the MATLAB integration of stiff differential equations (ode15s, ode23s) with standard parameters (absolute tolerance of 1e-6, and relative tolerance of 1e-3).

Algorithms for Parameter Estimation

Parameter estimation used a particle swarm search [Vaz and Vicente, 2007] combined with the Nelder-Mead-Simplex algorithm [Flannery, 2007]. The first algorithm offers the possibility to find a global minimum of the optimization functional because of its stochastic nature. For the same reason, its convergence to the minimum is very slow and therefore enhanced by the simplex algorithm. Both algorithms are implemented to consider box constraints of the parameters. It was not possible to impose constraints on the states. The estimation is carried out only for the maximal rates and the substrate yield within a box of $10^{-3}\ mM$ to $10^{3}\ mM$. Michaelis-Menten constants were set to $1\ mM$ and inhibitory constants were set to $1\ M$.

Approach for Simulation of Mutation Experiments

Experiments in which promoters were replaced by a new promoter are mapped to the model by replacing the transcript level profiles of the original promoter with the corresponding profiles of transcripts behind the new promoter. This *profile transfer* is feasible under the assumption that the newly integrated promoter and the changed transcript expression does not profoundly affect the overall expression.

Knock-down and over-expression studies are simulated by assuming that the dynamics of the transcript-data persist, but its levels are changed. This is done by alternating the maximal conversion rate $k_{s|p}$.

For deletion experiments the corresponding rate was decreased by three orders of magnitude because low values of activities are reported even for deletion mutants. For up-regulation experiments the double of the corresponding rate was assumed, as [Mann and Luetke-Eversloh, 2013] indeed reported.

Approach for Disturbance Analysis

For estimation of parameter certainty a disturbance analysis is carried out. The initial parameter-set is perturbed by 10% and re-estimated using the Nelder-Mead-simplex algorithm. This is done several times, here $n = 200$, and the resulting parameter-distributions after re-estimation are evaluated.

Approach for Cross-Validation

A cross-validation by using batch and continuous data is carried out. The corresponding models, BM and CM, are re-estimated within a 20% range of the original parameter set. Only the substrate yield Yp_{Glc} was permitted to re-adjust freely, assuming that different media compositions strongly affect glucose consumption.

List of Goals for Validation

The following list of facts is tested for validation of the model and its simulation results.

1. Levels and dynamics of measured metabolomic data of wild-type cultures [Jones et al., 2008, Grimmler et al., 2011] are predicted.

2. Levels and dynamics of measured metabolomic data of mutant-type cultures [Green et al., 1996, Lehmann et al., 2012a, Lehmann et al., 2012b, Mann and Luetke-Eversloh, 2013] are predicted.

3. Cross-Validation of continuous data and batch data leads to comparable results.

4. In batch culture in complex medium butyryl-CoA and butyryl-phosphate show twin-peaks, acetyl-phosphate shows one peak, corresponding to acid uptake [Zhao et al., 2005, Amador-Noguez et al., 2011].

5. Acetyl-CoA and butyryl-CoA concentrations decrease during the shift in continuous culture [Grupe and Gottschalk, 1992].

6. In batch culture the amounts of acetoacetyl-CoA, 3-hydroxybutyryl-CoA, crotonyl-CoA are less than $21\mu M$, $11\mu M$, $10\mu M$, respectively [Boynton et al., 1994].

7. The following metabolites are higher concentrated in the mid-exponential phase than in the solventogenic phase: acetyl-phosphate, acetyl-CoA, butyryl-CoA, 3-hydroxybutyryl-CoA. Acetyl-CoA and butyryl-CoA profiles are similar [Amador-Noguez et al., 2011].

8. Acetate concentrations and acetone production are correlated but not butyrate concentrations, the ATP balance is favoured via acetate phosphorylation [Desai et al., 1999, Lehmann et al., 2012b].

9. The CoA-transferase preferably acts on acetate. Acetate kinase is favoured in the reverse direction. Phosphotransbutyrylase and thiolase have the highest activity [Vasconcelos et al., 1994].

10. Activity of phosphotransbutyrylase is seen earlier than butyrate kinase in batch fermentations. Thiolase and β-hydroxybutyryl-CoA dehydrogenase peak in mid-exponential phase [Hartmanis and Gatenbeck, 1984].

11. The pools of acetyl-CoA and acetate are highly interchangeable because of the rapid reversibility of phosphotransacetylase [Amador-Noguez et al., 2011].

12. Parallel activity of solvent and acid pathways in one organism are unfavourable and indicate a mixed population [Clarke et al., 1988].

4.4.2 Parameter Estimations and Validation

Results of the Batch Model

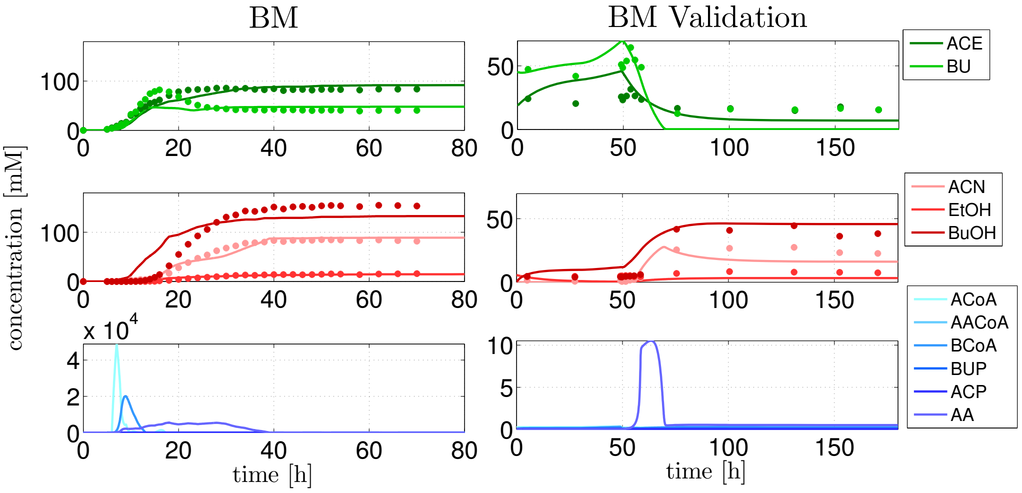

Figure 4.3: Parameter estimation of the batch culture metabolome data using the BM and cross-validation to the CM and continuous culture data. Model results (left) are split into acids(above), solvents (center) and intracellular metabolites (below). The validation (right) is split accordingly. Dotted values represent data, lines represent model results.

Metabolite profiles are shown in figure 4.3 and reaction profiles in figure 4.4. The simulation of concentrations of acetic acid and butyric acid in the BM follow the dynamics of the data. Levels of simulated butyric acid are too low compared to measurement data and the peak at 18h cannot be reached. The simulated solvent levels of acetone and ethanol correspond to measurements, however the butanol production starts too early and does not reach the maximal level. The intracellular metabolites acetyl-CoA, acetoacetate and butyryl-CoA reach an unrealistic high level in the simulation. Acetoacetyl-CoA shows a short spike to 1 M around 8 h. The simulated concentrations of acetyl-phosphate and butyryl-phosphate reach a plateau around 0.5 M and 0.25 M respectively during the exponential phase.
The sum of the CoA-transferase reactions $r_{3|7}$ is highest at 18 h, the onset of solventogenesis. The major contribution to this reaction is given by $r_{63|47}$ not by $r_{23|17}$. Activity of butanol dehydrogenase $r_{4|10}$ shows a similar profile to acetoacetate formation, however there is an earlier peak. Acetone formation is bimodal, one peak at 18h, the second between 30 h and 40 h. Ethanol formation is very low overall.
$r_{4|5}$ and $r_{5|4}$ are the fastest reaction, they have an equal and again bimodal behaviour: one peak at 18 h and a second 4 h later. $r_{1|3}$ has a peak at the same time, the further processing via $r_{3|4}$ is very low, with a maximum earlier

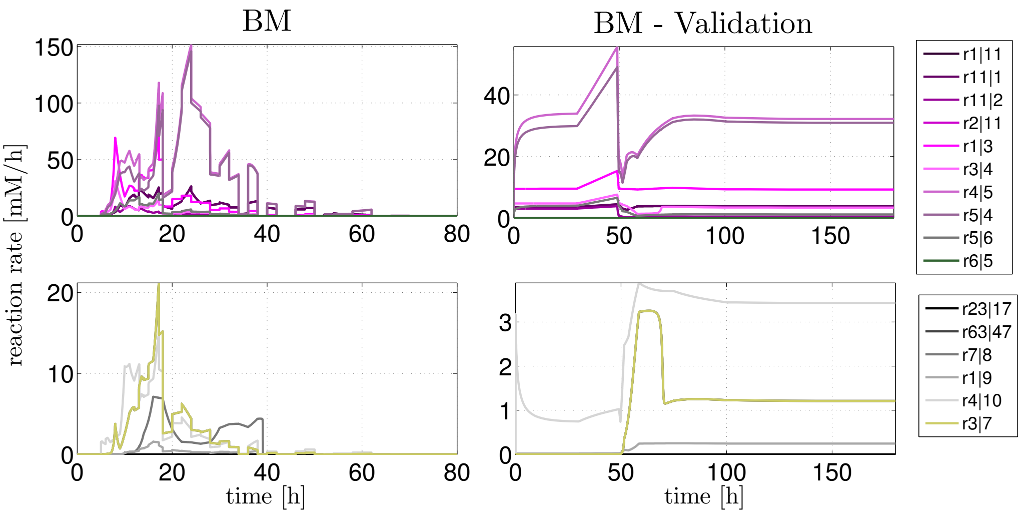

Figure 4.4: Parameter estimation of the batch culture data metabolome using the BM and cross-validation to the CM and continuous culture data. Model results (left) are split into reactions known to be active during acidogenesis (above) and during solventogenesis (below). The validation (right) is split accordingly.

around 8 h. Both reverse kinase reactions $r_{6|5}$ and $r_{2|11}$ are not detectable. The forward direction of the kinases is similar during acidogenic conditions, only the acetate kinase reaction reappears during solventogenic conditions. The phosphotransacetylase reaction performs equally well in both directions ($r_{1|11}$ and $r_{11|1}$).

Results of the Continuous Model

Metabolite profiles are shown in figure 4.5 and reaction profiles in figure 4.6.
Estimation of acids and solvents concentrations in the CM follow the dynamics of the pH-shift. A sharp decrease to zero and almost zero is calculated for butyrate and acetate concentrations, respectively. Accordingly, the solvents increase to the levels given by the measurement data. Acetoacetate concentration is highest with a peak of more than 2 mM when butyrate uptake stops. Its level continue to be high around 1 mM during solventogenesis. Butyryl-CoA and butyryl-phosphate concentrations show an elevation during the shift to 1 mM and to 0.25 mM respectively. The acetyl-CoA concentrations is shortly decreased during the shift and regains its previous value of around 0.25 mM.
As in the batch model, the acetoacetate production $r_{3|7}$ is dependent only on $r_{63|47}$ and not on $r_{23|17}$. The decarboxylation $r_{7|8}$ has the same velocity as $r_{3|7}$. Production of alcohols occurs in the expected split ratio 1:3 approximately.
As before, $r_{4|5}$, $r_{5|4}$ are the fastest reactions, they are decreasing during the shift

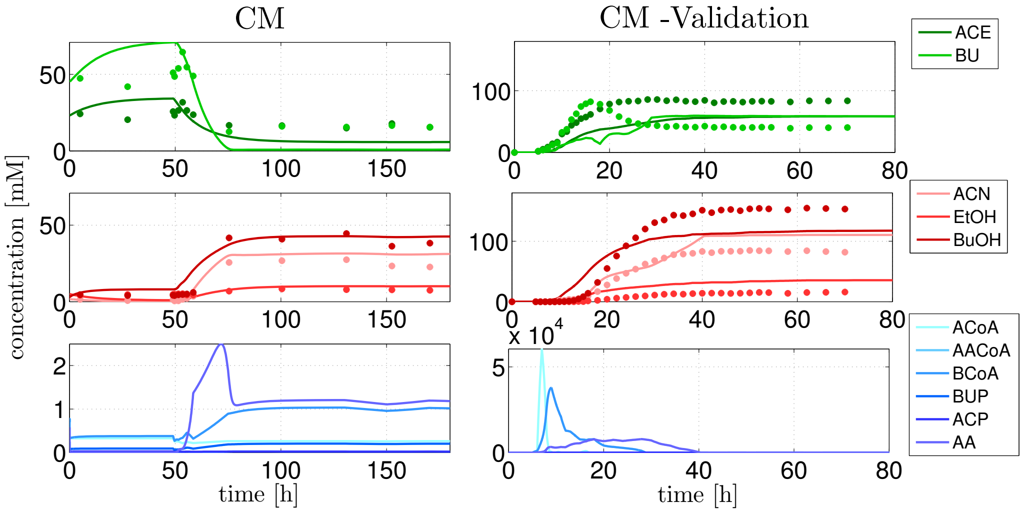

Figure 4.5: Parameter estimation of the continuous culture metabolome data using the CM and validation using the BM. Model results (left) are split into acids(above), solvents (center) and intracellular metabolites (below). The validation (right) is split accordingly. Dotted values represent data, lines represent model results.

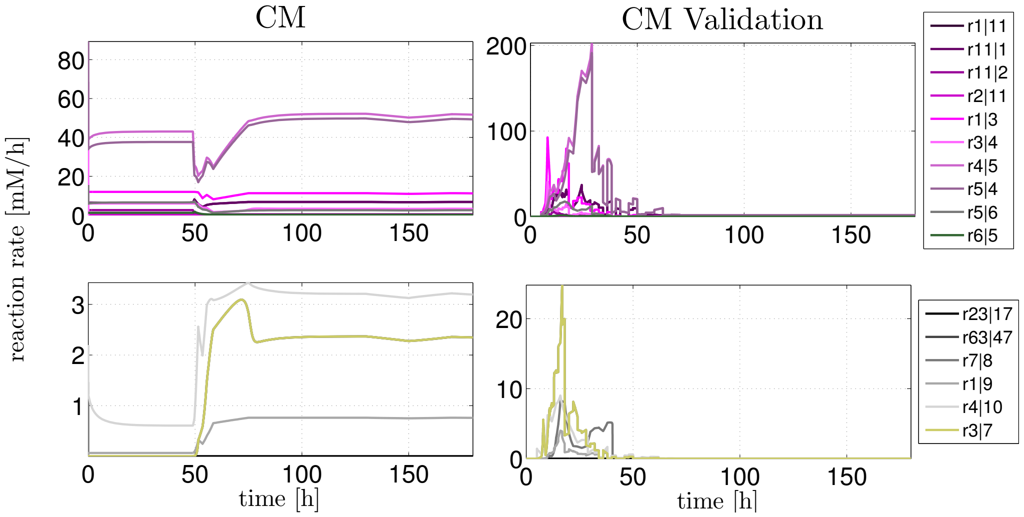

Figure 4.6: Parameter estimation of the continuous culture metabolome data using the CM and validation using the BM. Model results (left) are split into reactions known to be active during acidogenesis (above) and during solventogenesis (below). The validation (right) is split accordingly.

and increase afterwards. This behaviour is seen also for $r_{1|3}$, $r_{11|1}$,$r_{1|11}$, the others decline after the shift. $r_{5|6}$ is stronger than $r_{6|5}$.

Validation of the Batch Model

Metabolite profiles are shown in figure 4.3 and reaction profiles in figure 4.4. The simulation results of the BM are comparable to the CM calculations: Acid and solvent concentrations are met, with the exception of acetone. The dynamics of solvent formation are also met. For the intracellular metabolites, again a huge acetoacetate signal is seen, however, this time there is no accumulation of butyryl-CoA, but still a twin-peak of 0.3 mM height. The acetoacetyl-CoA concentration is step-wise increasing, which is in contrast to the CM. The reason for this is the CoA-transferase reaction $r_{3|7}$, that has reduced activity during solventogenesis compared to the CM. The other solventogenic reactions look similar. During acidogenesis $r_{4|5}$ and $r_{5|4}$ are still the fastest reactions. The other reaction rates are comparable to the rates in the CM.

Validation of the Continuous Model

Metabolite profiles are shown in figure 4.5 and reaction profiles in figure 4.6. The dynamics of solvent production are met by the validation model, however levels for butanol are underestimated and for acetone and ethanol overestimated. The dynamics of acid concentrations are well met for acetate but levels are largely underestimated. The dynamics of butyrate concentrations are not met and levels are underestimated. The validation model shows the same huge spikes of intracellular metabolites as the BM. The other intracellular concentrations and dynamics are similar to the BM, the plateau of butyryl-phosphate is more elongated until 40 h. The reactions involved in solventogenesis have similar dynamics and levels as the CM. The twin-peaked reactions $r_{4|5}$ and $r_{5|4}$ are appearing as a single peak with its maximum higher than in the CM.

Parameter Sets and Disturbance Analysis

Considering the parameters (table 4.4), it is obvious that the conversaion rate for acetyl-CoA dehydrogenation to acetaldehyde and ethanol ($k_{1|9}$) is two orders of magnitude lower than for butyryl-CoA ($k_{4|10}$), still there is accumulation of butyryl-CoA in the CM. $k_{1|3}$, $k_{3|4}$, $k_{4|5}$, $k_{5|4}$ have the highest specific rates and the reverse rate $k_{4|5}$ is stronger than the forward rate in every model. This is similar for the acetate production branch: $k_{11|1}$ is higher than $k_{1|11}$. Not surprisingly from the simulation curves $k_{63|47}$ is large and the values of $k_{23|17}$, $k_{6|5}$ and $k_{11|2}$ are close to zero. The major difference between the two parameter sets of CM and BM is the activity of acetate kinase, the reverse direction $k_{6|5}$ is active in the continuous culture but not in the batch culture. Not surprisingly, many of the

Table 4.4: Parameter-Estimations of batch culture data and continuous culture data with the corresponding cross-validation of CM and BM and vice-versa within a 20% margin of the parameters

parameter	CM parameters	cross-validation of CM to BM	BM parameters	cross-validation of BM to CM
Yp_{Glc}	6.07E-01	1.80E+00	1.42E+00	4.80E-01
$k_{1\|11}$	3.30E+01	3.95E+01	2.72E+01	2.17E+01
$k_{11\|1}$	3.29E+02	2.64E+02	7.36E+01	8.84E+01
$k_{11\|2}$	8.69E+01	1.04E+02	6.27E+01	5.01E+01
$k_{2\|11}$	1.00E-04	8.03E-05	1.74E-05	2.09E-05
$k_{23\|17}$	1.00E-04	1.20E-04	1.64E-05	1.97E-05
$k_{1\|3}$	2.02E+02	2.55E+02	1.91E+02	2.29E+02
$k_{3\|4}$	9.61E+02	8.31E+02	8.28E+02	6.62E+02
$k_{4\|5}$	1.31E+02	1.05E+02	1.65E+02	1.97E+02
$k_{5\|4}$	3.72E+02	4.47E+02	6.19E+02	4.95E+02
$k_{5\|6}$	9.78E+01	7.83E+01	8.45E+01	1.01E+02
$k_{6\|5}$	1.49E+00	1.19E+00	1.26E-05	1.49E-05
$k_{63\|47}$	1.97E+02	2.36E+02	1.94E+02	1.93E+02
$k_{7\|8}$	6.08E+00	4.86E+00	4.17E+00	5.00E+00
$k_{1\|9}$	3.89E+00	3.11E+00	1.21E+00	1.46E+00
$k_{4\|10}$	2.94E+02	3.53E+02	1.03E+03	8.23E+02

estimated parameters are highly uncertain, as a disturbance analysis indicates (figure 4.7): The major part of parameters shows an approximately 10% variance after the disturbance, the parameters $k_{63|47}, k_{3|4}$ have a lower variance, and the parameters $k_{1|9}, k_{6|5}, Yp_{\mathrm{glu}}$ the lowest variance. Asymmetries in the parameter sets' variances indicate that the objective function is asymmetrically shaped, e.g. $k_{3|4}$.

Validation by Mutant Experiments

Promoter experiments are shown in figure 4.8, deletion studies are given in figure 4.9.

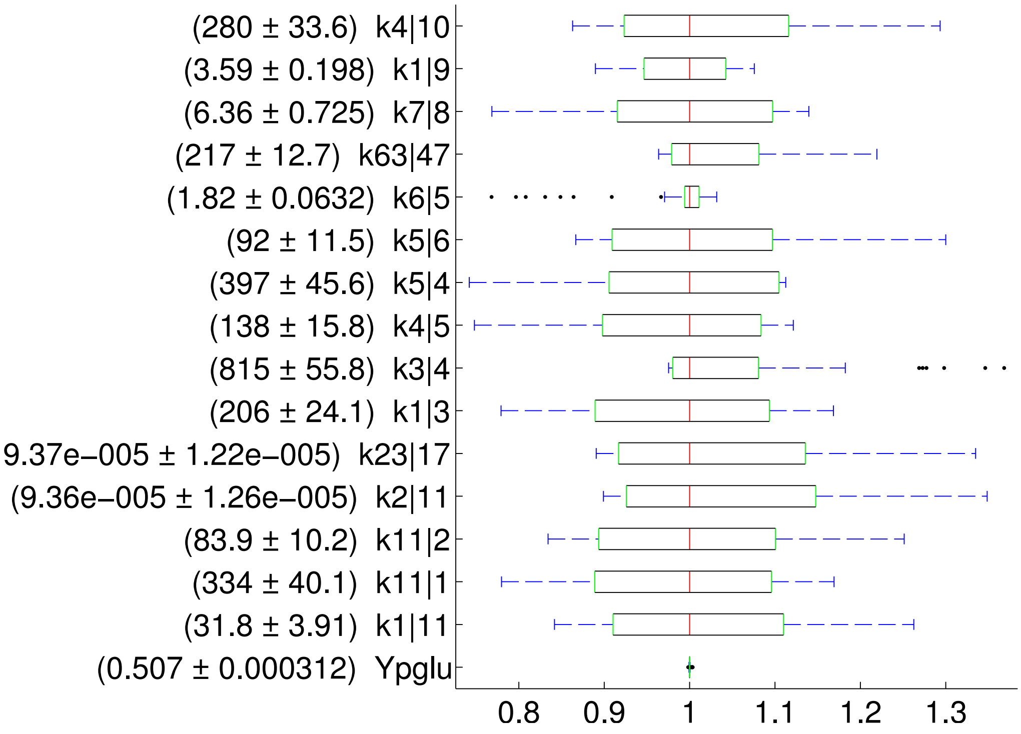

Figure 4.7: Boxplot of the parameter uncertainties in the CM after a 10% disturbance of the initial parameter set. Parameters are scaled to their mean value.

