

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

On metabolic networks and multi-omics integration

Rasmus Ågren

Systems and Synthetic Biology
Department of Chemical and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2013

On metabolic networks and multi-omics integration

Rasmus Ågren

ISBN 978-91-7385-839-7

© Rasmus Ågren, 2013.

Doktorsavhandlingar vid Chalmers tekniska högskola

Ny serie nr 3520

ISSN 0346-718X

Department of Chemical and Biological Engineering

Chalmers University of Technology

SE-412 96 Gothenburg

Sweden

Telephone +46 (0)31-772 1000

Cover:

Workflow for identification of transcriptionally regulated reactions. See Figure 4-4 on page 29 for details.

Printed by Chalmers Reproservice

Gothenburg, Sweden 2013

On metabolic networks and multi-omics integration

Rasmus Ågren

Systems and Synthetic Biology

Department of Chemical and Biological Engineering

Chalmers University of Technology

Abstract

Cellular metabolism is a highly complex chemical system, involving thousands of interacting metabolites and reactions. The traditional approach to understanding metabolism has been that of reductionism; by isolating and carefully measuring the involved components, the goal has been to understand the whole as the sum of its parts. This reductionist approach has successfully identified most of the components of metabolism but, unfortunately, it fails to capture the long-range and complex interactions that are essential for the functionality. Systems biology is an emerging research field which uses high-throughput data generation and mathematical modelling in order to apply a holistic, or network-centric, view on metabolism. One type of modelling framework, which is in line with this thinking, is genome-scale metabolic modelling. These models, called GEMs, represent very valuable resources, but their applications have been limited due to the large manual effort required to reconstruct them. In this project, we have developed algorithms and software for streamlining the reconstruction process, as well as for novel applications of GEMs. More specifically, we here present: the RAVEN Toolbox, a software suite for automated reconstruction and quality control; the INIT algorithm, an algorithm for inferring GEMs for human cell types; an algorithm which integrates fluxomics and transcriptomics data in order to identify transcriptionally controlled metabolic reactions.

The methods and software were used in a number of case studies to address real biological questions. These studies were: 1) Metabolic engineering of *Saccharomyces cerevisiae* for succinic acid overproduction. The predictions from the modelling were successfully validated experimentally. 2) Study of metabolic regulation in *S. cerevisiae*. This led to the identification of a small number of transcription factors and enzymes which were predicted to be controlling central parts of metabolism. 3). Penicillin production in *Penicillium chrysogenum*. This led to the reconstruction of the first GEM for *P. chrysogenum*, an important resource in itself, and to identification of metabolic engineering targets for more efficient production of penicillin. 4) Human cancer metabolism. This led to the identification of metabolic subnetworks which were predicted to be significantly more active in cancers, and to identification of potential drug targets for treatment. 5) Lipid metabolism in obesity. This led to new insights into the large-scale metabolic rearrangements associated with obesity, and to identification of possible therapeutic strategies. 6) Metabolism in non-alcoholic fatty liver disease. This led to the identification of serine deficiency as a central aspect of the disease, and to proposed therapeutic strategies for remedying it.

The work put forward in this thesis has resulted in improvements on several important aspects of genome-scale metabolic modelling, and it has shown how the framework can be applied to gain novel biological insights. As such, it can contribute to further increase the role of the framework in modelling of human health and disease.

Keywords: genome-scale metabolic model; penicillin; succinate; NASH; cancer; metabolic engineering; systems biology; metabolism; reconstruction; flux balance analysis

List of publications

This thesis is based on the following publications, referred to as Paper I to VI in the text:

- I. **Agren, R.**, Otero, J.M. and Nielsen, J. (2013) Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production, *J Ind Microbiol Biot*, doi:10.1007/s10295-013-1269-3.
- II. Bordel, S., **Agren, R.** and Nielsen, J. (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes, *PLoS Comput Biol*, 6(7), p. e1000859.
- III. **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), p. e1002980.
- IV. **Agren, R.**^{*}, Bordel, S.^{*}, Mardinoglu, A., Pornputtpong, N., Nookaew, I. and Nielsen, J. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, *PLoS Comput Biol*, 8(5), p. e1002518.
- V. Mardinoglu, A., **Agren, R.**, Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., Walley, A.J., Froguel, P., Carlsson, L.M., Uhlen, M., Nielsen, J. (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte, *Mol Syst Biol*, 9, p. 649.
- VI. Mardinoglu, A.^{*}, **Agren, R.**^{*}, Kampf, C., Uhlen, M. and Nielsen, J. (2013) Genome-scale metabolic modeling of hepatocytes leads to identification of serine deficiency in non-alcoholic fatty liver disease, (Submitted).

Additional publications not included in this thesis:

- VII. **Agren, R.**^{*}, Mardinoglu, A.^{*}, Kampf, C., Uhlen, M. and Nielsen, J. (2013) Drug discovery through the use of personalized genome-scale metabolic models for liver cancer, (Submitted).
- VIII. Thiele, I., Swainston, N., Fleming, R.M., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., Thorleifsson, S.G., **Agren, R.**, Bolling, C., Bordel, S., . . . Palsson, B.O. (2013) A community-driven global reconstruction of human metabolism, *Nat Biotechnol*, doi:10.1038/nbt.2488.
- IX. Caspeta, L., Shoaie, S., **Agren, R.**, Nookaew, I. and Nielsen, J. (2012) Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials, *BMC Syst Biol*, 6, p. 24.
- X. Pabinger, S., Rader, R., **Agren, R.**, Nielsen, J. and Trajanoski, Z. (2011) MEMOSys: Bioinformatics platform for genome-scale metabolic models, *BMC Syst Biol*, 5, p. 20.
- XI. Liu, L., **Agren, R.**, Bordel, S. and Nielsen, J. (2010) Use of genome-scale metabolic models for understanding microbial physiology, *FEBS Lett*, 584(12), pp. 2556-2564.
- XII. Cvijovic, M.^{*}, Olivares-Hernandez, R.^{*}, **Agren, R.**^{*}, Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K.R., Nielsen, J. (2010) BioMet Toolbox: genome-wide analysis of metabolism, *Nucleic Acids Res*, 38, pp. W144-149.

^{*} Authors contributed equally

Contribution summary

- I. Performed the modelling, assisted in the design of the fermentation experiments, analysed results, prepared and submitted the paper.
- II. Formulated and implemented the sampling algorithm, participated in the analysis of results, assisted on preparation and submission of the paper.
- III. Created the software, reconstructed the *Penicillium chrysogenum* model, performed the modelling, prepared and submitted the paper.
- IV. Formulated and implemented the INIT algorithm, participated in the creation of the HMR database, reconstructed the cell type-specific models, participated in the analysis of results, participated in the preparation and submission of the paper.
- V. Participated in the reconstruction of the adipocyte model, performed the modelling, participated in the analysis of results, assisted on preparation and submission of the paper.
- VI. Participated in the reconstruction of the hepatocyte model, performed the modelling, participated in the analysis of results, participated in the preparation and submission of the paper.
- VII. Formulated and implemented the tINIT algorithm, performed the modelling, participated in the analysis of results, participated in the preparation and submission of the paper.
- VIII. Participated in the reconstruction of the model.
- IX. Supervised the work, assisted in the analysis of results, assisted in the preparation and submission of the paper.
- X. Participated in the design of the software, generated data to populate the database, assisted in the preparation and submission of the paper.
- XI. Participated in the preparation and submission of the paper.
- XII. Participated in the design and creation of the software, participated in the preparation and submission of the paper.

Table of contents

Abstract	iii
List of publications	v
Contribution summary	vi
Table of contents	vii
Lists of figures and tables	viii
Preface	x
Abbreviations and symbols	xi
1 Introduction	1
1.1 Thesis structure	2
2 Background	3
2.1 Metabolic engineering	3
2.2 Systems biology	5
2.3 Constraint-based modelling	6
3 Applications of genome-scale metabolic models	11
3.1 Constraint-based modelling using GEMs	11
3.2 Data integration using GEMs	13
3.3 Reconstruction of GEMs	15
3.3.1 Automated reconstruction of microbial GEMs	18
3.3.2 Reconstruction of cell type-specific GEMs	19
4 Results and discussion	21
4.1 GEMs applied to metabolic engineering of fungi	21
4.1.1 Paper I: Genome-scale modelling enables metabolic engineering of <i>Saccharomyces cerevisiae</i> for succinic acid production	24
4.1.2 Paper II: Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes	28
4.1.3 Paper III: The RAVEN Toolbox and its use for generating a genome-scale metabolic model for <i>Penicillium chrysogenum</i>	32
4.2 GEMs applied to human health and disease	37
4.2.1 Paper IV: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT	39
4.2.2 Paper V: Global analysis of human adipocyte metabolism in response to obesity	44
4.2.3 Paper VI: Identification of serine deficiency in non-alcoholic fatty liver disease through genome-scale metabolic modelling	49
5 Conclusions and future perspectives	56
5.1 Conclusions	56
5.2 Future perspectives	58
Acknowledgements	60
References	61

Lists of figures and tables

Figure 2-1. Examples of pathway modifications.	4
Figure 2-2. Metabolic engineering of <i>E. coli</i> for production of DHAP.	5
Figure 2-3. The systems biology cycle.	6
Figure 2-4. Principles of FBA.	8
Figure 3-1. Introduction to GEMs.	11
Figure 3-2. The Reporter metabolites / Reporter subnetworks algorithms.	15
Figure 3-3. The GEM reconstruction process.	16
Figure 4-1. Overview of applications of fungal GEMs.	23
Figure 4-2. Comparison between experimental and simulated fermentation data.	25
Figure 4-3. Experimental and simulated data for reference strain, <i>Δoac1</i> , <i>Δmdh1</i> , and <i>Δdic1</i> strains.	26
Figure 4-4. Workflow for identification of transcriptionally regulated reactions.	29
Figure 4-5. The RAVEN Toolbox.	33
Figure 4-6. Evidence level for the <i>P. chrysogenum</i> metabolic network.	34
Figure 4-7. Integrative analysis of a high and a low producing strain.	36
Figure 4-8. Overview of applications of GEMs in human health and disease.	38
Figure 4-9. Principle of the INIT algorithm.	41
Figure 4-10. Metabolic subnetwork identified as being significantly more prominent in cancer tissues compared to their corresponding healthy tissues.	42
Figure 4-11. Schematic illustration of how a GEM for adipocytes may provide links between molecular processes and subject phenotypes.	45
Figure 4-12. Summary of the capabilities of iAdipocytes1809.	46
Figure 4-13. Simulated lipid droplet and acetyl-CoA production.	47
Figure 4-14. iHepatocytes2260 – a consensus GEM for hepatocytes.	51
Figure 4-15. Results from Reporter Metabolites analysis.	52
Figure 4-16. Results from Reporter Subnetworks analysis.	53
Table 3-1. Databases and resources for reconstruction of GEMs.	18
Table 3-2. Available omics types for reconstruction of cell type-specific GEMs.	20
Table 4-1. Some industrial applications of fungi.	22
Table 4-2. Top scoring enzymes for transcriptional, post-transcriptional and metabolic regulation for changes in carbon source.	30
Table 4-3. Comparison between the RAVEN Toolbox and some other software for automated GEM reconstruction.	34

To my parents, for all that has been
To Evelina, for all that will be

“No offense, son, but that's some weak-ass thinking.
You equivocating like a motherfucker.”
Reginald “Bubbles” Cousins, *The Wire* (2004)

Preface

This dissertation is submitted for the partial fulfilment of the degree of doctor of philosophy. It is based on work carried out between 2009 and 2013 in the Systems and Synthetic Biology group, Department of Chemical and Biological Engineering, Chalmers University of Technology under the supervision of Professor Jens Nielsen. The research was funded by the Knut and Alice Wallenberg Foundation, Sandoz, Vetenskapsrådet, and the Chalmers Foundation.

Rasmus Ågren

May 2013

Abbreviations and symbols

A. nidulans: *Aspergillus nidulans*

A. niger: *Aspergillus niger*

A. oryzae: *Aspergillus oryzae*

AKG: α -ketoglutarate

BLAST: Basic local alignment search tool

BLP: Bi-level programming

BMI: Body mass index

C-mol: Carbon mol

CBM: Constraint-based model / constraint-based modelling

CE: Cholesterol ester

CS: Chondroitin sulfate

DCW: Dry cell weight

E. coli: *Escherichia coli*

EFM: Elementary flux mode

EP: Extreme pathway

FA: Fatty acid

FAD⁺: Flavin adenine dinucleotide

FADH₂: Reduced flavin adenine dinucleotide

FBA: Flux balance analysis

FL: Fatty liver

GEM: Genome-scale metabolic model

GPR: Gene-protein-reaction

HDL: High-density lipoprotein

HMDB: Human metabolome database

HMM: Hidden Markov model

HMR: Human metabolic reaction database

HPA: Human protein atlas

HS: Heparan sulfate

HSPG: Heparan sulfate proteoglycans

IHC: Immunohistochemistry

INIT: Integrative network inference for tissues

KO: KEGG orthology

LD: Lipid droplet

LDL: Low-density lipoprotein

LP: Linear programming

LPL: Lipoprotein lipase

MCA: Metabolic control analysis

MILP: Mixed-integer linear programming

MFA: Metabolic flux analysis

MOMA: Minimisation of metabolic adjustment

NAFLD: Non-alcoholic fatty liver disease

NASH: Non-alcoholic steatohepatitis

NEFA: Non-esterified fatty acid

ORF: Open reading frame

P. chrysogenum: *Penicillium chrysogenum*

PEP: Phosphoenolpyruvate

PC: Phosphatidylcholine

PE: Phosphatidylethanolamine

PG: Proteoglycan

PS: Phosphatidylserine

QC: Quality control

QP: Quadratic programming

RAVEN: Reconstruction, analysis, and visualization of metabolic networks

ROOM: Regulatory on/off minimization

S. cerevisiae: *Saccharomyces cerevisiae*

SGR: Specific growth rate

SOP: Standard operating procedure

SSP: Serine synthesis pathway

TAG: Triacylglyceride

TCA: Tricarboxylic acid

VLDL: Very low-density lipoprotein

Nomenclature

Standard nomenclature is used for *S. cerevisiae* and *P. chrysogenum* for designating genes, proteins and gene deletions: *IDH1*, *Idh1p* and *Aidh1*, respectively, for isocitrate dehydrogenase 1 as an example. For *H. sapiens* it will be designated as *IDH1* and *IDH1* for genes and proteins, respectively.

1 Introduction

The first genome-scale metabolic model (GEM), for the bacteria *Haemophilus influenzae*, was published in the year 2000, nine years before I started my Ph.D. studies in 2009. During these years the models grew increasingly complex, with the first model for a eukaryote published in 2003 and the first human model in 2007. Extensive method development also took place during this period, particularly for applications in metabolic engineering and strain design, where dozens of algorithms were developed. Some of these algorithms were quickly forgotten, while others proved themselves to be highly useful. The early successes in the field led to an ever larger number of GEMs being reconstructed, also for less well characterized organisms. After working with a GEM during my master's studies I had identified a number of issues that I felt should be targets for further work.

- **Model reconstruction was labour-intensive and error-prone.** Some of the published models, particularly for non-model organisms, were of rather low quality. I therefore started working on what later became the RAVEN Toolbox (**Paper III**), a software suite with focus on speeding up the reconstruction process, while at the same time ensuring a high-quality model.
- **GEMs were underused as scaffolds for omics integration.** GEMs developed when the first genome projects were finished, which allowed for incorporating gene/transcript/protein/reaction relationships in metabolic networks. A GEM can therefore be viewed as a highly structured map of metabolism in a cell, from metabolites all the way up to the genes. This would make the GEM very well suited as a scaffold for integrating and interpreting omics data of different sorts. I started working on methods for integrating transcriptomics data and flux data from fermentations into GEMs. This resulted in an algorithm for finding reactions which are likely to be transcriptionally regulated (**Paper II**).
- **There were difficulties associated with modelling of complex organisms.** My long-term goal when starting my studies was to model human metabolism and interactions between different tissues. This was associated with some issues that were not seen for simple prokaryotic organisms. Firstly, eukaryotes have their metabolism divided across subcellular compartments, and this information is not readily available. Secondly, most human cells do not actively divide, which makes the assumption of optimization of biomass yield, commonly used for microbial cells, unrealistic. Thirdly, different cell types have very different phenotypes even though they share the same genotype. It is therefore not possible to use the same GEM for all cell types, nor is it possible to reconstruct cell type-specific models only from the genome sequence. While attempting to deal with these issues I developed an algorithm for assigning sub-cellular localization in GEMs (**Paper III**) and the INIT algorithm for reconstruction of cell type-specific GEMs based on a multitude of omics types (**Paper IV**).

The software and algorithms mentioned above were then used to study succinate production in *Saccharomyces cerevisiae* (**Paper I**), regulation of metabolism in *S. cerevisiae* (**Paper II**), penicillin production in *Penicillium chrysogenum* (**Paper III**), cancer metabolism in human (**Paper IV**), adipocyte metabolism in obesity (**Paper V**), and hepatocyte metabolism in non-alcoholic fatty liver disease (**Paper VI**). Since the biological problems studied in this thesis span several areas there is no common background section. Rather, each problem is introduced in the corresponding results section. The background section represents a more

general review of the applications of genome-scale metabolic models and constraint-based modelling.

1.1 Thesis structure

This thesis represents a summary of a number of published scientific essays, a so-called compilation thesis. The thesis is divided into two parts: an extended summary and a compilation of research articles. Part one first puts the work in a larger scientific context by describing the field in which it is carried out, and the problems being studied in the field (sections 2.1 and 2.2). It then describes and explains the history and formulation of the main methodological framework underlying the work (section 2.3). After that follows an extensive examination and evaluation of the literature within areas relating to the work (section 3). This section does not address work carried out within the Ph.D. project. Part one is then concluded with a summary of the articles which form the basis for the thesis, and the results obtained in them (section 4). Lastly, some concluding remarks and future perspectives are presented (section 5). Part two contains the original research articles. The order of the articles follows the order in which the work was performed, rather than the order of publication.

2 Background

Part of the work described in this thesis deals with methods for optimizing the genetic composition of microbial organisms in order to enable production of industrially relevant metabolites. It can therefore be said to belong to the field of metabolic engineering. Other parts of the work are about utilizing a holistic view of metabolism in order to integrate and understand large-scale data sets, which would put it closer to the field of systems biology. The following two sections briefly introduce those research fields in order to outline the foundation upon which the work is based. Section 2.3 describes and explains the methodological framework underlying most of the work.

2.1 Metabolic engineering

Metabolic engineering is about analysing and modifying metabolic pathways in order to achieve some objective; normally efficient production of industrially relevant compounds. Attempts to change the metabolism of microorganisms to suit our purposes have been carried out for a long time in the biotechnology industry, for example for amino acids and antibiotics production. These early attempts utilized chemical mutagenesis to speed up the evolution process, and creative selection techniques to steer evolution in the desired direction (Stephanopoulos *et al.*, 1998). Such approaches can be very successful, as shown in the case of penicillin production, where the titre could be increased from 3-6 μM for the original strain identified by Fleming in 1928 to $>75 \text{ mM}$ in 1977 (Nielsen, 1995). However, the drawback with these techniques was that the underlying reasons for the change in phenotype remained unknown, which made it difficult to identify relevant constraints and pathways. In the late 1980s and early 1990s molecular biology tools for making genetic modifications became available; enabling targeted modifications of metabolic pathways. This became known as metabolic engineering (Bailey, 1991).

Mathematical concepts developed at about this time enabled quantification of the control each enzyme in a pathway had on the overall flux through the pathway, and for calculation of the theoretical production yields of metabolites in complex metabolic networks (see section 2.3). The general methodology in metabolic engineering, at least in its most traditional sense, is to apply mathematical modelling in order to identify constraints which could be limiting the production of the compound of interest. Such constraints could relate to for example substrate specificity of enzymes, product inhibition, medium composition, or redox balances. Once identified, genetic engineering is used to modify the genetic makeup of the organism in order to relax the constraints, leading to increased production of the compound of interest. Possible strategies include expression of heterologous enzymes, overexpression of endogenous enzymes, deletion of genes or modulation of enzymatic activity, transcriptional or enzymatic deregulation, and optimization of medium composition (Stephanopoulos *et al.*, 1998). But metabolic engineering also has a broader ambition. It tries to answer questions such as: How can the most important parameters which define the phenotype be identified? How can that information be contextualized in the control architecture of the metabolic network? How can we redirect cellular metabolism towards something that is often detrimental or even toxic to the cell? How can the network structure and dynamics be used to propose rational genetic engineering targets? In order to provide answers to questions such as these, metabolic engineering takes a holistic view of metabolism and looks at the integrated metabolic pathways rather than at individual reactions in isolation. This mind-set is closely related with

systems biology (see section 2.2). Figure 2-1 shows some types of pathway manipulations which are commonly attempted in metabolic engineering.

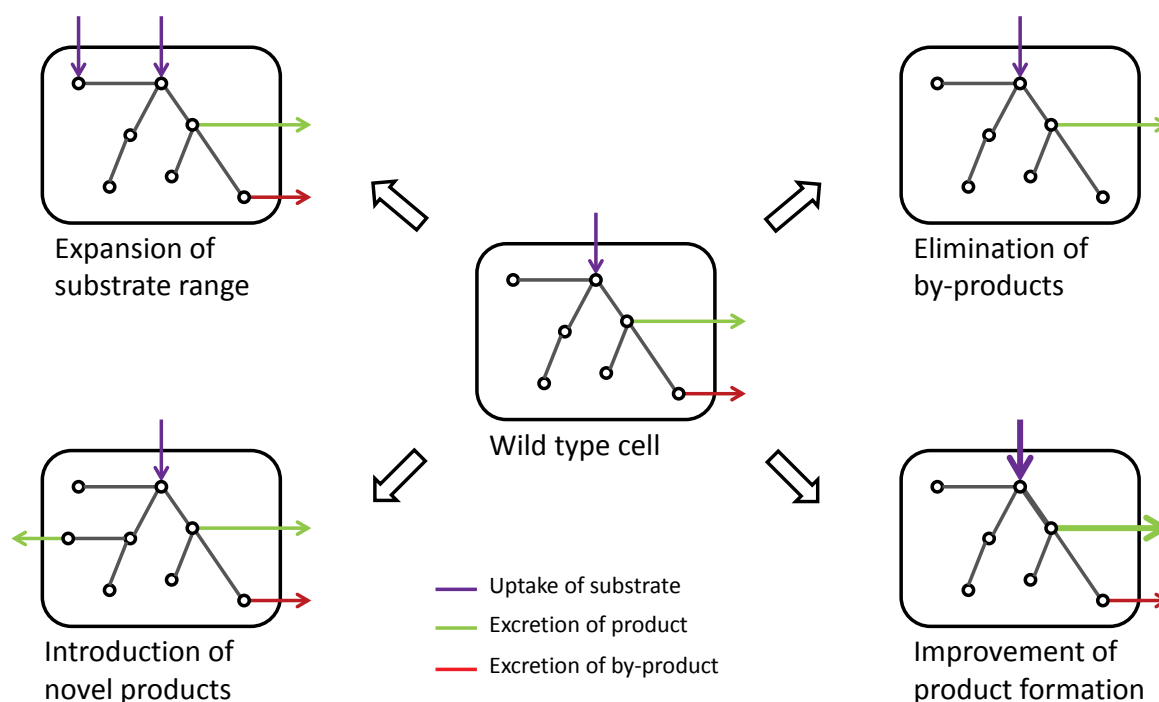


Figure 2-1. Examples of pathway modifications.

An early example of aromatic amino acid production using recombinant *Escherichia coli* can serve as a case study of the metabolic engineering workflow. *E. coli* and many other microorganisms synthesize aromatic amino acids through the condensation reaction between phosphoenolpyruvate (PEP) and erythrose 4-phosphate to form 3-deoxy-D-arabinoheptulosonate 7-phosphate (DAHP). Figure 2-2a shows the synthesis pathway of DHAP from glucose in *E. coli*. The first attempt to increase the production was made by screening of mutants which had deregulated product inhibition (Aiba *et al.*, 1980). This was followed by overexpression of the enzymes DHAP synthase (Forberg and Haggstrom, 1987) and then also transketolase (Draths *et al.*, 1992). However, the yields were still low. Forberg *et al.* (1988) then used a small metabolic model in order to identify the optimal flux distribution for DHAP production. This distribution can be seen in Figure 2-2a. It was predicted that 3 units of DHAP could be produced from 7 units of glucose. They identified the PEP phosphotransferase system, responsible for uptake of glucose, to be a suitable target. This system results in the conversion of PEP to pyruvate, which leads to a limitation in PEP for DHAP production. They further simulated the effect of introducing PEP synthase, which would regenerate the consumed PEP. The results can be seen in Figure 2-2b. The predicted yield would then rise to the double; 6 units of DHAP per 7 units of glucose. Patnaik and Liao (1994) then performed the suggested modification. They observed an almost two-fold increase in DHAP formation, from 52% yield to 90% yield, in excellent agreement with the predictions. This short example shows how powerful mathematical modelling and genetic engineering can be when used together. The field has developed tremendously since these early years, and both the dry lab and wet lab methods are now much more complex. The underlying thinking, however, remains the same.

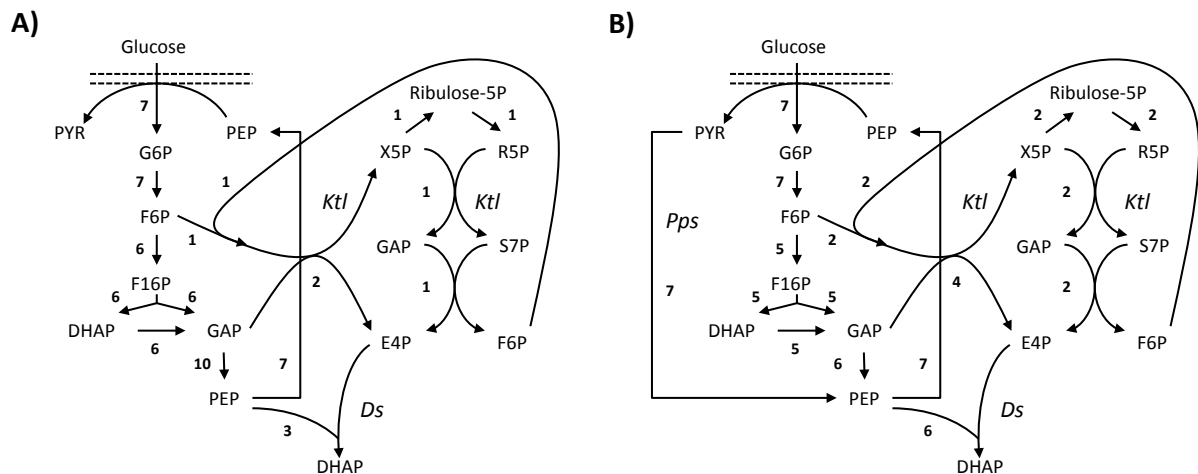


Figure 2-2. Metabolic engineering of *E. coli* for production of DHAP. **A)** Theoretical maximal yield of DHAP on glucose for the wild-type strain. **B)** Theoretical maximal yield of DHAP on glucose for a genetically modified strain with recombinant PEP synthase inserted. PYR: pyruvate, PEP: phosphoenolpyruvate, G6P: glucose 6-phosphate, F6P: fructose 6-phosphate, F16P: fructose 1,6-bisphosphate, DHAP: dihydroxyacetone phosphate, GAP: glyceraldehyde 3-phosphate, X5P: xylulose 5-phosphate, R5P: ribose 5-phosphate, S7P: sedoheptulose 7-phosphate, DAHP: 3-deoxy-D-arabinoheptulose 7-phosphate, Pps: PEP synthase, Tkt: transketolase, Ds: DAHP synthase. Adapted from Patnaik and Liao (1994).

2.2 Systems biology

The term systems biology has dual meanings. It can refer to an inter-disciplinary field which studies the interactions between components of complex biological systems. A central aspect in that interpretation is the concept of *emergent properties*. Emergence is the way complex patterns or behaviours arise from multiple relatively simple interactions. A classical example would be swarming. Each member of the swarm makes its decisions based on a few simple inputs, such as the proximity and speed relative to its neighbours, but the overall behaviour of a swarm can be awe-inspiring. Or in the words of physicist Dooyne Farmer: “It’s not magic...but it *feels* like magic” (Waldrop, 1992).

Systems biology claims to be particularly well posed to identify and study emergent properties, owing to its network-centric view. The underlying assumption is that biology itself has a strong link to emergence; due to the role evolution has in the growth of complexity in the natural world.

“[In] evolutionary processes, causation is iterative; effects are also causes. And this is equally true of the synergistic effects produced by emergent systems. In other words, emergence itself... has been the underlying cause of the evolution of emergent phenomena in biological evolution; it is the synergies produced by organized systems that are the key.” (Corning, 2012)

The second interpretation of the term systems biology is as a paradigm within biological sciences. It is then normally presented as the antithesis of the classical reductionist paradigm in the scientific method. The perceived limitations of the reductionist view has been described by Sauer *et al.* (2007) .

“The reductionist approach has successfully identified most of the components and many interactions but, unfortunately, offers no convincing concepts and methods to comprehend how system properties emerge.”

In practice, the systems biology approach, whether as a research field or as a scientific paradigm, often boils down to measuring multiple components simultaneously and then integrating the data with mathematical models. The field is therefore reliant on high-throughput measuring techniques such as metabolomics, transcriptomics and proteomics, as well as on methods from bioinformatics and computational biology.

As with most research fields there is no clear time point at which to put its birth. The ancestors of systems biology include the study of enzyme kinetics in the early 1900s (Michaelis *et al.*, 2011) and the application of control theory to biological systems in the 1960s and 1970s (Heinrich *et al.*, 1977). Denis Noble, who developed the first mathematical model of the working heart in 1960, is considered to be an early pioneer in the field. However, it was not until the 1990s, when the completion of the first genome projects resulted in large amounts of high quality data while at the same time computational power exploded, that the field really took off (Tomita *et al.*, 1997). Much of the work being done in systems biology is data driven rather than hypothesis driven, although this is by no means a requirement. Figure 2-3 illustrates the workflow commonly referred to as the systems biology cycle.

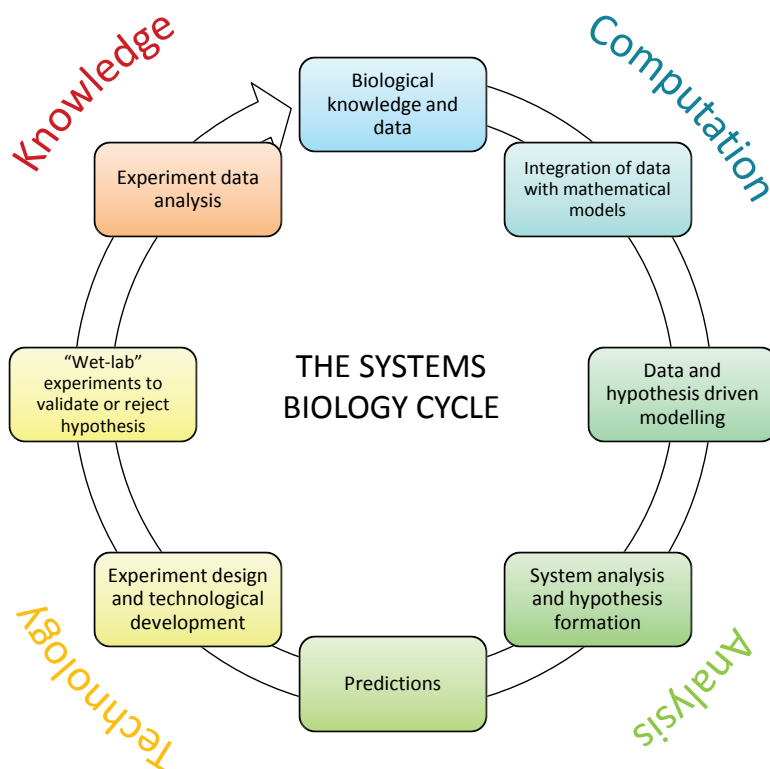


Figure 2-3. The systems biology cycle. Based on Kitano (2002a, b).

2.3 Constraint-based modelling

Mathematical modelling has been used to study metabolism for at least 100 years, since Michaelis and Menten derived their famous equation for enzyme kinetics in 1913. When the kinetic parameters for a large enough number of enzymes had been estimated, it was possible to formulate small models which could describe the basic metabolic functions of a living cell (Othmer, 1976). However, the data on kinetic parameters was fragmented, and the models

could be sensitive with respect to measurement errors (not to mention the problems associated with using *in vitro* measurements to estimate *in vivo* kinetics). There was a need to develop a mathematical framework to deal with uncertainty in data, and to quantify the control each of the enzymes had in the model. This led to the development of metabolic control analysis (MCA) (Heinrich *et al.*, 1977).

The framework was mainly applicable to small networks, and the availability of kinetic parameters continued to be limiting. If only the steady state metabolic fluxes inside the cell were of interest (rather than the dynamic change in metabolite pools) then those could be estimated in a method called metabolic flux analysis (MFA) (Aiba and Matsuoka, 1979). The method relies on measuring the rates of production/consumption of metabolites (called exchange fluxes) in the growth medium. If the set of possible enzymatic conversions is known, then the internal fluxes can be fitted from the exchange fluxes by linear regression. However, this requires that enough exchange fluxes are measured so that the resulting equation system is determined. Another issue was the determinability of fluxes in parts of the metabolism where there were cyclic or parallel reactions. By using isotope labelled substrates it was possible to track each atom, rather than each metabolite, through the metabolic network. This allowed for better resolution and more comprehensible models, but the fundamental limitations of the method remained (Wiechert, 2001). In a review paper entitled “Flux analysis of underdetermined metabolic networks: the quest for the missing constraints” Bonarius *et al.* (1997) describe how additional constraints, such as co-factor balancing or reaction reversibility, were incorporated in order to reduce the degrees of freedom and have a determined model. A large step forward was taken when the models were constrained to be optimal with respect to some cellular objective (Fell and Small, 1986). This formed the foundation for constraint-based modelling (CBM) of metabolism.

If the traditional approach to metabolic modelling is to describe the components of a model in such detail that the model correctly represents the phenotype, then the constraint-based approach is rather to impose increasingly detailed constraints on the solution space so that only relevant phenotypes are feasible. The term CBM is, at least when applied to metabolic modelling, largely synonymous to flux balance analysis (FBA), although FBA is a more narrow term. There are multiple excellent reviews describing the assumptions and mathematical formulation behind FBA (Varma and Palsson, 1994b; Edwards *et al.*, 2001; Price *et al.*, 2003). The following section will describe the methodology by using a small hypothetical metabolic network (see Figure 2-4a).

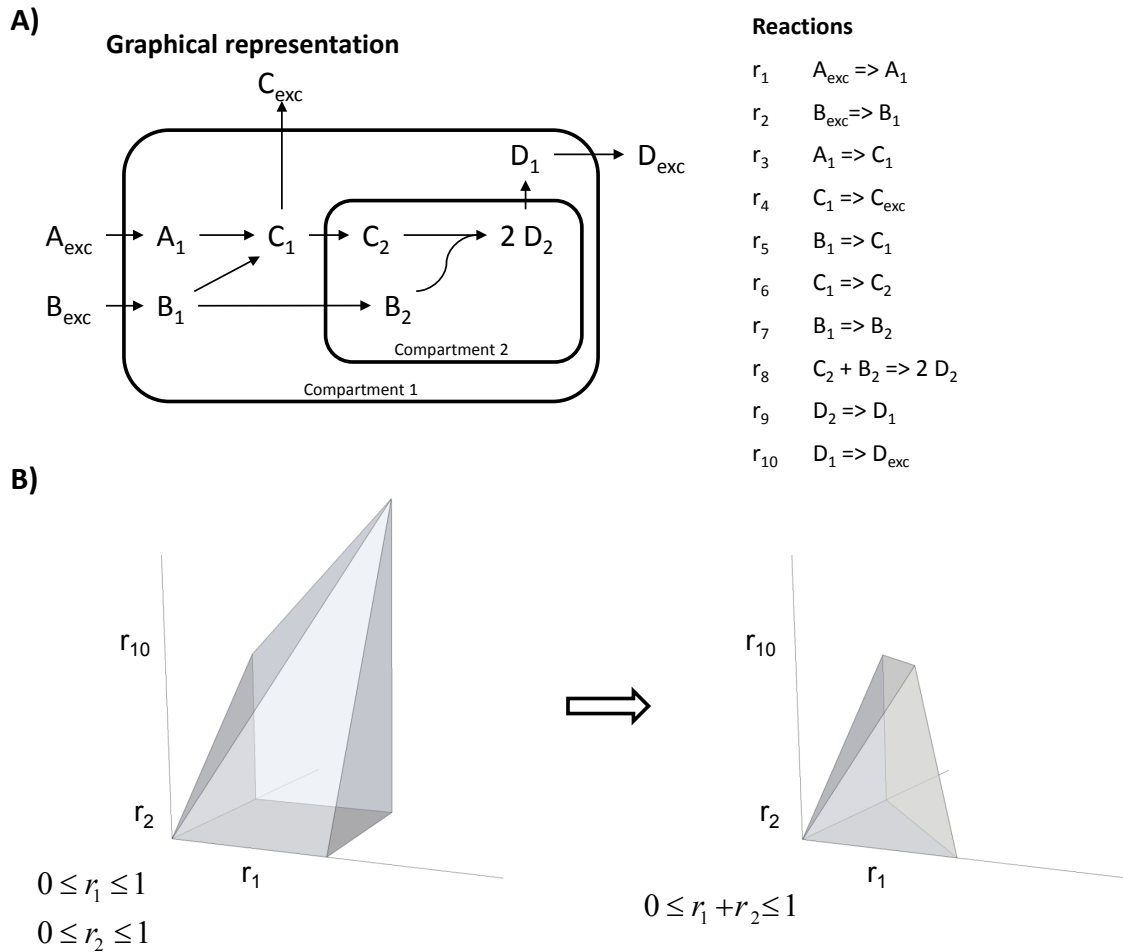


Figure 2-4. Principles of FBA. **A)** A small metabolic network. The network is comprised of 10 reactions (out of which 6 are internal), 12 metabolites (out of which 8 are internal) and it contains two compartments. The subscripts of the metabolites indicate which compartment they belong to. r_1 , r_2 , r_4 and r_{10} are exchange reactions. r_6 , r_7 and r_9 are reactions which transport metabolites between compartments. Note that while, for example, C_1 and C_2 represent the same chemical compound they are regarded as unique metabolites for modelling purposes. Also note that the stoichiometries of the enzymatic conversions are expressed in the network (see r_8). **B)** The feasible solution space is shown for the reactions r_1 , r_2 , and r_{10} . All points in the blue cone represent feasible solutions given the constraints. As additional constraints are imposed the solution space becomes narrower.

A mass balance over a metabolite can be expressed in the general form:

$$Accumulation = Input - Output + Generation - Consumption - Dilution \quad (1)$$

Or in a more mathematical form as:

$$\frac{dx_i}{dt} = v_{in,i} - v_{out,i} + v_{generation,i} - v_{consumption,i} - \mu x_i \quad (2)$$

In (2) the rate of accumulation of metabolite x_i is defined as the rate by which it is taken up ($v_{in,i}$), minus the rate by which it is excreted ($v_{out,i}$), plus the rate by which it is generated ($v_{generation,i}$) and minus the rate by which it is consumed ($v_{consumption,i}$). The dilution term (μx_i) accounts for the decrease in concentration that comes from the fact that a cell expands as it grows. Because the intracellular concentrations of most metabolites are very low compared to the fluxes affecting them, the dilution term can generally be neglected (Stephanopoulos *et al.*, 1998). This gives:

$$\frac{dx_i}{dt} = v_{in,i} - v_{out,i} + v_{generation,i} - v_{consumption,i} \quad (3)$$

For metabolite D_2 in the figure above, for example, the equation would then read:

$$\frac{dD_2}{dt} = 2r_8 - r_9 \quad (4)$$

This relationship can be expressed in a matrix notation to represent the mass balances for all metabolites

$$\frac{d\mathbf{x}}{dt} = \mathbf{S} \cdot \mathbf{v} \quad (5)$$

In (5) \mathbf{S} is a matrix which contains the stoichiometric coefficients that define the metabolic network. This matrix is referred to as the stoichiometric matrix. \mathbf{v} is a vector with the rate for each reaction and \mathbf{x} is a vector with the resulting changes in concentrations with respect to time for each of the internal metabolites. FBA is based on the assumption that the time scale for changes in the internal metabolite pools (typically seconds or minutes) is much faster than the time scale for growth or for changes in the environment (typically minutes or hours). It is therefore reasonable to assume that the internal metabolites are in steady state (meaning that their change in concentration is 0) (Varma and Palsson, 1994b). Equation (5) then simplifies to:

$$\mathbf{0} = \mathbf{S} \cdot \mathbf{v} \quad (6)$$

For the small network in Figure 2-4a this would look like:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{matrix} A_1 \\ B_1 \\ C_1 \\ C_2 \\ B_2 \\ D_2 \\ D_1 \end{matrix} \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \end{bmatrix} \quad (7)$$

In MFA the objective would now have been to measure a sufficiently large number of fluxes to have a determined model. With 10 variables (the unknown fluxes) and 7 equations (mass balances around the internal metabolites) the system has $10-7=3$ degrees of freedom (if all reactions were linearly independent). 3 fluxes would therefore have to be measured. In FBA the objective is instead to constrain the system to narrow the set of feasible flux distributions. One fundamental constraint is imposed by the thermodynamics (e.g. effective reversibility or irreversibility of reactions). In the example network all reactions are irreversible and it therefore holds that:

$$\mathbf{v} \geq 0 \quad (8)$$

For FBA to be effective three criteria have to be met: 1) the metabolic network should correctly describe the metabolic capabilities of the organism being studied, 2) the constraints should correctly describe the physiological limitations that the system operates under, 3) the objective function should correctly describe the objective which the cell strives to achieve. The first point is discussed in detail in section 3.3. The second and third points are discussed below.

Figure 2-4b shows the effect of imposing additional constraints on the model. In the left panel the uptake rates of metabolites A_1 and B_1 are constrained to be ≤ 1 . This defines a feasible cone of solutions, here shown in blue. Note that not all combinations of values for r_1 , r_2 and r_{10} are allowed, since their relationships are defined by the stoichiometry of the reactions. Constraints on uptake or excretion rates are the most widely used type and they are commonly based on experimentally measured fluxes. In the right panel an additional constraint has been imposed; that the sum of r_1 and r_2 should be ≤ 1 . This cuts the cone and further reduces the set of allowed flux distributions. There have been many attempts to define constraints that are biologically relevant and which do not require expensive and difficult in vivo measurements of enzymatic capabilities. Examples include: physical constraints such as diffusion rates; a general upper limit on enzyme capability and molecular crowding constraints (Beg *et al.*, 2007); binary regulatory constraints on which enzymes can be active under a given condition (Covert *et al.*, 2001); energy balancing to exclude thermodynamically infeasible solutions (Beard *et al.*, 2002); thermodynamic constraints based on the standard Gibbs free energies of formation (ΔG_f^0) for metabolites (Henry *et al.*, 2006).

How does a cell adjust its intracellular fluxes given the constraints that it is under? In FBA it is assumed that cell metabolism functions according to some objective, and that such an objective can be defined as a linear combination of the reaction rates.

$$\begin{aligned}
 & \text{Maximize} && \mathbf{c}^T \mathbf{v} \\
 & \text{Subject to} && \mathbf{0} = \mathbf{S} \cdot \mathbf{v} \\
 & && \text{lower bounds} \leq \mathbf{v} \leq \text{upper bounds}
 \end{aligned} \tag{9}$$

In (9) \mathbf{c} is a vector with coefficients for each of the reactions. The expression $\mathbf{c}^T \mathbf{v}$ then becomes the product of the flux and the objective coefficient, summed over all reactions. The system defined in (9) can be efficiently solved, also for very large problems, by using linear programming (Karp, 2008). There have been many studies on what constitutes a good objective function. Some of the first suggestions were rather basic, such as to maximize the NADPH production or minimize the ATP production (Bonarius *et al.*, 1997). When the molecular composition of biomass could be quantified in sufficient detail it was possible to use maximization of growth as an objective (Varma and Palsson, 1994a). This objective proved to be a very good approximation, and still remains by far the most commonly used objective for modelling of microbial cells. More complex objectives, such as maximization of entropy production (Henry *et al.*, 2006) or combinations of several of the objectives mentioned here (Schuetz *et al.*, 2012) have been proposed since then.

The idea that a complex system such as a living cell can be modelled from a small set of physiological constraints and some general objective remains very appealing. It is therefore likely that the quest for ever more detailed constraints and more predictive objective functions will continue also in the future.

3 Applications of genome-scale metabolic models

In the previous chapter, a small metabolic model with 10 reactions was used to show the principles behind FBA and CBM. However, the models used in practice are anything but small; rather, they contain thousands of reactions and metabolites. Ever since genome sequencing took off in the 1990s it has, at least in theory, been possible to identify each enzyme that exists in an organism, and thereby infer a metabolic network which describes the full metabolic capabilities of the organism (Schilling *et al.*, 1999). These models have therefore come to be known as genome-scale metabolic models (GEMs).

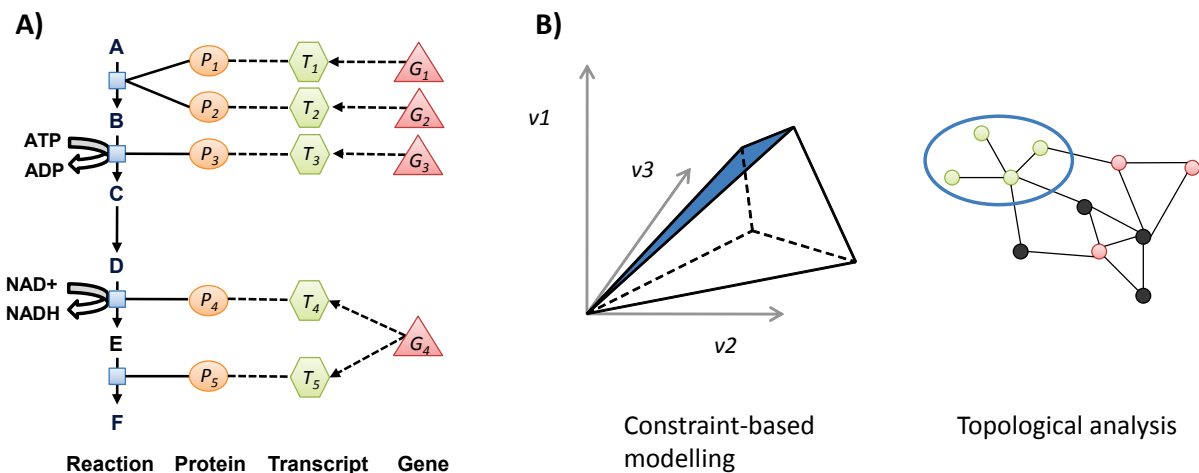


Figure 3-1. Introduction to GEMs. **A)** The layout of a genome-scale metabolic model (GEM). A GEM can be viewed as a highly structured map of how metabolism is controlled at different levels. At the bottom are the metabolic reactions and the metabolites which they involve. Each reaction can then be catalysed by zero or more enzymes. The enzymes are further linked to the corresponding transcripts, which in turn are linked to the corresponding genes. **B)** The two main applications of GEMs. In CBM the GEM is viewed as an equation system which describes how metabolism in a cell operates. In topological analysis the GEM is viewed as a map of how components in a cell interact with each other.

Figure 3-1a describes the general layout of GEMs. GEMs can be viewed as detailed maps of connections between the different levels of metabolism. Thereby they can provide a mechanistic description all the way from the metabolites, via reactions, enzymes and transcripts, up to the genes. Model elements at all levels can be extensively annotated so that GEMs can serve as highly structured databases. Figure 3-1b shows the two main application categories for GEMs. The following two sections will discuss how GEMs can be used for CBM (section 3.1) and for topological analysis/data integration (section 3.2). Section 3.3 deals with the reconstruction of GEMs; a very complex and time-consuming process.

3.1 Constraint-based modelling using GEMs

More than 100 algorithms for constraint-based modelling using GEMs have been published, as described in detail in an excellent review by Lewis *et al.* (2012). An in depth description of these algorithms would be outside the scope of this thesis, but has been extensively covered elsewhere (Price *et al.*, 2004; Durot *et al.*, 2009). Instead, this section will describe one algorithm each for a number of optimization frameworks, in order to illustrate the breadth of the available methods. The categories will be: 1) linear programming (LP), 2) quadratic

programming (QP), 3) mixed-integer linear programming (MILP), 4) bi-level programming (BLP), and 5) heuristic methods.

Linear programming. Linear programming represents the most fundamental of the optimization frameworks; so much that to many people CBM is synonymous to linear programming. The most widely used LP application is flux balance analysis (FBA) (see section 2.3). The approach relies on defining a linear objective function to be optimized, and then finding one solution (among the many) that is optimal with respect to the objective. An important advantage compared to the frameworks below is that LP problems can be solved to optimality very efficiently, even for large models. The objective function used for microorganisms is normally the maximization of the specific growth rate, which is consistent with the evolutionary advantage of the fastest growing species (Edwards et al., 2001). When used for strain design in metabolic engineering, the general approach is to iteratively remove enzymes from the GEM and then observe if the model produces the compound of interest as an effect of maximization of the growth rate. An example of this approach is in Lee *et al.* (2005), where the authors used FBA to suggest gene knockouts in *E. coli* with the purpose to overproduce succinic acid. The suggested modification involved a triple deletion to reduce the flux from PEP to pyruvate. When validated experimentally it resulted in a sevenfold increase in production of succinic acid.

Quadratic programming. Quadratic programming is similar to LP, but with the possibility of having quadratic terms in the objective function. This allows for minimization or maximization of the difference between fluxes, i.e. *minimize* $(v_i - v_j)^2$. The most widely used application of this optimization framework is Minimization of metabolic adjustment (MOMA) (Segre *et al.*, 2002). The underlying assumption in MOMA is that following a perturbation, such as deletion of a gene, the cell strives to minimize the distance from its flux distribution to the flux distribution of the non-perturbed cell. In a fascinating study, Wintermute and Silver (2010) used MOMA to study synthetic mutualism in auxotrophic *E. coli* mutants, and how they can complement one another's growth by cross-feeding of essential metabolites.

Mixed-integer linear programming. Mixed-integer linear programming (MILP) is based on LP, with the additional feature that variables can be constrained to only take integer values. MILP has found extensive use in algorithms for model reconstruction and gap filling (see section 3.3), since it is possible to formulate problems where a variable takes the value 1 if a reaction is included in a model and a value 0 if it is excluded. An algorithm which uses MILP for strain design is Regulatory on/off minimization (ROOM) (Shlomi *et al.*, 2005). The underlying assumption is similar to MOMA, but rather than minimization of the distance between the flux distributions, the cell is assumed to strive to minimize the difference in which reactions are active/passive. ROOM has been shown to give slightly better predictions when compared to MOMA, but at the cost of being significantly more computationally intensive (Shlomi *et al.*, 2005).

Bi-level programming. Bi-level programming represents optimization problems where one problem is embedded in another one. In this context it normally means optimization of some objective while the model is constrained to be optimal with respect to some other cellular objective. The first implementation of this optimization framework, OptKnock, has proven to be a very powerful tool for strain design (Burgard *et al.*, 2003). In OptKnock, the objective function is maximization of production of some relevant compound. In order to achieve this, the algorithm removes reactions so that production is stoichiometrically linked to optimal growth. For a small number of gene deletions this could be iteratively tested for in a brute-force approach, as described in the linear programming section. The strength of OptKnock is

that also combinations of relatively large numbers of gene deletions can be evaluated. At its foundation OptKnock is implemented as a MILP problem. OptKnock was used by Fong *et al.* (2005) in a study on lactate overproduction in *E. coli*. The algorithm suggested a triple deletion strategy resulting in 1) disabled ethanol and acetate production, 2) increased production of pyruvate and NADH, which are the precursors for lactic acid, 3) coupled uptake of glucose to the conversion of PEP to pyruvate. This deletion strategy, after a round of adaptive evolution, resulted in lactate titres of 0.87 to 1.75 g/L when the cells were grown in 2 g/L glucose.

Heuristic methods. Heuristic optimization methods are used for quickly finding an approximate solution; trading away optimality, completeness, accuracy, and/or precision. These methods also have the strength that they can make use of more general objective functions, not only the linear or quadratic forms described above. One example of a heuristic algorithm for strain design is OptGene (Patil *et al.*, 2005). It is based on randomly introducing perturbations to a population of GEMs, and then letting them compete and mate with each other based on their fitness. This approach is called evolutionary programming. OptGene has been applied to suggest metabolic engineering strategies for sesquiterpene production in *Saccharomyces cerevisiae* (Asadollahi *et al.*, 2009).