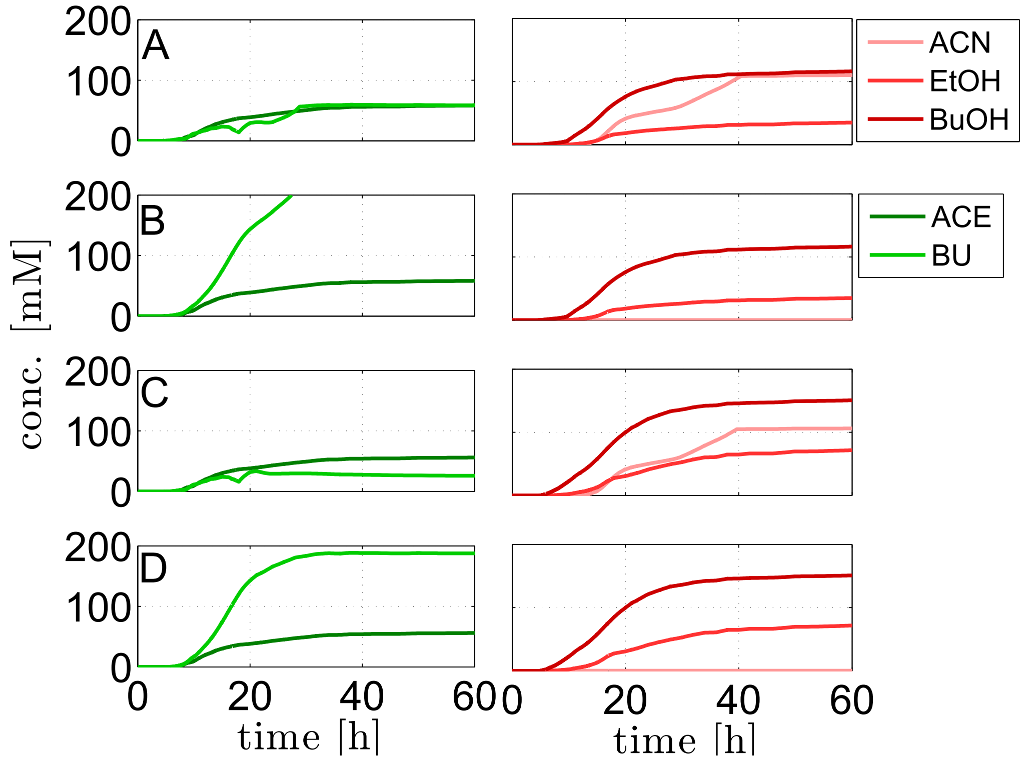

Figure 4.8: Model results for published batch experiments of mutant cultures whose transcripts are expressed behind a different promoter:
(A): wild type results
(B): silenced ctfB1 [Tummala et al., 2003b]. The authors report low butanol and acetone titers and high butyrate titers. Here, butanol titers are unchanged and indeed butyrate titers are elevated.
(C): alcohol dehydrogenase under the phosphotransbutyrylase promoter [Sillers et al., 2009]. The authors report enhanced butanol and ethanol yields, which is covered by the model. However, acetate does not accumulate as expected.
(D): both modification from (B) and (C) [Sillers et al., 2009]. Author report the highest solvent yields in this mutant and best re-uptake of butyrate. Here butyrate is taken up better than in (B) but still elevated levels are seen. Solvent yields are indeed a little higher.

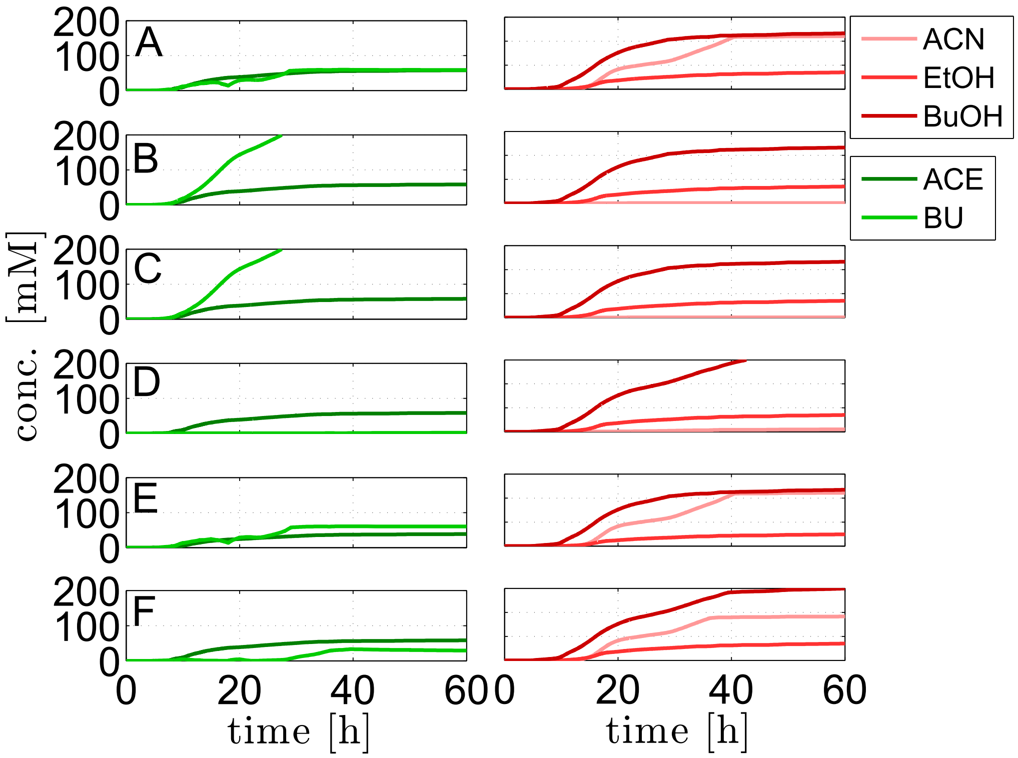

Figure 4.9: Model results for several published batch experiments with deletion mutants and optimised enzyme activities:
(A): phosphotransacetylase deletion [Lehmann et al., 2012a]. Authors report no drastic change is provoked through, here ethanol yields are increased.
(B): CoA-transferase and acetoacetate decarboxylase deletion [Lehmann et al., 2012a]. The authors report high amounts of acetate in the fermentation broth, this cannot be recovered. Here, butyrate is accumulating and the solvent production is similar to (A).
(C): CoA-transferase, acetoacetate decarboxylase and phosphotransacetylase deletion [Lehmann et al., 2012a]. The authors report drastically decreased acetate concentrations, this cannot be recovered. Here, solvent production is similar to (A).
(D): phosphotransbutyrylase deletion [Lehmann et al., 2012b]. The authors report elevated ethanol and butanol titers, this is covered by the model.
(E): enhanced thiolase activity [Mann and Luetke-Eversloh, 2013]. The authors report elevated ethanol and butanol titers by 50% to 19% respectively. Butanol titers are unchanged here, but ethanol elevated.
(F): butyrate kinase knock-down [Green et al., 1996]. The authors report reduced and delayed butyrate formation and increased butanol production. This is recovered by the model.

4.4.3 Validation-Results

1. Levels and dynamics of the measured data are very well covered by the CM, they are well covered by the BM. This expectation is met.

2. Mutant culture experiments can be partly mapped without further model adjustment. This expectation is partly met.

3. Cross-validation is able to cover the dynamics but not the levels of the metabolite data. This expectation is partly met.

4. Plateaus not peaks of butyryl-phosphate and acetyl-phosphate are observed in the BM. Since these plateaus coincide with the overflow peak of acetyl-CoA, there may be saturation of the reactions. This expectation may be met.

5. Effectively, acetyl-CoA and butyryl-CoA concentrations do decrease during the shift in the CM. This expectation is met.

6. The amount of acetoacetyl-CoA in the BM is larger than expected by approximately one magnitude. On the one hand, this could be explained to the overflow of acetyl-CoA, on the other hand, overnight incubation of cell pellets may have disrupted this intermediate [Grupe and Gottschalk, 1992], as intracellular metabolites usually have a high turn-over and rapid disruption [Schaub, 2005]. This expectation may be met.

7. The acetyl-CoA pool in the BM is very similar to the butyryl-CoA pool, although they are one hour apart. In the CM, intracellular metabolites are approximately equal concentrated in both phases. This expectation is met.

8. Butyrate concentrations are uniquely correlated to acetone production in both models, and acetate concentrations are only diverted via the kinase pathway. In order to overcome this drastic difference it is necessary to incorporate ATP generation and consumption into the model. This would allow to valorise the more profitable acetyl-phosphate generation on expense of the butyryl-phosphate production. This expectation is not met.

9. The reverse phosphotransferase reaction is favoured in both models, however the reverse kinase reaction can not be seen. This expectation is not met.

10. A high peak of thiolase activity and a smaller peak for 3-hydroxybutyryryl-CoA dehydrogenase can be effectively seen in the BM. The kinases and phosphotransferases however appear in parallel. This expectation is partly met.

11. High interchangeability of pools is given for the butyrate production from butyryl-CoA. This may be due as well to the symmetric structure of the model and be overcome integrating ATP. This expectation is not met.

12. Parallel occurrences of solventogenic and acidogenic pathways can be seen in both, the CM and the BM. This expectation is met when a mixed culture were present.

4.5 Sensitivity Analysis

The parametrised dynamic model will be used to identify bottle necks from which optimisation approaches are derived. Sensitivity analysis (SA) has aided in the identification of key factors in several biological and chemical models [Saltelli et al., 2000, Cho et al., 2003, Bentele et al., 2004, Lebedeva et al., 2012, Pagel et al., 2013]. The local SA (LSA) focusses on a single point in parameter space, the global SA (GSA) on parameter regions. A complete introduction to both approaches is given elsewhere [Cacuci, 2003].
Local methods provide a high level of detail with less extensive calculations, whereas global methods are best suited to handle highly variable parameters at the cost of higher calculation demand [Rabitz et al., 1983]. The possibility of LSA to only vary one parameter at a time neglects the opportunity to study the non-linear effects of the different parameters amongst each other [van Riel, 2006], through this the LSA is only able to capture few characteristics of the system [Zi et al., 2005]. It is indeed reported that solutions of the LSA appear as subsets of solutions of the GSA [Lebedeva et al., 2012], which usually varies several parameters at the same time. GSA is used for model simplification and it aids in parameter estimation, it is better suited to problems in which uncertainties are in the order of magnitudes and a strong non-linearity exists. Nevertheless, high numbers of input variables and parameters are difficult to treat and a focus on a specific model problem is recommended [Saltelli et al., 2000].

4.5.1 Local Sensitivity Analysis (LSA)

Given is a system of ordinary differential equations, with states $\mathbf{x}$ and parameters $\mathbf{p}$

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\big(\mathbf{x}, \mathbf{p}, t\big), \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{4.24}$$

Local sensitivities indices are calculated by evaluating the partial derivatives s_{nq} defined and normalised as follows:

$$s^*_{nq} = \frac{p_q}{x_n} \cdot \frac{\delta x_n(t)}{\delta p_q}, \qquad s_{nq}(0) = 1 \tag{4.25}$$

A state x_n is called *sensitive* when its sensitivity is large, it is insensitive when the sensitivity is close to zero. Since it is assumed, that all states are sensitive at $t = 0$, insensitive states will decline towards zero over time.

Implementation

Several mathematical methods exist to calculate these equations [Rabitz et al., 1983]. The direct differential method is implemented here, because its calculation is facilitated through the Symbolic Toolbox of MATLAB and SBTOOLBOX2. A script was programmed to integrate all sensitivity index equations into a SBTOOLBOX2 model (B.3.4).

LSA of the Continuous Model

The pathways are given in figure 4.1. A computation of the local sensitivity indices of the CM shows that the sensitivity of acetate high for the phosphotransacetylase forward-reaction ($r_{1|11}$) and the acetate kinase ($r_{11|2}$) during acidogenesis but not during solventogenesis. No other metabolite is sensitive to these two reactions (figure 4.10). Similarly, no metabolite is sensitive to the CoA-transferases ($r_{63|47}$, $r_{23|47}$), the thiolase ($r_{1|3}$) as shown in figure 4.11, or the lumped reaction $r_{3|4}$ and the transbutyrylase ($r_{4|5}$, $r_{5|4}$) as shown in figure 4.12. Butyrate and butanol are sensitive to the reverse reaction of the butyrate kinase ($r_{6|5}$) during solventogenesis as shown in figure 4.13 and ethanol is highly sensitive to the dehydrogenases ($r_{1|9}$) as shown in figure 4.14.
This finding suggests that on the one hand that only $k_{6|5}$ and $k_{1|9}$ are sensitive and an effect from variation of these parameters may be expected in the close proximity. On the other hand, this finding corresponds to the uncertainty analysis carried out earlier (figure 4.7) and suggests that these two parameters are the most certain ones.

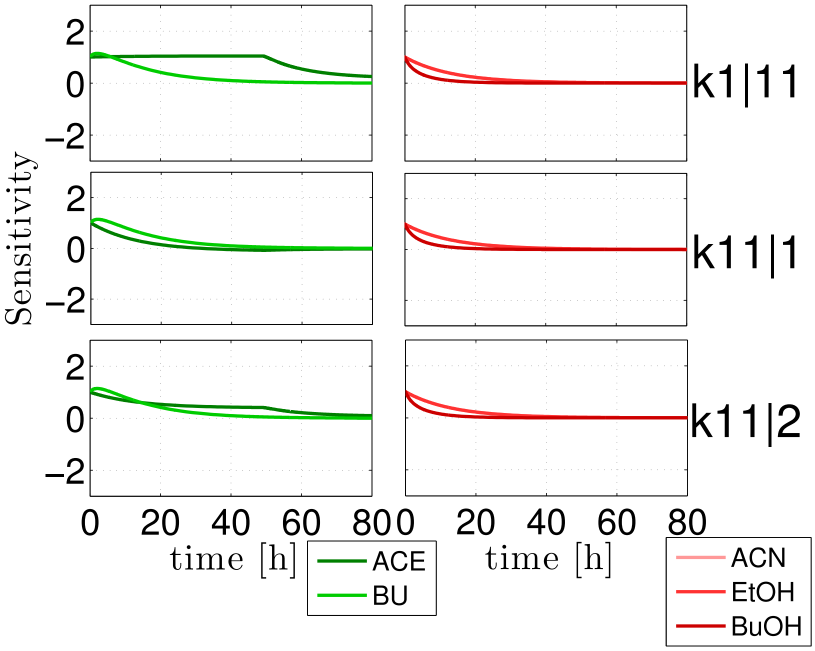

Figure 4.10: LSA of the upper branch of acid formation. Sensitivities of acids (left) and solvents (right) are shown as function of $k_{1|11}$ (above), $k_{11|1}$ (middle) and $k_{11|2}$ (below).

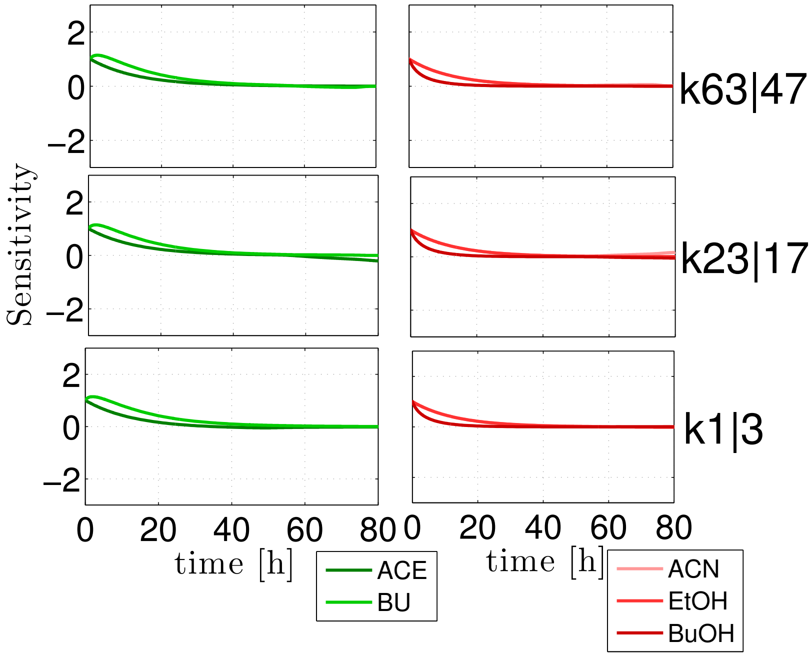

Figure 4.11: LSA of the CoA-transferase and the thiolase. Sensitivities of acids (left) and solvents (right) are shown as function of $k_{63|47}$ (above), $k_{23|17}$ (middle) and $k_{1|3}$ (below).

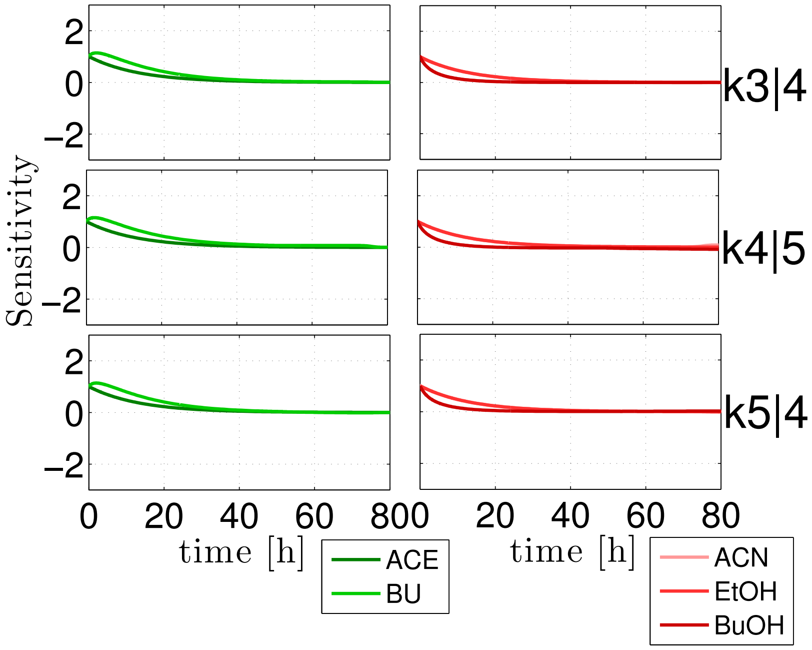

Figure 4.12: LSA of the lower branch of acid formation.Sensitivities of acids (left) and solvents (right) are shown as function of $k_{3|4}$ (above), $k_{4|5}$ (middle) and $k_{5|4}$ (below).

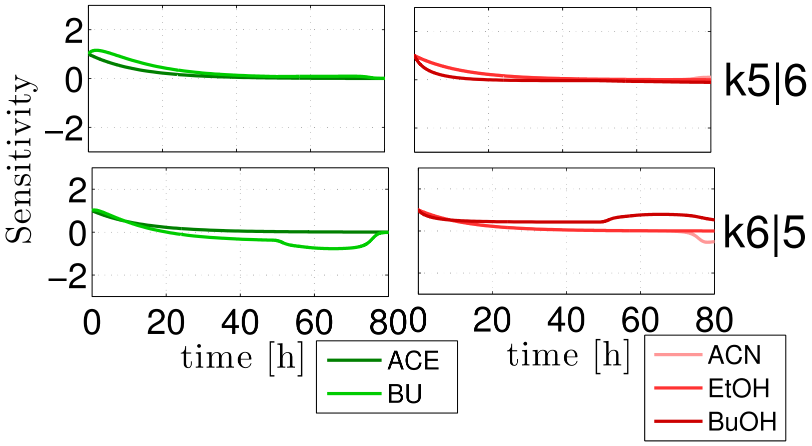

Figure 4.13: LSA of butyrate kinase. Sensitivities of acids (left) and solvents (right) are shown as function of $k_{5|6}$ (above) and $k_{6|5}$ (below).

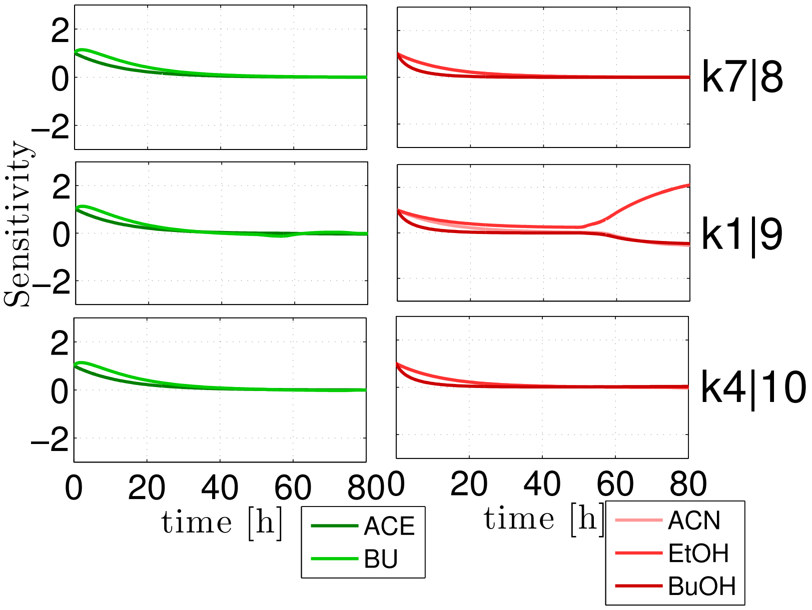

Figure 4.14: LSA of dehydrogenases. Sensitivities of acids (left) and solvents (right) are shown as function of $k_{7|8}$ (above), $k_{1|9}$ (middle) and $k_{4|10}$ (below).

4.5.2 Global Sensitivity Analysis (GSA)

The parameter-set is uncertain as was shown earlier, also the LSA indicates that the sensitivity of all states to all parameters is small. In the close proximity of this parameter set there is no possibility to enhance the productivity of the cell or to alleviate a possible bottle-neck. In order to estimate the global behaviour of the model, a huger parameter region is now considered. Such evaluation is possible by several methods that are all suited for GSA.

Methods and Implementation

Published methods for global sensitivity analysis are:

- Fourier Amplitude Sensitivity Test (FAST) [Saltelli et al., 1999]
- Partial Rank Correlation Coefficient (PRCC) [Bentele et al., 2004]
- Sobols Method [Sobol, 2001].

Applications of these three methods can be found in [Zheng and Rundell, 2006, Marino et al., 2008]. PRCC answers the question how much a model output is dependent on the parameter, while FAST indicates which parameter uncertainty has the highest influence on the model variance. Sobols method and FAST are comparable. In the scope of discrete transcriptomic data that induces steps into the model, FAST seems the best suited algorithm because it is suited also for non-monotonic systems. PRCC is not accurate for such systems [Marino et al., 2008].

The FAST Algorithm

This section is summarising [Saltelli et al., 1999].
Sensitivities by FAST represent fractions of the variance D_p caused by varying a parameter to the overall variance D.
The variance is the second moment of a summary statistic over the N_p-dimensional parameter space $K^{N_p} = (\mathbf{p}|0 \leq p_q \leq 1; q = 1, ..., N_p)$. More generally, the rth moment of the ODE-system f is given by:

$$< x^{(r)} >= \int_{K^n} f^r(\mathbf{p}) P(\mathbf{p}) d\mathbf{p} \tag{4.26}$$

where P is a probability distribution function over the parameters. The first step is the calculation of such a statistic, by exploring the parameter space K^{N_p} such that a filling curve $p_q(s) = G_q(\sin \omega_q s)$ with $-\infty \leq s \leq \infty$ comes close to any point within. G_q is a transformation and the ω_q are properly selected frequencies. K^{N_p} is filled entirely if only the frequencies are *incommensurate*:

These frequencies cannot be obtained by a linear integer combination of the other frequencies:

$$\sum_{q=1}^{N_p} r_q \omega_q \neq 0, -\infty < r_q < +\infty \tag{4.27}$$

Having distributed the parameters identically and uniformly, the integral for the rth moment (equation 4.26) is simplified and evaluated along the filling curves p_q:

$$\bar{x}^{(r)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f^r(\mathbf{p}(s))ds \tag{4.28}$$

However incommensurate frequencies cannot be achieved due to numeric precision. Hence, there is a T for which $f(s) = f(s+T)$ and it was shown that if ω_q are positive integers, $T = 2\pi$.

The total variance D of the model is therefore given by

$$D = \bar{x}^{(2)} - \left(x^{\bar{(1)}}\right)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^2(s)ds - \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} f(s)ds\right)^2 \tag{4.29}$$

Expanding $f(s)$ in a Fourier series over the domain of integer frequencies j with its spectrum

$$f(s) = \quad \textstyle\sum_{j=-\infty}^{+\infty} A_j \cos(js_p) + B_j \sin(js) \tag{4.30}$$

$$A_j = \quad \tfrac{1}{2\pi} \textstyle\int_{-\pi}^{\pi} f(s_p) \cos(js)ds \tag{4.31}$$

$$B_j = \quad \tfrac{1}{2\pi} \textstyle\int_{-\pi}^{\pi} f(s_p) \sin(js)ds \tag{4.32}$$

$$\Lambda_j = \quad A_j^2 + B_j^2 \tag{4.33}$$

$f(s)$ is real valued so that the variance attributed to the fundamental frequency ω_q and its higher harmonics $h\omega_q$ can be written as

$$D_q = 2 \sum_{h=1}^{+\infty} \Lambda_{h\omega_q} \tag{4.34}$$

The ratio $\dfrac{D_p}{D}$ is the estimate of the main effect of p_q on x.

Implementation

FAST, PRCC and Sobols method are readily implemented in SBTOOLBOX2. However, all three methods do not allow a temporal resolution of the sensitivities. The corresponding scripts were adopted in order to allow the calculation of sensitivity indices over time (B.3.4).

The algorithm was set to calculate the sensitivities in the parameter cube centred around the original parameter set with an upper-boundary two-fold larger and a lower boundary only half of the original parameter set. Sensitivities are calculated over time-intervals, here the first interval was chosen to last until 40 h and then hourly steps until 80 h were calculated until then the last interval took until 100h.

GSA of the Continuous Model

Acetate concentrations are weakly sensitive to the transacetylase ($k_{1|11}$, $k_{11|1}$) and acetate kinase ($k_{11|2}$) but to no other metabolite (figure 4.15).
Increasing sensitivity indices of butyrate and acetone towards $k_{63|47}$ during solventogenesis indicate that only butyrate-uptake but not acetate-uptake can be influenced via the CoA-transferase pathway. Ethanol is weakly sensitive to thiolase ($k_{1|3}$) (figure 4.16).
Acetone sensitivity is decreasing for $k_{3|4}$ from high values in acidogenesis to small values in solventogenesis, the profile of acetate is increasing during solventogenesis to medium values. Sensitivity to the transbutyrylase ($k_{4|5}$, $k_{5|4}$) is decreasing for butyrate from medium values in acidogenesis to small values in solventogenesis, it is increasing for acetone to medium values in solventogenesis, it is constant for butanol again at low values (figure 4.17).
The forward reaction of butyrate kinase ($k_{5|6}$) has a weak influence on butyrate during acidogenesis and on butanol for the whole fermentation. Again, sensitivity of acetone is increasing for the forward-reaction of the kinase but not the reverse direction during solventogenesis (figure 4.18).
Sensitivity of ethanol to $k_{1|9}$ is largest, suggesting that ethanol yields are easily influenced while the sensitivity of butanol towards $k_{4|10}$ is production rate is only average throughout the entire fermentation. Sensitivity of acetate peaks weakly for the same parameter at $60h$ (figure 4.19).

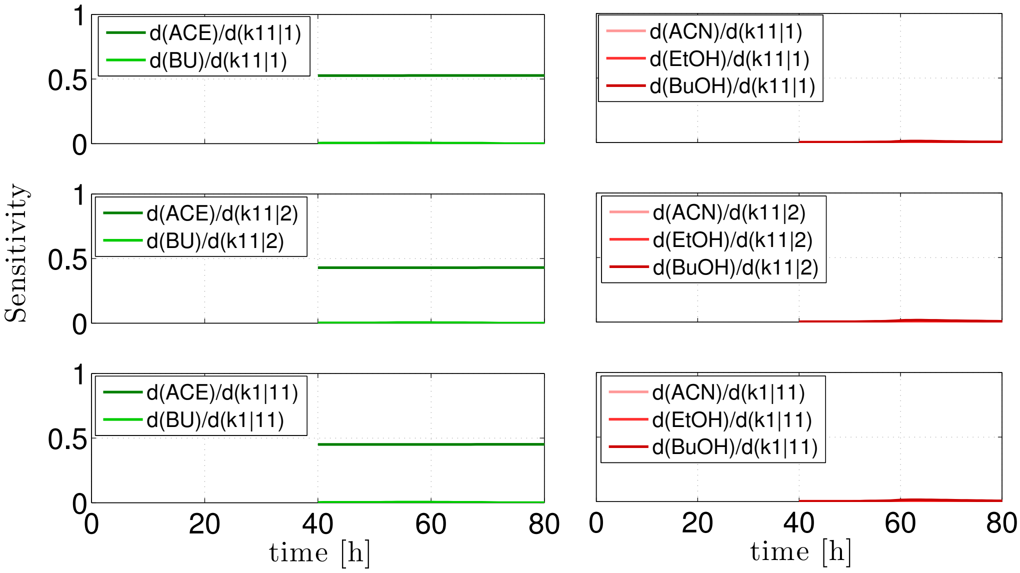

Figure 4.15: GSA of the upper branch of acid formation.

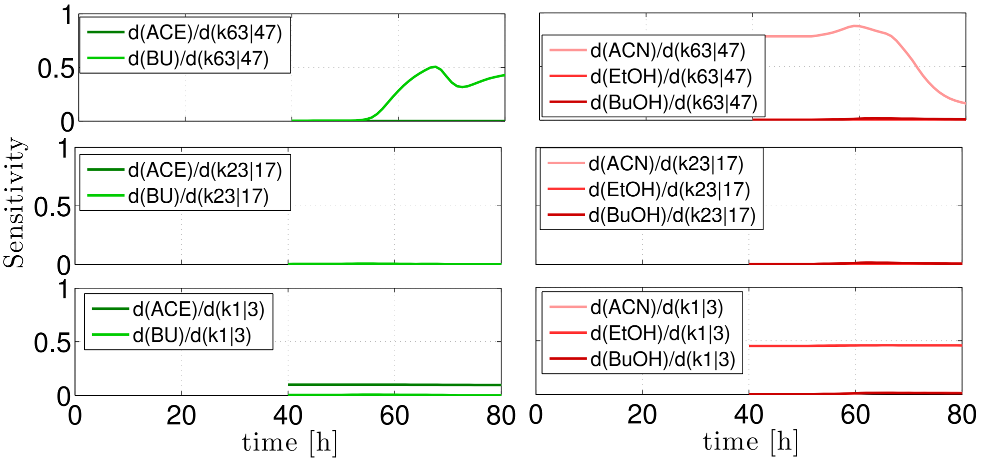

Figure 4.16: GSA of the CoA-transferase and thiolase.

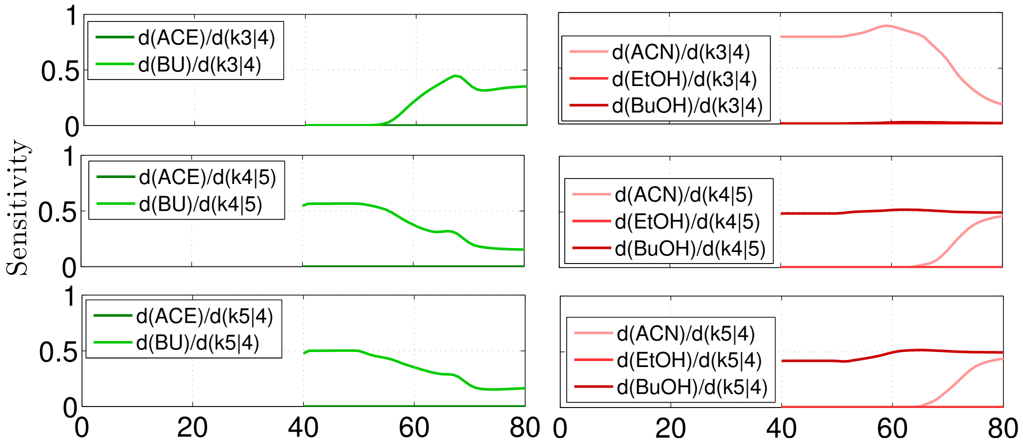

Figure 4.17: GSA of the lower branch of acid formation.

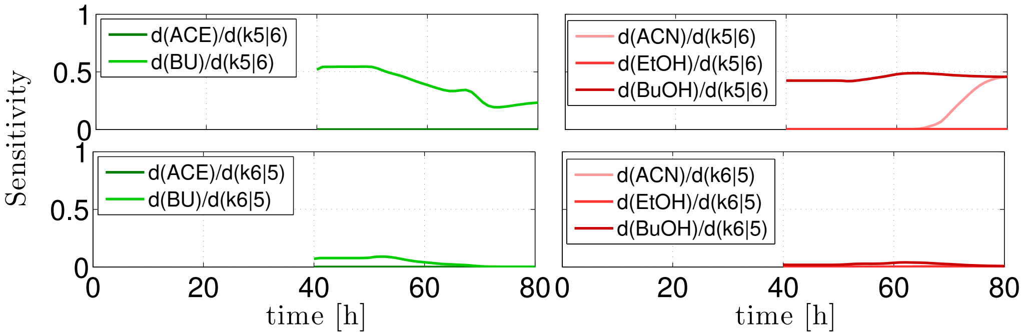

Figure 4.18: GSA of the butyrate kinase.

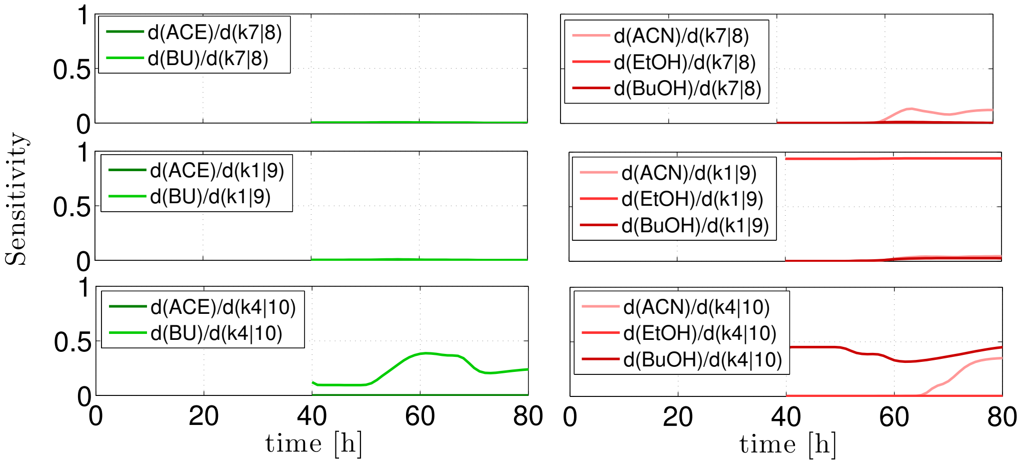

Figure 4.19: GSA of the dehydrogenases.

4.5.3 Summary

Not surprisingly, LSA and GSA obtained different results from the CM: While the LSA focusses on a single point in parameter space, the global analysis summarises the behaviour of the model in a parameter cube.

In both simulations ethanol is highly sensitive to $k_{1|9}$ indicating that it is easy to shift the strain to an ethanol producer, which is proven by experiments [Lehmann and Luetke-Eversloh, 2011, Lehmann et al., 2012a]. Acetate concentrations are not sensitive to the activity of CoA-transferases. While the local sensitivity points out the sensitivity of the reverse butyrate kinase reaction ($k_{5|6}$) alone, the global sensitivity analysis shows that cycling of butyrate through the CoA-transferase and the transbutyrylase/kinase pathway influences the model.

4.6 Final Conclusions

In this chapter a dynamical model was presented that integrated transcriptome data with the biochemical reaction network of solvent production. Two different experiments, a continuous culture experiment under phosphate limiting conditions and a batch culture experiment in complex medium were used for parameter estimation and cross-validation. Biological data and mutation experiments were qualitatively mapped to the model and then two sensitivity analyses, the local and the global sensitivity analysis, were conducted to identify bottle necks and engineering strategies. High expectations are imposed on *in silico* models [Lee et al., 2008b]. Universal applicability and predictive capability are only two of them.

Universal Applicability of the Dynamic Model

The proceeding of model development and evaluation in this thesis was undertaken semi-automatically. Download of relevant information from KEGG and the creation of a standard format for SBTOOLBOX2 allows the automated integration of any reaction network with any data, not only transcriptome but also proteome data. The selection of desired sub-networks is manual work. As soon as published curated reactomes become available as download [Kumar et al., 2012, Agren et al., 2013] entire automation is possible with this script or the SBML interface of SBTOOLBOX2.

The model was able to cover two entirely different experimental proceedings (batch culture versus continuous culture, complex medium versus minimal medium, highly resolved transcriptome data versus low resolved transcriptome data in the temporal dimension) and some mutation experiment findings.

Universality arises from the description of these very different experiments and through the integration of transcriptome data: The evaluation of mutation experiments is conveniently done by integration of new transcriptome data in the existing model or by *profile transfer* as was shown for data of [Sillers et al., 2009]. Achieved parameters from the estimations are comparable to each other within a 20% margin and only the substrate yield required large adjustment. The model meets one third of the proposed expectations completely, one third only partly and it fails for the last third. To overcome the mismatch to biological findings, e.g. the importance of butyrate up-take via the CoA-transferase pathway despite the coupling of acetate up-take, the integration of further data is necessary, e.g. energy metabolism data. Online annotations of energy and redox influencing metabolites and reactions are available, for growth the stoichiometry of growth and biomass maintenance, linear models are already developed [Papoutsakis, 1984].

Predictive Power and Usability of the Dynamical Model

The derivation of meaningful predictions is a challenging task and may often be unsuccessful because of many parameters involved [Jamshidi and Palsson, 2008, Durot et al., 2009]. 35 Parameters, thereof 13 maximal rates are many parameters to describe five metabolome measurement time series.
Discrepancies of the model to the experimental findings are expected for the prediction of mutant culture behaviours, as those generally show a different growth behaviour and different regulatory mechanisms [Tummala et al., 2003b].
It is not surprising, that all parameters, despite the reverse butyrate kinase reaction and the ethanol dehydrogenase reaction are uncertain. This robustness is a desired feature of a biological model [van Riel, 2006], however several mutation experiments show that it is not that robust. Here the predictive power of the model fails. This robustness can also be interpreted as high uncertainty of the parameters, as the calculated sensitivity analyses hinted.
Still, the results of a sensitivity analysis could be transferred into an engineering strategy if the maximal conversion velocity were directly accessible. Such an engineering is possible but imposes several problems, as was recently shown for the clostridial thiolase [Mann and Luetke-Eversloh, 2013]. Still, SA did not produce suitable engineering strategies, since ethanol is not in the scope of this work and also the increase of the reverse-reaction of butyrate kinase will not directly increase butanol yields because of the present forward-reaction.
However, this model format allows a second approach that will be introduced in the next chapter, from which other strategies can be derived.

Problems of Dynamic Models for Integration of Omics

Currently efforts are spend on creating mechanistic models of transcription and translation via bayesian networks and differential equations [Chen et al., 1999, Dhaeseleer et al., 2000, Arnold, 2002, Vijesh et al., 2013], [Ingalls, 2013, chapter 7]. Implementations are reported for many networks that use Flux Balance Analysis (FBA) [Lee et al., 2008a, Senger and Papoutsakis, 2008] and many softwares integrate FBA-methods, e.g. YanaSquare [Schwarz et al., 2007b] or CellNetAnalyzer [Klamt et al., 2007].
Other ways of integration transcriptome and metablome are not frequently encountered [Joyce and Palsson, 2006, Jamshidi and Palsson, 2008] and thereby several problems concerning implementation and computation needed to be solved here.
First, only two softwares were found that were able to allow integration of the model formalism into their architecture and all following analyses needed to be build on this integration. Missing SBML support for this model is a good indication that this model type is not widely spread, although a work-around solution

may be present by using SED-ML [*] [Waltemath et al., 2011]. SBTOOLBOX2 was made the tool of choice because of two reasons: Codes are publicly available and the text-based format of the model allowed easy automation of model creation and evaluation. This is necessary to leverage one computational problem of parameter estimation: Stiffness of the differential equations is increased by the integration of transcript expression profiles. In steep profiles only many small time steps of the integrator can lead to convergence, hence a huge number of calls for each interpolation of transcript levels is necessary. Computational speed is proportional to these calls if the number of integrated profiles is large enough. One solution to this problem will be presented in the next chapter. This problem and its solution are unique to this type of model and represent a new area of research for the modelling science.

[*] Personal communication, Dagmar Waltemath, Rostock

Chapter 5

Principal Component Analysis In Modelling

That honey is sweet I refuse to assert;
that it appears sweet I fully grant.
Timon of Phlius

Principal Component Analysis (PCA) is widely used, this chapter proposes three independent applications to this method. After its introduction (5.1), its historical use as compression method to alleviate simulation effort in the dynamical model from interpolations of transcript expression data (5.2). Very closely related to this, is a optimisation tool based on PCA to alternate transcript expression data under the side condition of preserved dynamic (5.3). The third application then discusses the uses of this approach for a novel clustering method (5.4). A conclusion is given afterwards (5.5).

5.1 Introduction of PCA

Mathematical Background

An introduction to principal component analysis (PCA) and its mathematical derivation is given by [Abdi and Williams, 2010].
The set $\mathcal{D}$ is the set of all transcript expression data in a matrix format in which columns represent time and rows represent transcripts:

$$\mathcal{D} = \{x_j(t_i), j = 1...N_J, i = 1...N_T\}. \tag{5.1}$$

Usually $N_T \ll N_J$. A PCA of the data consists in a dimensionality reduction of $\mathcal{D}$ into few representative basis functions, called principal components (PC) with respective coefficients (c). Each of the PCs is sorted according to their contribution to the data's variance. If indeed the data's variance is low dimensional then PCA is able to compress the data. It is reported that the first principal components are usually representative for the data, they are optimal for compression [Janes and Yaffe, 2006]. Conversely, higher components may bear structural information of the data [Yeung and Ruzzo, 2001].
Here, it will be assumed that only the first $N_P < N_T$ components are relevant and further components hold information about uncertainties. This happens on the expense of low abundant pattern of the data that are furthermore neglected. The basis representation of $x_j(t)$ in terms of principal components is then given by:

$$x_j(t) = \sum_{p=1}^{N_P} c_{jp}\mathrm{PC}_p(t) \tag{5.2}$$

Uses of PCA

Besides its compression abilities [Janes and Yaffe, 2006], PCA has a manifold of other uses. Perturbed states are distinguished via PCA at the metabolome and transcriptome level from the unperturbed states [Dutta et al., 2009]. Reverse engineering approaches were possible by calculation of a PC representation of gene expression data [Yeung et al., 2002]. It furthermore allows the imputation of missing data-points in a less error-prone way than regression models [Troyanskaya et al., 2001]. Correlations of the data to the principal components help in the observation of genome-wide effects [Alter et al., 2000]. Evaluation of global sensitivity results was recently shown to be facilitated when the principal components where calculated [Sumner et al., 2012]. The use of PCA is reported to minimise effects of measurement noise prior to clustering in several publications [Brown et al., 2005, Janes and Yaffe, 2006]. Other authors argue that it may degrade cluster quality [Yeung and Ruzzo, 2001]. Uses of PCA for clustering are only reported within a side-note [Holter et al., 2000].

5.2 Data Compression in the Dynamical Model

Preliminaries

The dynamic model of the previous chapter suffers severe difficulties when it comes to parameter estimations. The global sensitivity analysis shows on the one hand that the parameters are not accurately estimated. On the other hand the criterions to stop the estimation are not met rapidly enough. Interrupting the simulation manually results in a highly non-reproducible parameter set. In the scope of automation and reproducibility the convergence speed is a serious problem.

Transcript Expression Data Problems

One reason why the dynamical model is computationally demanding is the integration of highly dynamic transcript expression profiles into the equations. Changes of expression levels in the continuous culture consist up to two orders of magnitudes.

Noise of these data further impedes the estimation speed. It is hardly accessible due to the missing replicates of the available data. Highly expressed genes during the solventogenic chemostat can be averaged to estimate noise, which is around 50% on the nominal scale, or 0.5 on the log scale.

Solutions

Dynamics of data cannot be reduced, however the number of dynamic profiles can be reduced by choosing a suitable data-model, e.g. transcript within the same open reading frame usually have similar dynamics and levels. Such a data-model further needs to eliminate noise from the data.

Clustering offers one possibility to reduce size of data and create regulatory assumptions, as seen before (3.5).

The performance of alternative representations in terms of non-linear regression is greatly known in the biological community. The right choice of regressors or basis functions is a non-trivial problem that requires in-depth knowledge of underlying data structure. Also such a representation should be useful in a greater scope than simple representation, e.g. the unravelling of regulatory dependencies. For automation both these facts are critical.

Other possibilities of convergence amelioration by introducing system control theory approaches as semi-quantitative measurements, Kalman filtering or network modular topology increase estimation quality and convergence [van Riel, 2006].

The principal component representation of the data is a self-contained representation of the data and does not require prior knowledge, it also discriminates between data and data-noise.

PCA-assisted Data Compression

Equation 5.2 and the observation that usually $N_P < N_T \ll N_J$, it is easy to conclude that in a model in which all N_J transcript data levels are calculated, the major computation time is used for the interpolation. However, all the information necessary for calculation is stored in the N_T components of the PC expansion. For the CM, there are 14 different profiles and $N_T = 5$, hence it is always better to calculate only the N_T PCs instead of all 14 profiles. The necessary $5 \cdot 14 = 70$ coefficient c_{jp} can be integrated to the model prior calculation (B.3.2). The calculation effort of the original model is $N_T \cdot N_J = 70$ versus $N_T^2 = 25$ of this compressed model. This corresponds to a theoretical improvement of speed by 63%. Computing 200 independent consecutive runs of the CM and its compressed format, took 61 s to 29 s, respectively. This is a gain of 53%.