3.2 Data integration using GEMs

A number of algorithms have been developed for the purpose of using GEMs as scaffolds for data integration and interpretation. An in depth description of these algorithms would be outside the scope of this thesis, but has been extensively covered elsewhere (Joyce and Palsson, 2006; Durot *et al.*, 2009). Instead, this section will describe one or two algorithms each for three important types of omics data in order to illustrate the concept.

Fluxomics. Measured intracellular fluxes, for example using ^{13}C labelled substrates, represent a data type that is directly applicable to integration with GEMs. In order to go from measured labelling patterns in metabolites to fluxes, an atom mapping model is used. These models are similar to GEMs in that they are stoichiometric models of metabolism, but they can track each atom through the network, rather than each metabolite. They are traditionally rather small models, and only built for central carbon metabolism. This causes some problems, for example with co-factor balancing. GEMs have therefore been used to expand and complement atom mapping models, in order to also take more peripheral metabolism into account (Suthers *et al.*, 2007). As discussed in section 2.3, a driving force in CBM is the hunt for ever more precise objective functions. Burgard and Maranas (2003) developed an algorithm, ObjFind, which makes use of fluxomics data to try to infer the cellular objectives that could have given rise to the phenotype. They found that regardless of the growth condition, maximization of growth was the objective that best fitted the data. This is good news for FBA; since the method is based on that there are simple objectives which hold for a wide range of conditions.

Metabolomics. Large-scale quantification of internal metabolites has been made possible thanks to developments in mass spectrometry and NMR technology. Since GEMs are based on the assumption that metabolite pools are in quasi-steady state, the concentrations of metabolites are not immediately possible to integrate into GEMs. Instead, the main use of metabolomics data has been to evaluate the capabilities of the GEM (can the model produce the detected metabolites?). This can then lead to directed search for the missing functions,

thereby generating new biological knowledge as well as improving on the model (Oh *et al.*, 2007). Metabolomics has also been used together with estimated Gibbs free energies of formation for metabolites in order to predict reaction directionality (Kummel *et al.*, 2006a).

Transcriptomics. As discussed in section 3.3.2, there is not a good correlation between transcript level and flux, owing to the several layers of regulation between them (Akesson *et al.*, 2004). It is therefore difficult, or impossible, to directly use the expression levels to modify the model constraints, although several attempts have been made. The most common use of transcriptomics data is therefore to classify genes in a binary fashion; either as expressed or non-expressed (see the part about GIMME in section 3.3.2). An alternative approach, developed by Patil and Nielsen (2005), uses the data in a different way. Rather than looking at the expression level, they look at the significance of differential expression between two conditions, and make use of the network topology to analyse the data. The method works by first converting the metabolic network into a bipartite graph. In a bipartite graph the metabolites are connected to genes based on the reactions in which they participate. A meta-analysis is then performed for each metabolite by testing if the genes that it is associated to are differentially expressed when taken as a group (known as gene set analysis). If so, then the metabolite is classified as a Reporter. Reporter metabolites can be said to represent “hot spots” in metabolism around which transcriptional changes occur. Reporter subnetworks, presented in the same paper, is an algorithm with a somewhat similar mind-set. Both algorithms are described in Figure 3-2. The most well-known use of metabolic network topology is elementary flux modes (EFMs) (Schuster *et al.*, 1999) and its cousin extreme pathways (EPs) (Schilling and Palsson, 2000). These are minimal sets of reactions which can operate in steady state in a metabolic network. A critical drawback is that the enumeration of EFMs or EPs is very computationally demanding, and the method is therefore only applicable for medium-sized networks. In an approach conceptually similar to Reporter subnetworks, Schwartz *et al.* (2007) used EFMs to aid in interpretation of transcriptomics data. Rather than calculating a p-value for the set of enzymes in a subnetwork, they calculated it for the set of enzymes in each EFM. The approach was then applied to study stress responses in *S. cerevisiae*.

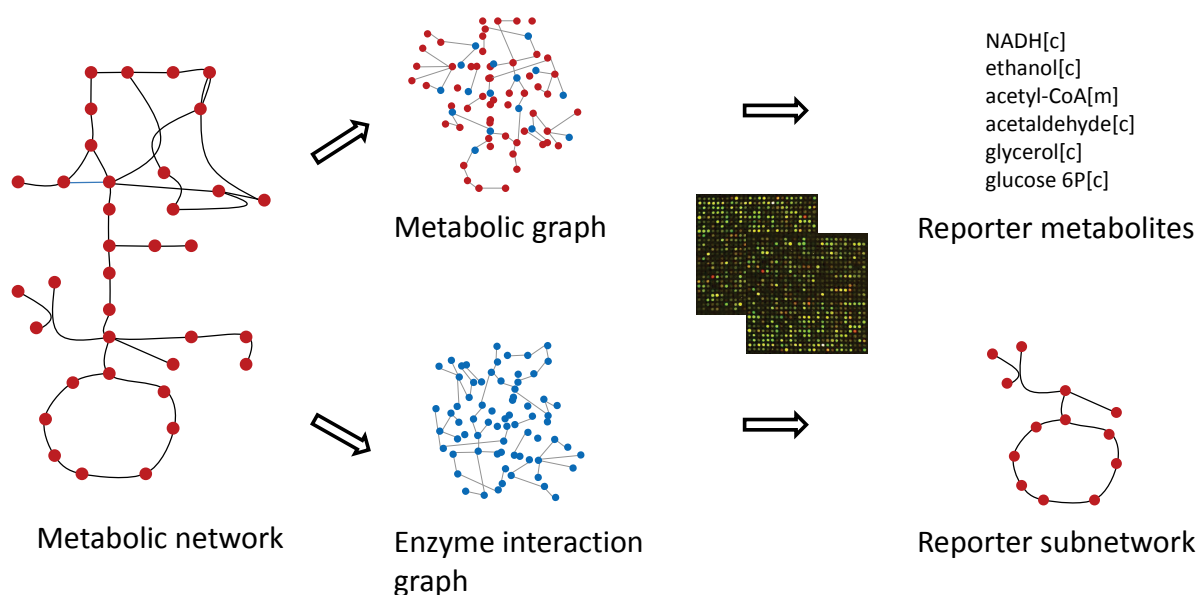


Figure 3-2. The Reporter metabolites / Reporter subnetworks algorithms. For Reporter metabolites, the metabolic network is converted into a bipartite graph, where each metabolite is connected to the genes for the reactions it participates in. A metabolite can then be scored based on the p-values for differential expression for the genes connected to it. If there is an overall significant change then the metabolite is a Reporter. For Reporter subnetworks, the metabolic network is converted into a unipartite enzyme interaction graph. A simulated annealing algorithm is then applied in order to find sets of connected enzymes which exhibit an overall significant change in expression. The metabolic network involving those enzymes can then be reconstructed from the original metabolic network. These are called Reporter subnetworks. Adapted from Patil and Nielsen (2005).

3.3 Reconstruction of GEMs

The reconstruction process of GEMs is traditionally very labour- and time-intensive, spanning from several months for a well-studied bacteria to several years for a human model (Duarte *et al.*, 2007). The very aspects that make GEMs so powerful, their scope and multi-level structure, are also what makes the reconstruction process so complex. This section gives an overview of the traditional approach, where models are manually reconstructed from genomic and bibliomic data in a bottom-up manner. The two subsections describe top-down approaches for microbial and cell type-specific models, respectively.

There has been a multitude of published descriptions of the reconstruction process, but one has had a particularly large impact on the field. In a review in Nature Protocols, Thiele and Palsson (2010) collected existing reconstruction practices and summarized them in a 96 step standard operating procedure (SOP). Figure 3-3 depicts the most important steps of the reconstruction process.

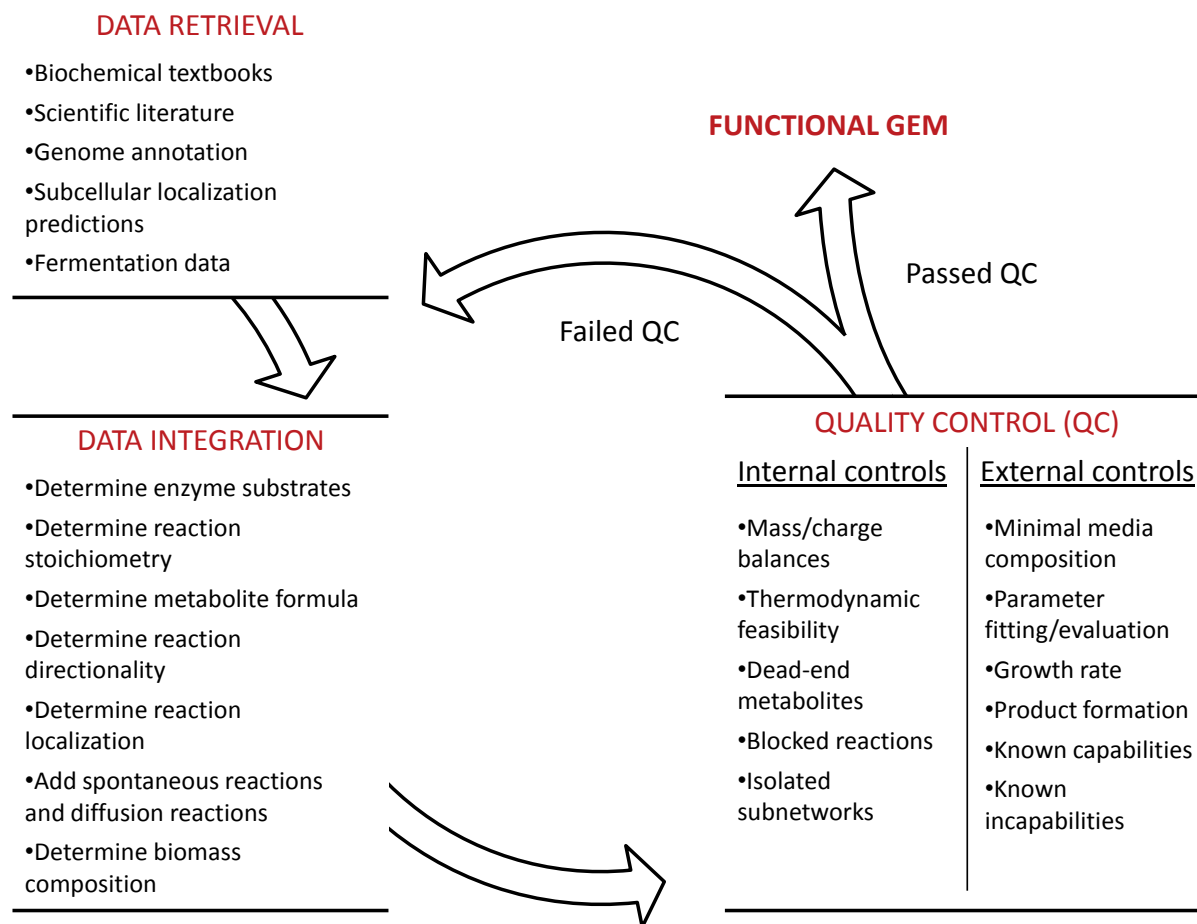


Figure 3-3. The GEM reconstruction process.

The level of detail required for reconstruction of a GEM is in a sense rather low, only the reaction stoichiometries and directionalities and the associated enzymes catalysing each reaction. This can be contrasted to the large number of kinetic parameters required for reconstruction of dynamic models. The reconstruction starts with the retrieval and organizing of the necessary input data. Depending on the organism, the availability and amount of data differs. The minimal required input can be said to be a sequenced genome and some amount of known physiological data, such as growth conditions. In general, the better the availability of physiological, biochemical and genetical data, the better the predictive ability of the model. Some data types which can be very valuable if they are available include: ^{13}C fluxomics data, measured subcellular localizations, or gene knockout libraries. The next step is the generation of a draft model. That starts by going through each of the annotated genes and deciding whether its function is within the scope of the reconstruction. Normally, only metabolic enzymes are included. There are also grey areas, such as DNA methylation, protein phosphorylation or complex glycan metabolism, which are often excluded from the GEM, even though they can be viewed as metabolic functions. The enzymes are then mapped to their respective metabolic reactions. This step is often performed via EC numbers, followed by retrieval of reactions from reaction databases such as KEGG (Ogata *et al.*, 1999) or RHEA (Alcantara *et al.*, 2012). It is important to note that EC numbers refer to the type of chemical transformation in a reaction, not the enzyme class which performs the conversion. For example, both ethanol dehydrogenase and choline dehydrogenase have EC number 1.1.1.1, since they both act on primary alcohols and use NAD^+ as a co-factor. This mapping is therefore not exact, and should be viewed as a first draft. The process described above is for

an annotated genome. If the genome for the organism of interest is not annotated, a first step would then be to use one of the many genome annotation pipelines which have been developed (Stein, 2001).

The network at this stage will most likely have a large number of issues which must be addressed. Potential issues include, but are not limited to: unclear metabolite naming, generic metabolites such as “an alcohol” or “fatty acid”, missing spontaneous conversions, wrongly assigned directionality of reactions, or generic stoichiometries such as “starch(n) + H₂O => starch(n-1) + glucose”. Careful manual evaluation and modification of the reactions is used to clear up the reaction list. The reactions are then partitioned into the relevant subcellular compartments, and transport reactions are added based on literature and genomic evidence. Lastly, the chemical composition of biomass is determined based on measurements or literature.

A number of quality controls are then performed in order to validate the model. In Figure 3-3 these controls are categorized as internal or external, where internal relates to the inner workings of the network and outer to the capabilities for predicting the cellular phenotype. As GEMs are fundamentally mass balance models, it is critical that the reactions are elementally and charge balanced. This is trivial if the elemental composition is known, but that is not always the case, for example in the case of polymers. The model can then be tested for “stoichiometric inconsistencies”, meaning that there are reaction sets such as $A \rightarrow B$ and $A \rightarrow B + C$ (Gevorgyan et al., 2008). Other tests include thermodynamic and redox feasibility, so that the model cannot produce high energy compounds from low energy compounds or reduced compounds from oxidized compounds (Kummel *et al.*, 2006b). There are a number of methods for identifying reactions which cannot carry flux or metabolites which cannot be produced, and to suggest strategies for connecting them (Mahadevan and Schilling, 2003; Reed *et al.*, 2006; Satish Kumar *et al.*, 2007; Kumar and Maranas, 2009; Brooks *et al.*, 2012). This process is referred to as gap filling, and it is of central importance for the quality of the reconstruction.

The metabolic capabilities of the model must then be evaluated by comparison to the known capabilities of the organism in question. This is referred to as external controls in Figure 3-3. Such controls include that the model can grow on media that the organism in question can grow on, that it can produce products that are known to be produced by the organism, and that it can perform other known metabolic functions of the organism. It is equally important to control that the model is not too flexible, so that it can perform functions that are known to not occur in the organism, or that it can grow faster than what is seen experimentally. Lastly, there are a small number of parameters which need to be fitted from experimental data. These include ATP maintenance costs and the P/O ratio. These parameters have to be validated to be within reasonable bounds. If all quality controls pass, then the model is functional and the reconstruction complete. If any of them fails, then the metabolic network has to be further modified. This is an iterative process, where the model is modified, validated, and modified again until a functional and high-quality model is achieved. This interplay between annotation, verification, and testing is a valuable process, as it results in the refinement of both the genome annotation and the reaction network. Table 3-1 lists some databases and resources that are widely used for reconstruction of GEMs.

Table 3-1. Databases and resources for reconstruction of GEMs.

	EMBL	GenBank	SEED	BRENDA	ENZYME	UniProt	TransportDB	PSORTdb	CheBI	Pub chem	HMDB	LipidMaps	Reactome	KEGG	BioCyc	UniPathway	PubMed	RHEA	BIGG	BioModels
Biochemical activities	X	X	X	X		X	X						X	X	X	X	X	X		
Enzyme specificity				X	X	X	X										X	X		
Subcellular localization				X			X	X					X		X		X			
Reaction equation				X	X		X						X	X	X	X	X	X		
Reaction direction				X			X						X	X	X		X	X		
Metabolite annotation				X					X	X	X	X		X	X		X	X		
GPR association	X	X	X			X	X						X	X	X	X	X	X		
GEM repositories																			X	X

Based on Durot *et al.* (2009) and Mardinoglu and Nielsen (2012).

3.3.1 Automated reconstruction of microbial GEMs

There have been a number of methods published for automating parts of the reconstruction process. Most or all of these methods aim primarily at automating the annotation step, but there are also methods that integrate parts of the quality control process as well. These methods make use of reaction databases and the connections between EC numbers and genes within such databases. One of the earliest such methods was Pathway Tools; a software for generating organism-specific databases from a general database (Karp *et al.*, 2002). The input to the software is a set of genes annotated with EC numbers. Pathway Tools was not developed with GEM reconstruction in mind, but rather as a more general resource. Other methods which also matched enzymes to reactions, but without the ambition to reconstruct GEMs, were IDENTICS (Sun and Zeng, 2004), which attempted to simultaneously annotate predicted ORFs and link them to reactions by using BLAST to match known metabolic genes to a non-annotated genome, and metaSHARK (Pinney *et al.*, 2005), which used PSI-BLAST profiles rather than BLAST for the same purpose. The first software dedicated for GEM reconstruction was GEM System (Arakawa *et al.*, 2006). GEM System first identified ORFs by using GLIMMER (Delcher *et al.*, 1999), then matched the ORFs to known enzymes using BLAST. A metabolic network was then generated based on mapping to an internal database. GEM System also contained a simple algorithm for gap filling. AUTOGRAPH (Derrien *et al.*, 2007) is a software for inferring GEMs based on previously reconstructed GEMs for other species. The software that has had the highest impact by far is the Model SEED resource (DeJongh *et al.*, 2007; Henry *et al.*, 2010). Model SEED builds on the gene calling and annotation pipeline in SEED, and then uses an internal reaction database to map annotated genes to reactions. It also contains an automatic gap filling feature, where a generic biomass equation is assumed, after which the software applies a gap filling algorithm in order to ensure that the model can form biomass.

The main advantage of using software like the ones described above is that it speeds up the reconstruction process. However, it is important to note that it comes at the cost of decreased control and insight over how and why different elements are included in the model.

3.3.2 Reconstruction of cell type-specific GEMs

The cells of multicellular organisms can have very different phenotypes even though they share the same genotype. For example, the longest neural cells in a human can be more than a meter long (Fletcher and Theriot, 2004), while one of the smallest cell types, the erythrocytes, only measure 7-8 μm (Fabry *et al.*, 1981). In order to model cellular metabolism, it is therefore necessary to reconstruct a GEM specifically for the relevant cell type. In practise, this is done by starting from a generic network for the organism in question, and then manually or algorithmically select a subset of enzymes which are thought to be present in the specific cell type. In 2007 two such generic GEMs were published for human: Recon 1 (Duarte *et al.*, 2007) and EHMN (Ma *et al.*, 2007). A number of cell type-specific models have been manually reconstructed by using these models as scaffolds, including for liver (Gille *et al.*, 2010), brain (Lewis *et al.*, 2010), alveolar macrophage (Bordbar *et al.*, 2010), and a multi-tissue model for hepatocytes, adipocytes and myocytes (Bordbar *et al.*, 2011). The manual reconstruction process follows the same workflow as described in Figure 3-3. These examples are either rather small GEMs or for well-studied cell types, where there is a wealth of physiological literature available.

In parallel to this there have also been a number of algorithms developed which aim at reconstructing cell type-specific GEMs in an automated manner based on high-throughput data. Note that the problem of inferring a cell type-specific network from a generic model is closely related to the problem of inferring an organism-specific network from a generic reaction database (as described in section 3.3.1). The difference is in the input data. While the organism-specific models are reconstructed based on protein homology, the cell type-specific models have to be reconstructed from omics data. Table 3-2 lists some relevant omics types and their respective advantages and disadvantages. The first of these algorithms was GIMME (Becker and Palsson, 2008). GIMME takes transcriptome data as input and removes reactions for which the expression levels for the genes are below some threshold. It then constrains the model to perform some function, after which it uses a gap filling algorithm to reinsert the required reactions so that the model can satisfy the constraints. The state of the art algorithm, MBA (Jerby *et al.*, 2010), adds another layer of complexity by dividing the reactions with supporting evidence into two groups; one with reactions which must be included, and one with reactions that should be included. The algorithm then uses a gap filling algorithm to include as many reactions as possible from the second group while using as few reactions as possible for gap filling. These algorithms have been applied to reconstruct GEMs for liver (Jerby *et al.*, 2010), kidney (Chang *et al.*, 2010) and a generic cancer model (Folger *et al.*, 2011).

Table 3-2. Available omics types for reconstruction of cell type-specific GEMs.

	Type	Advantages	Disadvantages
Transcriptomics from DNA microarrays	Relative	Cheap, widely available, high throughput	Low correlation between gene expression and protein level
Transcriptomics from RNA-Seq	Semi-quantitative	As above, but with the added benefit that the measurements are semi-quantitative	As above
Proteomics	Semi-quantitative	Direct evidence for the presence/absence of enzymes	Expensive, not all proteins/cell types are currently covered
Metabolomics	Quantitative	Detection of a metabolite indicates that the cell must possess the metabolic capabilities to synthesize it (or be able to import it from its surroundings)	Detection of a metabolite says nothing about which pathways it was synthesized in, or about the fluxes involving it
Fluxomics	Quantitative	Direct evidence for intracellular metabolic fluxes	A metabolic model has to be used for fitting the fluxes, only available for central carbon metabolism
Bibliomics	Categorical	Can be very reliable if based on high-quality experimental data	Time- and labour-intensive to retrieve and organize the data

4 Results and discussion

In the following sections I will summarize the publications underlying the thesis and discuss their contribution to the field. The publications could broadly be divided to deal with GEMs applied to metabolic engineering of fungi (section 4.1) and GEMs applied to human health and disease (section 4.2), although there are plenty of links between them. Each section is preceded by a short review of research previously carried out in the field.

4.1 GEMs applied to metabolic engineering of fungi

Fungi have been used by humans since ancient time for production of cheese, bread, beer, wine and soy sauce. Today they are used in many industrial processes, such as the production of enzymes, vitamins, polysaccharides, alcohols, pigments, lipids, and glycolipids. Fungal secondary metabolites, in particular antibiotics, are extremely important to our health and nutrition and have tremendous economic impact (Adrio and Demain, 2003). The industrial production of β -lactam antibiotics by the mold *Penicillium chrysogenum* is one of the success stories of biotechnology, and represents one of the largest biotechnological products in terms of value, with dosage form sales of about USD 15 billion (Elander, 2003). Table 4-1 lists some important industrial applications of fungi and the species used.

Much work has gone into metabolic engineering of fungi, partly owing to their large industrial relevance and partly because several fungi are important model organisms. In terms of genome-scale metabolic modelling most of the efforts have been centred on the yeast *S. cerevisiae*. To date there are ten published *S. cerevisiae* GEMs with different scopes and applications (Osterlund *et al.*, 2012). GEMs have also been developed for the industrially relevant yeasts *Pichia pastoris* (Sohn *et al.*, 2010), *P. stipitis* (recently renamed to *Scheffersomyces stipitis*) (Caspeta *et al.*, 2012), *Candida glabrata* (Xu *et al.*, 2013), and *Yarrowia lipolytica* (Loira *et al.*, 2012) as well as for the model yeast *Schizosaccharomyces pombe* (Sohn *et al.*, 2012). Three models have been reconstructed for filamentous fungi in the *Aspergillus* genus; namely *A. niger* (Andersen *et al.*, 2008), *A. oryzae* (Vongsangnak *et al.*, 2008) and *A. nidulans* (David *et al.*, 2008).

Table 4-1. Some industrial applications of fungi.

Process	Organism
Pre-modern products	
Ang-kak	<i>Monascus purpurea</i>
Miso	<i>Aspergillus oryzae</i>
Ontjam	<i>Neurospora crassa</i>
Soy sauce	<i>A. oryzae</i> , <i>A. sojae</i>
Tempeh	<i>Rhizopus niveus</i>
Brewing and baking	<i>Saccharomyces cerevisiae</i> , <i>S. carlbergensis</i>
Mold-ripened cheeses	<i>Penicillium roqueforti</i> , <i>P. camembetii</i>
Pharmaceuticals	
Penicillins	<i>P. chrysogenum</i>
Cephalosporin	<i>Cephalosporium acremonium</i>
Cyclosporin	<i>Tolypocladium inflatum</i>
Ergot alkaloids	<i>Claviceps purpurea</i>
Griseofulvin	<i>P. griseofulvin</i>
Mevalonin	<i>A. terreus</i>
Statins	<i>P. brevicompactum</i> , <i>P. citrinum</i> , <i>M. ruber</i> , <i>A. terreus</i>
Taxol	<i>Taxomyces andreanae</i>
Proteins	
α -Amylases	<i>A. niger</i> , <i>A. oryzae</i>
Cellulase	<i>Humicola insolens</i> , <i>P. funiculosum</i> , <i>Trichoderma viride</i>
Glucoamylases	<i>A. phoenicis</i> , <i>R. delemar</i> , <i>R. niveus</i>
Glucose oxidase	<i>A. niger</i>
Invertase	<i>A. niger</i> , <i>A. oryzae</i>
Laccase	<i>Coriolus versicolor</i>
Pectinase	<i>A. niger</i> , <i>A. oryzae</i> , <i>H. insolens</i>
Proteinases	<i>A. oryzae</i> , <i>A. melleus</i>
Recombinant enzymes	<i>M. miehei</i> , <i>M. Pusillus</i> , <i>Pichia pastoris</i> , <i>S. cerevisiae</i>
Organic acids	
Citric acid	<i>A. niger</i>
Itaconic acid	<i>A. terreus</i>
Solvents and fuels	
Ethanol	<i>S. cerevisiae</i>

Based on Bennett (1998); Adrio and Demain (2003); Choi *et al.* (2003).