5.3 Optimisation Approach via PCA

Time-resolved steering of promoter activity was a project goal within COSMIC2. It can be expected that the directed change of transcript levels over time will be in research focus soon. Models require to take this challenge and to guide the experimentalist through the outcomes of different strain designs and temporal profile designs.
The presented dynamical model contains transcript expression data and can be used for such a prediction, if only a suitable scheme of alternation of these profiles can be found.

Bootstrapping of Transcriptome Data Does not Influence the Metabolic Spectrum

The most general approach consists in the random generation of artificial transcript expression profiles. However, this easily results in numerical problems and unrealistic model results.
A more directed approach than random design is toggling existing measurement data, also known as bootstrapping [Pattengale et al., 2010]. The transcript expression data profile is considered a measurement curve that is the result of a signal diverted by additive noise. For transcript expression profiles this approach is feasible since the sample space is low populated in the temporal dimension and an original profile can be parametrised for bootstrapping. Using again global sensitivity analysis a huge number of model simulations and their effects were studied. For the CM only the central point of the pH-shift data is sensitive (results not shown). However, changes at this point showed no changes in productivity.

Directed Approach to Optimisation - Dynamic Features

Here, a novel approach is suggested. It includes an intelligent choice of profiles. The data structure itself is not known but the use of PCA identifies the directions of maximal variance, such direction will be called *dynamic feature*. Identifying the PC with the dynamic aspects of the data [Holter et al., 2000] offers a possibility to vary the dynamics of the data. This is referred to as *dynamic features optimisation*. The study of combinations of dynamic features to construct new data restrains the amount of possible profiles to only *dynamically equal* profiles (refer to 5.4).

Implementation

A script for alternation (B.4.1) was used to compute the new profiles. After caluclation of the PCs, the scores are alternated by multiplication with 2, 1 or 0.5. Every combination of alternated scores gives a new profile from which a new model is constructed and simulated.

Primary Target - Butanol Dehydrogenase

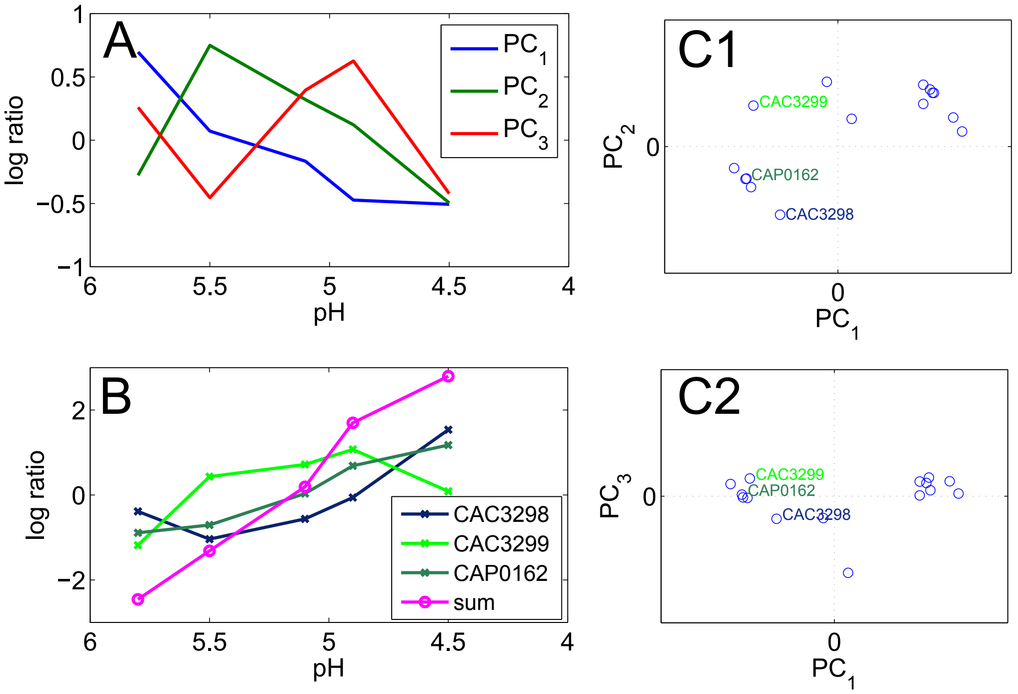

Figure 5.1: PCA Analysis of the transcript levels of the reaction $r_{4|10}$.
(A): Three principal components are calculated from the entire transcriptome data $\mathcal{D}$ over the pH-range of the pH-shift.
(B): The time-courses of the three transcripts and their sum as responsible for butanol synthesis from butyryl-CoA in the CM.
(C): PC coefficients. The position of the three transcripts in the three dimensional PC-space shows that all transcript levels have a negative contribution of the first PC. The second PC introduces major dynamic differences between CAC3299 and the other transcripts. Contributions of the third PC are not as important.

From GSA it was reckoned, that butanol dehydrogenase is weakly sensitive. This is also the primary choice of all strategies.
The PCA transformation from the entire data $\mathcal{D}$ gives the three PC, as shown in figure 5.1,A. The profiles of the three relevant transcripts of reaction $r_{4|10}$, CA_{C3298}, CA_{C3299}, CA_{P0162} add approximately to a straight line (figure 5.1,B) - their coefficients are marked in green in figure 5.1,C1 and C2. As expected, the profiles of CA_{C3298} and CA_{C3299} are very similar in their first PC. Their major difference is found in the second PC, CA_{P0162} and CA_{C3298} have a negative contribution, CA_{C3299} a positive contribution. The contribution of the third PC is nearly negligible for all three transcript levels.
Alternation of the PCA coefficients according to the implemented scheme, yields

a set of curves with different dynamic impacts. The maximum is attained always at pH 4.5, the dynamics are shifted to even lower values during acidogenesis (figure 5.2). For each profile, metabolite spectra are calculated (figures 5.3 and 5.4). Neither profile from the dynamic features increases the pull from glycolysis to generate butanol, only butyrate is converted during acidogenesis to butanol. Despite the earlier start of butanol synthesis in these scenarios, no further increase of butanol yield is visible.

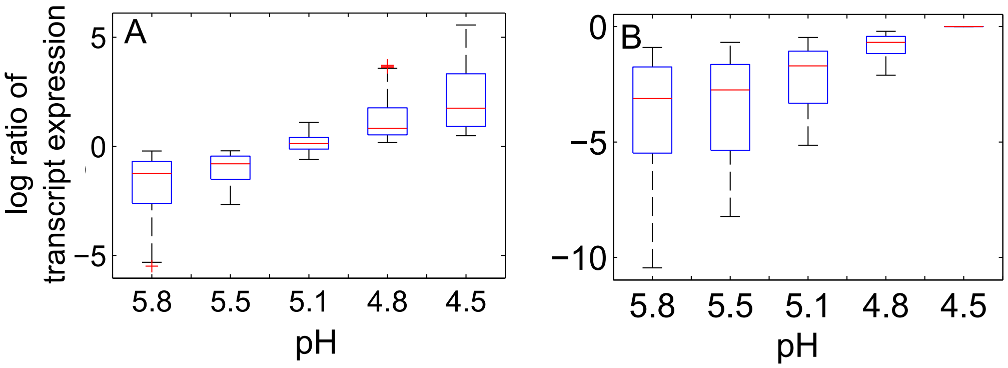

Figure 5.2: Profile design for $r_{4|10}$ from dynamic features. Coefficients of the individual profiles were alternated by either doubling or halving. Profiles prior to normalisation (A) and after normalisation to the maximum (B) shows that the maximal amount of enzyme, corresponding the maximal velocity is reached for all profiles after the shift. Dynamics are unaltered since the profiles intersect at the same level during the pH-shift.

Other Targets

The calculation of dynamic features of butyryl-CoA synthesis from acetoacetyl-CoA (r_{34}) shows that this reaction is not limiting, alternation of the transcript profiles does not yield an improvement. Down-regulation of the enzymes accumulates the substrate and thereby leads to emptying the butyryl-phosphate and butyryl-CoA pools. Interestingly, no changes on the upper-branch of acid or ethanol synthesis is seen. Apparently, these reactions are limited. The surplus availability of acetoacetyl-CoA leads to increased acetone formation that crashes when no butyrate is present in the medium anymore.

The phosphtransbutyrylase reaction (r_{45}) and the butyrate-kinase reaction (r_{56}) both have an identical influence. Butanol yields are increased by 15 mM and acetone is reduced by 10 mM at the end of the fermentation (figure 5.5).

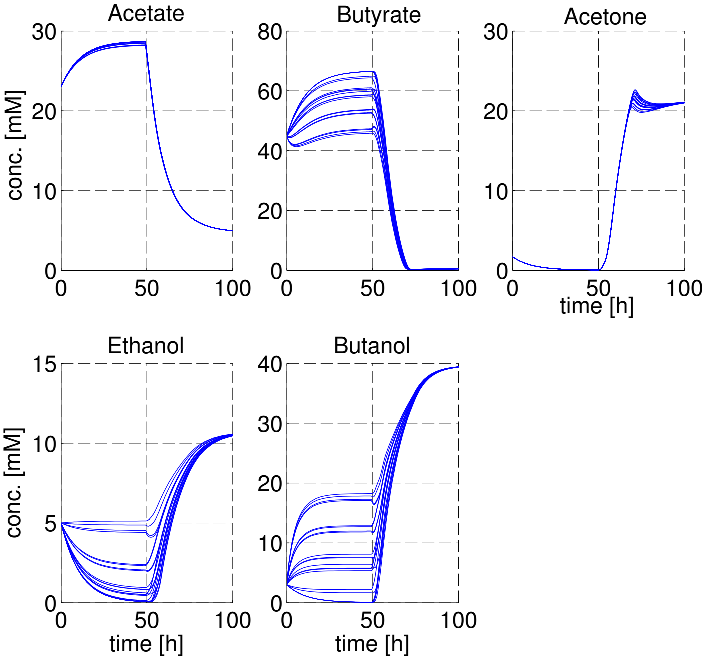

Figure 5.3: Metabolic spectrum of the transcript optimization of $r_{4|10}$.

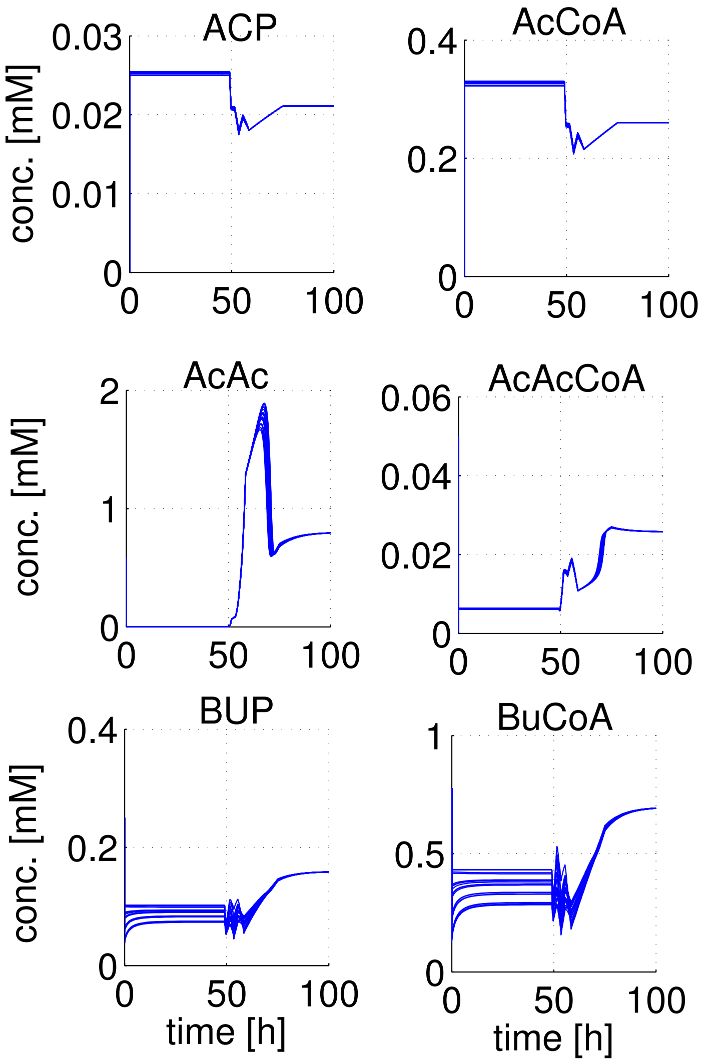

Figure 5.4: Metabolic spectrum of the transcript optimisation of $r_{4|10}$.

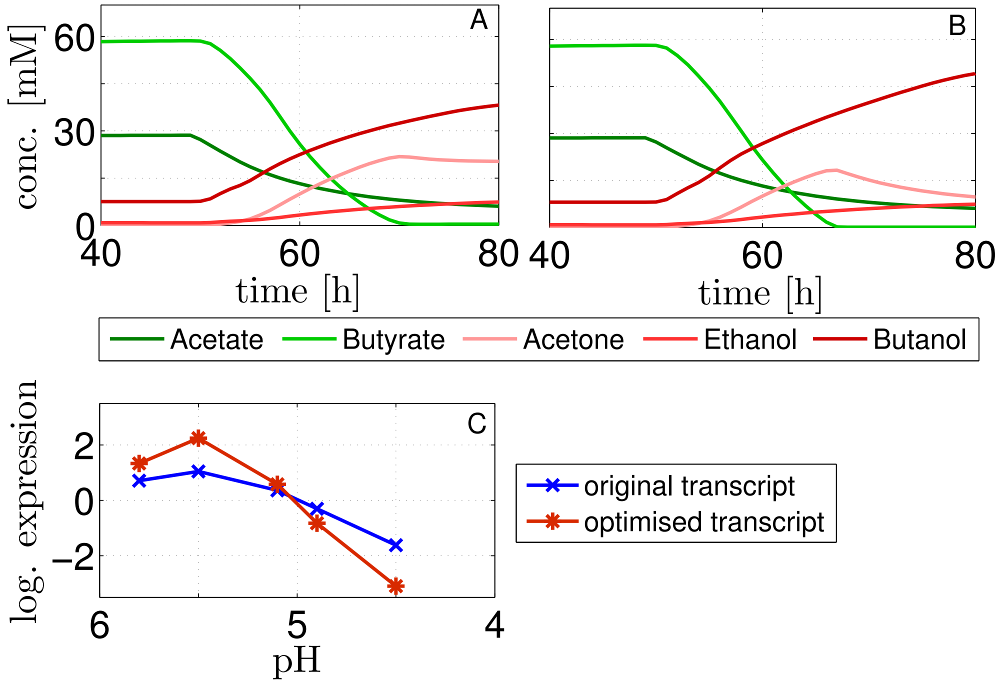

Figure 5.5: Profile design for $r_{4|5}$ from dynamic features.
(A): original product spectrum
(B): new product spectrum
(C): original and new time-course of CA$_{C3076}$ prior to normalisation

5.4 Clustering from Principal Components

For this section it is necessary to introduce some preliminary thoughts on current clustering techniques, because PCA will be used in a somewhat different way of thought than current techniques.

5.4.1 Introduction

The Current Paradigm in Transcriptome Analysis

Many problems in Omics are classification problems [Nobeli and Thornton, 2006]. A class or a cluster is a set items that share the same properties. In the scope of transcript level profiles this classification is based on a similarity evaluation, like the euclidean distance of two profiles or their co-variance: From the combination of the correct metric with an expectation of measurement noise, the clustering algorithm derives groups that share a low inner-cluster variance and a large intra-cluster variance [Lukashin and Fuchs, 2001]. This algorithm requires user-input on how much measurement noise is expected or how many clusters constitute the data. Through this, a cluster is a function of the chosen metric, the chosen cluster algorithm and a set of necessary parameters. The number of employable metrics and clustering algorithms is extremely large [Jiang et al., 2004b, Brown et al., 2005, Janes and Yaffe, 2006].

Critique of the Current Paradigm

The current paradigm has helped in the interpretation of many data since first large data sets became first available: The original paper by Eisen et al. [Eisen et al., 1998] was cited 13587 times [*]. However, what are the limitations of transcriptional analysis by clustering?

It is well known, that not all transcripts are equally well transcribed during the amplification reaction. From a stochastic point of view the multiplication of very low abundant transcripts is a rare event because of the meeting probability of enzyme and transcript. Factors like the affinity between enzyme and transcript, stress factors create an additional bias that is variable also in time. This leads to the conclusion that quantities of co-transcribed genes are not equal [You and Yin, 2000, Feder and Walser, 2005].

Further, clustering is known to reveal operons because the genes behind an operon are concomitantly transcribed. However, metabolism-wide events will effect multiple open reading frames. These frames are not all transcribed behind promoters of identical strength, the dynamics of expression will be similar but not the amounts. These limitations mostly attack similarity metrics that are based on

[*] 15th August, 2013

distances. Correlation metrics are better suited tools regarding these limitations, they are, however, more severely affected by measurement noise.
Finally, there is the question of information between clusters: With any metric it is only possible to make a binary decision whether two profiles are contained within the same cluster or not, nothing more.

5.4.2 Geometric Approach to Clustering

PCA offers a possibility to achieve a geometrical dissection of the data in a similar way as a correlation metric but without loosing relations between the different geometric objects, and it dissects the measurement noise.
The N_T dimensional vector of PCA coefficients c_{jp} will be furthermore named *trait* of the pth PC to focus on the fact that the PCs are universal within the data - they span a space - and the appearance, the *phenotype* of a transcript level profile depends on the impact of each individual PC. Indeed, all possible dynamics of the treated experiment are coded within the PCs [Holter et al., 2000].

Conventional Clustering of Traits

In a first attempt the use of published clustering methods like k-nearest neighbours or hierarchical clustering to identify agglomerations of c_{jp} in the spanned space was carried out. This attempt is difficult for the beforehand mentioned reasons: First, the determination which amount of distance is due to measurement noise can not be carried out, because noise is encoded in the higher dimensional components and it is not transferred to the traits. Second, the traits are sorted according to their importance in the total data variance. This imposes the necessity to intelligible weighing of each dimension of PCs. One such weight could be the variance contribution of each PC to the data. Since structural information may be present in higher components [Yeung and Ruzzo, 2001], and not the entire variance is considered, this approach is discouraged here.

Clustering by Tiling of the PC-Space

In a second attempt, it seems intelligent to start with new assumptions and to try a new approach for making sense of traits with the help of a suitable tiling that neither requires a clustering algorithm, nor a distance metric.
The space spanned by the principal components, is the space of *all achievable dynamics* $\mathcal{C}_{\text{PC}}$:

$$\mathcal{C}_{\text{PC}} := \{\text{PC}_1, \text{PC}_2, ...\text{PC}_{N_P}\}$$

In order to construct a tiling of this space, a meaningful concept for combination of dynamics needs to be introduced. The trivial tiling is a partition into half-spaces according to the sign of the respective components [†]. By pure visual inspection any agglomeration of points within one partition may be considered a tiling, as in figure 5.6,A. This is the established approach. Now, consider a straight line from the origin that connects to any point. Mathematically, all points on a line in this space do share the same proportions of each PCs.Such transcripts are called here *dynamically equal* with respect to that line. In order to account for uncertainties and to be of practical use, this line-object requires a thickness. Because there are several possibilities to define this thickness, assumptions are required to define this new geometric *object*:

1. A maximum principle: The line-object should not be larger than its containing half-space.

2. An equality principle: Two line-objects are non overlapping and equally sized.

3. A geometrical principle:

   a) either the half-spaces are bisected into equal parts

   b) or the half-spaces are trisected into equal parts.

The first assumption accounts for the impact of signs of traits on the phenotype a change of sign may have significant effect on the overall phenotype. The second assumption makes sure that no transcript expression profile is present in two different tilings. The third assumption is the core of angular traits, because it represents the human factor that classifies the relation between the coefficients whether one is stronger present in the data than the other. Choosing a trisection helps in classifying the traits that are strongly different to each other (sectors 1 and 3 in figure 5.6,C) by assuming that equally sized dynamic aspects are of interest. By choosing the bisection, equal coefficients are put out of focus. Due to the curse of dimensionality, this is the approach taken here. The number of cones per two coefficients will be named n_{angle}. For two principal components, $n_{\text{angle}} = 8$ cones are constructed (figure 5.6, B), by increasing the number to four principal components, this number is increasing to $n_{\text{angle}} = 8^3 = 512$ in the case of trisection 12 cones are constructed from two principal components, and $n_{\text{angle}} = 12^3 = 1728$ from four principal components.

Both tilings further provides a possibility to define pairs of co-regulated and anti-regulated transcript expression data: For a chosen cone, its point-reflection is anti-regulated because all traits are reversed in their signs (figure 5.6,B, cones 2 and 5). A trait within such a cone will be called *angular trait*, it is defined as $\alpha_{jp} = \arctan\left(\frac{c_{jp}}{c_{j1}}\right)$.

[†] There are 2^{N_P} half-spaces.

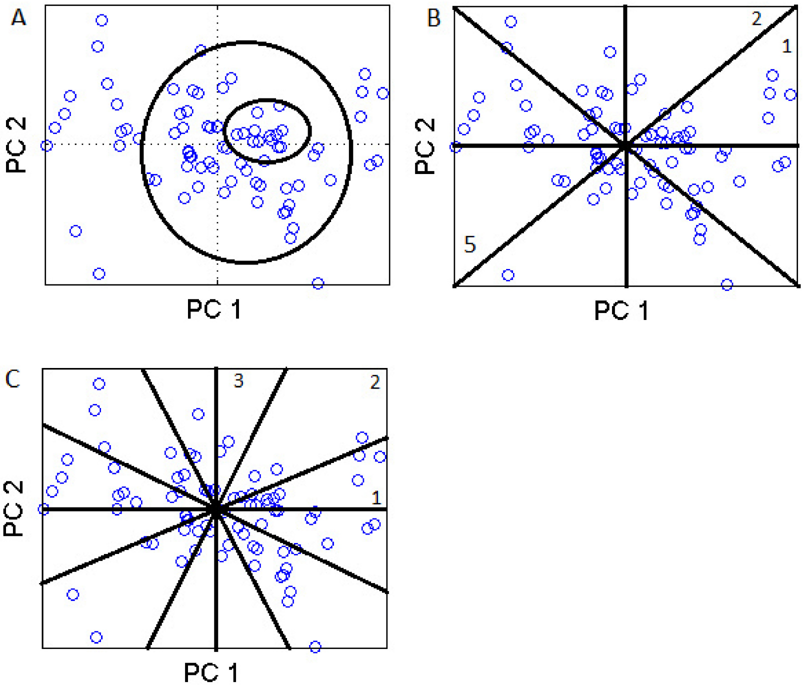

Figure 5.6: Tiling concept of PC-space from the continuous data. The first two components from 100 randomly selected profiles are plotted against each other to illustrate the concept.
(A): Partition from clustering approaches as hierarchical clustering or k-nearest neighbours.
(B): Partition by bisection of the half-spaces.
(C): Partition by trisection of the half-spaces.

Proof of Concept for Tiling and Angular Traits

Yeung et al. used PCA as pre-processing for data, they give an example where observed pattern are color-coded (figure 5.7). Band-like and cone-like structures can be clearly observed but pass through unremarked in the paper. A similar coloring scheme from clustering was observed by other authors, cone-like structures can be seen but they were not discussed [Holter et al., 2000].