In an excellent review Osterlund *et al.* (2012) summarized the published applications of *S. cerevisiae* GEMs since the first model was made available in 2003. They classified its applications in four categories: 1) guidance for metabolic engineering and strain improvement, 2) biological interpretation and discovery, 3) applications of novel computational frameworks, and 4) evolutionary elucidation. Figure 4-1 builds on their

classification, but also includes non-*Saccharomyces* yeasts and filamentous fungi, as well as publications from after 2010. As can be seen, the first phase was completely dominated by work on *S. cerevisiae*. Around 2007 there seems to have been an increased interest in filamentous fungi and during the last couple of years there have been a number of studies on non-*Saccharomyces* yeasts. Another trend is that up until about 2009 many of the papers are about novel computational frameworks (category 3) and about evolutionary elucidation/comparative genomics (category 4). It is not until more recently that GEMs are widely used for strain engineering (category 1). This can indicate that the field has matured, and that the mathematical methods developed in the early years are starting to prove themselves and are now being used to solve concrete problems.

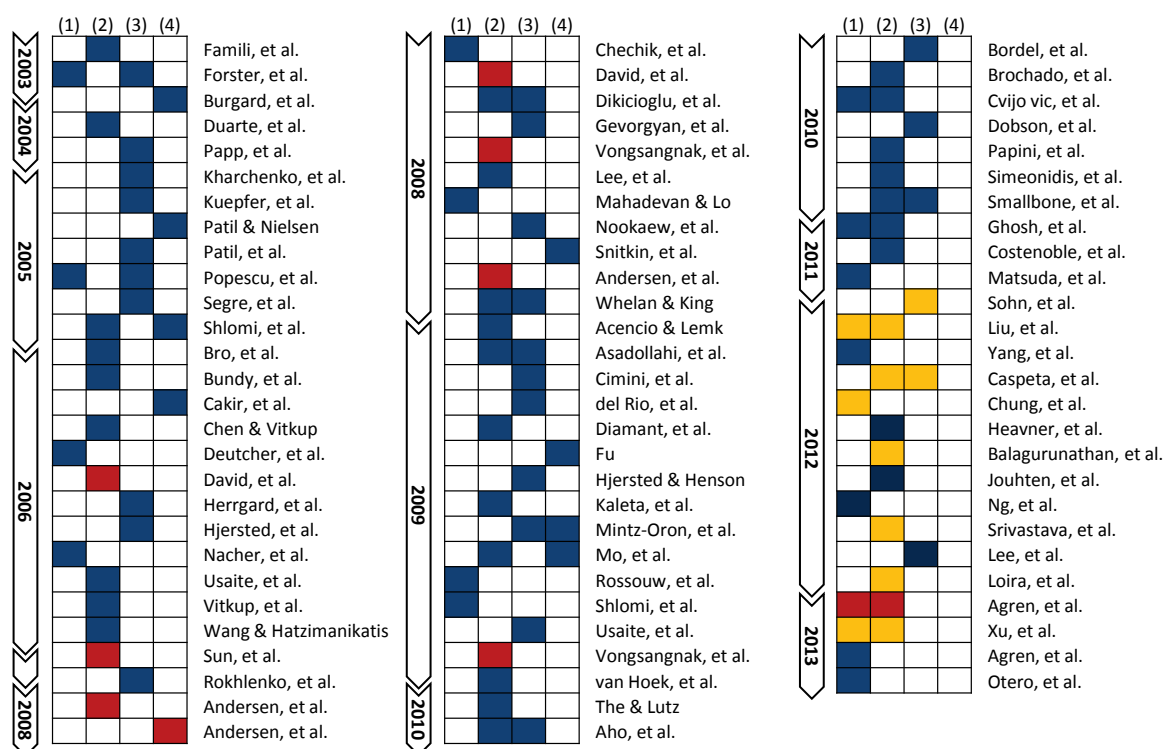


Figure 4-1. Overview of applications of fungal GEMs. The available publications which make use of GEMs to study fungal metabolism were categorized as follows: 1) guidance for metabolic engineering and strain improvement, 2) biological interpretation and discovery, 3) applications of novel computational framework, and 4) evolutionary elucidation. Blue: *S. cerevisiae*, Yellow: Non-*Saccharomyces* yeasts, Red: Filamentous fungi. Partly adapted from Osterlund *et al.* (2012).

In this section I present some of my work about genome-scale metabolic modelling applied to fungal metabolism. In **Paper I** we used genome-scale metabolic modelling to suggest knock-out targets in *S. cerevisiae* for the purpose of succinic acid production. The targets were then validated experimentally. In **Paper II** we developed an algorithm for identifying transcriptionally regulated reactions, with the aim of aiding metabolic engineering. In **Paper III** we developed a software suite for automatic reconstruction of GEMs and used this software to reconstruct a GEM for *Penicillium chrysogenum*. The model was then used to suggest metabolic engineering targets which could improve penicillin production.

4.1.1 Paper I: Genome-scale modelling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production

The US Department of Energy has identified succinic acid as an added-value chemical building block, with an estimated 15,000 t/year world-wide demand. The demand was predicted to grow to 270,000 t/year, representing a >2 billion USD annual market (Willke and Vorlop, 2004; McKinlay *et al.*, 2007). Succinic acid is currently used industrially in a variety of applications, such as surfactant, ion chelator, and as additive in the pharmaceutical and food industries. Currently, the only succinic acid derived from fermentation is that which is used in the food market, while the bulk is petrochemically produced from butane through maleic anhydride (McKinlay *et al.*, 2007).

Several biotechnology and metabolic engineering efforts have focused on overproduction of succinic acid in prokaryotes (Song and Lee, 2006). These bacterial hosts all grow at neutral pH, which results in secretion of the salt form, succinate, rather than the acid form. A costly acidification and precipitation step is then required in order to produce succinic acid, which is the desired product. This is a general concern when using bacterial cells for production of organic acids (Sauer *et al.*, 2008). One way to approach this issue could be to use the yeast *Saccharomyces cerevisiae* as a host. *S. cerevisiae* is a well-established, generally regarded as safe, and robust industrial production host. It is capable of growth on a variety of carbon sources, both aerobic and anaerobic, and it has a large pH operating range (3.0-6.0). Since it is capable of growth at such low pH it could be used to produce the desired acid form directly and thereby circumventing the acidification step. However, unlike the bacterial hosts described above succinate does not natively accumulate in *S. cerevisiae*.

Succinate is a TCA cycle intermediate produced from the oxidation of succinyl-CoA by succinyl-CoA synthase, or from isocitrate in a reaction catalysed by isocitrate lyase in the glyoxylate shunt. It is then further oxidized to fumarate by succinate dehydrogenase, resulting in the formation of FADH₂. Only limited work has been done on metabolic engineering of *S. cerevisiae* for production of succinate. The most successful work to date has been by Raab *et al.* (2010). They pursued an oxidative production route by a quadruple deletion of *SDH1*, *SDH2*, *IDP1*, and *IDH1*. This led to an interrupted TCA cycle and flux being redirected through the glyoxylate cycle instead, thereby resulting in succinate production. They could demonstrate a 0.07 C-mol/C-mol glucose succinate yield following this approach.

In **Paper I** we used FBA to propose gene deletion strategies for succinate overproduction. The main strategy was to optimize for biomass formation under constrained glucose uptake, and then observe the resulting succinate yield under a variety of conditions. Unlike the previously mentioned study, we focused primarily on anaerobic fermentation conditions, since it is a large advantage to be able to run industrial fermentations anaerobically. The three most promising single gene deletion strategies, identified under anaerobic glucose fermentation conditions, were experimentally evaluated. These strains were then physiologically and transcriptionally characterized in order to gain further knowledge into the C4 acid production by *S. cerevisiae*.

Figure 4-2 shows a comparison of the simulated specific growth rate and specific productivities compared to data generated by using the reference *S. cerevisiae* CEN.PK113-7D and BY4741 under aerobic and anaerobic glucose batch fermentations. As can be seen, the growth rate is predicted well, as is the ethanol production. However, the model is unable to capture the glycerol formation under aerobic conditions, and it underestimates the formation also under anaerobic conditions. This is due to the inability of the model to describe the Crabtree effect, as discussed earlier by Akesson *et al.* (2004).

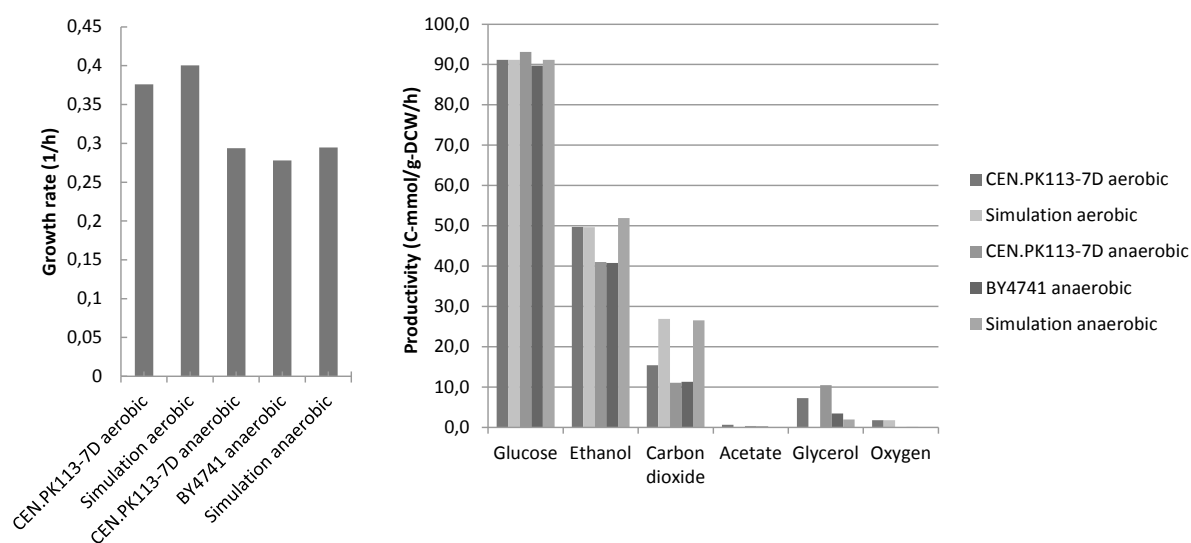


Figure 4-2. Comparison between experimental and simulated fermentation data. Comparison of the specific growth rate and specific productivities for simulated data and experimental data generated using the reference *S. cerevisiae* CEN.PK113-7D and BY4741 under aerobic and anaerobic glucose batch fermentations. For the conditions, simulation aerobic¹ and simulation anaerobic, the rO₂ was constrained to 1.8 mmol-O₂/g-DCW/h and 0.016 mmol-O₂/g-DCW/h, respectively. For aerobic experimental data the specific glucose uptake rate was 91.2 C-mmol/g-DCW/h for CEN.PK113-7D. For anaerobic experimental data the specific glucose uptake rate was 93.1 C-mmol/g-DCW/h for CEN.PK113-7D and 89.7 C-mmol/g-DCW/h for BY4741. For all simulation conditions the glucose uptake rate was constrained to 91.2 C-mmol/g-DCW/h.

Overproduction of succinate was simulated using the conditions previously described. Under aerobic conditions there were no single gene deletions which resulted in succinate production. Double gene deletions resulted only in minor improvement of succinate production. Figure 4-3 presents the top single gene deletions for succinate overproduction under anaerobic conditions. There is a small but significant predicted yield on substrate for the gene deletions $\Delta oac1$, $\Delta mdh1$, and $\Delta dic1$ (0.033 C-mol/C-mol glucose). The increase in succinate yield resulted in nearly no impact on growth rate (0.28h^{-1} vs. 0.30h^{-1} , single gene deletion vs. reference case simulation, respectively). The strains $\Delta oac1$, $\Delta mdh1$, and $\Delta dic1$ are viable null mutants, and their annotation is well known, encoding for an inner mitochondrial membrane transporter (Oac1p), malate dehydrogenase (Mdh1p), and an inner dicarboxylate mitochondrial transporter (Dic1p), respectively (Cherry et al., 1998).

In order to investigate if the single gene deletions identified in silico resulted in more succinate production, the corresponding strains were cultivated anaerobically in 2L well controlled fermenters. A comparative analysis between simulation and experimental results are presented in Figure 4-3. There is a fair agreement between model predictions and experimental data. Focusing more closely on the specific succinate productivity, the reference case, $\Delta mdh1$, and $\Delta oac1$ experimentally determined yields are significantly lower than expected based on model simulations. The $\Delta dic1$ case, however, demonstrated a significantly higher yield of succinate compared to the reference case (0.02 vs. 0.00 C-mol/C-mol glucose,

¹ The condition here referred to as “aerobic” corresponds to the condition referred to as “semi-aerobic” in **Paper I**. The original “aerobic” was for totally unconstrained oxygen uptake; a condition which badly represented the experimental data and which therefore is not discussed further here.

Adic1 vs. reference, respectively), and was in line with the in silico prediction (0.02 vs. 0.03 C-mol/C-mol glucose, *Adic1* experimental vs. *Adic1* simulation, respectively). This represents a significant improvement in succinate productivity based exclusively on a novel in silico prediction.

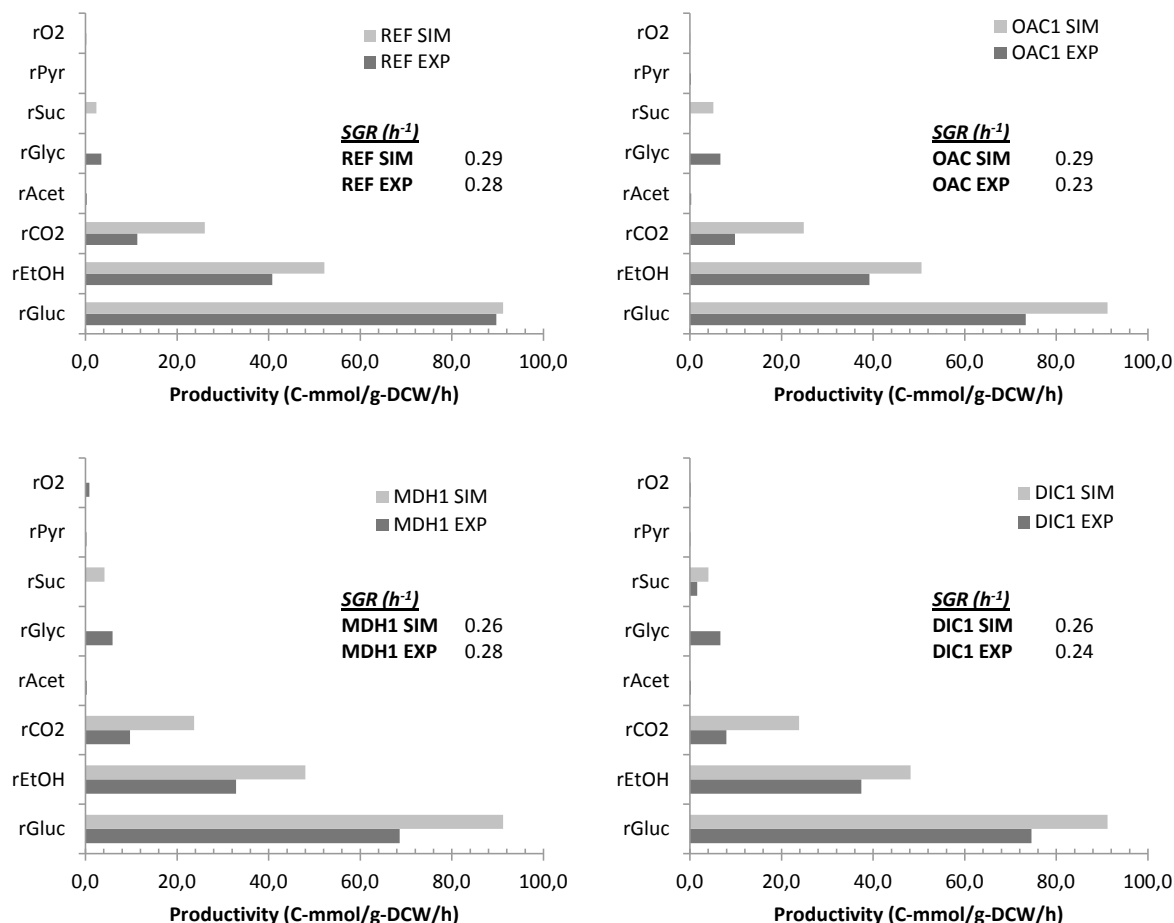


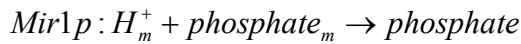
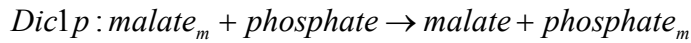
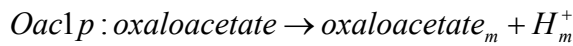
Figure 4-3. Experimental and simulated data for reference strain, *Δoac1*, *Δmdh1*, and *Δdic1* strains. Summary of the specific growth rate (SGR) and specific consumption/productivity for major products (glucose, ethanol, carbon dioxide, acetate, glycerol, succinate, pyruvate, and oxygen) for both experimentally determined data of anaerobic batch glucose fermentations and the corresponding anaerobic simulation data. The panels show the BY4741 reference strain and the single gene deletion strains *Δmdh1*, *Δdic1*, and *Δoac1*. In general, the simulated growth rates are in very good agreement with the experimental values. The *Δdic1* mutant results in some succinate production.

To gain further insight into the performance of each strain, DNA microarray profiling was performed for the anaerobic batch glucose fermentations. The complete list of differentially expressed genes² for *Δdic1* and *Δmdh1* were submitted for metabolic pathway annotation using the SGD Pathway Expression Viewer and Reactome databases (Paley and Karp, 2006; Matthews *et al.*, 2009). Only a rather small number of metabolic genes were identified in

² The number of differentially expressed genes for the *Δoac1* strain compared to the reference strain was very low, and consequently suggests that deletion of *Δoac1* causes virtually no transcriptional, and consequently, physiological differences compared to the reference BY4741 strain. No further analysis of the transcriptional data was performed for this strain.

Δdic1 and *Δmdh1* compared to the reference; a total of 10 and 20 genes respectively. Perhaps more striking is that there is an overlap of 9 metabolic genes between *Δdic1* and *Δmdh1*. The only differentially expressed gene present in the *Δdic1* condition, not present in the *Δmdh1* condition, is *DIC1*.

The identified metabolic engineering strategies through *Δdic1*, *Δmdh1*, and *Δoac1*, suggest a common mechanism. Mitochondrial redox balance must be maintained, and while respiratory metabolic activity under anaerobic conditions is reduced compared to aerobic conditions, some activity is required to support glutamate/glutamine metabolism from α -ketoglutarate (Camarasa *et al.*, 2003; Camarasa *et al.*, 2007). This leads to the accumulation of mitochondrial NADH. During anaerobic metabolism, NAD⁺ regeneration occurs via the following pathways according to our simulations (where the subscript m denotes mitochondrial):

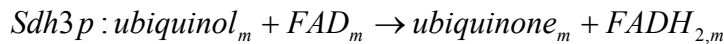


Net reaction stoichiometry: $oxaloacetate + NADH_m \rightarrow malate + NAD_m^+$

In the cytosol, malate is then converted to oxaloacetate, and the resulting NADH is converted to NAD⁺, resulting in the production of glycerol. The *Δdic1* strategy, relying on deletion of the mitochondrial dicarboxylate carrier *DIC1*, catalyses the following transport reaction:



Assuming *DIC1* deletion, then the resulting simulated pathway is:



Net reaction stoichiometry: $fumarate + NADH_m \rightarrow succinate + NAD_m^+$

The *Δdic1* strategy relies heavily on the subcellular localization and function of *Frds1p*, soluble mitochondrial fumarate reductase, which continues to be poorly understood. However, recent work has suggested that a double deletion *S. cerevisiae* mutant, *Δosm1Δfrds1*, failed to grow under batch glucose anaerobic conditions. During anaerobic growth, *FRDS1* expression in the wild-type was two to eight times higher than that of *OSM1*, suggesting that formation of succinate is strictly required for the re-oxidation of FADH₂ and that its expression may be oxygen-regulated (Camarasa *et al.*, 2007). There was a strong upregulation of *CYC1* in both the *Δdic1* and *Δmdh1* mutants. *CYC1* facilitates electron transfer from ubiquinone cytochrome C oxidoreductase to cytochrome C oxidase. This direction, which is the normal oxidative route and ends in reduction of O₂, would not be possible under fully anaerobic conditions. The upregulation can therefore be viewed as a coping strategy to deal with the stress of redox imbalance. Deletion of *CYC1* could therefore be a way to ensure that all NAD⁺ regeneration is coupled to succinate production. The strategies proposed here rely on the capacity for reductive TCA cycle activity under anaerobic conditions, and more specifically, the catalysis

of fumarate to succinate via fumarate reductase. There is evidence suggesting that *S. cerevisiae* can exhibit this metabolic state (Camarasa *et al.*, 2003; Camarasa *et al.*, 2007).

In conclusion, a GEM was used to predict single and double gene deletion strategies which could lead to increased succinate production under a variety of conditions. Three of these strategies, all utilizing anaerobic fermentation conditions, were validated *in vivo* and one, *Adic1*, was identified to lead to a significant improvement in succinate yield, in close agreement with the model prediction. However, the yield was not as high as what was reported by Raab *et al.* (2010) using a quadruple gene deletion strategy and the oxidative route (0.02 C-mol/C-mol glucose vs. 0.07 C-mol/C-mol glucose). Furthermore, physiological characterization and transcriptome analysis were used to aid in interpretation of the simulations and provide a mechanistic explanation of the results. The proposed mechanisms rely heavily on compartmental transport reactions and mitochondrial redox balancing. Transcriptional profiling suggests that succinate formation is coupled to mitochondrial redox balancing, and more specifically, reductive TCA cycle activity. While far from industrial titers, this proof-of-concept suggests that *in silico* predictions coupled with experimental validation can be used to identify novel and non-intuitive metabolic engineering strategies.

4.1.2 Paper II: Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes

Metabolic fluxes are the result of a complex interplay involving metabolite concentrations, enzyme kinetics, gene expression, and translational regulation. Due to these multiple layers of regulation, there is generally not a clear correlation between mRNA levels, enzyme levels, and the metabolic fluxes (Akesson *et al.*, 2004). Only a rather small fraction of enzymes show a clear positive correlation between their transcription levels and the rates of the reactions that they catalyse. These reactions are said to have transcriptional regulation.

In **Paper II** we propose a method which allows for identification of enzymes that show transcriptional regulation, and therefore represent suitable targets for metabolic engineering (up- or downregulation). The method is based on integration of gene expression data with flux data by transforming quantitative flux data into a genome-scale set of statistical scores analogous to those obtained from transcriptional profiling. This works by constraining a set of experimentally determined exchange fluxes in a GEM for the organism being studied. This is then done for each of the studied conditions, or for each of the strains investigated. A random sampling algorithm is then used to generate a set of internal flux distributions which all satisfy the experimentally determined exchange fluxes. By this approach it is possible to obtain means and standard deviations for each flux in the GEM, and from there it is then possible to derive p-values for the significance of flux change between conditions (Mo *et al.*, 2009; Schellenberger and Palsson, 2009). These values can then be compared to the significance of change in gene expression for the corresponding enzymes. The comparison of flux change and gene expression allows for identification of enzymes showing a significant correlation between flux change and expression change (transcriptional regulation) as well as reactions whose flux change is likely to be driven only by changes in the metabolite concentrations (metabolic regulation). This workflow is described in more detail in Figure 4-4.

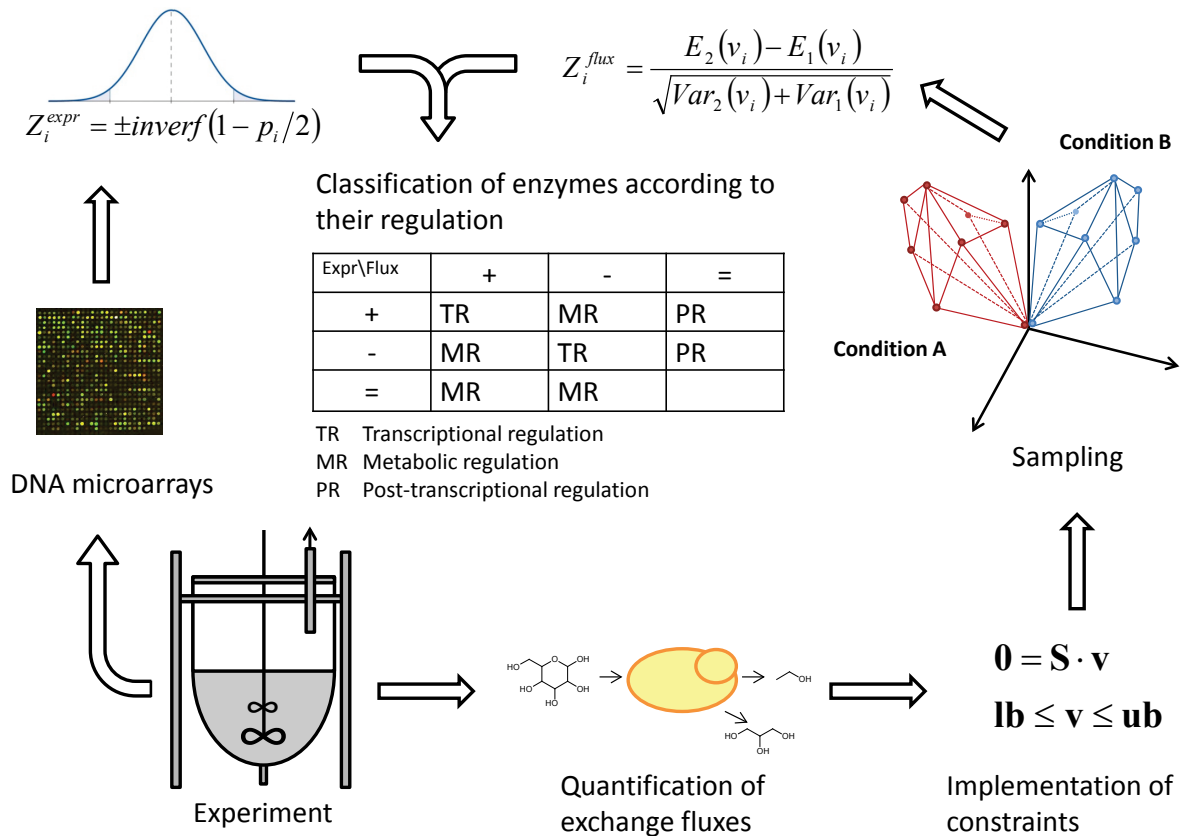


Figure 4-4. Workflow for identification of transcriptionally regulated reactions. The figure illustrates how the method can be applied to identify transcriptionally regulated reactions. Two fermentations experiments are performed; one for the test case and one reference. Gene expression levels are measured for the two cases using, for example, DNA microarrays. A statistical test is used for each of the genes to calculate the Z-score for differential expression between the cases. In parallel to this, key exchange fluxes are quantified. The corresponding exchange fluxes in two GEMs are then constrained to what was seen experimentally for each of the cases. Due to the high dimensionality of GEMs, and the redundancy in cell metabolism, there are many different internal flux distributions which all result in the measured exchange fluxes. A random sampling algorithm is applied to sample many such solutions from each of the two models. A Z-score for differential flux can then be estimated for each of the reactions from the difference in average flux divided by the square root of the sum of the variances. The Z-scores for differential expression/flux are then transformed into probabilities of change by using the cumulative Gaussian distribution. These probabilities are then used to calculate probabilities of having correlated gene expression and flux changes for the corresponding reaction. The enzymes in the network can then be classified as: 1) enzymes that have a significantly correlated change both in flux and expression level (reactions showing transcriptional regulation); 2) enzymes that show a significant change in expression but not in flux (post-transcriptional regulation); 3) enzymes that show significant changes in flux but not a change in expression (metabolic regulation).