5.4.3 Application of Angular Traits

Three examples of data are evaluated using this new clustering approach, the already discussed transcriptome data from continuous and batch fermentations and a real-time PCR data-set [‡]. The overview of data variances is given in figure 5.8. Increasing N_P and increasing n_{angle}, the fraction of occupied to unoccupied

[‡] that was generously contributed by Susanne Wickert, AG Prowe, Beuth Hochschule, Berlin

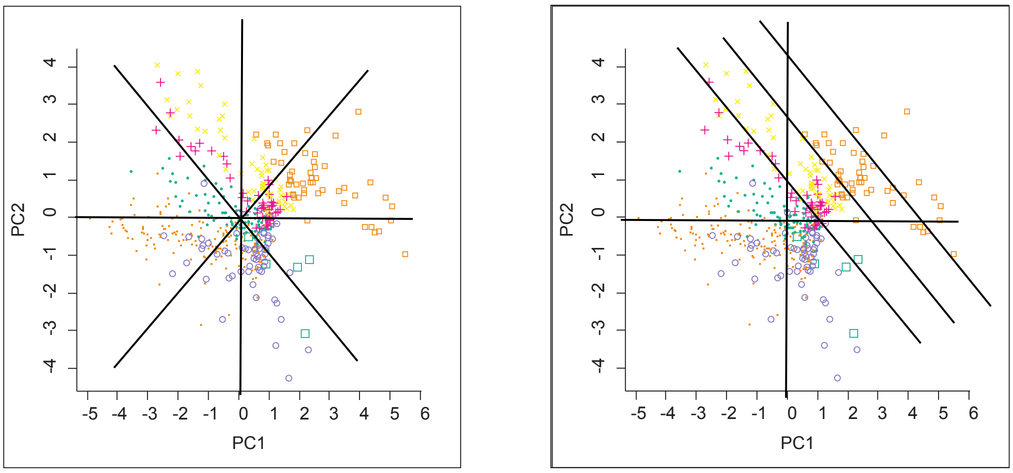

Figure 5.7: Pre-clustered data from cell cycle microarray data of yeast in a PC-representation. Clusters are represented as colors. Cone-like and beam-like structure are readily observed. Lines are manually drawn into the graphs. Adaption from [Yeung and Ruzzo, 2001].

sectors rapidly converges close to zero (table 5.1). There is no best choice of N_P known [Yeung and Ruzzo, 2001]. Examples of clusters are shown in figure 5.9.

Table 5.1: Overview of angular traits. The fraction of occupied to all sectors is presented as function of N_P and n_{angle}. In the continuous culture N_T=4, $N_J = 3807$ while in the batch culture $N_T = 25$, $N_J = 1862$ and in the RT-PCR experiment $N_T = 91$, $N_J = 181$.

experiment	n_{angle}	$N_P = 2$	$N_P = 3$	$N_P = 4$	$N_P = 5$	$N_P = 6$	$N_P = 7$
conti. cult.	8	1	0.5	0.22	-	-	-
batch. cult.	8	1	0.5	0.25	0.1157	0.034	0.0064
RT-PCR	8	0.5	0.2031	0.0684	0.0146	0.0026	0

Transcript data from batch culture

This data-set shows a rich dynamic. Four PCs are needed to model more than 80% of the data and sectors are the most occupied compared to the other two data-sets. One can expect to find many dynamic features. The approach allows to cover dynamically very similar and very close profiles whose levels spread during the end of fermentation when sporulation of the cells and degradation of transcripts occur.

In figure 5.9, examples of co-regulated and anti-regulated clusters are shown in A1 and A2 respectively.

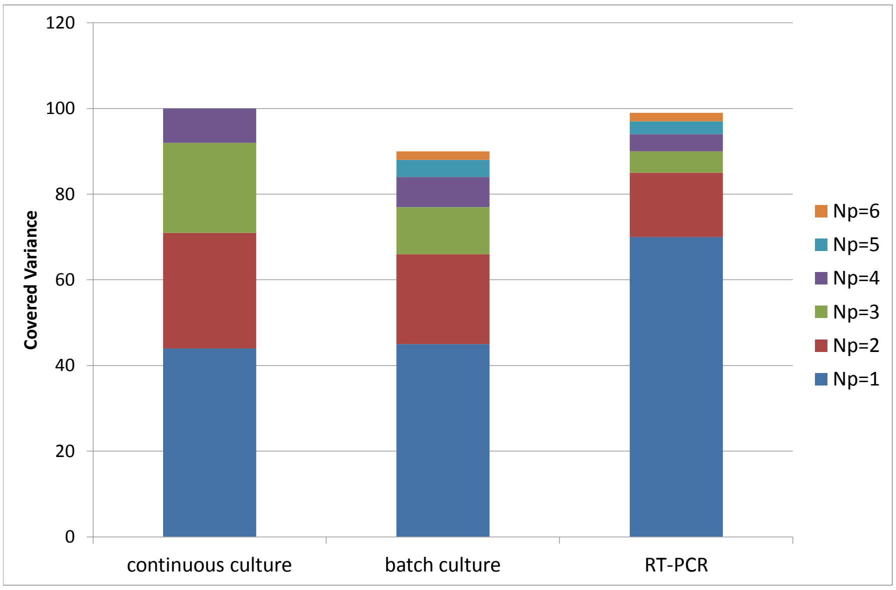

Figure 5.8: Variance covered by the principal component representation of transcriptome experiments from continuous culture, from batch culture and from RT-PCR data from *B.subtilis*.

Transcript data from continuous culture

This data-set shows a low dynamic and the smallest in temporal size. Three components are needed to cover more than 80% of the variance. Due to the lack of temporal resolution, clusters sizes are very large.
The example in 5.9, B shows again the co-regulated and anti-regulated clusters.

RT-PCR

This data-set is the smallest size although it has the largest resolution in time. Clustering from PCs shows that only two of four half-spaces are occupied by the loadings of the second PC. This takes into account that this data is only positive and no anti-regulated clusters can be expected. Further, two components suffice to model more than 80% of the data's variance.
The example in 5.9,C shows the typical profiles that contain a spike.

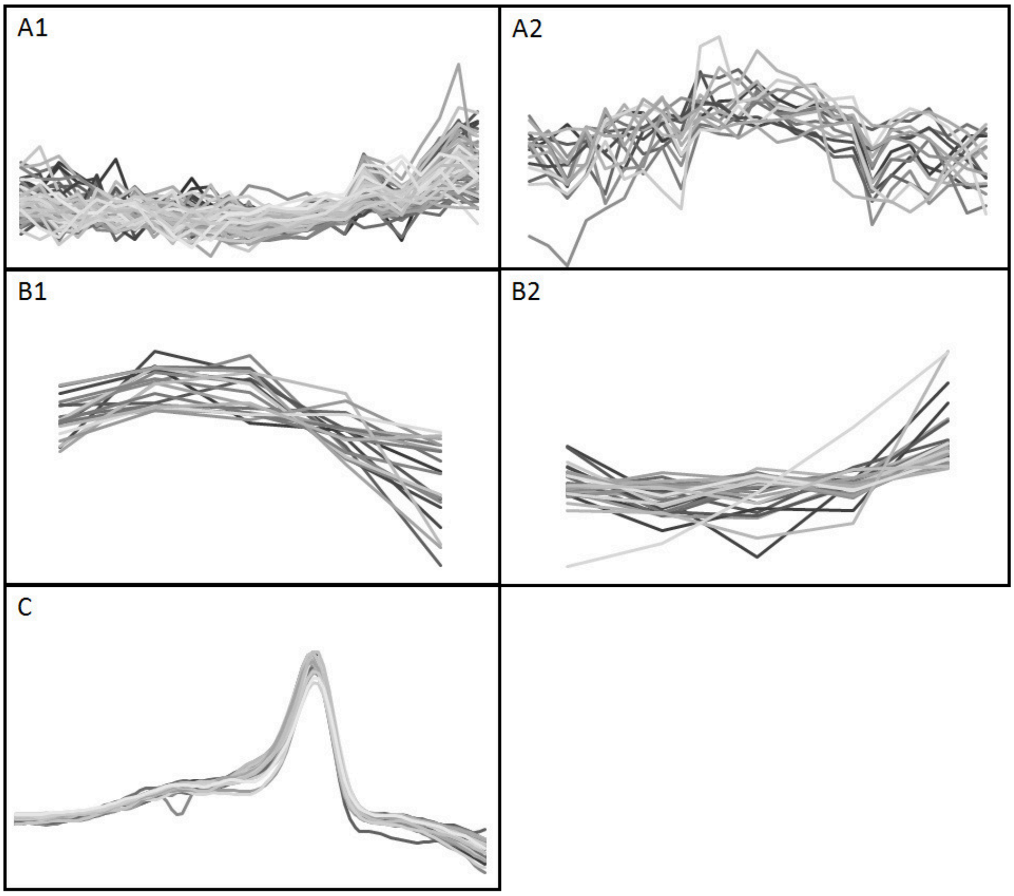

Figure 5.9: Angular traits examples from three data-sets. Examples are drawn by random selection from the whole data for better visualisation. Axes labels are omitted for simplicity.
(A1 & A2): batch data, $N_P = 4$, co-regulated and anti-regulated
(B1 & B2): continuous data, $N_P = 4$, co-regulated and anti-regulated
(C): RT-PCR data, $N_P = 4$

5.5 Final Conclusions

PCA seems to encounter a revival in literature [§], methods that use PCA are widely accepted and sophisticated alterations, e.g. rotation of components are applicable [Abdi and Williams, 2010]. The compression function of PCA is greatly known since the famous eigen-faces have been published [Turk and Pentland, 1991]. Recently, a report was published on the use of PCA to simplify sensitivity analysis outputs in a meaningful way [Sumner et al., 2012]. The here presented analyses add three further application of PCA in the biological sciences.

[§] more than 21000 articles were published as of 16th august 2013, in Google Scholar, one year earlier only 2800 were published

First, a compression of transcript expression level profiles was calculated to increase computational power of the parameter estimation. This is not treated because such dynamic model is not published. Second, a model-assisted optimisation approach was suggested from PCA that allowed the directed alternation of transcript expression data dynamics. This is also new because of the novel model structure. Finally, the use of principal components for clustering was elucidated in a hitherto untreated way for the identification of dynamically equal profiles.

A Novel Clustering Approach Was Introduced

The here presented algorithm for clustering is to best knowledge novel in literature. The uses of PCA are manifold and its use in clustering was limited to preprocessing of data [Yeung and Ruzzo, 2001]. Here, two concepts, *angular traits* and *angular clusters* make use of the information stored in the principal components.
Four challenges were defined by [Jiang et al., 2004b] to a good clustering algorithm:

1. No dependance on prior knowledge:
   No such knowledge is necessary here and parameter numbers are low.

2. High fidelity to filter signal from noise:
   Noisy components of the data are filtered by PCA in the higher dimensional components.

3. Possibility to build hierarchical structures:
   The number of sectors n_{angle} and the angular closeness approach can be used to build a hierarchical structure.

4. Possibility to retrieve relationships between clusters:
   Anti-regulation and co-regulation are retrieved by comparison of sectors.

Validation of clusters is usually carried out by assessing several similarity criteria [Jiang et al., 2004b], this is not possible here since by definition of sectors, genes within one sector are considered dynamically equal. A recurrence to other similarity metrics like the euclidean space is not possible. A second possibility is given by comparing the given clustering with a master clustering and to compare the sorting of both. One similar method to this clustering is the Pearson correlation to assess co-expression [Carrera et al., 2009], the building of a coherent cluster [Jiang et al., 2004a] using Pearson correlation requires the definition of a threshold. Such definition was avoided here by introducing a qualitative factor, the geometric assumptions.

Not All Data is Equally Well Represented by PCA

The use of PCA in knowledge generation is known [Alter et al., 2000], the here presented algorithm can be considered as a complementary approach to already existing ones, because it emphasises on structure evaluation. Since the number of considered principal components is directly related to the structure properties, angular traits becomes difficult when a large number of components is possible but only few data are present ($N_T \approx N_J$), which is the case for the RT-PCR data. Data curves that are degenerated in the sense of having a unique profile compared to the others are passing through undistinguished when N_P is chosen too small. When in contrast in such data N_P is chosen too large, major pattern are split into small pieces. The here presented algorithm works under the assumption that all noise is equally spread in components with numbers larger than N_P. In the RT-PCR data this is not true. An identification of the correct N_P would require to measure the information content of angular traits per dimensions. Ultimately, this leads to the question, if the proposed equality principle must be relaxed.

PCA for Modelling

The use of PCA for enhancing calculation speed by reducing interpolation effort was proven. Its use for optimisation also proposes beneficial outcomes. Here, the effect of such a change is critical. Large changes of the solvent spectrum lead to large effect on transcriptome [Tummala et al., 2003a], the change of transcript expression pattern should be soft. In order to investigate the effect of these changes further, more experimental data is required.

Summary

1. Workflows were programmed that allowed automatic harvesting of the KEGG-database. Compound information, Pfam-motif annotation of various organisms, gene-enzyme-reaction-reaction pair mapping were achieved and created the basis for automated creation of static and dynamic models.

2. A formalism for the integration of pathway information and transcript expression data was proposed. Transcriptome data was organised in a coherent way and it was visualised successfully. Inspection of the pathway models and of comparative maps to *B. subtilis* allowed hypothesis generation of a novel pathway that requires the activity of an unannotated 3-hydroxybutyrate dehydrogenase.

3. Application of the domain-grammar hypothesis inspired the formulation of an algorithm for Pfam-motif collection of 3-HBDH from 750 organisms from a frequentist point of view. Integration of experimental data by assuming parallel regulation in clusters constructed hypotheses that can be experimentally verified. The ranking of hypotheses was enabled by weighting the results. This weight needs re-thinking in order to facilitate the ranking procedure. The CA_{C3335} gene was suggested to contain a 3-HBDH activity during solventogenesis.

4. Automated creation and integration of transcript data into KEGG-derived dynamic models was programmed. This model type succeeded the representation of the pH-shift experiment without containing a pH-dependency. Cross-validation between batch and continuous data was successful and the parametrised model was qualitatively validated by simulating published mutation experiments. Parameter certainty was discussed and possible extensions of the model were proposed.

5. Local and global sensitivity analysis were performed for bottleneck identification. GSA showed that butanol dehydrogenase cannot pull enough the carbon from other reactions like the butyrate production that was shown to

be cycling between the kinase and the CoA-transferase pathway. The model was shown to be robust against the major parameter changes.

6. Speed optimisation of the dynamic model was performed using PCA. This approach inspired a novel algorithm for model evaluation based on the dynamics of transcript level data. An optimal transcript profile for butanol dehydrogenase could not be found from the data but for butyrate transacetylase.

7. Furthermore, a clustering algorithm was constructed from PCA by changing the perspective from parametric clustering to geometric assumptions. The characterisation of data according to their angular traits and angular similarity gives a promising novel way of performing informative clustering. Co-regulated and anti-regulated genes are easily computed by this algorithm.

Bibliography

[Abdi and Williams, 2010] Abdi, H. and Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–59. [cited at p. 120, 135]

[Agren et al., 2013] Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS Comput Biol*, 9(3):e1002980. [cited at p. 29, 116]

[Aittokallio and Schwikowski, 2006] Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–55. [cited at p. 32, 34, 78]

[Akesson et al., 2004] Akesson, M., Förster, J., and Nielsen, J. (2004). Integration of gene expression data into genome-scale metabolic models. *Metabolic engineering*, 6(4):285. [cited at p. 36]

[Aksenov et al., 2005] Aksenov, S., Church, B., Dhiman, A., Georgieva, A., Sarangapani, R., Helmlinger, G., and Khalil, I. (2005). An integrated approach for inference and mechanistic modeling for advancing drug development. *FEBS Letters*, 579(8):1878 – 83. [cited at p. 3]

[Alsaker and Papoutsakis, 2005] Alsaker, K. and Papoutsakis, E. (October 15, 2005). Transcriptional program of early sporulation and stationary-phase events in clostridium acetobutylicum. *Journal of Bacteriology*, 187(20):7103–18. [cited at p. 21]

[Alsaker et al., 2010] Alsaker, K., Paredes, C., and Papoutsakis, E. (2010). Metabolite stress and tolerance in the production of biofuels and chemicals: Gene-expression-based systems analysis of butanol, butyrate, and acetate stresses in the anaerobe clostridium acetobutylicum. *Biotechnology and Bioengineering*, 105(6):1131–47. [cited at p. 21, 22, 35, 39, 46]

[Alsaker et al., 2004] Alsaker, K., Spitzer, T., and Papoutsakis, E. (2004). Transcriptional analysis of spo0a overexpression in clostridium acetobutylicum and its effect on the cell's response to butanol stress. *Journal of Bacteriology*, 186(7):1959–71. [cited at p. 14, 20, 21]

[Alter et al., 2000] Alter, O., Brown, P., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–6. [cited at p. 120, 137]

[Altschul et al., 1997] Altschul, S., Madden, T., Schaeffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–402. [cited at p. 64]

[Alves et al., 2006] Alves, R., Antunes, F., and Salvador, A. (2006). Tools for kinetic modeling of biochemical networks. *Nature biotechnology*, 24(6):667–72. [cited at p. 3, 92]

[Amador-Noguez et al., 2011] Amador-Noguez, D., Brasg, I. A., Feng, X., Roquet, N., and Rabinowitz, J. D. (2011). Metabolome remodeling during the acidogenic-solventogenic transition in clostridium acetobutylicum. *Applied and Environmental Microbiology*, 77(22):7984–97. [cited at p. 17, 20, 95]

[Andrade and Vasconcelos, 2003] Andrade, J. C. and Vasconcelos, I. (2003). Continuous cultures of clostridium acetobutylicum: culture stability and low-grade glycerol utilisation. *Biotechnology Letters*, 25:121–5. [cited at p. 16]

[Arnold, 2002] Arnold, S. (2002). *Kinetic Modelling of Gene Expression.* PhD thesis, University of Stuttgart, Germany. [cited at p. 89, 117]

[Badr et al., 2001] Badr, H., Toledo, R., and Hamdy, M. (2001). Continuous acetone-ethanol-butanol fermentation by immobilized cells of clostridium acetobutylicum. *Biomass and Bioenergy*, 20(2):119 – 32. [cited at p. 18]

[Bahl et al., 1982a] Bahl, H., Andersch, W., Braun, K., and Gottschalk, G. (1982a). Effect of ph and butyrate concentration on the production of acetone and butanol by clostridium acetobutylicum grown in continuous culture. *Applied Microbiology and Biotechnology*, 14:17–20. [cited at p. 13, 16]

[Bahl et al., 1982b] Bahl, H., Andersch, W., and Gottschalk, G. (1982b). Continuous production of acetone and butanol by clostridium acetobutylicum in a two-stage phosphate limited chemostat. *Applied Microbiology and Biotechnology*, 15:201–5. [cited at p. 16]

[Bajad et al., 2006] Bajad, S., Lu, W., Kimball, E., Yuan, J., Peterson, C., and Rabinowitz, J. (2006). Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *Journal of Chromatography A*, 1125(1):76 – 88. [cited at p. 20]

[Balodimos et al., 1988] Balodimos, I., Kashket, E., and Rapaport, E. (1988). Metabolism of adenylylated nucleotides in clostridium acetobutylicum. *Journal of Bacteriology*, 170(5):2301–2305. [cited at p. 17]

[Barabási and Oltvai, 2004] Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Review Bioinformatics*, 5:101–13. [cited at p. 32]

[Barbeau et al., 1988] Barbeau, J., Marchal, R., and Vandecasteele, J. (1988). Conditions promoting stability of solventogenesis or culture degeneration in continuous fermentations of clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 29(5):447–55. [cited at p. 10]

[Barker et al., 1978] Barker, H., Jeng, I., Neff, N., Robertson, J., Tam, F., and Hosaka, S. (1978). Butyryl-coa:acetoacetate coa-transferase from a lysine-fermenting clostridium. *Journal of Biological Chemistry*, 253(4):1219–25. [cited at p. 66]

[Bassett Jr et al., 1999] Bassett Jr, D., Eisen, M., and Boguski, M. (1999). Gene expression informatics its all in your mine. *Nature Reviews*, 21:51–5. [cited at p. 37]

[Beesch, 1952] Beesch, S. (1952). Acetone-butanol fermentation of sugars. *Industrial & Engineering Chemistry*, 44(7):1677–82. [cited at p. 8, 16]

[Bentele et al., 2004] Bentele, M., Lavrik, I., Ulrich, M., Ster, S., Heermann, D., Kalthoff, H., Krammer, P., and Eils, R. (2004). Mathematical modeling reveals threshold mechanism in cd95-induced apoptosis. *The Journal of Cell Biology*, 166(6):839–51. [cited at p. 106, 110]

[Bonnet et al., 2013] Bonnet, E., Calzone, L., Rovera, D., Stoll, G., Barillot, E., and Zinovyev, A. (2013). Binom 2.0, a cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Systems Biology*, 7(1):18. [cited at p. 3]

[Borden and Papoutsakis, 2007] Borden, J. and Papoutsakis, E. (2007). Dynamics of genomic-library enrichment and identification of solvent tolerance genes for clostridium acetobutylicum. *Applied and Environmental Microbiology*, 73(9):3061–8. [cited at p. 18]

[Bowles and Ellefson, 1985] Bowles, L. and Ellefson, W. (1985). Effects of butanol on clostridium acetobutylicum. *Applied and Environmental Microbiology*, 50(5):1165–70. [cited at p. 17]

[Boynton et al., 1994] Boynton, Z. L., B, G. N., and Rudolph, F. B. (1994). Intracellular concentrations of coenzyme a and its derivatives from clostridium acetobutylicum atcc 824 and their roles in enzyme regulation. *Applied and Environmental Microbiology*, 60(1):39–44. [cited at p. 13, 95]

[Brekke, 2007] Brekke, K. (2007). Butanol - an energy alternative. *Ethanol Today*, pages 36–9. [cited at p. 7]

[Brohee and van Helden, 2006] Brohee, S. and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488. [cited at p. 76]

[Brown et al., 2005] Brown, M., Dunn, W., Ellis, D., Goodacre, R., Handl, J., Knowles, J., O'Hagan, S., Spasic, I., and Kell, D. (2005). A metabolome pipeline: from concept to data to knowledge. *Metabolomics*, 1(1):39–51. [cited at p. 120, 129]

[Brown et al., 2000] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–7. [cited at p. 76]

[Buday et al., 1990] Buday, Z., Linden, J., and Karim, M. (1990). Improved acetone-butanol fermentation analysis using subambient hplc column temperature. *Enzyme and Microbial Technology*, 12(1):24 – 7. [cited at p. 20]

[Cacuci, 2003] Cacuci, D. G. (2003). *Sensitivity and uncertainty analysis*. Chapman & Hall. [cited at p. 106]

[Cakir et al., 2006] Cakir, T., Raosaheb Patil, K., Ilsen Önsan, Z., Özergin Ülgen, K., Kirdar, B., and Nielsen, J. (2006). Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular Systems Biology*, 2:1–11. [cited at p. 37, 53]