To evaluate our method we used data for *S. cerevisiae*. Data from growth on four different carbon sources (glucose, maltose, ethanol and acetate) in chemostat cultures and from five deletion mutants grown in batch cultures were used. This summary will only focus on the different carbon sources, and not on the deletion mutants. Table 4-2 shows the top scoring enzymes in the different categories for each of the carbon source shifts.

Table 4-2. Top scoring enzymes for transcriptional, post-transcriptional and metabolic regulation for changes in carbon source. The 10 top scoring enzymes in each group are shown (or fewer when less than 10 enzymes had a probability larger than 0.95).

Carbon source shift	Enzymes showing transcriptional regulation	Enzymes showing post-transcriptional regulation	Enzymes showing metabolic regulation
Glucose/Maltose	<ul style="list-style-type: none"> • α-glucosidase MAL32 • Low-affinity glucose transporter HXT4 	<ul style="list-style-type: none"> • Mevalonate kinase • Inosine-5'-monophosphate dehydrogenase IMD2 • Asparagine synthetase 1 • Glycerol-3-phosphatase 1 • Uncharacterized deaminase • Nicotinate-nucleotide pyrophosphorylase • Mevalonate kinase • Nicotinate-nucleotide pyrophosphorylase • Glycerol-3-phosphate dehydrogenase 1 	<ul style="list-style-type: none"> • Acetate transport via proton symport
Glucose/Ethanol	<ul style="list-style-type: none"> • Phosphoenolpyruvate carboxykinase • Fructose-1,6-bisphosphatase • Isocitrate lyase • Malate dehydrogenase • Citrate synthase • Ribose-5-phosphate isomerise • Low-affinity glucose transporter HXT4 • External NADH-ubiquinone oxidoreductase 2 • Glucose-6-phosphate isomerase 	<ul style="list-style-type: none"> • Formate dehydrogenase 2 • ATP-NADH kinase • Sulfate permease 1 • Formate dehydrogenase 1 • Dicarboxylate transporter • NADP-specific glutamate dehydrogenase 2 • Uncharacterized deaminase • phosphogluconolactonase 3 • 6-phosphofructo-2-kinase 2 • Nucleoside diphosphate kinase 	<ul style="list-style-type: none"> • Fructose-bisphosphate aldolase • Triosephosphate isomerase • Pyruvate dehydrogenase E1 subunit alpha • α-ketoglutarate dehydrogenase • Succinyl-CoA ligase subunit beta • Malate synthase 2 • Glucose-6-phosphate 1-dehydrogenase • Cytochrome b-c1 subunit Rieske • Adenylate kinase
Glucose/Acetate	<ul style="list-style-type: none"> • Fumarate hydratase • Phosphoenolpyruvate carboxykinase • Fructose-1,6-bisphosphatase • Isocitrate dehydrogenase • Succinate-semialdehyde dehydrogenase • Citrate synthase • Isocitrate dehydrogenase subunit 1 • Pyruvate kinase 2 • Low-affinity glucose transporter HXT4 	<ul style="list-style-type: none"> • Phospho-2-keto-3-deoxyheptonate aldolase • Ribonucleoside-diphosphate reductase large chain 1 • 6-phosphofructo-2-kinase 1 • Glutamine-dependent NAD synthetase • Ribose-phosphate pyrophosphokinase 4 • ATP-dependent permease AUS1 • Fructose-2,6-bisphosphatase • Nicotinate-nucleotide pyrophosphorylase • Squalene monoxygenase 	<ul style="list-style-type: none"> • Fructose-bisphosphate aldolase • Triosephosphate isomerase • Ribose-5-phosphate isomerase • Inorganic pyrophosphatase • Adenylate kinase • Glutamate decarboxylase • 4-aminobutyrate aminotransferase • Tricarboxylate transport protein • Prephenate dehydrogenase

In the glucose to maltose transition, only two enzymes showed transcriptional change correlated with their flux. The α -glucosidase Mal32p, which is responsible for the breakdown of maltose into glucose, was upregulated and the glucose transporter Hxt4p was

downregulated. Only very minor adjustments could be seen in terms of fluxes, and only enzymes directly related with the substrate uptake and utilization were detected. The changes in gene expression were also few and only 11 metabolic enzymes were significantly perturbed (without significant flux changes).

The glucose to ethanol and the glucose to acetate shifts showed much more widespread changes in flux and expression. They therefore represent more interesting cases studies. In the glucose-ethanol transition 19 enzymes showed transcriptional regulation and 22 enzymes changed in expression but not in flux. For the glucose-acetate shift the same numbers were 33 and 23, respectively. Among the enzymes showing transcriptional regulation, 14 were shared between the glucose-ethanol and glucose-acetate transitions. However, there was no overlap between the sets of enzymes which do not change in flux. Metabolic regulation was observed in 21 reactions for each case, out of which 8 overlap.

The enzymes with transcriptional regulation clearly show a downregulation of enzymes involved in the uptake and utilization of glucose (e.g. glucose transporter Hxt4p or hexokinase 2) and the upregulation of enzymes involved in gluconeogenesis (e.g. fructose-1,6-biphosphatase) or the TCA cycle (e.g. succinate dehydrogenase or citrate synthase). Acetyl-CoA synthetase 2, responsible for providing acetyl-CoA to the TCA cycle, is also identified as transcriptionally upregulated, as well together with ATP synthetase and external NADH-ubiquinone oxidoreductase 2, which provide the necessary NAD^+ needed for oxidation of ethanol or acetate in the cytoplasm and thereby maintaining the redox balance in the cell. Isocitrate lyase, a key component of the glyoxylate cycle, is also transcriptionally upregulated. This allows for net formation of malate, which can then be further converted to phosphoenolpyruvate (via oxaloacetate) in order to fuel gluconeogenesis. All these changes in fluxes are consistent with what is known about the changes in metabolism between growth on glucose to C2 carbon sources like ethanol and acetate. Interestingly, not all reactions associated with this shift in flux distributions are transcriptionally regulated. Rather, the cell has selected a small number of key reactions to regulate at the transcriptional level.

We also performed an enrichment test in order to compare the transcription factors involved in the expression of the enzymes classified as having transcriptional regulation and the enzymes showing changes in expression but not in flux. We identified three transcription factors which were strongly overrepresented in the metabolic genes showing transcriptional regulation. In the glucose-ethanol transition, the transcription factors Gcr1p and Gcr2p both appeared in 11 transcriptionally regulated genes and in none of the other genes, whereas the transcription factor Hap4p appeared in 11 transcriptionally regulated genes and 5 of the other regulated genes. For the glucose-acetate transition these numbers were 15-0, 11-0 and 15-0 for the same transcription factors. This can be interpreted as if certain transcription factors are particularly involved in the transcriptional regulation of metabolic fluxes. This implies that there is global regulation of major flux alterations, which is in agreement with what has been shown experimentally for galactose metabolism (Ostergaard *et al.*, 2001).

The top scoring metabolically regulated reactions, both for the glucose-ethanol and glucose-acetate shifts, are fructose biphosphate aldolase and triosephosphate isomerase. These reactions are known to operate close to their equilibrium and they are therefore sensitive to changes in the metabolite pools, which is consistent with metabolic regulation of the fluxes. In the two shifts the direction of these reactions is inverted. This can only be explained by a decrease in the fructose-1,6-diphosphate pool and an increase in the glyceraldehyde-3-phosphate and dihydroxyacetone pools. This hypothesis is supported by the fact that in

chemostat cultures there is not found to be any correlation between the glycolytic flux and the expression of the genes encoding for these two enzymes (Daran-Lapujade *et al.*, 2007).

In conclusion, the combined use of random sampling in GEMs and expression data allows for global identification of reactions which are either transcriptionally or metabolically regulated. The reactions exhibiting transcriptional regulation form a set of putative metabolic engineering targets, where enzyme overexpression or downregulation is likely to influence the flux through these reactions. The reactions exhibiting metabolic regulation points to parts of metabolism where the metabolite pools are possibly increasing or decreasing in connection with transcriptional changes, and thereby counteracting possible changes in enzyme concentration. This knowledge can be used to identify whether one should target changes in enzyme concentration (v_{\max} changes), e.g. through overexpression, or changes in enzyme affinity (K_m changes), e.g. through expression of heterologous enzymes, in order to alter the fluxes.

4.1.3 Paper III: The RAVEN Toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*

In **Paper III** we developed a software suite named the RAVEN Toolbox (Reconstruction, Analysis, and Visualization of Metabolic Networks), which aims at automating parts of the GEM reconstruction process in order to allow for faster and easier reconstruction of high-quality GEMs. The software was then used for reconstructing a model of the ascomycetous fungi *Penicillium chrysogenum*, the organism used for industrial production of penicillin, and an important microbial cell factory. The resulting model was validated in an extensive literature survey and by comparison to fermentation data. The model was then used together with the algorithm presented in **Paper II** in order to identify transcriptionally regulated metabolic bottlenecks in penicillin fermentations. These bottlenecks can then be targets for metabolic engineering.

The RAVEN Toolbox has three main foci: 1) automatic reconstruction of GEMs based on protein homology, 2) network analysis, modelling and interpretation of simulation results, 3) visualization of GEMs using pre-drawn metabolic network maps. Figure 4-5 summarizes the capabilities of the RAVEN Toolbox.

Previously published GEMs represent a solid basis for metabolic reconstruction of models for new organisms, in particular if the organisms are closely related and therefore share many metabolic capabilities. The main advantage of using existing models compared to reaction databases, such as KEGG or BRENDA (Schomburg *et al.*, 2002), is that they contain information which can be difficult to obtain in an automated manner, in particular directionality and compartmentalization. GEMs are also typically constructed for modelling purposes, which is not the case for reaction databases. The downside is that only reactions present in the template models can be included. The RAVEN Toolbox therefore contains two approaches for automatic generation of draft models; one which relies on the metabolic functions represented in previously published models, and one method which uses the KEGG database for automatic identification of new metabolic functions that are not included in the published models.

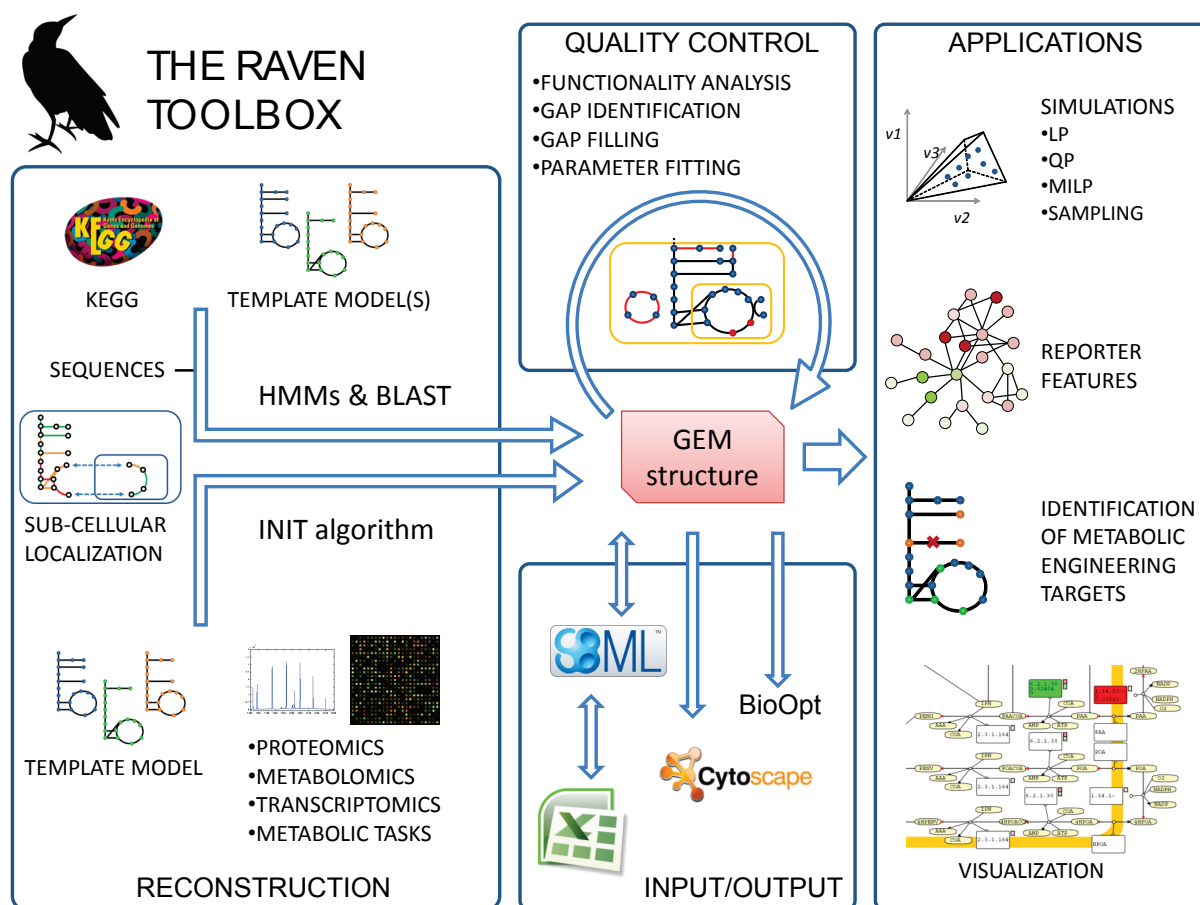


Figure 4-5. The RAVEN Toolbox. The software allows for reconstruction of GEMs based on template models or on the KEGG database. Subcellular localizations of reactions can be estimated based on signal peptides and other characteristics of the catalysing enzymes. The resulting models can be exported to a number of formats, or they can be used for various types of simulations. The RAVEN Toolbox has a strong focus on quality control. Visualization of simulation result and/or integration of other types of data can be performed by overlaying information on pre-drawn metabolic maps. The software also implements the INIT algorithm, which is a powerful approach for reconstruction of tissue-specific models (**Paper IV**). Figure taken from Agren *et al.* (2013a).

In the first approach the software generates a draft model based on protein orthology using bi-directional BLASTp (Altschul *et al.*, 1990). The second approach is also based on protein homology but requires no template models. Instead it relies on the information on protein sequences and on the assigned metabolic reactions that is available in the KEGG database. The method makes use of the KEGG Orthology (KO) IDs, which are manually annotated sets of genes that encode some specified metabolic function. Each KO is associated with a number of metabolic reactions. The aim of the method is then to assign genes to these KOs based on the consensus protein sequence. This works by generating a hidden Markov model based on the sequences for each KO using HMMER (Eddy, 1998). The final step is the querying of the set of HMMs with the protein sequences of the organism of interest. If a gene has a significant match to one KO, the reactions associated to that KO are added to the model together with the corresponding gene.

The approach proposed above will facilitate and accelerate the generation of a draft metabolic network reconstruction. However, the automated reconstruction can lead to some loss of control compared to a stricter manual, bottom-up approach. It is therefore important to identify and fill gaps in the model to ensure that the network is functioning as required. The RAVEN Toolbox therefore contains a number of novel methods to support the gap filling

process. Table 4-3 shows a comparison between the RAVEN Toolbox and some other software with similar functionalities.

Table 4-3. Comparison between the RAVEN Toolbox and some other software for automated GEM reconstruction.

	RAVEN	Model SEED ^a	AUTOGRAPH ^b	IdentiCS ^c	GEM System ^d
Includes general network	X	X		X	X
Generates functional models	X	X			
Assigns subcellular localization	X				
Can use user defined models	X		X		
Integrates gap filling	X	X			X
Offline software	X			X	
Includes visualization	X			X	X
Gene prediction		X		X	X

^aHenry *et al.* (2010); ^bNotebaart *et al.* (2006); ^cSun and Zeng (2004); ^dArakawa *et al.* (2006). Taken from Agren *et al.* (2013a).

The *P. chrysogenum* metabolic network was reconstructed based on a combination of the automated reconstruction approaches in the RAVEN Toolbox, manual curation, and an extensive bibliomic survey. Three GEMs for closely related filamentous fungi, *Aspergillus nidulans* iHD666 (David *et al.*, 2006), *A. niger* iMA871 (Andersen *et al.*, 2008), and *A. oryzae* iWV1314 (Vongsangnak *et al.*, 2008), were used as template models for the reconstruction of the *P. chrysogenum* model. The model comprises 1471 unique metabolic reactions in four subcellular compartments; extracellular, cytosolic, mitochondrial, and peroxisomal. 1006 ORFs are associated to the reactions, 89 of which participate in one of 35 protein complexes. In parallel to the automatic reconstruction, an extensive literature study was performed. In total 440 cited articles provide experimental evidence for the majority of the reactions. The model was validated with respect to 76 metabolic functions known to occur in *P. chrysogenum*.

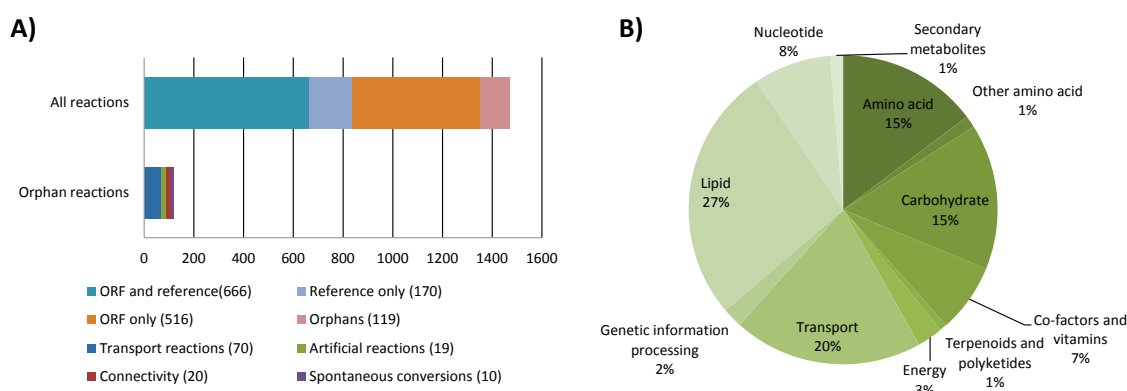


Figure 4-6. Evidence level for the *P. chrysogenum* metabolic network. **A)** Properties of the reconstructed network. The top bar shows the support for the 1471 unique reactions (not counting exchange reactions) sorted by the type of evidence. The bottom bar shows the orphan reactions; reactions inferred without supporting ORFs or literature references. **B)** ORF classification. The ORFs in the model are classified into broad groups based on KEGG classification.

Figure 4-6 summarises the literature support for the reactions in the model and shows a classification of the ORFs in the model based on KEGG Pathways. To illustrate the metabolic network, and to aid in interpretation of gene expression data and simulation results, a map of

the full model was drawn and annotated so as to be compatible with the visualization functions in the RAVEN Toolbox.

We then used the GEM in a study of penicillin yields and in particular the relative importance of ATP and NADPH provision during penicillin production. In a second study we showed how the model can be used to integrate fermentation data with transcriptome data using the method developed in **Paper II**.

The industrial *P. chrysogenum* strains have been subjected to 50 years of directed evolution to increase the yields and titers of penicillin, with great cost reduction and productivity gain, but the yields are still far from the theoretical maximum (Thykaer and Nielsen, 2003). Penicillin production is associated with an increased requirement of energy in the form of ATP; in the condensation of the three precursor amino acids to form the tripeptide ACV; in the reduction of sulfate; and when a side chain (the precursor molecule which is supplied to the media and which differs depending on the type of penicillin produced) is activated by ligation to coenzyme A. Penicillin production is also associated with a large requirement of NADPH, primarily needed for the reduction of sulfate, but also in the biosynthesis of valine and homoserine from α -ketobutyrate. Elucidating the impact increased ATP requirements have compared to the NADPH requirements is useful when choosing among possible metabolic engineering strategies.

The maximum theoretical yield of penicillin on glucose with sulfate as the sulfur source was calculated to be 0.42 mol penicillin/mol glucose using the reconstructed GEM. To investigate the effect of ATP, an artificial reaction was included that allowed for ATP production from ADP without any energetic costs. This resulted in a yield of 0.52 mol penicillin/mol glucose, using sulfate as the sulfur source. The conclusion is that ATP availability has a relatively small effect on the yield. The shadow prices (how much the penicillin production can increase if the availability of a metabolite were to increase by a small amount) were calculated to be 0.015 mol penicillin/mol ATP, 0.040 mol penicillin/mol NADPH, and 0.037 mol penicillin/mol NADH.

NADPH and NADH are similar when it comes to energy content, but have different roles in the metabolism, where NADPH serves primarily anabolic roles and NADH primarily catabolic roles. NADPH is mainly produced in the pentose phosphate pathway, which makes NADPH somewhat more energetically expensive to regenerate compared to NADH. In order to investigate the relative importance of NADH and NADPH an artificial reaction was included that allowed for production of NADPH from NADH to simulate a potential increase of the NADPH availability. Simulations were then carried out maximizing first for growth and then for penicillin production. The resulting flux through the artificial reaction was 8.5 times larger when maximizing for penicillin than when maximizing for growth. This demonstrates that the cells will have a much higher NADPH demand at high penicillin yields compared to normal growth conditions. Redirecting a higher flux through the pentose phosphate pathway and/or introducing NADH-dependent versions of NADPH-consuming enzymes could therefore be potential metabolic engineering strategies for achieving higher penicillin yields.

For the direct identification of possible metabolic engineering targets a gene deletion analysis was performed by searching for sets of gene deletions that resulted in an increased yield of penicillin, and which would stoichiometrically couple penicillin production to growth. This was performed using FBA, and combinations of up to three gene deletions were evaluated (MoMA was also applied and gave similar results). The only targets which could be identified were the deletion of any of the genes responsible for breakdown of phenylacetic acid

(homogentisate 1,2-dioxygenase, maleylacetoacetate isomerase, or fumarylacetoacetase). Deletion of any of these genes resulted in a predicted 21% increase in penicillin production.

Penicillin biosynthesis

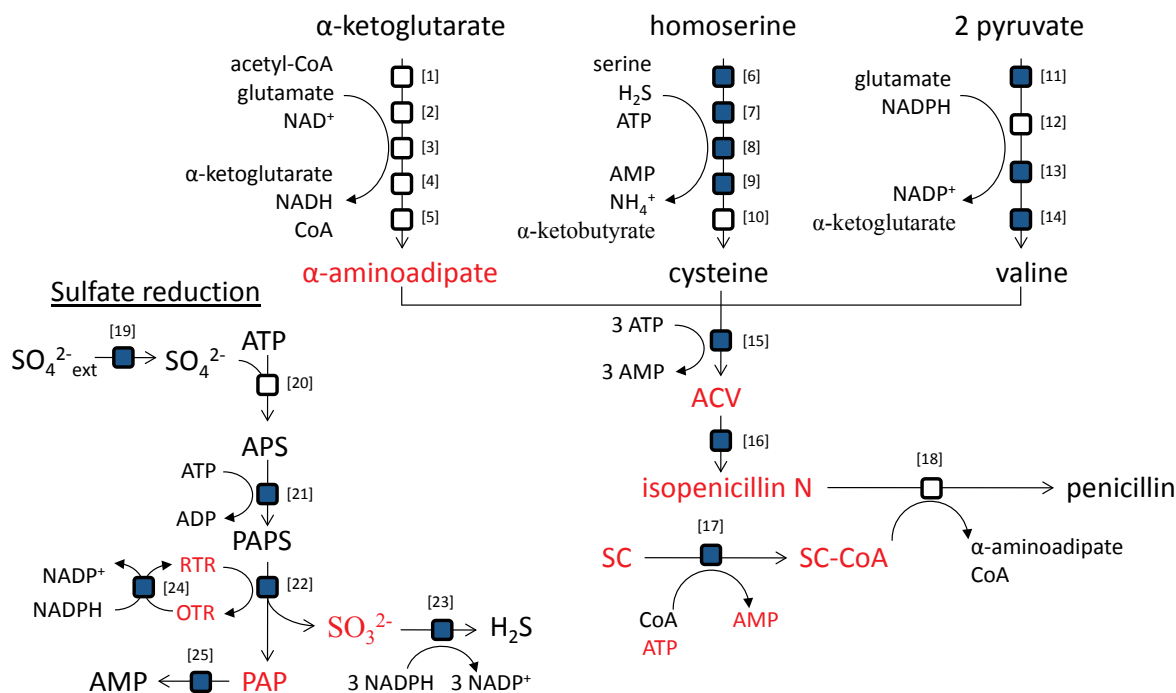


Figure 4-7. Integrative analysis of a high and a low producing strain. Depicts synthesis pathways of penicillin and important precursors. Blue boxes correspond to reactions identified as being transcriptionally controlled and upregulated by the algorithm presented in **Paper II**. Metabolites around which significant transcriptional changes occur compared to a low producing strain are coloured red. SC: side chain (e. g. the precursor molecule phenylacetic acid). The biosynthesis of penicillin starts with the condensation of the three amino acids α -aminoadipate (an intermediate in the L-lysine biosynthesis pathway), L-cysteine, and L-valine to form the tripeptide ACV. ACV is further converted to isopenicillin N. For the industrially relevant types of penicillin a side-chain is supplied to the media. This side-chain is activated by ligation to coenzyme A. In the last step of penicillin biosynthesis an acyl transferase exchanges the α -aminoadipate moiety of isopenicillin N with the side-chain, thereby generating penicillin and regenerating α -aminoadipate. Since L-cysteine is a sulfur-containing amino acid penicillin production is also tightly associated with sulfur metabolism. [1] homogitrate synthase; [2] homogitrate dehydrase; [3] homoaconitate hydrase; [4] homoisocitrate dehydrogenase; [5] α -aminoadipate aminotransferase; [6] homoserine transacetylase; [7] O-acetylhomoserine sulfhydrylase; [8] cystathione- β -synthase; [9] cystathione- γ -lyase; [10] acetate CoA ligase; [11] acetolactate synthase; [12] ketol-acid reductoisomerase; [13] dihydroxy acid dehydrase; [14] branched chain amino acid transferase; [15] ACV synthase; [16] isopenicillin N synthase; [17] acyl CoA ligase (side chain dependent, reaction is for phenylacetate CoA ligase); [18] isopenicillin N N-acyltransferase; [19] sulfate permease; [20] sulfate adenyl transferase; [21] adenyl sulfate kinase; [22] phosphoadenyl sulfate reductase; [23] sulfite reductase; [24] thioredoxin reductase; [25] 3'(2'),5'-bisphosphate nucleotidase.

In order to identify transcriptionally regulated metabolic bottlenecks we applied the method from **Paper II** and compared the high producing industrial strain DS17690, which has been developed by DSM, and the low producing reference strain Wis 54-1255 (Harris *et al.*, 2006). Flux data and gene expression levels for aerobic, glucose-limited chemostat fermentation of DS17690 and Wis 54-1255 were used as input to the algorithm (van den Berg *et al.*, 2008). 58 fluxes were found to be significantly changed between the high and low production strains ($p < 0.05$) and 612 genes were differentially expressed ($p < 0.005$). 36 reactions were identified as having significantly higher flux and upregulated genes, i.e. they are likely to have transcriptional regulation of their fluxes. Figure 4-7 shows some of the most important reactions in penicillin biosynthesis together with the responsible enzymes. Reactions that

were identified as being transcriptionally controlled and upregulated are highlighted. In addition, the Reporter Metabolites algorithm was used to identify metabolites around which significant transcriptional changes occurred (Patil and Nielsen, 2005). These metabolites are highlighted in Figure 4-7 as well.

As can be seen in Figure 4-7, a large proportion of the reactions identified as being a transcriptionally controlled are directly involved in penicillin metabolism (15 out of 38). This indicates that the capabilities of the industrial strain to produce penicillin to a large extent depend on the reactions closely related to penicillin metabolism, rather than more peripheral effects. Among these reactions are many of the reactions responsible for the synthesis of the amino acids that are precursors for ACV, as well as the two penicillin producing reactions isopenicillin N synthase and ACV synthase. This is consistent with a study on the gene copy-number effect on penicillin production (Theilgaard *et al.*, 2001). The phenylacetate:CoA ligase is high ranking but the acyl-CoA:isopenicillin N acyltransferase is absent, which is consistent with measurements of high activities of this enzyme and the low flux control estimated for this enzyme (Jorgensen *et al.*, 1995b; Nielsen and Jorgensen, 1995). Several of the reactions involved in sulfate reduction are present, as well as the sulfate permease. It is interesting to note that none of the reactions in the pentose phosphate pathway are identified even though there is an increased demand for NADPH.

We also found that the pathway from α -ketobutyrate to succinate is identified to have both increased flux and increased gene expression. α -ketobutyrate is a by-product of cysteine production via the transsulfuration pathway, and it is used for isoleucine biosynthesis. Under normal growth conditions the demand for cysteine is less than that for isoleucine, meaning that all α -ketobutyrate is converted into isoleucine. However, during high-level penicillin production the cysteine production far exceeds the need for isoleucine, requiring an alternative route for α -ketobutyrate consumption. This route involves the decarboxylation of α -ketobutyrate to yield propionyl-CoA, which then goes into the methylcitrate pathway, eventually resulting in succinate (Jorgensen *et al.*, 1995a). Several of the reactions in this pathway are identified as transcriptionally controlled by the algorithm (2-methylcitrate synthase, 2-methylcitrate dehydratase, 2-methylisocitrate dehydratase, and methylisocitrate lyase). This finding strongly supports that the transsulfuration pathway is the dominating pathway for cysteine biosynthesis, even though the enzymes for the energetically more efficient direct sulfhydrylation pathway have been identified in *P. chrysogenum* (Ostergaard *et al.*, 1998).

In conclusion, the RAVEN Toolbox enables rapid reconstruction of high-quality models, which is not possible using a traditional manual approach. It is the first software which can be used to drive the model reconstruction process and which also contains extensive functions for simulations and analysis of results. The software and workflow was validated by reconstructing the first GEM for the industrially important fungi *P. chrysogenum*. This GEM was then used to gain novel insights in penicillin biosynthesis, and for suggesting metabolic engineering targets for increased penicillin yield.

4.2 GEMs applied to human health and disease

Abnormal metabolic states are at the origin of many diseases, such as diabetes, hypertension, heart diseases and cancer. Cancer and coronary diseases are the two main causes of death in the developed countries. It is expected that by 2030 close to 200 million persons (33% of the

total population) will be obese in the EU alone, and many of these will have one or more of the following co-morbidities: diabetes, hypertension, heart disease and increased risk of cancer. The direct (medical treatment) and indirect (inability to work) costs are estimated to amount to more than €100 billion per year (Rokholm *et al.*, 2010; Caveney *et al.*, 2011). These are all complex diseases, in the sense that they are the result of a large number of interacting molecular factors. This speaks in favour of taking a holistic approach rather than a more traditional reductionist one. Genome-scale metabolic modelling can therefore be a promising methodology for study of this type of diseases, but there are still obstacles to overcome.

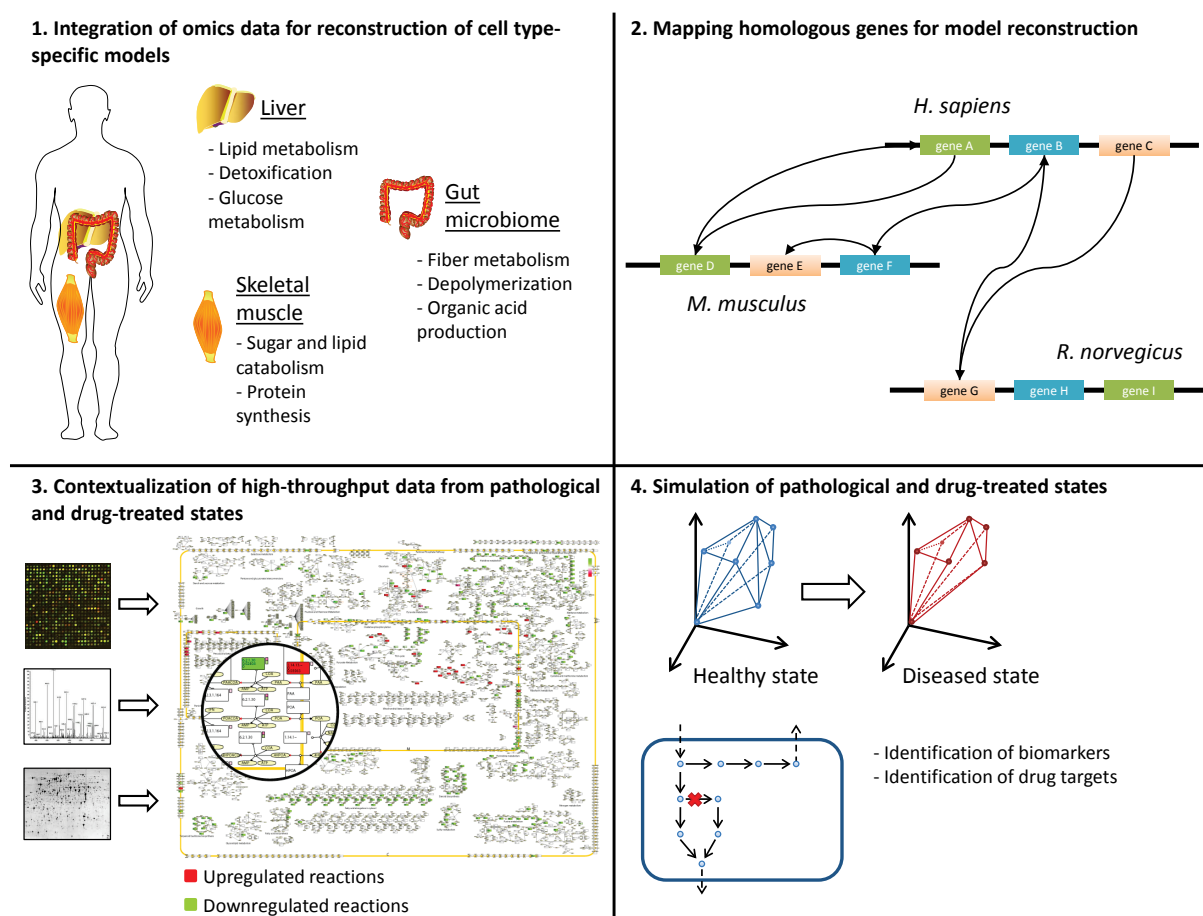


Figure 4-8. Overview of applications of GEMs in human health and disease. The available publications which make use of GEMs to study human health and disease can be categorized as follows: **1)** Integration of high-throughput data for model construction. Several algorithms have been developed which use different types of omics data to reconstruct cell type-specific models as subsets from a generic model (Becker and Palsson, 2008; Shlomi *et al.*, 2008; Wang *et al.*, 2012). There have also been attempts to integrate human GEMs with microbial GEMs, either pathogens or gut microbiota (Bordbar *et al.*, 2010; Heinken *et al.*, 2013). **2)** Mapping homologous genes for model construction. Much of medical research is carried out on mammals other than human. The high degree of homology between mammalian genes allows for reconstruction of GEMs for other mammals based on human GEMs (Sheikh *et al.*, 2005; Seo and Lewin, 2009). **3)** Contextualization of high-throughput data from pathological and drug-treated states. The work in this category uses GEMs as scaffolds for data analysis, rather than for predictive simulations. One such example is a study of the interactions between the tuberculosis bacteria and its host cell (Bordbar *et al.*, 2010). **4)** Simulation of pathological and drug-treated states. The work in the category uses GEMs for modelling, for example using FBA. Examples include simulations of hereditary metabolic diseases (Shlomi *et al.*, 2009) and simulations of the effect of potential cytostatic drugs (Folger *et al.*, 2011). Partly adapted from Bordbar and Palsson (2012).

Many of the methods developed for microorganisms, such as FBA, rely on the definition of a cellular objective. This is normally to grow as fast as possible given the available substrates. Since human cells do not grow uncontrollably those methods are not directly applicable. This

has led to that much of the work on genome-scale modelling of human cells has been focused on cancer; where the cells actually do grow uncontrollably (Folger *et al.*, 2011; Shlomi *et al.*, 2011). Another issue, which is extensively discussed in section 3.3.2, is that different cell types have different phenotypes even though they share the same genotype. There is still no clear workflow for how to generate cell type-specific models and to ensure that they are of high quality. Because of these issues, and others, the application of genome-scale metabolic modelling to human health and disease is a less mature science compared to when applied to microbial systems. Much of the work is therefore still centred on method development. Bordbar and Palsson (2012) categorized the publications which use GEMs to study human metabolism into the following four categories: 1) integration of high-throughput data for model construction, 2) mapping homologous genes for model construction, 3) contextualization of high-throughput data from pathological and drug-treated states, 4) simulation of pathological and drug-treated states. Figure 4-8 describes these categories in more detail.

In this section I present some of my work about genome-scale metabolic modelling applied to human health and disease. In **Paper IV** we developed an algorithm to reconstruct cell type-specific active metabolic networks based on different types of omics data. We then generated a large number of GEMs for cancers and their corresponding healthy cell types, and performed a statistical analysis to identify metabolic subnetworks which were more prominent in cancers. In **Paper V** we reconstructed a GEM for adipocytes and used it to study metabolic features associated to obesity. In **Paper VI** we reconstructed a GEM for hepatocytes and used it to study metabolic features associated to non-alcoholic fatty liver disease.

4.2.1 Paper IV: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT

In **Paper IV** we describe the generation of genome-scale active metabolic networks for 69 different cell types and 16 cancer types using the INIT (Integrative Network Inference for Tissues) algorithm. We first built a generic database of human metabolism by merging and curating previously available GEMs and databases. We then developed an algorithm which integrates several types of omics data in order to generate active metabolic networks, catalogues of the metabolic reactions that are likely to be active in a given cell type, from this database. These networks represent a resource which can form the basis for simulation of metabolic interactions between organs, or act as scaffolds for interpretation of high-throughput data. Lastly, we used these networks for a comparative analysis between cancer types and healthy cell types. This allowed for identification of cancer-specific metabolic features which constitute potential drug targets for cancer treatment.

In order to provide a reliable and up to date GEM template for our tissue/cell type-specific metabolic networks, we first constructed the Human Metabolic Reaction (HMR) database, containing the elements of the previously published generic genome-scale human metabolic models (Duarte *et al.*, 2007; Ma *et al.*, 2007), as well as the HumanCyc and KEGG (Ogata *et al.*, 1999; Romero *et al.*, 2005) databases. The HMR database has a hierarchical structure in which the genes are at the top and are linked to information about their tissue-specific expression profiles reported by Su *et al.* (2004) via BioGPS (Wu *et al.*, 2009). Each gene is linked to its different splicing variants, which in turn are linked to their corresponding proteins. Each protein is linked to its tissue specific abundances in the Human Protein Atlas (HPA) (Berglund *et al.*, 2008) and to the reactions they catalyse. The reactions are linked to

metabolites, which themselves are linked to their tissue specific information collected from the Human Metabolome Database (HMDB) (Wishart *et al.*, 2007). In order to have an unambiguous characterization of metabolites and reactions, KEGG and InChI identifiers were used for standardization. Each reaction was assigned to one or several of the eight compartments included in the HMR database: nucleus, cytosol, endoplasmic reticulum, Golgi apparatus, peroxisomes, lysosomes, mitochondria and extracellular. In cases where the subcellular localization was absent from the template models it was inferred from immunohistochemical (IHC) staining in the HPA. For enzymes that were not in the HPA, Swiss-Prot and GO were used to infer localization. The HMR database was used to generate a fully connected generic human GEM, which contained 4,137 metabolites (3,397 unique), 5,535 reactions (4,144 unique), and 1,512 metabolic genes.

As previously discussed, there have been several algorithms for reconstruction of cell type specific GEMs published (see section 3.3.2). The INIT algorithm was tailored to use data from the HPA as the main evidence source for assessing the presence or absence of metabolic enzymes in each of the human cell types that are present in the HPA. In the HPA project, cell type specific high quality proteomic data is being generated based on immunohistochemistry (Uhlen *et al.*, 2005; Berglund *et al.*, 2008; Uhlen *et al.*, 2010). Tissue specific gene expression (Su *et al.*, 2004) was used as an additional source of evidence. Figure 4-9 show how INIT is used to select a subset of reactions from a generic model. The problem was formulated so that all reactions in the resulting model are able to carry flux. Instead of imposing the steady state condition for all the internal metabolites, as it is usually done, we allowed for a small positive net accumulation rate. The reason for this was that we preferred to have a network able to synthesize molecules such as NADH or NADPH, rather than only being able to use them as cofactors. If a metabolite was present in a cell type (according to the HMDB) a positive net production of this metabolite was imposed on the network in order to assure that all the reactions necessary for its synthesis are included in the cell type-specific model. A distinct advantage of the INIT algorithm compared to existing approaches such as GIMME (Becker and Palsson, 2008) or MBA (Jerby *et al.*, 2010) (see section 3.3.2) is that it makes no predetermined classification of enzymes as either present or absent in the resulting model. In order to validate the output of the algorithm, the automatically generated hepatocyte model was compared with HepatoNet1 (Gille *et al.*, 2010), a manually curated and functional model of hepatocyte metabolism.

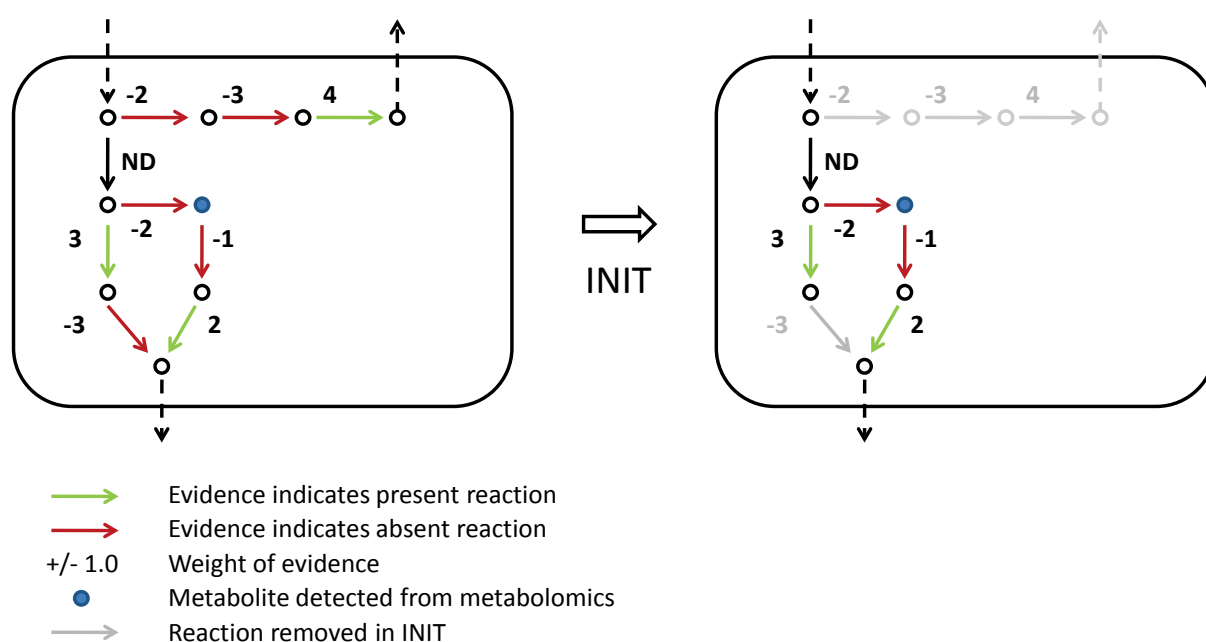


Figure 4-9. Principle of the INIT algorithm. Depending on the evidence for presence/absence of a given enzyme/gene in a cell type, a score is calculated for the reaction(s) catalysed by that enzyme. In the example above, the evidence indicates that it should exist in the cell type are coloured green. The opposite is true for the reactions coloured red. The aim of the algorithm is to find a subnetwork in which the involved genes/proteins have strong evidence supporting their presence, but at the same time maintaining a connected and functional network. This is done by maximizing for the sum of evidence scores under the constraint that all the included reactions should be able to carry a flux, and that all the metabolites observed experimentally (metabolite coloured blue in the example above) should be synthesized from precursors that the cell is known to take up. This is then implemented as a mixed-integer linear programming problem (MILP). In the example above, the three top reactions are excluded by INIT; despite that the last of them has evidence strongly supported its presence. This is because two other reactions would have to be included in order for it to be connected, and the net score would then be negative ($4-3-2<0$). The path via the blue metabolite to the end product at the bottom is also negative ($2-1-2<0$). However, since the blue metabolite is detected by metabolomics to exist in the cell it has to be synthesized in at least one reaction. The remaining pathway from the blue metabolite is then positive ($2-1>0$), and should therefore be included. The RAVEN Toolbox (**Paper III**) was used to perform the optimization and generate the cell type specific active networks.

Since the Warburg effect was observed at the beginning of the 20th century, it has been known that cancer cells show characteristic metabolic features that make them different from healthy cells (Koppenol *et al.*, 2011). This supposed metabolic similarity between cancer cells justified the development of a generic cancer GEM, which was used to identify potential drug targets against cancer proliferation (Folger *et al.*, 2011). We used INIT to infer active metabolic networks for 16 different cancer types, which can be compared with the 24 healthy cell types that they come from (there are several healthy cell types for some of the tissues associated to the cancers) in order to identify metabolic features that are characteristic of cancer. A hypergeometric test was used to identify genes and reactions that tended to be present in most of the cancer-specific active metabolic networks and absent in most of the original healthy cell types. The p-values obtained from the hypergeometric test were used to identify Reporter Metabolites (Patil and Nielsen, 2005) that are significantly more involved in the metabolism of cancer cells. These lists of genes, reactions and metabolites are cancer-specific features that are likely to be playing a specific role in proliferation of cancer cells and could be potential drug targets. Our comparative analysis between two sets of active metabolic networks can be seen as a high-throughput hypothesis generation method. These hypotheses are not based on mere correlations between cancer and the presence of a particular protein, but being based on the underlying metabolic network structure, and hereby our analysis provides a mechanistic interpretation about the possible role of each identified feature on the proliferation of cancer.

One of the most significant results from the Reporter Metabolites analysis is a much more pronounced metabolism of polyamines (PAs) such as spermidine, spermine, and putrescine in cancer cells. PAs play a variety of roles, of which several are related to oxidative stress, prevention and suppression of necrosis (Eisenberg *et al.*, 2009). PAs have long been known to be of particular importance for rapidly proliferating cells, and as such its transport and synthesis have been thoroughly investigated as anti-cancer drug targets (Seiler, 2003b). Inhibition of single enzymes in the PA synthesis pathway has proved disappointing, due to extensive regulation of the system and use of exogenous PAs by the cancer cells. Second generation drugs instead work by targeting the transport system, by structural homology to the PAs themselves, or by linking other antineoplastic drugs to the PAs (Seiler, 2003a).

Another high-ranking target is the isoprenoid biosynthesis pathway, in particular the intermediate geranylgeranyl diphosphate. This metabolite has been shown to promote oncogenic events due to its role in prenylation of important cancer proteins such as Ras and Rho GTPases (Sebti and Hamilton, 2000). Several drugs have therefore been developed to target the prenylation process (Philips and Cox, 2007) or the biosynthesis of geranylgeranyl diphosphate (Dudakovic *et al.*, 2011).

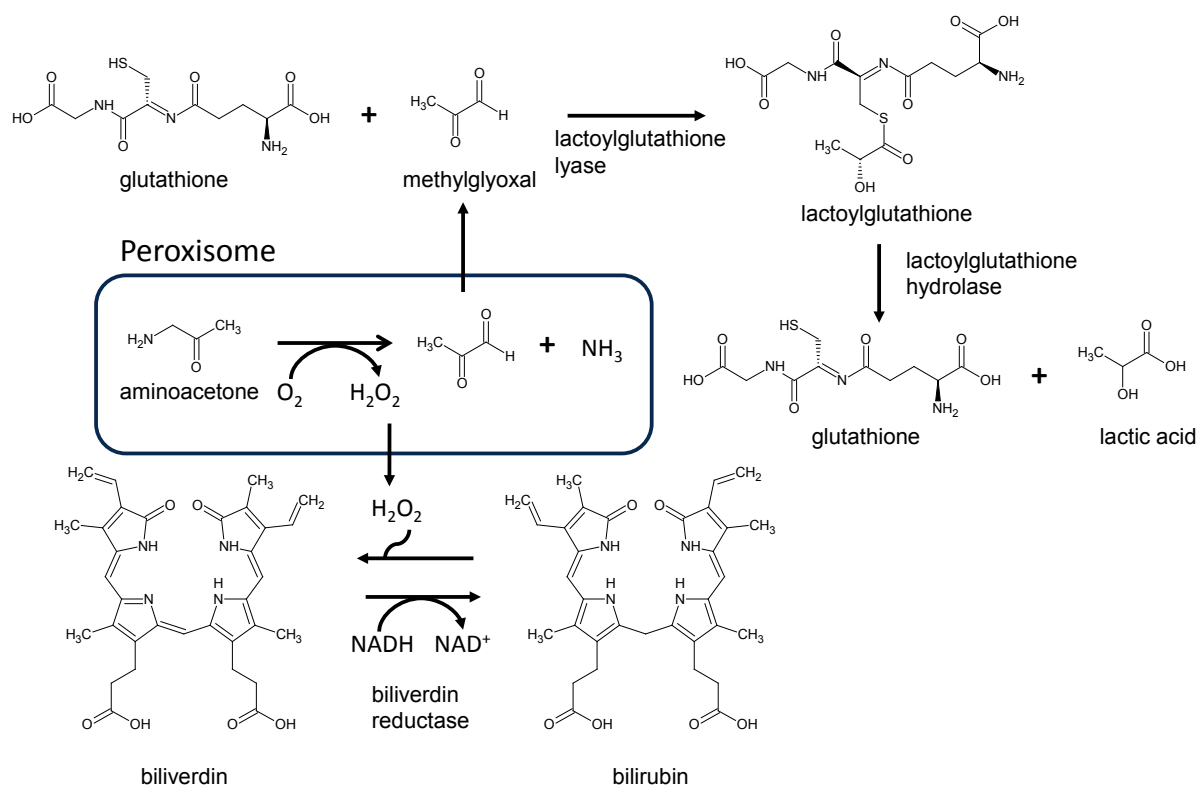


Figure 4-10. Metabolic subnetwork identified as being significantly more prominent in cancer tissues compared to their corresponding healthy tissues. Aminoacetone, which is a toxic by-product of amino acid catabolism, is converted to toxic methylglyoxal in a reaction that also results in hydrogen peroxide. The toxicity of methylglyoxal is relieved by two reaction steps involving ligation to glutathione and resulting in lactic acid. The generated hydrogen peroxide is taken care of by the enzyme biliverdin reductase. This is an example of how network-based analysis can lead to a more mechanistic interpretation of data. Figure taken from Agren *et al.* (2012).

A third prominent group among the Reporter Metabolites is prostaglandins and leukotrienes together with the intermediate HPETE. These autocrine compounds are synthesized from arachidonic acid and are elevated in connection with inflammation. They have been shown to aid in cancer progression by promoting metastasis and by influencing the immune system

(Schneider and Pozzi, 2011). Of particular interest is prostaglandin E2, where both the synthesis and degradation have been investigated as promising targets for drug development (Eruslanov *et al.*, 2009).