[Cammer et al., 2003] Cammer, S., Hoffman, B., Speir, J., Canady, M., Nelson, M., Knutson, S., Gallina, M., Baxter, S., and Fetrow, J. (2003). Structure-based active site profiles for genome analysis and functional family subclassification. *Journal of Molecular Biology*, 334(3):387 – 401. [cited at p. 75]

[Carrera et al., 2009] Carrera, J., Rodrigo, G., and Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic acids research*, 37(5):1–11. [cited at p. 81, 136]

[Chang et al., 2009] Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009). Brenda, amenda and frenda the enzyme information system: new content and tools in 2009. *Nucleic Acids Research*, 37(suppl 1):D588–D592. [cited at p. 38]

[Chen et al., 1999] Chen, T., He, H., and Church, G. (1999). Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, number 29 in 4, page 4. [cited at p. 89, 117]

[Chevenet et al., 2006] Chevenet, F., Brun, C., Banuls, A., Jacq, B., and Christen, R. (2006). Treedyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*, 7(1):439. [cited at p. 55]

[Cho et al., 2012] Cho, D., Shin, S., and Kim, Y. (2012). Effects of acetic and formic acid on abe production by clostridium acetobutylicum and clostridium beijerinckii. *Biotechnology and Bioprocess Engineering*, 17:270–5. [cited at p. 13]

[Cho et al., 2003] Cho, K., Shin, S., Kolch, W., and Wolkenhauer, O. (2003). Experimental design in systems biology, based on parameter sensitivity analysis using a monte carlo method: A case study for the tnf α-mediated nf-κ b signal transduction pathway. *Simulation*, 79(12):726–9. [cited at p. 106]

[Clarke et al., 1988] Clarke, K., Hansford, G., and Jones, D. (1988). Nature and significance of oscillatory behavior during solvent production by clostridium acetobutylicum in continuous culture. *Biotechnology and Bioengineering*, 32(4):538–44. [cited at p. 14, 95]

[Company, 2006] Company, T. D. C. (2006). Product safety assessment n-butanol. Internet. Form No. 233-00247-KC-0406. [cited at p. 6]

[Covert et al., 2001] Covert, M., Schilling, C., and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213(1):73 – 88. [cited at p. 36, 81]

[Crown et al., 2011] Crown, S., Indurthi, D., Ahn, S., Choi, J., Papoutsakis, E., and Antoniewicz, M. (2011). Resolving the tca cycle and pentose-phosphate pathway of clostridium acetobutylicum atcc 824: Isotopomer analysis, in vitro activities and expression analysis. *Biotechnology journal*, 6(3):300–5. [cited at p. 56]

[Dallas et al., 2005] Dallas, P., Gottardo, N., Firth, M., Beesley, A., Hoffmann, K., Terry, P., Freitas, J., Boag, J., Cummings, A., and Kees, U. (2005). Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time rt-pcr - how well do they correlate? *BMC Genomics*, 6(1):59. [cited at p. 21]

[Daub et al., 2003] Daub, C., Kloska, S., and Selbig, J. (2003). Metagenealyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics*, 19(17):2332–3. [cited at p. 90, 173]

[Dennis et al., 2003] Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3. [cited at p. 55]

[Desai et al., 1999] Desai, R. P., Harris, L., Welker, N. E., and Papoutsakis, E. (1999). Metabolic flux analysis elucidates the importance of the acid-formation pathways in regulating solvent production by clostridium acetobutylicum. *Metabolic Engineering*, 1(3):206 – 13. [cited at p. 13, 80, 95]

[Dhaeseleer et al., 2000] Dhaeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–26. [cited at p. 37, 64, 76, 117]

[Doris de Guzman, 2011] Doris de Guzman, I. (2011). Green chemicals: Growing number of chemical firms enter bio-butanol space. Internet. [cited at p. 7]

[Downer et al., 2006] Downer, J., Sevinsky, J., Ahn, N., Resing, K., and Betterton, M. (2006). Incorporating expression data in metabolic modeling: A case study of lactate dehydrogenase. *Journal of Theoretical Biology*, 240(3):464 – 74. [cited at p. 81]

[Duarte et al., 2007] Duarte, N., Becker, S., Jamshidi, N., Thiele, I., Mo, M., Vo, T., Srivas, R., and Palsson, B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–82. [cited at p. 36]

[Dudley, 2012a] Dudley, B. (2012a). BP energy outlook 2030. Technical report, British Petroleum. [cited at p. 7]

[Dudley, 2012b] Dudley, B. (2012b). BP statistical review of world energy june 2012. Technical report, British Petroleum. [cited at p. 6]

[Duerre, 2005] Duerre, P. (2005). *Handbook On Clostridia*. Taylor & Francis. [cited at p. 10, 11, 13, 15, 66]

[Duerre, 2007] Duerre, P. (2007). Biobutanol: An attractive biofuel. *Biotechnol. J*, 2:1525–34. [cited at p. 6, 7]

[Duerre et al., 2002] Duerre, P., Boehringer, M., Nakotte, S., Schaffer, S., Thormann, K., and Zickner, B. (2002). Transcriptional regulation of solventogenesis in clostridium acetobutylicum. *Journal of molecular microbiology and biotechnology*, 4(3):295. [cited at p. 14]

[Durot et al., 2009] Durot, M., Bourguignon, P., and Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS microbiology reviews*, 33(1):164–90. [cited at p. 28, 56, 78, 82, 117]

[Dusseaux et al., 2013] Dusseaux, S., Croux, C., Soucaille, P., and Meynial-Salles, I. (2013). Metabolic engineering of clostridium acetobutylicum atcc 824 for the high-yield production of a biofuel composed of an isopropanol/butanol/ethanol mixture. *Metabolic Engineering*, 18(0):1 – 8. [cited at p. 15]

[Dutta et al., 2009] Dutta, B., Kanani, H., Quackenbush, J., and Klapa, M. (2009). Time-series integrated omic analyses to elucidate short-term stress-induced responses in plant liquid cultures. *Biotechnology and bioengineering*, 102(1):264–79. [cited at p. 34, 120]

[Eckert and Schuegerl, 1987] Eckert, G. and Schuegerl, K. (1987). Continuous acetone-butanol production with direct product removal. *Applied Microbiology and Biotechnology*, 27:221–8. [cited at p. 18]

[Eisen et al., 1998] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–8. [cited at p. 129]

[Ezeji and Blaschek, 2008] Ezeji, T. and Blaschek, H. P. (2008). Fermentation of dried distillers grains and solubles (ddgs) hydrolysates to solvents and value-added products by solventogenic clostridia. *Bioresource Technology*, 99(12):5232 – 42. [cited at p. 17]

[Ezeji et al., 2010] Ezeji, T., Milne, C., Price, N., and Blaschek, H. (2010). Achievements and perspectives to overcome the poor solvent resistance in acetone and butanol-producing microorganisms. *Applied Microbiology and Biotechnology*, 85(6):1697–712. [cited at p. 18]

[Farrugia and Quigley, 2011] Farrugia, M. and Quigley, A. (2011). Effective temporal graph layout: A comparative study of animation versus static display methods. *Information Visualization*, 10(1):47–64. [cited at p. 55]

[Feder and Walser, 2005] Feder, M. and Walser, J. (2005). The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology*, 18(4):901–10. [cited at p. 129]

[Fiehn, 2001] Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2(3):155–68. [cited at p. 20]

[Flannery, 2007] Flannery, W. P. S. T. W. V. B. (2007). *Numerical Recipies in C.* Cambridge University Press; 3 edition. [cited at p. 94]

[Fond et al., 1986] Fond, O., Engasser, J., Matta-El-Amouri, G., and Petitdemange, H. (1986). The acetone butanol fermentation on glucose and xylose. ii regulation and kinetics in fed-batch cultures. *Biotechnology and Bioengineering*, 28(2):167–75. [cited at p. 16]

[Fond et al., 1984] Fond, O., Petitdemange, E., Petitdemange, H., and Gay, R. (1984). Effect of glucose flow on the acetone butanol fermentation in fed batch culture. *Biotechnology Letters*, 6(1):13–8. [cited at p. 20]

[Forsberg, 1987] Forsberg, C. W. (1987). Production of 1,3-propanediol from glycerol by clostridium acetobutylicum and other clostridium species. *Applied and Environmental Microbiology*, 53(4):639–43. [cited at p. 16]

[Forslund, 2011] Forslund, K. (2011). *The relationship between orthology, protein domain architectue and protein function*. PhD thesis, Stockholm University, Sweden. [cited at p. 56, 64, 74]

[Freeman et al., 2007] Freeman, R., Goldovsky, L., Brosch, M., van Dongen, S., Maziere, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol*, 3(10):e206. [cited at p. 36, 37, 76]

[Gama-Castro et al., 2008] Gama-Castro, S., Jimnez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H., Bonavides-Martinez, C., Abreu-Goodger, C., Rodriguez-Penagos, C., Miranda-Rios, J., Morett, E., Merino, E., Huerta, A., Trevino-Quintanilla, L., and Collado-Vides, J. (2008). Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Research*, 36(suppl 1):D120–D124. [cited at p. 54]

[Gheshlaghi, 2009] Gheshlaghi, R. (2009). Metabolic pathways of clostridia for producing butanol. *Biotechnology advances*, 27(6):764–81. [cited at p. 12, 17, 82]

[Golding et al., 2005] Golding, I., Paulsson, J., Zawilski, S., and Cox, E. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36. [cited at p. 37, 53, 89]

[Gottwald and Gottschalk, 1985] Gottwald, M. and Gottschalk, G. (1985). The internal ph of clostridium acetobutylicum and its effect on the shift from acid to solvent formation. *Archives of Microbiology*, 143(1):42–6. [cited at p. 13]

[Götz and Reuss, 2009] Götz, P. and Reuss, M. (2009). Generation of regulatory hypotheses by descriptive modeling and data integration. *New Biotechnology*, 25, Supplement(0):S338. ¡ce:title¿Abstracts of the 14th European Congress on BiotechnologyBarcelona, Spain 1316 September, 2009¡/ce:title¿. [cited at p. 81]

[Green et al., 1996] Green, E., Boynton, Z., Harris, L., Rudolph, F., Papoutsakis, E., and Bennett, G. (1996). Genetic manipulation of acid formation pathways by gene inactivation in clostridium acetobutylicum atcc 824. *Microbiology*, 142(8):2079–86. [cited at p. 14, 20, 65, 66, 95, 103]

[Grimmler et al., 2011] Grimmler, C., Janssen, H., Krausse, D., Fischer, R., Bahl, H., Duerre, P., Liebl, W., and Ehrenreich, A. (2011). Genome-wide gene expression analysis of the switch between acidogenesis and solventogenesis in continuous cultures of clostridium acetobutylicum. *J Mol Microbiol Biotechnol*, 20:1–15. [cited at p. 10, 22, 39, 47, 66, 68, 93, 95]

[Grupe and Gottschalk, 1992] Grupe, H. and Gottschalk, G. (1992). Physiological events in clostridium acetobutylicum during the shift from acidogenesis to solventogenesis in continuous culture and presentation of a model for shift induction. *Applied and Environmental Microbiology*, 58(12):3896–902. [cited at p. 13, 95, 104]

[Gustafsson et al., 2009] Gustafsson, M., Hörnquist, M., Lundström, J., Björkegren, J., and Tegner, J. (2009). Reverse engineering of gene networks with lasso and nonlinear basis functions. *Annals of the New York Academy of Sciences*, 1158(1):265–75. [cited at p. 88]

[Harris et al., 2000] Harris, L., Desai, R., Welker, N., and Papoutsakis, E. (2000). Characterization of recombinant strains of the clostridium acetobutylicum butyrate kinase inactivation mutant: Need for new phenomenological models for solventogenesis and butanol inhibition? *Biotechnology and Bioengineering*, 67(1):1–11. [cited at p. 14, 20]

[Harris et al., 2002] Harris, L., Welker, N. E., and Papoutsakis, E. (2002). Northern, morphological, and fermentation analysis of spo0a inactivation and overexpression in clostridium acetobutylicum atcc 824. *Journal of Bacteriology*, 184(13):3586–97. [cited at p. 14]

[Hartmanis and Gatenbeck, 1984] Hartmanis, M. and Gatenbeck, S. (1984). Intermediary metabolism in clostridium acetobutylicum: Levels of enzymes involved in the formation of acetate and butyrate. *Applied and Environmental Microbiology*, 47(6):1277–83. [cited at p. 95]

[Hartmanis et al., 1984] Hartmanis, M., Klason, T., and Gatenbeck, S. (1984). Uptake and activation of acetate and butyrate in clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 20(1):66–71. [cited at p. 66]

[Haus et al., 2011] Haus, S., Jabbari, S., Millat, T., Janssen, H., Fischer, R., Bahl, H., King, J., and Wolkenhauer, O. (2011). A systems biology approach to investigate the effect of ph-induced gene regulation on solvent production by clostridium acetobutylicum in continuous culture. *BMC Systems Biology*, 5(1):10. [cited at p. 13, 81]

[Heap et al., 2007] Heap, J., Pennington, O., Cartman, S., Carter, G., and Minton, N. (2007). The clostron: A universal gene knock-out system for the genus clostridium. *Journal of Microbiological Methods*, 70(3):452 – 464. [cited at p. 14]

[Hernandez, 2004] Hernandez, O. (2004). n-butyl alcohol. Technical report, UNEP. [cited at p. 7]

[Holter et al., 2000] Holter, N., Mitra, M., Maritan, A., Cieplak, M., Banavar, J., and Fedoroff, N. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–14. [cited at p. 120, 123, 130, 132]

[Huang et al., 2009] Huang, D., Sherman, B., and Lempicki, R. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13. [cited at p. 53, 55]

[Huang et al., 1985] Huang, L., Gibbins, L., and Forsberg, C. (1985). Transmembrane ph gradient and membrane potential in clostridium acetobutylicum during growth under acetogenic and solventogenic conditions. *Applied and Environmental Microbiology*, 50(4):1043–7. [cited at p. 13, 17]

[Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., the rest of the SBML Forum:, Arkin, A., Bornstein, B., Bray, D., Cornish-Bowden, A., Cuellar, A., Dronov, S., Gilles, E., Ginkel, M., Gor, V., Goryanin, I., Hedley, W., Hodgman, T., Hofmeyr, J., Hunter, P., Juty, N., Kasberger, J., Kremling, A., Kummer, U., Le Novere, N., Loew, L., Lucio, D., Mendes, P., Minch, E., Mjolsness, E., Nakayama, Y., Nelson, M., Nielsen, P., Sakurada, T., Schaff, J., Shapiro, B., Shimizu, T., Spence, H., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531. [cited at p. 3]

[Huesemann and Papoutsakis, 1988] Huesemann, M. and Papoutsakis, E. (1988). Solventogenesis in clostridium acetobutylicum fermentations related to carboxylic acid and proton concentrations. *Biotechnology and Bioengineering*, 32(7):843–52. [cited at p. 13]

[Huesemann and Papoutsakis, 1990] Huesemann, M. and Papoutsakis, E. (1990). Effects of propionate and acetate additions on solvent production in batch cultures of clostridium acetobutylicum. *Applied and Environmental Microbiology*, 56(5):1497–500. [cited at p. 12]

[Hull et al., 2006] Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(suppl 2):W729–W732. [cited at p. 29, 167]

[Ingalls, 2013] Ingalls, B. (2013). *Mathematical Modelling in Systems Biology: An Introduction*. Internet. [cited at p. 117]

[Inui et al., 2008] Inui, M., Suda, M., Kimura, S., Yasuda, K., Suzuki, H., Toda, H., Yamamoto, S., Okino, S., Suzuki, N., and Yukawa, H. (2008). Expression of clostridium acetobutylicum butanol synthetic genes in escherichia coli. *Applied Microbiology and Biotechnology*, 77(6):1305–16. [cited at p. 15]

[Izak et al., 2008] Izak, P., Schwarz, K., Ruth, W., Bahl, H., and Kragl, U. (2008). Increased productivity of clostridium acetobutylicum fermentation of acetone, butanol, and ethanol by pervaporation through supported ionic liquid membrane. *Applied Microbiology and Biotechnology*, 78:597–602. [cited at p. 18]

[Jamshidi and Palsson, 2008] Jamshidi, N. and Palsson, B. (2008). Formulating genome-scale kinetic models in the post-genome era. *Molecular systems biology*, 4(1). [cited at p. 117]

[Janes and Yaffe, 2006] Janes, K. and Yaffe, M. (2006). Data-driven modelling of signal-transduction networks. *Nature Reviews Molecular Cell Biology*, 7(11):820–28. [cited at p. 120, 129]

[Jang et al., 2012] Jang, Y., Lee, J., Lee, J., Park, J., Im, J., Eom, M., Lee, J., Lee, S., Song, H., Cho, J., Seung, D., and Lee, S. (2012). Enhanced butanol production obtained by reinforcing the direct butanol-forming route in clostridium acetobutylicum. *mBio*, 3(5). [cited at p. 15]

[Janssen et al., 2010] Janssen, H., Doering, C., Ehrenreich, A., Voigt, B., Hecker, M., Bahl, H., and Fischer, R. (2010). A proteomic and transcriptional view of acidogenic and solventogenic steady-state cells of clostridium acetobutylicum in a chemostat culture. *Applied Microbiology and Biotechnology*, 87(6):2209–26. [cited at p. 13, 21]

[Janssen et al., 2012] Janssen, H., Grimmler, C., Ehrenreich, A., Bahl, H., and Fischer, R. (2012). A transcriptional study of acidogenic chemostat cells of clostridium acetobutylicum - solvent stress caused by a transient n-butanol pulse. *Journal of Biotechnology*, 161(3):354 – 65. [cited at p. 17, 21]

[Jarzebski et al., 1992] Jarzebski, A., Goma, G., and Soucaille, P. (1992). Modelling of continuous acetonobutylic fermentation. *Bioprocess and Biosystems Engineering*, 7:357–361. [cited at p. 80]

[Jiang et al., 2004a] Jiang, D., Pei, J., Ramanathan, M., Tang, C., and Zhang, A. (2004a). Mining coherent gene clusters from gene-sample-time microarray data. In *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. University, West. [cited at p. 136]

[Jiang et al., 2004b] Jiang, D., Tang, C., and Zhang, A. (2004b). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–86. [cited at p. 129, 136]

[Jiang et al., 2009] Jiang, Y., Xu, C., Dong, F., Yang, Y., Jiang, W., and Yang, S. (2009). Disruption of the acetoacetate decarboxylase gene in solvent-producing clostridium acetobutylicum increases the butanol ratio. *Metabolic Engineering*, 11(4-5):284 – 91. [cited at p. 65, 78]

[Jones and Woods, 1986] Jones, D. and Woods, D. (1986). Acetone-butanol fermentation revisited. *Microbiological Reviews*, 50(4):484–524. [cited at p. 7]

[Jones et al., 2008] Jones, S., Paredes, C., Tracy, B., Cheng, N., Sillers, R., Senger, R., and Papoutsakis, E. (2008). The transcriptional program underlying the physiology of clostridial sporulation. *Genome Biology*, 9(7):R114. [cited at p. 21, 39, 43, 68, 93, 95]

[Joyce and Palsson, 2006] Joyce, A. and Palsson, B. (2006). The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210. [cited at p. 20, 31, 32, 77, 78, 117]

[Junghans, 2008] Junghans, M. (2008). Visualization of hyperedges in fixed graph layouts. Master's thesis, Brandenburg University of Technology, Cottbus. [cited at p. 54]

[Junker et al., 2006] Junker, B., Klukas, C., and Schreiber, F. (2006). Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109. [cited at p. 29]

[Junne, 2010] Junne, S. (2010). *Stimulus response experiments for modelling product formation in Clostridium acetobutylicum fermentations*. PhD thesis, Berlin Institute of Technology, Germany, Berlin. [cited at p. 12, 13, 17, 80, 81, 82]

[Kansiz et al., 2001] Kansiz, M., Gapes, J., McNaughton, D., Lendl, B., and Schuster, K. (2001). Mid-infrared spectroscopy coupled to sequential injection analysis for the on-line monitoring of the acetone-butanol fermentation process. *Analytica Chimica Acta*, 438(1-2):175 – 86. [cited at p. 20]

[Katsnelson, 2013] Katsnelson, A. (2013). Momentum grows to make personalized medicine more precise. *Nature Medicine*, 19(3):249. [cited at p. 78]

[Khatri et al., 2012] Khatri, P., Sirota, M., and Butte, A. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375. [cited at p. 34, 53, 77, 78]

[Kim et al., 1984] Kim, B., Bellows, P., Datta, R., and Zeikus, J. (1984). Control of carbon and electron flow in clostridium acetobutylicum fermentations: Utilization of carbon monoxide to inhibit hydrogen production and to enhance butanol yields. *Applied and Environmental Microbiology*, 48(4):764–70. [cited at p. 17]

[Klamt and Gilles, 2004] Klamt, S. and Gilles, E. D. (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234. [cited at p. 78]

[Klamt et al., 2007] Klamt, S., Saez-Rodriguez, J., and Gilles, E. (2007). Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Systems Biology*, 1(1):2. [cited at p. 31, 117]

[Klipp et al., 2007] Klipp, E., Liebermeister, W., Helbig, A., Kowald, A., and Schaber, J. (2007). Standards in computational systems biology. [cited at p. 3]

[Klipp et al., 2004] Klipp, E., Liebermeister, W., and Wierling, C. (2004). Inferring dynamic properties of biochemical reaction networks from structural knowledge. *Genome Informatics Series*, pages 125–37. [cited at p. 32, 78]

[Koepke et al., 2011] Koepke, M., Mihalcea, C., Bromley, J., and Simpson, S. (2011). Fermentative production of ethanol from carbon monoxide. *Current Opinion in Biotechnology*, 22(3):320 – 5. [cited at p. 47]

[Koestler et al., 2010] Koestler, T., von Haeseler, A., and Ebersberger, I. (2010). Fact: Functional annotation transfer between proteins with similar feature architectures. *BMC Bioinformatics*, 11(1):417. [cited at p. 56, 64]

[Kuit et al., 2012] Kuit, W., Minton, N., Lopez-Contreras, A., and Eggink, G. (2012). Disruption of the acetate kinase (ack) gene of clostridium acetobutylicum results in delayed acetate production. *Applied Microbiology and Biotechnology*, 94(3):729–41. [cited at p. 15, 20]

[Kumar et al., 2012] Kumar, A., Suthers, P., and Maranas, C. (2012). Metrxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, 13(1):6. [cited at p. 28, 38, 116]

[Lebedeva et al., 2012] Lebedeva, G., Sorokin, A., Faratian, D., Mullen, P., Goltsov, A., Langdon, S., Harrison, D., and Goryanin, I. (2012). Model-based global sensitivity analysis as applied to identification of anti-cancer drug targets and biomarkers of drug resistance in the erbb2/3 network. *European Journal of Pharmaceutical Sciences*, 46(4):244 – 58. [cited at p. 106]

[Lee and Lee, 2009] Lee, B. and Lee, D. (2009). Protein comparison at the domain architecture level. *BMC Bioinformatics*, 10(Suppl 15):S5. [cited at p. 64, 71, 75]