The fact that so many of the identified targets correspond to well known and used drug targets indicates that the method is able to generate biologically relevant hypotheses. Of particular interest are therefore the Reporter Metabolites that are currently not targeted in cancer treatment. Among the top-scoring Reporter Metabolites we identified biliverdin and bilirubin (see Figure 4-10). Biliverdin reductase and the reactions catalysed by this enzyme also appear among the genes and reaction most enriched in the cancer networks. Biliverdin reductase is known to be a major physiologic cytoprotectant against oxidative stress (Baranano *et al.*, 2002). Cancer cells are known to be exposed to high oxidative stress resulting from the hydrogen peroxide generated during the oxidation of polyamines and other products of amino acid breakdown taking place in the peroxisome. Bilirubin is oxidized to biliverdin by hydrogen peroxide and subsequently reduced back to bilirubin by biliverdin reductase. This mechanism has been proven to be a major relief system for oxidative stress and could be considered a potential target against cancer proliferation. One of the hydrogen peroxide generating reactions taking place in the peroxisomes is the transformation of aminoacetone, which is an intermediate in the degradation of glycine, into methylglyoxal. Another source of methylglyoxal in cancer cells is from gluconeogenesis (Titov *et al.*, 2010). Methylglyoxal is known to be a toxic compound (Kalapos, 1994) that has been proven to induce apoptosis in some cancer cell lines (Kang *et al.*, 1996). Methylglyoxal also appeared among our top scoring reporter metabolites and both the gene coding for lactoylglutathione lyase (an enzyme that transforms methylglyoxal and glutathione into lactoylglutathione) and its associated reactions appear among the most enriched genes and reactions in the cancer active metabolic networks. Lactoylglutathione is further transformed into glutathione and lactic acid by the enzyme lactoylglutathione hydrolase (which also shows a significant enrichment in cancer metabolic networks with a p-value of $2e-3$). The mentioned two enzymes seem to be playing a relevant role in relieving the toxicity generated by methylglyoxal and could be potential drug targets against cancer proliferation. Targeting these enzymes would have the same effect on cancer cells as using methylglyoxal as a drug, but the advantage is that there would be no toxicity effects of methylglyoxal on healthy tissues.

In conclusion, the HMR database represents the most comprehensive generic human GEM (see **Paper VI**) to date and is an important resource in itself. The INIT algorithm was demonstrated to automatically generate active metabolic networks which were similar in scope compared to a high-quality manually reconstructed model. This was made possible by use of the enormous amount of proteomics data generated within the HPA project. The INIT algorithm was then applied to reconstruct GEMs for 69 cell types and 16 cancers; models which can form the basis for future work on modelling of human metabolism. The content of these models was analysed and we were able to identify a number of metabolic subnetworks which were significantly more prominent in cancers compared to their healthy counterparts. These subnetworks contained many known drug targets, but we were also able to identify a number of novel drug targets based on our analysis.

4.2.2 Paper V: Global analysis of human adipocyte metabolism in response to obesity

Adipose tissue dysfunction or overload of its lipid storage capacity can lead to wide range of diseases (e.g. immunological and inflammatory diseases), including metabolic diseases such as obesity (Lago *et al.*, 2007; Auffray *et al.*, 2009). An increased understanding of the mechanisms behind obesity and related diseases would provide valuable insights into their etiology and pathogenesis, and could lead to new treatment strategies. In **Paper V** we reconstructed a GEM for adipocytes based primarily on adipocyte specific proteome data generated within the Human Protein Atlas (HPA) project (Uhlen *et al.*, 2010).

We expanded the coverage of the HPA to include the protein profiles of adipocytes found in breast and two different soft tissues and examined the spatial distribution and the relative abundance of proteins encoded by 14,077 genes in these tissues. A total of 17,296 affinity-purified antibodies were generated and used for immunohistochemical staining of tissue micro array blocks. The proteome data was merged with previously published proteome data on adipocytes in order to increase the coverage. In total, we have proteome evidence for the presence/absence of proteins associated with 14,337 genes in adipocytes. The proteomics analysis resulted in evidence for presence of proteins associated with 7,340 genes.

As discussed in section 3.3, the subcellular localization of reactions has large implications on the functionality of GEMs, as only a portion of metabolites can be transported between compartments. Furthermore, compartments can be individually redox and/or energy balanced. The HPA includes subcellular profiling data using immunofluorescence-based confocal microscopy in three human cancer cell lines of different origin. Here, proteins were classified into eight different compartments following our HMA standard (**Paper IV**): cytosol, nucleus, endoplasmic reticulum (ER), Golgi apparatus (GA), peroxisome, lysosome, mitochondria and extracellular space. Reactions were assigned to compartments through their association with proteins in these different compartments.

In order to reconstruct a GEM for adipocytes, biochemical and genetic evidence was combined with data on protein expression and localization. HepatoNet1, a GEM for hepatocytes which is reconstructed based on the manual evaluation of the original scientific literature (Gille *et al.*, 2010), was used as a starting point for our reconstruction process and used to generate an initial candidate list of network components. Firstly, metabolism of lipids and lipoproteins in Reactome, a manually curated and peer-reviewed pathway database (Croft *et al.*, 2011), was merged into HepatoNet1. Secondly, the resulting network was combined with the evidence-based generic human models Recon1 (Duarte *et al.*, 2007) and the compartmentalized EHMN (Hao *et al.*, 2010). This combined reaction list resulted in an updated version of our Human Metabolic Reaction (HMR) database (**Paper IV**). The HMR database contains 6,049 metabolites in eight different compartments (3,162 unique metabolites), 8,107 reactions and 3,668 genes associated to those reactions. Thirdly, the existence of each protein coding gene associated to a reaction in HMR was assessed for the presence or absence in adipocytes using previously published and the here generated adipocyte-specific proteome data. This process provided us with a list of reactions that occur in adipocytes. Gaps in the resulting network were filled using the updated HMR database, public databases such as KEGG (Kanehisa *et al.*, 2010) and HumanCyc (Romero *et al.*, 2005) and manual evaluation of the literature about adipocyte metabolism. The RAVEN Toolbox (**Paper III**) was used for gap filling and quality control. This gap filling resulted in generation of iAdipocytes1809, which is a fully functional and connected GEM for adipocytes. In iAdipocytes1809, individual metabolites, rather than generic pool metabolites, for 59 fatty

acids (FAs) have been used. This allowed us to incorporate measured concentrations of different FAs in human plasma and adipocytes into the model. Figure 4-11 shows the reconstruction workflow.

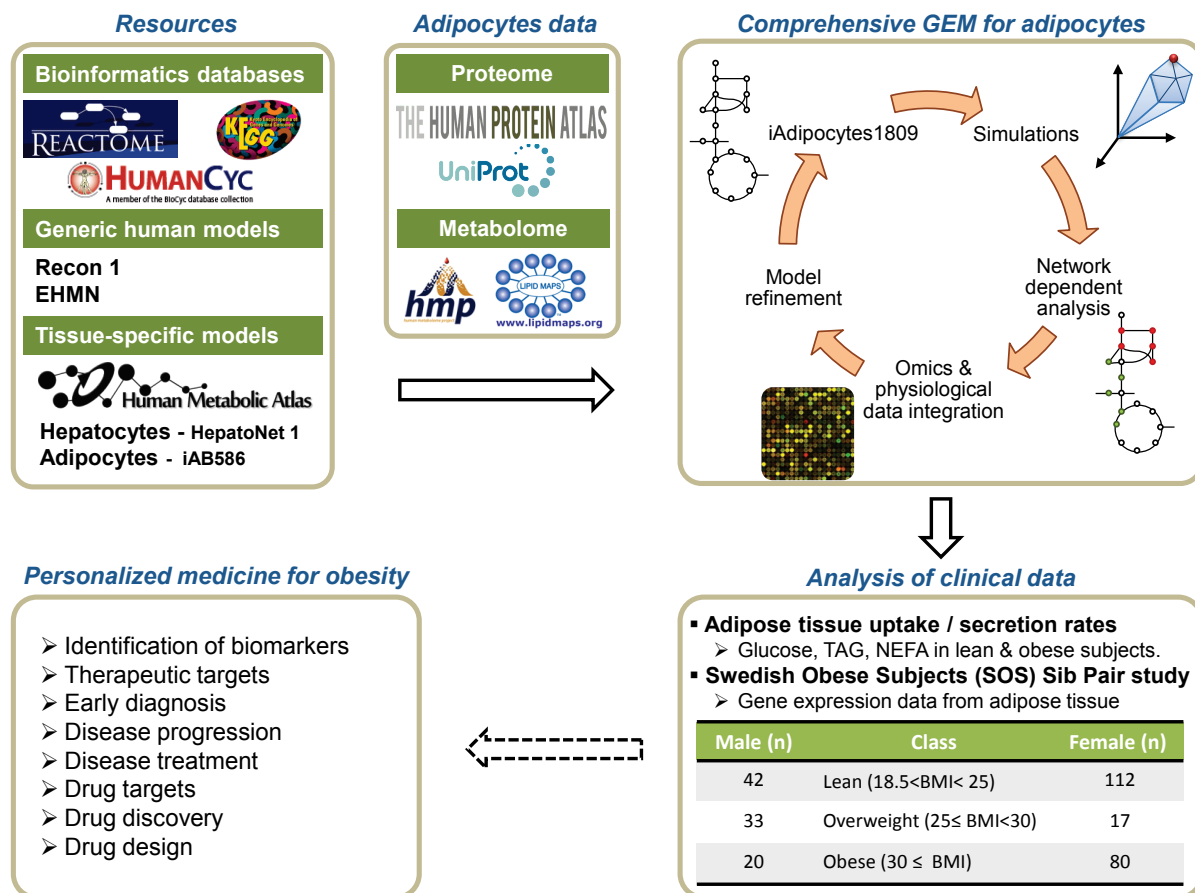


Figure 4-11. Schematic illustration of how a GEM for adipocytes may provide links between molecular processes and subject phenotypes. Here the GEM iAdipocytes1809 was reconstructed through the use of proteome, metabolome, lipidome and transcriptome data, literature based models (Recon 1, Edinburg Human Metabolic Network (EHMN) and HepatoNet1), and public resources (Reactome, HumanCyc, KEGG and the Human Metabolic Atlas). We first performed global protein profiling of adipocytes using antibodies generated within the Human Protein Atlas (HPA). We further used information on metabolome and lipidome data from the Human Metabolome Database (HMDB) and LIPID MAPS Lipidomics Gateway, respectively. The model was then used for the analysis of gene expression data obtained from subjects with different body mass indexes in the Swedish Obese Subjects (SOS) Sib Pair study and other adipose tissue relevant clinical data such as uptake/secretion rates in lean and obese subjects (see below).

In iAdipocytes1809, 59 different common long and very long chain FAs in human plasma can be taken up as NEFAs and lipoproteins. The lipid related functionality of iAdipocytes1809 is summarized in Figure 4-12. The GEM was subject to extensive quality control by using the RAVEN Toolbox (**Paper III**), following the workflow in Figure 3-3. Even a well-connected, thermodynamically correct and balanced model may not be able to perform all relevant metabolic functions, or it may be able to perform functions that it should not do (such as synthesis of essential amino acids or fatty acids). The model was therefore validated for 250 known metabolic functions of adipocytes, adapted from the definitions provided in connection with setting up HepatoNet1 (Gille *et al.*, 2010).

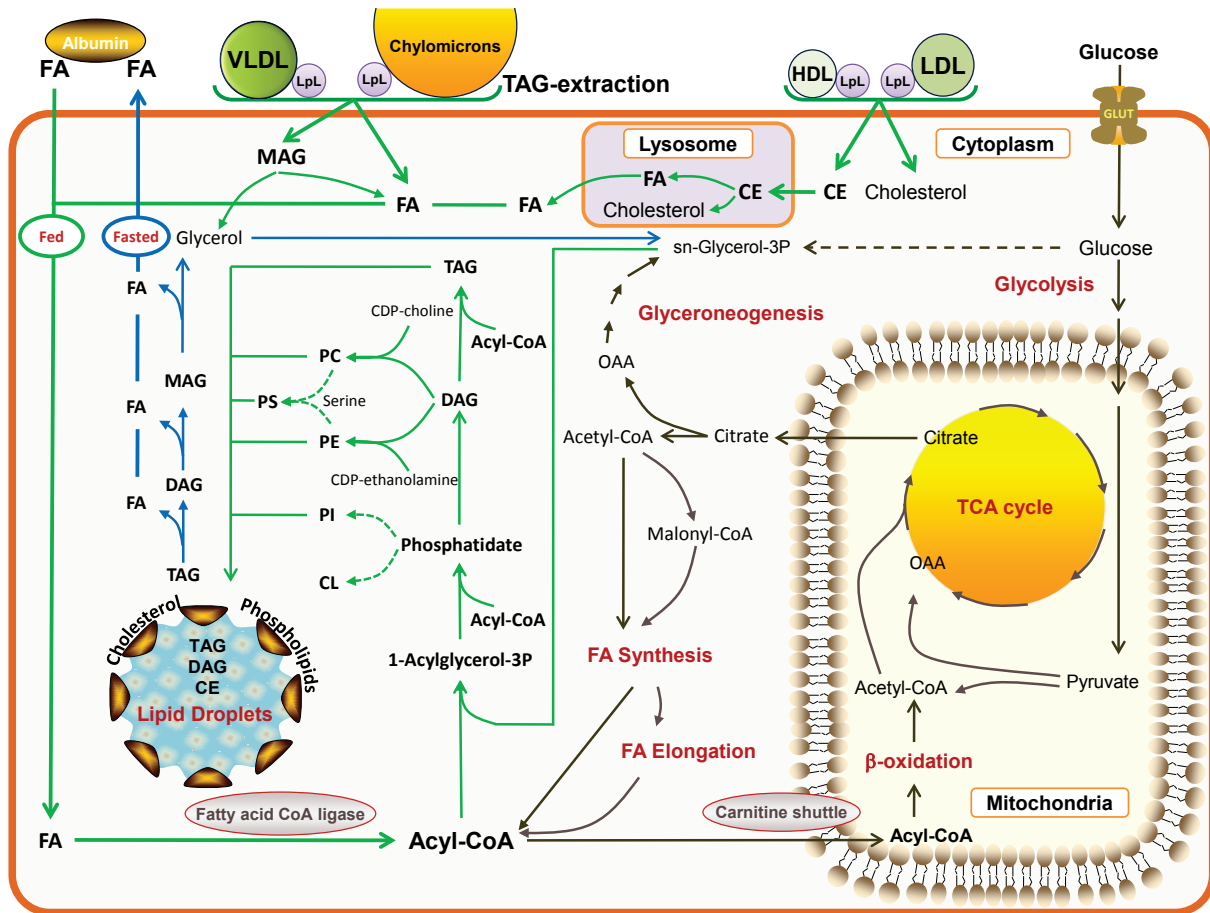


Figure 4-12. Summary of the capabilities of iAdipocytes1809. Adipocytes store lipid mainly in the form of triacylglycerols (TAGs) and cholesterol esters (CEs). They form lipid droplets (LDs) in the post-prandial state (green arrows) and release them by degrading LDs in the post-absorptive state (blue arrows) in order to provide energy for other tissues. The released fatty acids (FAs) from adipocytes are transported to other tissues by albumin. The FAs are taken up from non-esterified FAs (NEFA) and lipoproteins, including chylomicrons, very-low-density lipoprotein (VLDL) and CEs together with cholesterol are taken up with low-density lipoproteins (LDL) and high-density lipoproteins (HDL) through lipoprotein lipase (LPL). CEs taken up from lipoproteins are degraded to cholesterol and FAs in lysosomes and transported to the endoplasmic reticulum and cytosol to be stored in LDs. Adipocytes also take up glucose to be used in the de novo synthesis of FAs (black arrows) that occurs at low level in adipocytes. LDs are rich in TAGs, CEs and an unknown neutral lipid that migrated between CEs and TAGs, ether neutral lipid monoalk(en)yl diacylglycerol (MADAG). LDs also contain small amounts of free FAs, cholesterol and phospholipids including phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylinositol (PI), ether-linked phosphatidylcholine (ePC), ether-linked phosphatidylethanolamine (ePE), lyso phosphatidylcholine (LPC), lysophosphatidylethanolamine (LPE), phosphatidylserine (PS) and sphingomyelin (SM). Formation of ePC, ePE, LPC, LPE and SM is included in iAdipocytes1809 but is not shown in figure.

The function of iAdipocytes1809 was tested by estimating the formation of LDs based on clinical data for lean and obese subjects. Recently, McQuaid *et al.* (2011) measured the delivery and transport of FAs in adipose tissue using multiple and simultaneous stable-isotope FA tracers in lean and obese subjects groups over 24 hours period. Even though abdominally obese subjects have greater adipose tissue mass than control lean subjects, the rates of delivery of NEFAs were downregulated in obese subjects. Based on measurements of the uptake of glucose and TAG and the release of NEFAs over a 24 hour period we simulated the change in LD size. We found from our simulations that lean subjects have large dynamic changes in LD formation compared with obese subjects (see Figure 4-13c). Furthermore, we predicted a lower acetyl-CoA production in obese subjects (see Figure 4-13d).

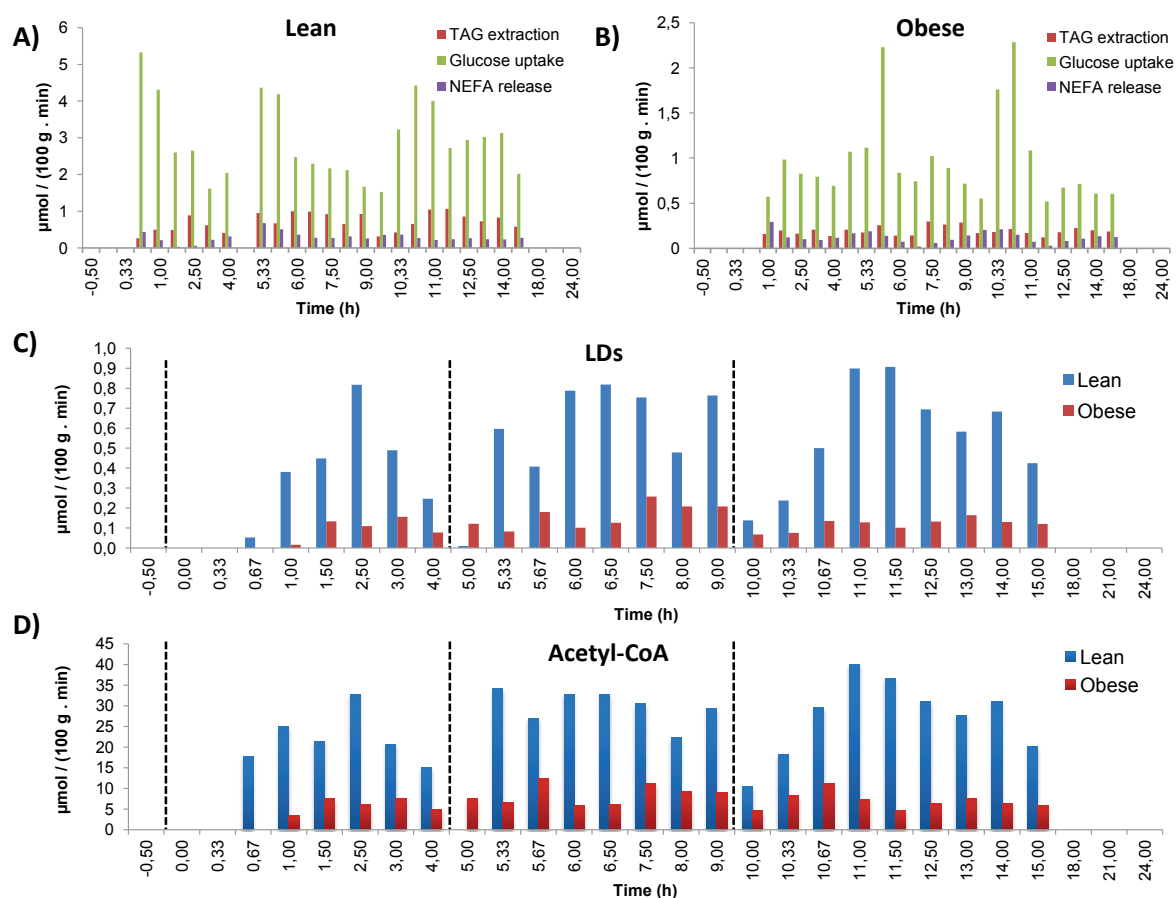


Figure 4-13. Simulated lipid droplet and acetyl-CoA production. Uptake rates for glucose and TAGs and release rates for NEFAs in adipocytes for lean (A) and obese (B) subjects were used as lower and upper bounds for input reactions (McQuaid *et al.*, 2011) together with amino acid uptake rates (Patterson *et al.*, 2002), and the amounts of LDs (C) and acetyl-CoA (D) were predicted over a 24 h period. The dashed lines at time 0, 5, 10 hours represent breakfast, lunch and dinner, respectively, for each participant of the study. Figure taken from Mardinoglu *et al.* (2013).

We then employed the iAdipose1809 GEM for the analysis of microarrays which profile the gene expression from subcutaneous adipose tissue of subjects from the Swedish Obese Subjects (SOS) Sib Pair Study. This study includes nuclear families with BMI-discordant sibling pairs (BMI difference $\geq 10 \text{ kg/m}^2$). Besides the gene expression data from the SOS Sib Pair Study, additional clinical data (e.g. plasma and WAT lipid concentrations) were also incorporated into the model. By integrating gene expression data and adipose tissue uptake/secretion rates with the reconstructed GEM, we identified metabolic differences between individuals with different BMIs by using the concept of Reporter Metabolites (Patil and Nielsen, 2005) and transcriptionally controlled reaction fluxes (**Paper II**).

The results from this analysis showed that the following pathway fluxes were transcriptionally downregulated in obese subjects: uptake of glucose, uptake of FAs, oxidative phosphorylation, mitochondrial and peroxisomal β -oxidation, FA metabolism, and TCA cycle. Furthermore, fluxes associated with beta-alanine metabolism were found to be transcriptionally downregulated in obese subjects. Previously it has been reported that blood flow, glucose uptake, release of NEFA and the extraction of TAG from plasma was significantly lower in abdominally obese subjects compared to lean subjects (McQuaid *et al.*, 2011). Most of these pathways are linked with mitochondrial dysfunction. Several therapeutic interventions, including antioxidants and chemical uncoupler treatments, have been shown to improve mitochondrial dysfunction (Kusminski and Scherer, 2012).

Mitochondrial acetyl-CoA plays a central role in different pathways in the mitochondria, where it reacts with oxaloacetate to form citrate. Citrate can then be transported from the mitochondria to the cytosol where it is participating in FA synthesis (Dean *et al.*, 2009). Acetyl-CoA derived through other principal sources, including degradation of amino acid and ketone bodies and fatty acid oxidation processes are insufficient for FA synthesis. Increasing the acetyl-CoA concentration and eventually FA synthesis in adipose tissue of obese subjects results in whole body regulation of metabolism as reported by Cao *et al.* (2008). We therefore propose to boost the metabolic activity of mitochondria in the adipocytes of obese subjects by aiming to increase the availability of mitochondrial acetyl-CoA.

As previously mentioned, beta-alanine metabolism came out as significantly changed between healthy and obese. The effect of beta-alanine as a dietary supplement was previously examined in football players and it is reported that it has effect on lean tissue accrument and body fat composition (Hoffman *et al.*, 2006). Furthermore rat studies reported that beta-alanine decreases the lipoprotein lipase (LPL) enzyme activity in adipose tissue which may help to decrease the uptake of FAs to be stored in adipocytes (Prabha *et al.*, 1988). Our results suggest that increasing the level of beta-alanine in obese subjects may help to decrease the fat composition in obese subjects.

Another high-ranking target for upregulated genes is ganglioside GM2. Gangliosides, one of the major glycosphingolipids in mammals, play major roles as mediators for cell to cell or cell to matrix recognition and regulate the transmembrane signal transducers and cell proliferation. Gangliosides in adipose tissues are also associated with insulin signalling mechanisms and it is reported that series of gangliosides GM2, GM1, and GD1a are dramatically increased in adipose tissues of obese mice (Tanabe *et al.*, 2009).

A third prominent group among the Reporter Metabolites for upregulated genes is the degradation products of heparan sulfate proteoglycans (HSPG) and keratan sulfate. These compounds are classified as glycosaminoglycans and attach to cell surface or extracellular matrix proteins. Keratan sulfate, a biomarker of proteoglycan degradation, can be expressed from stem cells in human SAT and its relevance with obesity has been reported earlier. It has been reported that catalytically active adipose tissue lipoprotein lipase (LPL) attaches to HSPG at the luminal surface of vascular endothelium (Olivecrona and Beisiegel, 1997; Lafontan, 2008) and hydrolyse the TAGs for uptake of FAs into the cell. The LPL moves between individual HSPG chains within the layer and this creates a high concentration of LPL along the surface layer of HSPG chains (Lookene *et al.*, 1996). In the presence of heparin more LPL is secreted and increased secretion was balanced by decreased degradation of LPL. There are special mechanisms that inhibit LPL and one mechanism is that LPL forms complexes with FAs (Bengtsson and Olivecrona, 1980). During the LPL hydrolysis and accumulation of FAs in the cells, the LPL is sequestered into enzyme FA complexes, lipolysis is reduced and eventually the binding of LPL to heparan sulfate is broken. If a high-affinity ligand (e.g. FAs, heparin, apoCII) is available, the LPL detaches from the cell surface to heparan sulfate chains and without ligand in the medium, the LPL recycles into the cells where it is degraded. Furthermore, several studies have reported that more sulfated polysaccharide chains increase the affinity for binding of LPL (Olivecrona and Olivecrona, 2009). One possible intervention strategy could therefore be to try to reduce the degradation rate of HSPG.

In conclusion, the first human GEM with extensive lipid metabolism was reconstructed. This was made possible by close collaboration with groups that generated large-scale proteomics data for adipocytes and transcriptomics data for healthy/obese siblings. The model could correctly capture the reduced dynamics of lipid droplet formation in obese subjects and we

could see that this was associated to mitochondrial dysfunction. The model was then used together with the algorithm developed in **Paper II** and the Reporter Metabolites algorithm to identify metabolic differences between healthy and obese siblings. This led us to hypothesize that obesity could be treated with interventions aiming at reactivating the mitochondria by increasing the availability of acetyl-CoA. An alternative approach was to target the degradation of heparan sulfate proteoglycans.