[Lee et al., 2012] Lee, J., Jang, Y., Choi, S., Im, J., Song, H., Cho, J., Seung, D., Papoutsakis, E., Bennett, G., and Lee, S. (2012). Metabolic engineering of clostridium acetobutylicum atcc 824 for isopropanol-butanol-ethanol fermentation. *Applied and Environmental Microbiology*, 78(5):1416–23. [cited at p. 15]

[Lee et al., 2008a] Lee, J., Yun, H., Feist, A., Palsson, B., and Lee, S. (2008a). Genome-scale reconstruction and in silico analysis of the clostridium acetobutylicum atcc 824 metabolic network. *Applied Microbiology and Biotechnology*, 80:849–62. [cited at p. 28, 36, 64, 80, 117]

[Lee et al., 2008b] Lee, S., Park, J., Jang, S., Nielsen, L., Kim, J., and Jung, K. (2008b). Fermentative butanol production by clostridia. *Biotechnology and Bioengineering*, 101(2):209–28. [cited at p. 6, 11, 14, 116]

[Lehmann et al., 2012a] Lehmann, D., Hoenicke, D., Ehrenreich, A., Schmidt, M., Weuster-Botz, D., Bahl, H., and Luetke-Eversloh, T. (2012a). Modifying the product pattern of clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 94(3):743–54. [cited at p. 15, 20, 56, 65, 78, 95, 103, 115]

[Lehmann and Luetke-Eversloh, 2011] Lehmann, D. and Luetke-Eversloh, T. (2011). Switching clostridium acetobutylicum to an ethanol producer by disruption of the butyrate/butanol fermentative pathway. *Metabolic Engineering*, 13(5):464 – 73. [cited at p. 13, 15, 21, 74, 115]

[Lehmann et al., 2012b] Lehmann, D., Radomski, N., and Luetke-Eversloh, T. (2012b). New insights into the butyric acid metabolism of clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 96(5):1325–39. [cited at p. 15, 56, 65, 78, 95, 103]

[Lehmann et al., 1999] Lehmann, T., Gonner, C., and Spitzer, K. (1999). Survey: Interpolation methods in medical image processing. *Medical Imaging, IEEE Transactions on*, 18(11):1049–75. [cited at p. 88]

[Levandovski and D, 1971] Levandovski, M. and D, W. (1971). Distance between sets. *Letters to Nature*, 234:34–5. [cited at p. 58]

[Li et al., 2010] Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M., Snoep, J., Hucka, M., Le Novere, N., and Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4:92. [cited at p. 3]

[Li et al., 2007] Li, C., Yang, C., and Shan, H. (2007). Maximizing propylene yield by two-stage riser catalytic cracking of heavy oil. *Industrial & Engineering Chemistry Research*, 46(14):4914–20. [cited at p. 6]

[Liebermeister and Klipp, 2006] Liebermeister, W. and Klipp, E. (2006). Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theoretical Biology and Medical Modelling*, 3(1):42. [cited at p. 81]

[Lin et al., 2006] Lin, K., Zhu, L., and Zhang, D. (2006). An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, 22(17):2081–6. [cited at p. 56, 58, 74]

[Luetke-Eversloh and Bahl, 2011] Luetke-Eversloh, T. and Bahl, H. (2011). Metabolic engineering of clostridium acetobutylicum: recent advances to improve butanol production. *Current Opinion in Biotechnology*, 22(5):634 – 47. [cited at p. 10, 14]

[Lukashin and Fuchs, 2001] Lukashin, A. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17(5):405–14. [cited at p. 129]

[Maddox et al., 2000] Maddox, I., Steiner, E., Hirsch, S., Wessner, S., Gutierrez, N., Gapes, J., and Schuster, K. (2000). The cause of "acid crash" and "acidogenic fermentations" during the batch acetone-butanol-ethanol(abe-) fermentation process. *Journal of Molecular Microbiology and Biotechnology*, 2(1):95–100. [cited at p. 13]

[Mann and Luetke-Eversloh, 2013] Mann, M. and Luetke-Eversloh, T. (2013). Thiolase engineering for enhanced butanol production in clostridium acetobutylicum. *Biotechnology and Bioengineering*, 110(3):887–97. [cited at p. 14, 94, 95, 103, 117]

[Mariano et al., 2011] Mariano, A., Qureshi, N., Filho, R., and Ezeji, T. (2011). Bioproduction of butanol in bioreactors: New insights from simultaneous in situ butanol recovery to eliminate product toxicity. *Biotechnology and Bioengineering*, 108(8):1757–65. [cited at p. 18]

[Marino et al., 2008] Marino, S., Hogue, I., Ray, C., and Kirschner, D. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254(1):178 – 96. [cited at p. 110]

[Mashego et al., 2007] Mashego, M., Rumbold, K., Mey, M., Vandamme, E., Soetaert, W., and Heijnen, J. (2007). Microbial metabolomics: past, present and future methodologies. *Biotechnology Letters*, 29(1):1–16. [cited at p. 20]

[Matthews et al., 2009] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl 1):D619–D622. [cited at p. 28, 78]

[Meyer et al., 1986] Meyer, C., Roos, J., and Papoutsakis, E. (1986). Carbon monoxide gasing leads to alcohol production and butyrate uptake without acetone formation in continuous cultures of clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 24(2):159–67. [cited at p. 13, 17]

[Meyer and Papoutsakis, 1989] Meyer, C. L. and Papoutsakis, E. (1989). Continuous and biomass recycle fermentations of clostridium acetobutylicum. *Bioprocess and Biosystems Engineering*, 4:1–10. [cited at p. 19]

[Millat et al., 2013a] Millat, T., Janssen, H., Bahl, H., Fischer, R., and Wolkenhauer, O. (2013a). Integrative modelling of ph-dependent enzyme activity and transcriptomic regulation of the acetonebutanolethanol fermentation of clostridium acetobutylicum in continuous culture. *Microbial Biotechnology*, 6(5):526–39. [cited at p. 81]

[Millat et al., 2013b] Millat, T., Janssen, H., Thorn, G., King, J., Bahl, H., Fischer, R., and Wolkenhauer, O. (2013b). A shift in the dominant phenotype governs the ph-induced metabolic switch of clostridium acetobutylicum in phosphate-limited continuous cultures. *Applied Microbiology and Biotechnology*, 97(14):6451–66. [cited at p. 14]

[Monot et al., 1984] Monot, F., Engasser, J., and Petitdemange, H. (1984). Influence of ph and undissociated butyric acid on the production of acetone and butanol in batch cultures of clostridium acetobutylicum. *Applied Microbiology and Biotechnology*, 19(6):422–6. [cited at p. 13]

[Monot et al., 1982] Monot, F., Martin, J., Petitdemange, H., and Gay, R. (1982). Acetone and butanol production by clostridium acetobutylicum in a synthetic medium. *Applied and Environmental Microbiology*, 44(6):1318–24. [cited at p. 16]

[Nacher et al., 2005] Nacher, J., Yamada, T., Goto, S., Kanehisa, M., and Akutsu, T. (2005). Two complementary representations of a scale-free network. *Physica A: Statistical Mechanics and its Applications*, 349(1-2):349 – 63. [cited at p. 30]

[Nair et al., 1999] Nair, R., Green, E., Watson, D., Bennett, G., and Papoutsakis, E. (1999). Regulation of the sol locus genes for butanol and acetone formation in clostridium acetobutylicumatcc 824 by a putative transcriptional repressor. *Journal of Bacteriology*, 181(1):319–30. [cited at p. 14]

[Nielsen et al., 2009] Nielsen, D., Leonard, E., Yoon, S., Tseng, H., Yuan, C., and Prather, K. (2009). Engineering alternative butanol production platforms in heterologous bacteria. *Metabolic Engineering*, 11(4-5):262 – 73. [cited at p. 15]

[Nobeli and Thornton, 2006] Nobeli, I. and Thornton, J. (2006). A bioinformatician's view of the metabolome. *BioEssays*, 28(5):534–45. [cited at p. 129]

[Noelling et al., 2001] Noelling, J., Breton, G., Omelchenko, M., Makarova, K., Zeng, Q., Gibson, R., Lee, H., Dubois, J., Qiu, D., Hitti, J., Wolf, Y., Tatusov, R., Sabathe, F., Doucette-Stamm, L., Soucaille, P., Daly, M., Bennett, G., Koonin, E., and Smith, D. (2001). Genome sequence and comparative analysis of the solvent-producing bacterium clostridium acetobutylicum. *Journal of Bacteriology*, 183(16):4823–4838. [cited at p. 10]

[Nolan et al., 2006] Nolan, T., Hands, R., and Bustin, S. (2006). Quantification of mrna using real-time rt-pcr. *Nature protocols*, 1(3):1559–82. [cited at p. 21]

[Oberhardt et al., 2009] Oberhardt, M., Palsson, B., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5(1). [cited at p. 77]

[Ofran et al., 2005] Ofran, Y., Punta, M., Schneider, R., and Rost, B. (2005). Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today*, 10(21):1475 – 82. [cited at p. 56]

[Özilgen, 1988] Özilgen, M. (1988). Kinetics of multiproduct acidogenic and solventogenic batch fermentations. *Applied Microbiology and Biotechnology*, 29(6):536–43. [cited at p. 80]

[Pagel et al., 2013] Pagel, H., Ingwersen, J., Poll, C., Kandeler, E., and Streck, T. (2013). Micro-scale modeling of pesticide degradation coupled to carbon turnover in the detritusphere: model description and sensitivity analysis. *Biogeochemistry*, pages 1–20. [cited at p. 106]

[Palsson and Zengler, 2010] Palsson, B. and Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature chemical biology*, 6(11):787. [cited at p. 3]

[Papoutsakis et al., 1987] Papoutsakis, E., Bussineau, C., Chu, I., Diwan, R., and Huesemann, M. (1987). Transport of substrates and metabolites and their effect on cell metabolism (in butyric-acid and methylotrophic fermentations)a. *Annals of the New York Academy of Sciences*, 506(1):24–50. [cited at p. 13, 65, 66]

[Papoutsakis, 1984] Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and Bioengineering*, 26(2):174–87. [cited at p. 80, 116]

[Paredes et al., 2005] Paredes, C. J., Alsaker, K. V., and Papoutsakis, E. T. (2005). A comparative genomic view of clostridial sporulation and physiology. *Nature Reviews Microbiology*, 3(12):969–78. [cited at p. 13]

[Patil and Nielsen, 2005] Patil, K. and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685–9. [cited at p. 36, 78]

[Pattengale et al., 2010] Pattengale, N., Alipour, M., Bininda-Emonds, O., Moret, B., and Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology*, 17(3):337–54. [cited at p. 123]

[Peguin et al., 1994] Peguin, S., Goma, G., Delorme, P., and Soucaille, P. (1994). Metabolic flexibility of clostridium acetobutylicum in response to methyl viologen addition. *Applied Microbiology and Biotechnology*, 42(4):611–6. [cited at p. 13]

[Peguin and Soucaille, 1995] Peguin, S. and Soucaille, P. (1995). Modulation of Carbon and Electron Flow in Clostridium acetobutylicum by Iron Limitation and Methyl Viologen Addition. *Appl. Environ. Microbiol.*, 61(1):403–5. [cited at p. 16]

[Pellegrini et al., 1999] Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–8. [cited at p. 64]

[Pfromm et al., 2010] Pfromm, P., Amanor-Boadu, V., Nelson, R., Vadlani, P., and Madl, R. (2010). Bio-butanol vs. bio-ethanol: A technical and economic assessment for corn and switchgrass fermented by yeast or clostridium acetobutylicum. *Biomass and Bioenergy*, 34(4):515 – 24. [cited at p. 7]

[Punta et al., 2012] Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S., Bateman, A., and Finn, R. (2012). The pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301. [cited at p. 56]

[Qureshi et al., 2006] Qureshi, N., Li, X., Hughes, S., Saha, B., and Cotta, M. (2006). Butanol production from corn fiber xylan using clostridium acetobutylicum. *Biotechnology Progress*, 22(3):673–80. [cited at p. 17]

[Qureshi and Maddox, 2005] Qureshi, N. and Maddox, I. (2005). Reduction in butanol inhibition by perstraction: Utilization of concentrated lactose/whey permeate by clostridium acetobutylicum to enhance butanol fermentation economics. *Food and Bioproducts Processing*, 83(1):43 – 52. [cited at p. 18]

[Rabitz et al., 1983] Rabitz, H., Kramer, M., and Dacol, D. (1983). Sensitivity analysis in chemical kinetics. *Annual Review of Physical Chemistry*, 34(1):419–461. [cited at p. 106]

[Rao and Mutharasan, 1987] Rao, G. and Mutharasan, R. (1987). Altered electron flow in continuous cultures of clostridium acetobutylicum induced by viologen dyes. *Applied and Environmental Microbiology*, 53(6):1232–35. [cited at p. 13]

[Ravagnani et al., 2000] Ravagnani, A., Jennert, K., Steiner, E., Gruenberg, R., Jefferies, J., Wilkinson, S., Young, D., Tidswell, E., Brown, D., Youngman, P., Morris, J., and Young, M. (2000). Spo0a directly controls the switch from acid to solvent production in solvent-forming clostridia. *Molecular Microbiology*, 37(5):1172–85. [cited at p. 14]

[Reardon and Bailey, 1989] Reardon, K. and Bailey, J. (1989). Effects of ph and added metabolites on bioconversions by immobilized non-growing clostridium acetobutylicum. *Biotechnology and Bioengineering*, 34(6):825–37. [cited at p. 18]

[Reardon and Bailey, 1992] Reardon, K. and Bailey, J. (1992). Activity regeneration in continuous clostridium acetobutylicum bioconversions of glucose. *Biotechnology Progress*, 8(4):316–26. [cited at p. 18]

[Reed et al., 2006] Reed, J., Famili, I., Thiele, I., and Palsson, B. (2006). Towards multidimensional genome annotation. *Nature Reviews*, 7:130–41. [cited at p. 34, 37, 38]

[Reuter, 2011] Reuter, A. (2011). Datengetriebene forschung - herausforderung fr die informatik. *Spektrum der Wissenschaft - Extra*, pages 6–9. [cited at p. 77]

[Rogers, 2002] Rogers, P. (2002). *Clostridia, Solvent Formation*. John Wiley & Sons, Inc. [cited at p. 10]

[Rost et al., 2003] Rost, B., Liu, J., Nair, R., Wrzeszczynski, K., and Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS*, 60(12):2637–50. [cited at p. 64, 74]

[Saltelli et al., 2000] Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Statistical Science*, pages 377–95. [cited at p. 106]

[Saltelli et al., 1999] Saltelli, A., Tarantola, S., and Chan, K. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56. [cited at p. 110]

[Schaedel and Franco-Lara, 2009] Schaedel, F. and Franco-Lara, E. (2009). Rapid sampling devices for metabolic engineering applications. *Applied Microbiology and Biotechnology*, 83(2):199–208. [cited at p. 20]

[Schaffer et al., 2002] Schaffer, S., Isci, N., Zickner, B., and Duerre, P. (2002). Changes in protein synthesis and identification of proteins specifically induced during solventogenesis in clostridium acetobutylicum. *Electrophoresis*, 23(1):110–21. [cited at p. 21]

[Schaub, 2005] Schaub, J. (2005). *Isotopisch instationäre 13C-Stoffflussanalyse in Escherichia coli*. PhD thesis, Universität Stuttgart, Germany. [cited at p. 20, 104]

[Schaub et al., 2006] Schaub, J., Schiesling, C., Reuss, M., and Dauner, M. (2006). Integrated sampling procedure for metabolome analysis. *Biotechnology Progress*, 22(5):1434–42. [cited at p. 20]

[Schmidt and Jirstrand, 2006] Schmidt, H. and Jirstrand, M. (2006). Systems biology toolbox for matlab: a computational platform for research in systems biology. *Bioinformatics*, 22(4):514–5. http://www.sbtoolbox2.org. [cited at p. 90]

[Schwarz et al., 2007a] Schwarz, K., Fiedler, T., Fischer, R., and Bahl, H. (2007a). A standard operating procedure (sop) for the preparation of intra- and extracellular proteins of clostridium acetobutylicum for proteome analysis. *Journal of Microbiological Methods*, 68(2):396 – 402. [cited at p. 21]

[Schwarz et al., 2012] Schwarz, K., Kuit, W., Grimmler, C., Ehrenreich, A., and Kengen, S. (2012). A transcriptional study of acidogenic chemostat cells of clostridium acetobutylicum - cellular behavior in adaptation to n-butanol. *Journal of Biotechnology*, 161(3):366 – 77. [cited at p. 17, 21]

[Schwarz et al., 2007b] Schwarz, R., Liang, C., Kaleta, C., Kuhnel, M., Hoffmann, E., Kuznetsov, S., Hecker, M., Griffiths, G., Schuster, S., and Dandekar, T. (2007b). Integrated network reconstruction, visualization and analysis using yanasquare. *BMC Bioinformatics*, 8(1):313. [cited at p. 29, 117]

[SE, 2008] SE, B. (2008). Technisches merkblatt. Internet. M 2084 d. [cited at p. 6]

[Secrier and Schneider, 2013] Secrier, M. and Schneider, R. (2013). Visualizing time-related data in biology, a review. *Briefings in Bioinformatics*. [cited at p. 55]

[Senger and Papoutsakis, 2008] Senger, R. and Papoutsakis, E. (2008). Genome-scale model for clostridium acetobutylicum: Part i metabolic network resolution and analysis. *Biotechnology and Bioengineering*, 101(5):1036–52. [cited at p. 28, 36, 80, 117]

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504. [cited at p. 3, 31]

[Shinto et al., 2007] Shinto, H., Tashiro, Y., Yamashita, M., Kobayashi, G., Sekiguchi, T., Hanai, T., Kuriya, Y., Okamoto, M., and Sonomoto, K. (2007). Kinetic modeling and sensitivity analysis of acetone-butanol-ethanol production. *Journal of Biotechnology*, 131(1):45 – 56. [cited at p. 81, 88]

[Siegel and Himmele, 1980] Siegel, H. and Himmele, W. (1980). Synthesis of intermediates by rhodium-catalyzed hydroformylation. *Angewandte Chemie International Edition in English*, 19(3):178–83. [cited at p. 6]

[Sillers et al., 2009] Sillers, R., Al-Hinai, M., and Papoutsakis, E. (2009). Aldehyde-alcohol dehydrogenase and/or thiolase overexpression coupled with coa transferase downregulation lead to higher alcohol titers and selectivity in clostridium acetobutylicum fermentations. *Biotechnology and Bioengineering*, 102(1):38–49. [cited at p. 14, 78, 102, 116]

[Sillers et al., 2008] Sillers, R., Chow, A., Tracy, B., and Papoutsakis, E. (2008). Metabolic engineering of the non-sporulating, non-solventogenic clostridium acetobutylicum strain m5 to produce butanol without acetone demonstrate the robustness of the acid-formation pathways and the importance of the electron balance. *Metabolic Engineering*, 10(6):321 – 32. [cited at p. 13, 15]

[Snoep and Olivier, 2002] Snoep, J. and Olivier, B. (2002). Java web simulation (jws); a web based database of kinetic models. *Molecular Biology Reports*, 29:259–63. [cited at p. 3]

[Sobol, 2001] Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271 – 80. [cited at p. 110]

[Srere, 1987] Srere, P. A. (1987). Complexes of sequential metabolic enzymes. *Annual Review of Biochemistry*, 56(1):89–124. [cited at p. 64]

[Srivastava and Volesky, 1991] Srivastava, A. and Volesky, B. (1991). Measurement and regulation of the culture reduction state in clostridium acetobutylicum. *Biotechnology and Bioengineering*, 38(2):181–90. [cited at p. 20]

[Stelling et al., 2002] Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Letters to Nature*, 420:190–3. [cited at p. 32]

[Sturn et al., 2002] Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–8. [cited at p. 68]

[Sumner et al., 2012] Sumner, T., Shephard, E., and Bogle, I. (2012). A methodology for global-sensitivity analysis of time-dependent outputs in systems biology modelling. *Journal of The Royal Society Interface*, 9(74):2156–66. [cited at p. 120, 135]

[Tanya Rezler, 2012] Tanya Rezler, M. L. (2012). Global n-butanol market to reach 4 mln tonnes by 2020. Internet. [cited at p. 6]

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–5. [cited at p. 64]

[Terracciano and Kashket, 1986] Terracciano, J. and Kashket, E. (1986). Intracellular conditions required for initiation of solvent production by clostridium acetobutylicum. *Applied and Environmental Microbiology*, 52(1):86–91. [cited at p. 13]

[Terracciano et al., 1988] Terracciano, J., Rapaport, E., and Kashket, E. (1988). Stress- and growth phase-associated proteins of clostridium acetobutylicum. *Applied and Environmental Microbiology*, 54(8):1989–95. [cited at p. 21]

[Theocharidis et al., 2009] Theocharidis, A., Dongen, S., Enright, A., and Freeman, T. (2009). Network visualization and analysis of gene expression data using biolayout express 3d. *Nature Protocols*, 4:1535–50. [cited at p. 31]

[Thormann et al., 2002] Thormann, K., Feustel, L., Lorenz, K., Nakotte, S., and Duerre, P. (2002). Control of butanol formation in clostridium acetobutylicum by transcriptional activation. *Journal of Bacteriology*, 184(7):1966–73. [cited at p. 14]

[Tomas et al., 2004] Tomas, C., Beamish, J., and Papoutsakis, E. (2004). Transcriptional analysis of butanol stress and tolerance in clostridium acetobutylicum. *Journal of Bacteriology*, 186(7):2006–18. [cited at p. 17]

[Tomas et al., 2003] Tomas, C., Welker, N., and Papoutsakis, E. (2003). Overexpression of groesl in clostridium acetobutylicum results in increased solvent production and tolerance, prolonged metabolism, and changes in the cell's transcriptional program. *Applied and Environmental Microbiology*, 69(8):4951–65. [cited at p. 20]

[Troyanskaya, 2005] Troyanskaya, O. (2005). Putting microarrays in a context: Integrated analysis of diverse biological data. *Briefings in Bioinformatics*, 6(1):34–43. [cited at p. 31, 38]

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–5. [cited at p. 120]

[Tummala et al., 2003a] Tummala, S., Junne, S., and Papoutsakis, E. (2003a). Antisense rna downregulation of coenzyme a transferase combined with alcohol-aldehyde dehydrogenase overexpression leads to predominantly alcohologenic clostridium acetobutylicum fermentations. *Journal of Bacteriology*, 185(12):3644–53. [cited at p. 14, 20, 137]

[Tummala et al., 2003b] Tummala, S., Junne, S., Paredes, C., and Papoutsakis, E. (2003b). Transcriptional analysis of product-concentration driven changes in cellular programs of recombinant clostridium acetobutylicumstrains. *Biotechnology and Bioengineering*, 84(7):842–54. [cited at p. 21, 102, 117]

[Turk and Pentland, 1991] Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–91. IEEE. [cited at p. 135]

[van Dongen, 2000] van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, Netherlands. [cited at p. 31]

[van Riel, 2006] van Riel, N. A. (2006). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings in Bioinformatics*, 7(4):364–374. [cited at p. 77, 106, 117, 121]

[Vasconcelos et al., 1994] Vasconcelos, I., Girbal, L., and Soucaille, P. (1994). Regulation of carbon and electron flow in clostridium acetobutylicum grown in chemostat culture at neutral ph on mixtures of glucose and glycerol. *Journal of Bacteriology*, 176(5):1443–50. [cited at p. 95]

[Vaz and Vicente, 2007] Vaz, A. and Vicente, L. (2007). A particle swarm pattern search method for bound constrained global optimization. *Journal of Global Optimization*, 39:197–219. 10.1007/s10898-007-9133-5. [cited at p. 94]

[Vijesh et al., 2013] Vijesh, N., Chakrabarti, S., and Sreekumar, J. (2013). Modeling of gene regulatory networks: A review. *Journal of Biomedical Science and Engineering*, 6:223 – 31. [cited at p. 117]

[Votruba et al., 1986] Votruba, J., Volesky, B., and Yerushalmi, L. (1986). Mathematical model of a batch acetone-butanol fermentation. *Biotechnology and Bioengineering*, 28(2):247–55. [cited at p. 80]

[W and BM, 2007] W, B. and BM, P. (2007). Personalized medicine in the era of genomics. *JAMA*, 298(14):1682–1684. [cited at p. 78]

[Waltemath et al., 2011] Waltemath, D., Adams, R., Bergmann, F., Hucka, M., Kolpakov, F., Miller, A., Moraru, I., Nickerson, D., Sahle, S., Snoep, J., and Le Novere, N. (2011). Reproducible computational biology experiments with sed-ml - the simulation experiment description markup language. *BMC Systems Biology*, 5(1):198. [cited at p. 3, 118]

[Wang et al., 2011] Wang, S., Zhang, Y., Dong, H., Mao, S., Zhu, Y., Wang, R., Luan, G., and Li, Y. (2011). Formic acid triggers the "acid crash" of acetone-butanol-ethanol fermentation by clostridium acetobutylicum. *Applied and Environmental Microbiology*, 77(5):1674–80. [cited at p. 13]

[Wiesenborn et al., 1989] Wiesenborn, D., Rudolph, F., and Papoutsakis, E. (1989). Phosphotransbutyrylase from clostridium acetobutylicum atcc 824 and its role in acidogenesis. *Applied and Environmental Microbiology*, 55(2):317–22. [cited at p. 12]

[Wrzodek et al., 2013] Wrzodek, C., Buchel, F., Ruff, M., Drager, A., and Zell, A. (2013). Precise generation of systems biology models from kegg pathways. *BMC Systems Biology*, 7(1):15. [cited at p. 29]

[Yang et al., 2003] Yang, H., Haddad, H., Tomas, C., Alsaker, K., and Papoutsakis, T. (2003). A segmental nearest neighbor normalization and gene identification method gives superior results for dna-array analysis. *Proceedings of the National Academy of Sciences*, 100(3):1122–7. [cited at p. 37, 53]

[Yang et al., 2005] Yang, Y., Xiao, Y., and Segal, M. (2005). Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21(7):1084–93. [cited at p. 78]

[Yerushalmi et al., 1983] Yerushalmi, L., Volesky, B., Leung, W., and Neufeld, R. (1983). Variations of solvent yield in acetone-butanol fermentation. *European journal of applied microbiology and biotechnology*, 18(5):279–86. [cited at p. 80]

[Yerushalmi et al., 1986a] Yerushalmi, L., Volesky, B., and Votruba, J. (1986a). Modelling of culture kinetics and physiology for c. acetobutylicum. *The Canadian Journal of Chemical Engineering*, 64(4):607–16. [cited at p. 80]

[Yerushalmi et al., 1986b] Yerushalmi, L., Volesky, B., and Votruba, J. (1986b). Systems analysis of the culture physiology in acetone-butanol fermentation. *Biotechnology and Bioengineering*, 28(9):1334–47. [cited at p. 11, 80]

[Yerushalmi et al., 1988] Yerushalmi, L., Volesky, B., and Votruba, J. (1988). Fermentation process diagnosis using a mathematical model. *Applied Microbiology and Biotechnology*, 29(2-3):186–97. [cited at p. 80]

[Yeung et al., 2001] Yeung, K., Haynor, D., and Ruzzo, W. (2001). Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–18. [cited at p. 71]

[Yeung and Ruzzo, 2001] Yeung, K. and Ruzzo, W. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–74. [cited at p. 120, 130, 133, 136]

[Yeung et al., 2002] Yeung, M., Tegnér, J., and Collins, J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–8. [cited at p. 120]

[You and Yin, 2000] You, L. and Yin, J. (2000). Patterns of regulation from mrna and protein time series. *Metabolic engineering*, 2(3):210–7. [cited at p. 129]

[Zhao et al., 2005] Zhao, Y., Tomas, C. A., Rudolph, F. B., Papoutsakis, E., and Bennett, G. N. (2005). Intracellular butyryl phosphate and acetyl phosphate concentrations in clostridium acetobutylicum and their implications for solvent formation. *Applied and Environmental Microbiology*, 71(1):530–7. [cited at p. 13, 95]

[Zheng and Rundell, 2006] Zheng, Y. and Rundell, A. (2006). Comparative study of parameter sensitivity analyses of the tcr-activated erk-mapk signalling pathway. *IEE Proceedings-Systems Biology*, 153(4):201–11. [cited at p. 110]

[Zi et al., 2005] Zi, Z., Cho, K., Sung, M., Xia, X., Zheng, J., and Sun, Z. (2005). In silico identification of the key components and steps in ifn-γ induced jak-stat signaling pathway. *FEBS Letters*, 579(5):1101 – 8. [cited at p. 106]

Appendices

Appendix A

Dynamic Model Equations

$$\begin{aligned}
\frac{dc^e_{\mathrm{Glc}}}{dt} &= -r_{0|1} \\
\frac{dc^i_{\mathrm{ACoA}}}{dt} &= \frac{\rho}{c_X}(r_{0|1}Y_{p,Glc} - r_{1|11} + r_{11|1} + r_{2,3|1,7} - r_{1|3} - r_{1|9}) - \mu \cdot c^i_{\mathrm{ACoA}} \\
\frac{dc^e_{\mathrm{ACE}}}{dt} &= r_{11|2} - r_{2|11} - r_{2,3|1,7} - D \cdot c^e_{\mathrm{ACE}} \\
\frac{dc^i_{\mathrm{AACoA}}}{dt} &= \frac{\rho}{c_X}(0.5r_{1|3} - r_{3|7} - r_{3|4}) - \mu \cdot c^i_{\mathrm{AACoA}} \\
\frac{dc^i_{\mathrm{BCoA}}}{dt} &= \frac{\rho}{c_X}(r_{3|4} + r_{6,3|4,7} - r_{4|5} + r_{5|4} - r_{4|10}) - \mu \cdot c^i_{\mathrm{BCoA}} \\
\frac{dc^i_{\mathrm{BUP}}}{dt} &= \frac{\rho}{c_X}(-r_{5|4} + r_{4|5} + r_{6|5} - r_{5|6}) - \mu \cdot c^i_{\mathrm{BUP}} \\
\frac{dc^e_{\mathrm{BU}}}{dt} &= r_{5|6} - r_{6|5} - r_{6,3|4,7} - D \cdot c^e_{\mathrm{BU}} \\
\frac{dc^i_{\mathrm{AA}}}{dt} &= \frac{\rho}{c_X}(r_{3|7} - r_{7|8}) - \mu \cdot c^i_{\mathrm{AA}} \\
\frac{dc^e_{\mathrm{ACN}}}{dt} &= r_{7|8} - D \cdot c^e_{\mathrm{ACN}} \\
\frac{dc^e_{\mathrm{ETOH}}}{dt} &= r_{1|9} - D \cdot c^e_{\mathrm{ETOH}} \\
\frac{dc^e_{\mathrm{BUOH}}}{dt} &= r_{4|10} - D \cdot c^e_{\mathrm{BUOH}} \\
\frac{dc^i_{\mathrm{ACP}}}{dt} &= \frac{\rho}{c_X}(r_{1|11} - r_{11|1}) - \mu \cdot c^i_{\mathrm{ACP}}
\end{aligned}$$

$$
\begin{aligned}
r_{1|11} &= f(\mathrm{C}_{1742}) \cdot c_X \cdot k_{1|11} \frac{c^i_{\mathrm{ACoA}}}{\mathrm{Km}_{1|11} + c^i_{\mathrm{ACoA}}} \\
r_{11|1} &= f(\mathrm{C}_{1742}) \cdot c_X \cdot k_{11|1} \frac{c^i_{\mathrm{ACP}}}{\mathrm{Km}_{1|11} + c^i_{\mathrm{ACP}}} \\
r_{11|2} &= f(\mathrm{C}_{1743}) \cdot c_X \cdot k_{11|2} \frac{c^i_{\mathrm{ACP}}}{\mathrm{Km}_{11|2} + c^i_{\mathrm{ACP}}} \\
r_{2|11} &= f(\mathrm{C}_{1743}) \cdot c_X \cdot k_{2|11} \frac{c^e_{\mathrm{ACE}}}{\mathrm{Km}_{2|11} + c^e_{\mathrm{ACE}}} \\
r_{2,3|1,7} &= f(\mathrm{P}_{0163,0164}) \cdot c_X \cdot k_{2,3|1,7} \frac{c^e_{\mathrm{ACE}}}{\mathrm{Km}_{2,3|1,7} + c^e_{\mathrm{ACE}}} \frac{c^i_{\mathrm{AACoA}}}{\mathrm{Km}_{2,3|1,7} + c^i_{\mathrm{AACoA}}} \\
r_{1|3} &= f(\mathrm{C}_{2873}) \cdot c_X \cdot k_{1|3} \frac{{c^i_{\mathrm{ACoA}}}^2}{{\mathrm{Km}_{1|3}}^2 + \mathrm{Kn}_{1|3} c^i_{\mathrm{ACoA}} + (c^i_{\mathrm{ACoA}})^2} \\
r_{3|4} &= f(\mathrm{C}_{2708,2711,2712}) \cdot c_X \cdot k_{3|4} \frac{c^i_{\mathrm{AACoA}}}{\mathrm{Km}_{3|4} + c^i_{\mathrm{AACoA}} + \frac{(c^i_{\mathrm{AACoA}})^2}{\mathrm{Ki}_{3|4}}} \\
r_{4|5} &= f(\mathrm{C}_{3076}) \cdot c_X \cdot k_{4|5} \frac{c^i_{\mathrm{BCoA}}}{\mathrm{Km}_{4|5} + c^i_{\mathrm{BCoA}} + \frac{c^i_{\mathrm{BUP}}}{\mathrm{Ki}_{4|5}}} \\
r_{5|4} &= f(\mathrm{C}_{3076}) \cdot c_X \cdot k_{5|4} \frac{c^i_{\mathrm{BUP}}}{\mathrm{Km}_{5|4} + c^i_{\mathrm{BUP}}} \\
r_{5|6} &= f(\mathrm{C}_{3075}) \cdot c_X \cdot k_{5|6} \frac{c^i_{\mathrm{BUP}}}{\mathrm{Km}_{5|6} + c^i_{\mathrm{BUP}}} \\
r_{6|5} &= f(\mathrm{C}_{3075}) \cdot c_X \cdot k_{6|5} \frac{c^e_{BU}}{\mathrm{Km}_{6|5} + c^e_{\mathrm{BU}}} \\
r_{7|8} &= f(\mathrm{P}_{0165}) \cdot c_X \cdot k_{7|8} \frac{c^i_{\mathrm{AA}}}{\mathrm{Km}_{7|8} + c^i_{\mathrm{AA}}} \\
r_{1|9} &= f(\mathrm{C}_{3298,3299}, \mathrm{P}_{0162}) \cdot c_X \cdot k_{1|9} \frac{c^i_{\mathrm{ACoA}}}{\mathrm{Km}_{1|9} + c^i_{\mathrm{ACoA}}} \\
r_{4|10} &= f(\mathrm{C}_{3298,3299}, \mathrm{P}_{0162}) \cdot c_X \cdot k_{4|10} \frac{c^i_{\mathrm{BCoA}}}{\mathrm{Km}_{4|10} + c^i_{\mathrm{BCoA}}} \frac{\mathrm{Ki}_{10|4,10}}{\mathrm{Ki}_{10|4,10} + c^e_{\mathrm{BUOH}}} \\
r_{3|10} &= r_{2,3|1,7} + r_{6,3|4,7} \\
r_{6,3|4,7} &= \frac{f(\mathrm{CA}_{P0163,P0164}) \cdot c_X \cdot kc^e_{\mathrm{BU}} \cdot c^i_{\mathrm{AACoA}}}{\mathrm{Km} + c^i_{\mathrm{AACoA}} \cdot c^e_{\mathrm{BU}} + \mathrm{KnA} c^i_{\mathrm{AACoA}} \cdot (1 + \frac{c^i_{\mathrm{AACoA}}}{\mathrm{KiA}}) + \mathrm{KnB} c^e_{\mathrm{BU}} \cdot (1 + \frac{c^e_{\mathrm{BU}}}{\mathrm{KiB}})}
\end{aligned}
$$

Appendix B

Scripts

B.1 Taverna Workflows

In this section the design and operation of work-flows with Taverna [Hull et al., 2006] is introduced.
Taverna is a free work-flow management system based on Java. It offers access to various web-services by third-parties through the Web Service Description Language (WSDL). Such parties are e.g. the European-Bioinformatics Institute from the European Molecular Biology Laboratory (EMBL-EBI), the National Center for Biotechnology Information (NCBI), Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioMart. Furthermore, local services, like Beanshell scripts, Java API, R scripts and Excel interaction, support the processing of automated database queries. A list of accessible services is given on biocatalogue.

Taverna Workflow Services

Access to web services and scripts is granted via Simple Object Access Protocols (SOAPs) [**?**] or Representational State Transfer (REST) services [**?**]. Both consist of a defined number of input ports, the processing of inputs and the output ports for the query results. A sequence of services is connected by linking the output and input ports and, if necessary, by introducing formatting services.

Obtaining The Reactome

Since the recent change of accessibility of the KEGG-API, the here treated work-flows needed entire re-structuring. Luckily, things got easier:
Download of the organism-specific maps from genes to enzymes, to reactions, to compounds are possible from the respective sites through copy paste. From these three lists the unique reaction identifiers serve as input for the *reaction-*

pair_mapping workflow. It extracts the reactions with a REST-service first, then it extracts the different RPAIR-identifiers for each reaction-identifier. Similar to this workflow, annotation of compounds and genes is performed (*annotation_genes*, *annotation_compounds*).

Phylogenetic Comparison

Given a specific enzyme number, the gene-identifiers for all annotated organism are retrieved in KEGG, this is done by the *enzyme_in_organism*-workflow. Due to formatting issues, the list of genes from this workflow is reorganised by a MATLAB-script, *enzymelistconversion.m*, before Cytoscape can be used.

B.2 Static Model Scripts

B.2.1 Model Creation

The *determineboundaryparameter.m* script serves for the data generation to determine the boundary parameter b that is used for integration of data to the reactome graph. It requires five arguments at maximum:

- the reactome database from KEGG, with reaction-IDs in the first, gene-IDs in the second, and reaction pairs (cpd:_cpd:) in the third column,
- a data matrix, with genes in the rows and experimental conditions in the columns,
- an optional string argument H-graph to evaluate the boundary parameter for the augmentation, otherwise the G-graph is calculated,
- an optional column-number which data column corresponds to state s_1, otherwise it is the first,
- an optional column-number which data column corresponds to state s_2, otherwise it is the last column.

The output is of this script a three dimensional data-cube, the two variable dimensions are values of b and time of the data. From these the number of active genes, the number of active reactions, the number of active metabolites, and the edges to nodes fraction are calculated.
figure_show_determination_b.m then provides a visualisation of the data-cube.

CreateFilteredGraphs.m

Integration of data into the reactome database is done by this script, it requires seven input parameters:

1. the output directory,
2. a data matrix as before,
3. the reactome database as before,
4. the column-number for the first state,
5. the column-number for the second state,
6. boundary parameter b,
7. a string, assessing how to augment:

- *both*, both states serve for augmentation
- *simple*, no augmentation
- *after*, the first state only serves for augmentation
- *before*, the second state only serves for augmentation.

The output creates several nodelists that are converted into graphs by Cytoscape: the graph in states s_1, s_2, the graph for all reaction occurring in neither state, and the graphs for all reactions activated or inactivated in both states.

B.2.2 Creation of a Comparison Database

Creation of the comparison database from chapter 3.4.1 requires the following lists from two organisms: a list of genes, the annotation databases as downloaded from taverna, and the mappings of genes to reactions. From these three, the Cytoscape maps are constructed. An example file is given in *bsu_cac_comp.m*.

B.2.3 Creation of 3-HBDH Candidates

This analysis requires one approach to compare pfam-motifs: A script to generate a map of pfam-motifs from the gene-annotation (*CreateGeneMotifMAP.m*). This map is suited for import into Cytoscape. A script counts the occurrences of genes and motifs in the map (*CountOccurrences.m*). Selection of the most frequent motifs or genes is carried out by *FindFrequentMotifs.m.* For the clustering it is necessary to find genes in the same clusters over different data (*IdentifySameClusterGenes.m*) and a method to create subsets of the reactome in terms of the candidate genes and their pfam-motifs (*SelectFromGeneMotifMap.m*). Finally the conversion into matrix-format allows saving the results (*SaveMapAsMatrix*).
Inputs for these methods are simple lists of genes that are derived from the analyses in other programmes.
The example script *Search_3HBDH.m* shows the automated candidate generation with the help of the beforehand mentioned scripts.

B.3 Dynamic Model Scripts

Model creation requires two computational steps, the conversion of the Reactome database into the SBTOOLBOX2-format (*ConvertKEGG2SBToolbox2.m*) and the supplying of the model with data (*ConvertStdModel2SBT.m*).

B.3.1 Converting The KEGG-Database Into The Standard Format

For conversion from KEGG into the standard format in the *KEGG2SBTOOLBOX2*-script three data-inputs are required:

- the local database as discussed earlier that contains the mapping of transcripts to reactions
- the model file that contains all desired reactions and compounds, this can be achieved either manually or by using Cytoscape.
- a list of extracellular compounds

The deposited script allows the computation of a complete standard model with all transcripts at place, Michaelis-Menten type kinetics for multiple substrates, if necessary and standardised parameters for the reactions, following the same scheme as for reaction-identifiers. It is not possible to manually create reactions that are not present in the database, if such a reaction is desired, it has to be in both files, the database and the model.

B.3.2 Integrating Data Into The Standard Format

Integration is performed by the *ConvertStdModel2SBT*-script. It requires four files

1. the standard-model file,
2. the transcriptome level timeseries data,
3. the glucose consumption profile,
4. the optical density profile

and a parameter that controls the integration of transcriptome level data that either is implemented directly or via simplified via PCA reduction (5.1).

B.3.3 Model Simulation

Parameter estimation of the models is readily done by SBPD and the SBPD file structure. For model comparison, several data were fed into the constructed models. For model validation, several parameter variations were performed.

B.3.4 Sensitivity Analysis

Local Sensitivity Analysis

The local sensitivity analyses requires the symbolic computing package from MATLAB. The *DeriveModel.m* file takes a SBTOOLBOX-model and calculates the derivatives of the maximal velocities, integrates them back into the model to allow simulation.

Global Sensitivity Analysis

The global sensitivity analysis is readily available through SBTOOLBOX-scripts. In order to allow temporally resolved indices, a calculation script (*MySensitivityAnalysis2.m*) was wrapped around this script to allow the analysis of time intervals and the SBTOOLBOX-scripts (*SBsensglobalfast.m*, *rel_sensglobaldefaultobjectiveSB.m*) were adapted. This adaption includes the calculation of the FAST-alternations for the whole time interval once. Then the calculation is split into intervals and the sensitivity indices calculated per interval.

B.4 Principal Component Analysis

B.4.1 Dynamic Features

The dynamic features extraction and calculation is performed by the *dynamic_features.m*-script. It is used to change three dynamical features by adding 150% or substracting 50% of the original parameter

B.4.2 Clustering

master_script.m

This is the main calling script of the clustering algorithm, all parameters are supplied in this module, the *input* and *output* directories, the *data input* and the *genes input* (in case not the whole data set is to be used). Then the data filter options, *nonan*, *nT*, how many missing values are admissible in a single transcript expression profile and how many time points are to be taken. Then the PCA-parameters, *ncoeff*, *ssec* and *lsec*, how many components are to be considered and how the PC-space $\mathcal{C}$ should be partitioned. As these two numbers are the numbers of partition in one half-space, they are dissected into $2^{\mathrm{ssec}} \leq 2^{\mathrm{lsec}}$ parts. Finally, the *percentage* of relevant data used to do the PCA is another input. Filtering will be applied on two levels, the maximal span of the data and the information content. Last but not least, the *name* of the annotation datafile is supplied.

myGetGenData.m

This file imports the data matrix, with gene-identifiers in the rows, and the temporal dimensions in the columns. A struct is handled back, it corresponds to the specific filtering options with maximal *nonan* missing values of length *nT*.

mySaveGenData.m

This routine saves the struct obtained from myGetGenData.m for the imputation or from the file obtained from the imputation back into a csv-file.

Imputation

Imputation is performed based on a web-service offered by the MPI-Potsdam called MetaGeneAlyse [Daub et al., 2003]. The output-file from mySaveGenData can be readily uploaded after manual replacement of occurring ”NaN” into ”NA”. This website offers the opportunity to impute missing points according to three different analyses based on Principal component analysis.

MyDataReduction.m

For meaningful data reduction we use a percentile approach. The lowest dynamics and the highest entropy content are filtered out from the respective distributions at a level of *perc*. These scripts are achieved using the Statistics Toolbox by MATLAB. Filtered out genes are written out in individual files.

myPCA2.m

This script is the core module where the evaluation of the data's properties happen.
It starts with the mapping of transcript expression profiles into the principal component space of size *ncoeff*. For small *ncoeff*, one can generate a typographic visualisation of different combinations of principal components, facilitating the overview, what type of dynamic behaviours can be achieved.
In the next script, *mycalculus.m*, the individual transcript expression profile coefficients are mapped on a sphere and the angles with respect to the first coefficient are calculated. Subsequently, these angles are grouped into the pre-defined sectors that represent the partition of the sphere into 2^{ssec} large parts up to 2^{lsec} small parts.
Transcript profiles with similar dynamics are grouped using the script *SimilarGenes.m* which compares the vectors of sector membership with each other and dependent and for *rco*=0 angular traits are calculated. For rco¿0 neighbouring sectors are taken into account.
For transcript profiles of inverse dynamics, the same script compares the vector of sector membership to the reflected sector.
These information are then evaluated in the script *SharedClusters.m*.

Clustering and Results

Finally, for any parameter *rco*, the program expects a cluster table file from BioLayoutExpress 3D, that is generated by using the MCL-algorithm. This file is then used to generate maps of cluster-identifiers to gene-identifiers to the data.