4.2.3 Paper VI: Identification of serine deficiency in non-alcoholic fatty liver disease through genome-scale metabolic modelling

Hepatocytes have a wide range of physiological functions, including production of bile and hormones, removal of toxic substances, homeostatic regulation of the plasma constituents and synthesis of most plasma proteins (Gille *et al.*, 2010). They are the most metabolically active cell types in human and play a major role in overall human metabolism. Deficiency or alterations in the metabolism of hepatocytes can lead to complicated disorders such as hepatitis, non-alcoholic fatty liver disease (NAFLD), cirrhosis and liver cancer, which are serious threats to public health (Baffy *et al.*, 2012). NAFLD is considered as the hepatic manifestation of obesity and metabolic syndrome, and encompasses a spectrum of pathological changes; ranging from simple fatty liver (FL) to non-alcoholic steatohepatitis (NASH) (Neuschwander-Tetri and Caldwell, 2003).

In **Paper VI** we reconstructed a consensus GEM for hepatocytes and applied it in order to suggest potential biomarkers and therapeutic targets for NAFLD. In parallel to this we built on the results from **Paper IV** and **Paper V** in order to reconstruct a generic human GEM.

Several generic (non-cell type-specific) GEMs for human metabolism have been previously constructed (as discussed in **Paper IV** and section 3.3.2). One such generic model is the HMR database presented in **Paper IV**. However, neither of these generic networks contain extensive lipid metabolism, which is necessary in order to study the effect of lipids on the underlying molecular mechanism of NAFLD. In **Paper V** we reconstructed a GEM for adipocytes with a strong focus on lipid metabolism. In this paper we presented HMR 2.0, in which we had integrated all published human GEMs, the original HMR database, and the lipid metabolism from the adipocyte GEM from **Paper V**. The HMR 2.0 database is the largest biochemical reaction database for human metabolism in terms of number of reactions/genes/metabolites, as well as in terms of which parts of metabolism that are covered. This represents an important step forward since lipids have major effects on the development of several important metabolic diseases (Newgard, 2012). The functionality of the model was tested using the RAVEN Toolbox (**Paper III**), in the same manner as previously described for the adipocyte GEM.

A draft hepatocyte GEM was then reconstructed from a subset of HMR 2.0 based on proteomics data from HPA and by using the tINIT algorithm (**Paper IV** and Agren *et al.* (2013b)). Previously, several GEMs for hepatocytes, including HepatoNET 1 (Gille *et al.*, 2010), iLJ1046 (Jerby *et al.*, 2010), iAB676 (Bordbar *et al.*, 2011) and iHepatocyte1154 (**Paper IV**), have been reconstructed. The draft model was then expanded to contain all of the protein coding genes and associated reactions in the previously published liver models (Figure 4-14a). In addition to the proteomics data and reactions from previously published models, protein coding genes were also included based on transcriptomics data and for connectivity reasons (Figure 4-14b). Lastly, additional clinical data for plasma and hepatocyte lipid

concentrations for individual FAs were incorporated into the model, resulting in the final iHepatocytes2260 GEM.

iHepatocytes2260 differs from previously published hepatocyte GEMs primarily in terms of coverage in lipid metabolism. Among the new lipid related functions are uptake of the remnants of lipoproteins (chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoproteins (LDL) and high-density lipoproteins (HDL)), the formation and degradation of lipid droplets (LDs) and secretion of synthesized lipoproteins (VLDL, LDL, HDL) (Figure 4-14c, see also the corresponding functionality for adipocytes in Figure 4-12). The model was validated by simulating 256 different biologically defined metabolic functions (e.g. the synthesis of FAs, amino acids, cholesterol and bile acids) that is known to occur in hepatocytes. Furthermore, the ability of the model for performing gluconeogenesis was demonstrated using experimentally measured secretion rates for glucose and albumin and uptake rates for glycerol, lactate, amino acids and FAs in primary rat hepatocytes (Chan *et al.*, 2003).

In the model reconstruction process, tINIT (Agren *et al.*, 2013b) identified 61 genes (out of the 3,673 genes in the HMR 2.0 database) which had to be integrated into the model in order to maintain the functionality, even though they had been reported to be non-expressed in hepatocytes according to the HPA. We then re-analysed the immunohistochemistry (IHC) data for these 61 proteins and found that 20 (33%) of these proteins actually show presence in hepatocytes. Initial discordant data were due to the suboptimal titration of the antibody, misinterpretation of weak IHC staining or due to interference with other cell types besides hepatocytes present in liver (e.g. kupffer cells and sinusoids). Nine (15%) of the investigated proteins showed more concordant results to the mathematic model when re-analysed using another antibody targeting the same protein. 15 proteins (25%) with negative IHC data were kept as negative in HPA since limited literature was available, and/or concordant results were seen in subsets of the remaining panel of tissues included in the HPA high-throughput set up. The remaining 17 proteins (28%) are believed to be inaccurately assessed by IHC due to technical issues, such as antigen recognition due to antigen conformational changes, fixation or sub optimal antibody. We think this is an excellent example of how a holistic view of metabolism can lead to biological insights, in this case as a targeted way to improve on the quality of experimental data.

NAFLD, and its most severe form NASH, is progressively diagnosed worldwide (Rector *et al.*, 2008). It is tightly associated with obesity, type 2 diabetes, insulin-resistance, and hypertension and represents a severe risk for development of cirrhosis and hepatocellular carcinoma (Ascha *et al.*, 2010). Despite its severe drawbacks, liver biopsy is still the most common procedure for diagnosing NASH (Machado and Cortez-Pinto, 2012). Thus, there is a need for identifying metabolic biomarkers to diagnose NASH, as well as to subcategorize the NAFLD patients without taking biopsies. A metabolic biomarker can be defined as a metabolite which is secreted to the blood where its level differs between two different states. We used the iHepatocytes2260 GEM as a scaffold for transcriptome analysis in an attempt to identify potential such biomarkers.

We retrieved liver gene expression data from 45 subjects out of which 19 were healthy, 10 steatotic, 9 had NASH with FL and 7 had NASH without FL (Fisher *et al.*, 2009; Lake *et al.*, 2011). We used the Piano package to compare gene expressions of NASH with and without FL to healthy subjects (Väremo *et al.*, 2013). The iHepatocyte2260 GEM was then used with the Reporter Metabolites algorithm in order to identify metabolites around which transcriptional changes occur between healthy and diseased subjects (Patil and Nielsen, 2005). A total of 50 statistically significant metabolites were such identified. In addition to several known subsystems involved in the progression of NASH, e.g. cholesterol biosynthesis, folate, vitamin B6, porphyrin, nucleotide, eicosanoid and amino acid metabolism (Greco *et al.*, 2008; Anstee and Day, 2012) we also identified several new Reporter Metabolites. These metabolites were involved in N-glycan metabolism and in the biosynthesis of the proteoglycan (PG) chondroitin sulfate (CS). PGs are composed of glycosaminoglycans, including CS and heparan sulfate (HS), and core proteins.

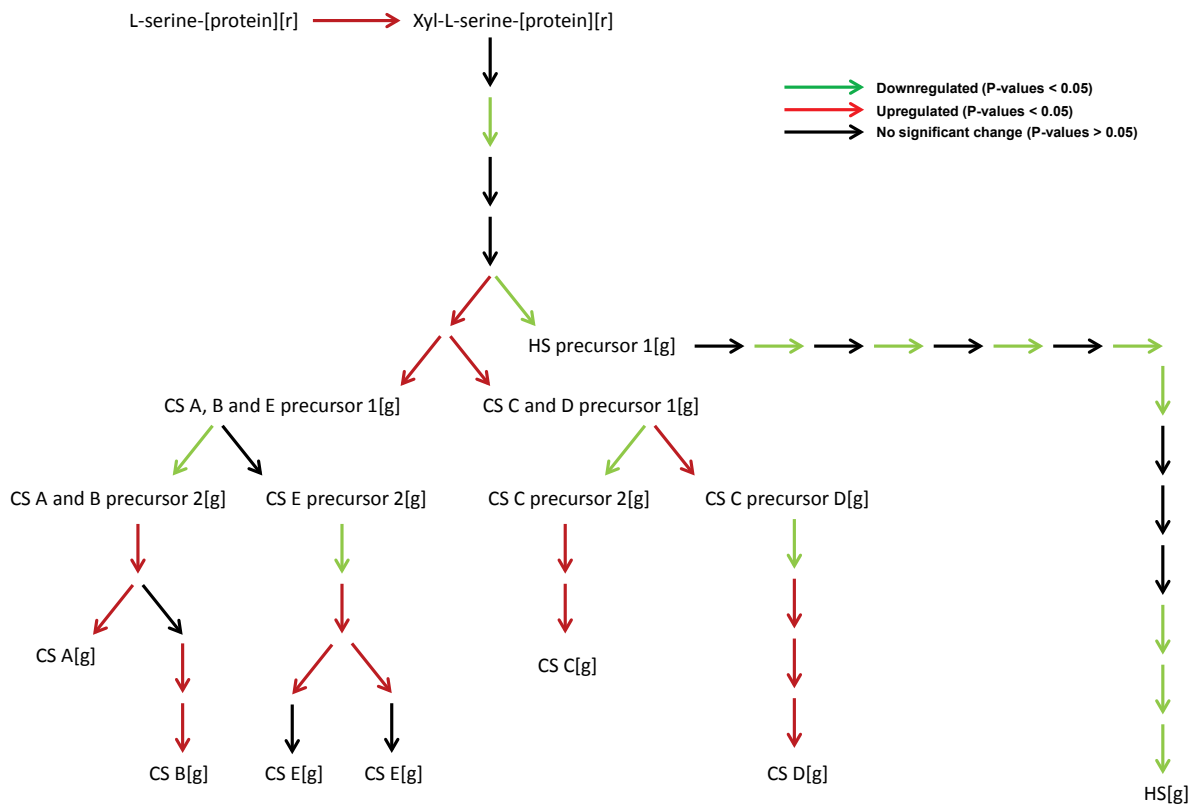


Figure 4-15. Results from Reporter Metabolites analysis. The figure shows the biosynthesis of chondroitin sulfate (CS) and heparan sulfate (HS) in Golgi apparatus, as formulated in iHepatocytes2260, together with the relative gene expression level of NASH vs. healthy samples. Red arrows indicate over-expression of a gene, whereas green arrows indicate under-expression. Non-significant changes ($p\text{-value} > 0.05$) is indicated with black arrows.

Figure 4-15 shows the reactions in the model which involves CS or HS and the relative gene expression for the corresponding genes. As can be seen, the genes responsible for the synthesis of CS are mainly upregulated in NASH subjects while the genes responsible for the synthesis of HS are mainly downregulated. CS and HS are implicated in cancer progression (Afratis *et al.*, 2012), one of the most severe outcomes of NASH. Because of this, and because of the clear upregulation of one branch and clear downregulation of the other, we therefore suggested that the blood concentration of the metabolites associated with these pathways

might change accordingly, and that they are therefore potential biomarkers for diagnosing NASH.

The Reporter Subnetworks algorithm was then applied to identify sets of metabolic reactions which exhibit transcriptional correlation after a perturbation (in this case NASH vs. healthy) (Patil *et al.*, 2005). Figure 4-16a show the resulting subnetwork. As can be seen, amino acid metabolism has a prominent role (the non-essential amino acids serine, glycine, glutamate, glutamine, aspartate, asparagine, alanine and the essential amino acids valine and methionine are all present). Several metabolites involved in folate metabolism (e.g. tetrahydrofolate (THF), 5-methyl-THF, 5-formyl-THF and 5,10-methenyl-THF, 5,10-methylene-THF) were also identified, and these metabolites are involved in the interconversion of serine, glycine and glutamate.

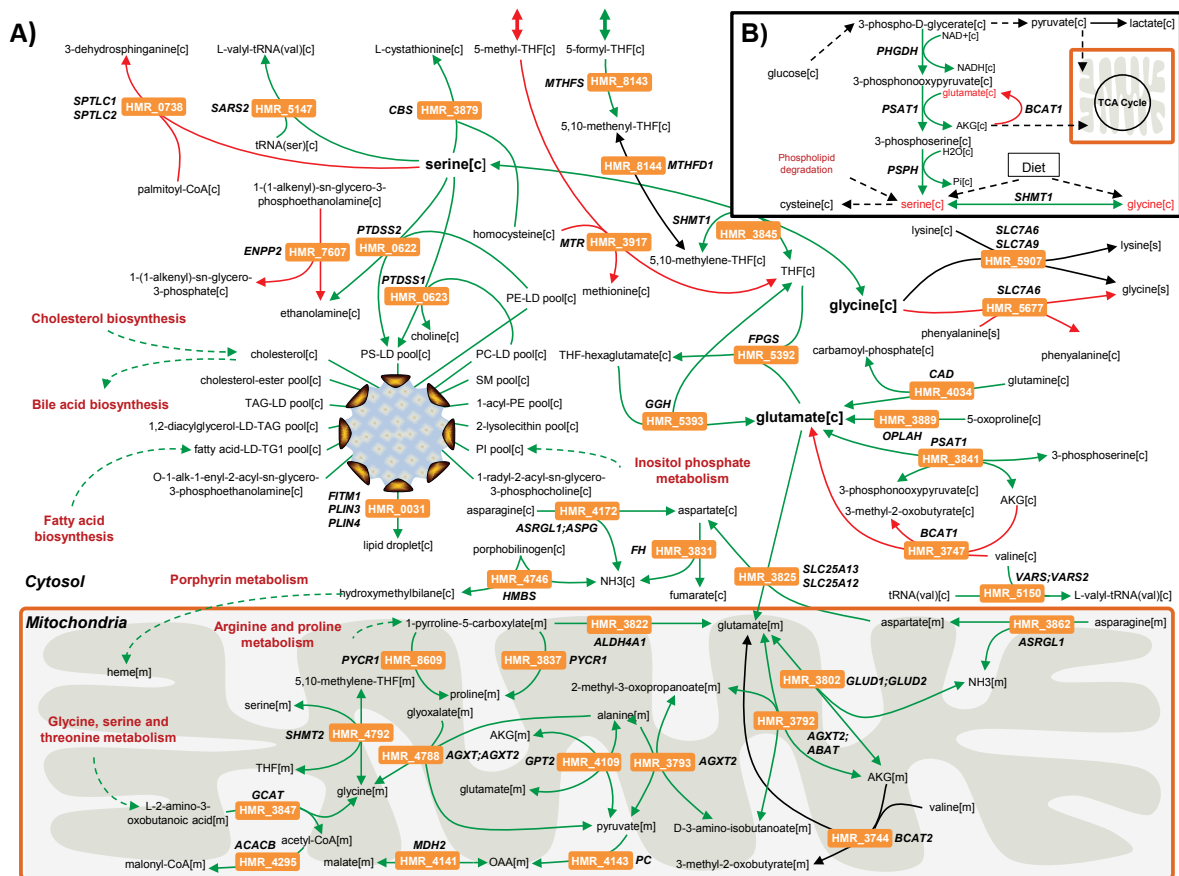


Figure 4-16. Results from Reporter Subnetworks analysis. A) The subnetwork identified using Reporter Subnetworks and gene expression data for NASH vs. healthy subjects. B) Some relevant reactions involved in serine biosynthesis and their corresponding change in expression. Red arrows indicate over-expression of the associated genes in NASH, whereas green arrows indicate under-expression. Non-significant changes (p-value>0.05) is indicated with black arrows.

Moreover, phosphatidylserine (PS), an essential component for formation of lipid droplets (LDs), was identified through our analysis. LDs have diverse roles in the cell, such as serving as storage for TAG and CEs or protecting the cell from excess lipids or lipophilic substances that may be toxic (Farese and Walther, 2009). The enzymes phosphatidylserine synthases (PTDSS1) and (PTDSS2) that catalyse the production of PS by condensation of phosphatidylcholine (PC) and phosphatidylethanolamine (PE), respectively, were significantly downregulated in NASH patients. The significant changes in the level of PS in

cirrhotic (severe stage of NASH) livers was previously reported in a study on changes in lipid species in subjects with cirrhotic livers compared with healthy controls (Gorden *et al.*, 2011). Given that PS is essential for hepatocytes, we hypothesize that decreased activity of these enzymes may be associated with a decrease in the endogenous level of serine, which is the second most connected node in our identified Reporter Subnetworks.

Serine is endogenously biosynthesized from a glycolytic intermediate, 3-phospho-D-glycerate. This three-step process is catalysed by phosphoglycerate dehydrogenase (*PHGDH*), phosphoserine aminotransferase 1 (*PSAT1*) and phosphoserine phosphatase (*PSPH*), as shown in Figure 4-16b. An alternative synthesis pathway is via the reversible interconversion with glycine through hydroxymethyltransferases (*SHMT1*) and (*SHMT2*). Serine can also be derived from the diet and the degradation of protein and/or phospholipids.

Through differential analysis of transcriptomics data from the NASH patients, it was also observed that gene expression of several enzymes that use serine, including *CBS* (cysteine synthesis), *SARS2* (aminoacyl-tRNA biosynthesis), *SHMT1* and *SHMT2* (glycine synthesis) were significantly downregulated (p-values < 0.05) whereas *SPTLC1* and *SPTLC2* (sphingosine synthesis) were significantly upregulated. Downregulation of *CBS* that catalyses the conversion of serine and homocysteine to L-cystathionine and upregulation of *MTR* that condensates homocysteine to methionine through the use of 5-methyl-THF indicate that there are metabolic changes around homocysteine in NASH patients. Notably, it has been earlier reported that the plasma homocysteine level can be used for diagnosing NASH and classifying steatosis and NASH patients (Gulsen *et al.*, 2005). It is not always straight forward to relate blood concentrations to gene expression levels of the involved enzymes, but our model-based analysis suggests a mechanistic explanation for this.

Taken together, the results suggest that the changes in the level of PS in liver (Gorden *et al.*, 2011) as well as the relative increase in the homocysteine blood level (Gulsen *et al.*, 2005) is caused by decreased level of endogenous serine. In order to test this hypothesis, we checked the expression level of enzymes that catalyse the biosynthesis of serine in the liver of NASH patients, and it was observed that the expression levels of *PHGDH*, *PSAT1*, *PSPH* in serine synthesis pathway (SSP) and *SHMT1* and *SHMT2* enzymes were significantly downregulated. Decreased levels of serine in NASH patients was supported by the plasma profiling of amino acids in NASH patients, and it was reported that the serine (15 % decrease, p-value=0.0568) level in the plasma is decreased (Kalhan *et al.*, 2011).

Equimolar amounts of serine and α -ketoglutarate (AKG) are synthesized in the SSP, and downregulation of reactions in SSP decrease the anaplerosis of glutamate to the TCA cycle in the form of AKG (Possemato *et al.*, 2011). Decreased level of serine also causes an accumulation of upstream glycolytic intermediates (Chaneton *et al.*, 2012), and a decreased flux of mitochondrial AKG is compensated by an increased flux of pyruvate to oxaloacetate in a healthy cell (see Figure 4-16b). In order to investigate the occurrence of this mechanism in NASH patients, we examined all mitochondrial reactions involving pyruvate as reactant in iHepatocytes2260. We found that the corresponding genes were downregulated for five out of seven such reactions. Furthermore, we investigated the expression of level of mitochondrial pyruvate carriers (*MPC1* and *MPC2*) and mitochondrial AKG/malate carrier (*SLC25A11*) and it was observed that their expression levels were downregulated in NASH patients. These indicate that the mitochondrial metabolic activity of hepatocytes is decreased in NASH patients compared to healthy subjects. This is in agreement with findings in **Paper V** where we investigated the metabolic changes in the case of fat accumulation in adipocytes in response to obesity.

Based on our analysis, increasing the serine level in hepatocytes through the uptake of serine as a dietary supplement could be beneficial for NASH patients. Activity loss of *PHGDH* in SSP in the brain, which causes low serine and glycine levels and affects neuronal function, is reversed by serine supplementation (de Koning *et al.*, 2004). The toxicity and the dosage of serine during its uptake through diet have been previously studied. Furthermore, long-term serine treatment decreased the homocysteine level in animal studies (Girard-Globa *et al.*, 1972) and in humans in a single dose situations (Verhoef *et al.*, 2004).

One other possible way to increase the serine level in order to offer the possibility for therapeutic interventions is activation of the enzymes in SSP or *SHMT1* and *SHMT2* that converts glycine to serine. Three different enzymes constitute the SSP and it is earlier reported that *PSPH* is the rate-controlling enzyme for the SSP in liver (Lund *et al.*, 1985). Activation of the SSP through the amplification of *PSPH* may also decrease the flux through pyruvate and lactate formation in cytosol since increased pyruvate and lactate levels were previously reported in NASH patients (Kalhan *et al.*, 2011).

In conclusion, the HMR 2.0 database published in this paper is the most comprehensive general human GEM, and the only one which incorporates extensive lipid metabolism. By applying the tINIT algorithm (Agren *et al.*, 2013b) in order to reconstruct a draft hepatocyte GEM, we found 61 proteins which the algorithm flagged as likely to have been misidentified in HPA. At least 51 (83%) of those proteins were indeed misidentified when the IHC stainings were re-evaluated and/or when tested for with different antibodies. This is an excellent example of how a network-centric analysis can pick up on targets which would not be possible to identify by other means. The reconstructed iHepatocyte2260 GEM was then applied to study metabolism in NASH. Our analysis suggests that it may be possible to diagnose NASH through identified metabolic biomarkers such as 5-methyl-THF, 5-formyl-THF, CS, and HS levels in blood. Furthermore, the development of therapeutics techniques based on the enhancement of endogenous serine and AKG levels may correct the underlying etiology of NASH. This could be achieved by activation (or elevated expression) of *PSPH* and *SHMT1* and inhibition of *BCAT1*.

5 Conclusions and future perspectives

5.1 Conclusions

In the introduction section I identified three key issues in need of more research. These were: 1) model reconstruction is very labour intensive and error prone, 2) GEMs are underused as scaffolds for omics integration, 3) difficulties associated with modelling of complex organisms.

In **Paper I**, genome-scale metabolic modelling was applied to succinic acid production in *S. cerevisiae*. The modelling was used to suggest single gene deletions, out of which three were validated experimentally. A central aspect in the study is the effect of oxygen on succinate production, where the simulations suggest that fully anaerobic conditions are necessary. One of the gene deletions, *Adic1*, led to a significant succinate yield, in close agreement with the model predictions. However, the yield was not as high as what had previously been achieved by utilizing a quadruple deletion strategy and aerobic conditions (0.02 C-mol/C-mol glucose vs. 0.07 C-mol/C-mol glucose). A distinct advantage over that strategy is that anaerobic fermentations are preferred industrially. Both these strategies result in far less succinate than what is possible in bacterial hosts, and the result can therefore be seen mainly as proof of concept. The most interesting result, in my opinion, is rather that the study provides some clues on the roles of Frds1 and reductive TCA cycle in mitochondrial NAD⁺ regeneration under anaerobic conditions, which is still not fully elucidated.

Paper I represents an excellent example of how powerful the systems biology cycle (see Figure 2-3) can be. Fermentation data for the wild-type was used to constrain a GEM, which was then used to study product formation under different conditions. The resulting flux distributions were visualized and analysed, and the simulation parameters were adjusted until the model correctly predicted the phenotype. Simulations were used to form hypotheses and predict the yields of product following the suggested perturbations. The predictions were then validated experimentally. Lastly, the experimental results were used to suggest further studies and identify parts of the model that might need to be revised.

One of the key issues that warranted further investigation, as identified in the introduction, was that the potential of GEMs to act as scaffolds for data integration was being underused. This gave the motivation for us to develop the algorithm in **Paper II**, which aims at integrating fermentation data with gene expression data, with the purpose of identifying transcriptionally controlled reactions. Such reactions could then form suitable targets for metabolic engineering. The algorithm was applied to study shifts in carbon sources in *S. cerevisiae* as a validation case. We identified three transcription factors which specifically regulated enzymes in transcriptionally controlled reactions. This implies that there is a global regulation of major flux alterations, which is highly relevant for metabolic engineering purposes.

Any textbook or review on metabolic engineering will have a figure on how modelling and experimental efforts interact and feed off the results from each other (as does this one, see Figure 2-3). However, this represents an ideal case and not necessarily how it works in practice. Instead, much of the modelling is based on pre-existing data from literature. A powerful aspect of the algorithm is therefore that it relies on data that is very widely available. The flexibility of the algorithm was shown in **Paper III**, **Paper V**, and **Paper VI**.

In **Paper III** we reconstructed a GEM for the filamentous fungi *Penicillium chrysogenum* and used it to study penicillin biosynthesis. The algorithm from **Paper II** was applied for comparison of an industrial high-producing strain and the wild-type strain in order to identify potential targets for increasing the penicillin yield. 36 reactions were identified as being transcriptionally controlled and upregulated in the industrial strain. They are therefore potential targets for further overexpression. We also identified three single gene deletions which were predicted to result in a 21% increase in penicillin production. In addition, we found strong evidence that in the industrial strain the transsulfuration pathway is the dominating pathway for cysteine biosynthesis, even though the enzymes for the energetically more efficient direct sulfhydrylation pathway have been identified in *P. chrysogenum*. This also represents an interesting target, as cysteine synthesis can be a limiting step in penicillin production.

The GEM for *P. chrysogenum* was reconstructed and validated using the RAVEN Toolbox. The software aims at automating parts of the GEM reconstruction process in order to allow for faster reconstruction of high-quality GEMs. It was developed to address the first issue identified in the introduction; model reconstruction is very labour intensive and error-prone. The RAVEN Toolbox has three main foci: 1) automatic reconstruction of GEMs based on protein homology and integrated quality control, 2) network analysis, modelling and interpretation of simulation results, 3) visualization of GEMs using pre-drawn metabolic network maps. It contains a number of novel approaches for gap filling, assignment of subcellular localization of reactions, and mapping of genes based on homology. This software represents by far the largest single part of the work carried out during my Ph.D. studies. As described in section 3.3, a number of other software and algorithms have been published which are partly overlapping with the RAVEN Toolbox in terms of functionality. The fundamental difference between the RAVEN Toolbox and those software is that it is not just a software for automatic reconstruction; it is a software for working with reconstruction.

In **Paper IV** we built a very comprehensive database of human metabolism, and then reconstructed cell type-specific models as subsets of this generic database. We did this by developing an algorithm which integrates different omics types, such as proteomics, transcriptomics and metabolomics, and then generates models which are in agreement with the data. The algorithm, INIT, was tailored to use large-scale proteomics data generated within the HPA project. The workflow was validated by an extensive comparison to a manually published high-quality model. We then applied the algorithm to reconstruct models for 69 cell types and 16 cancers. The models were then analysed in order to identify subnetworks which are more prominent in cancers. Such networks can be potential targets for treatment. The solutions contained several well-known targets and a few novel ones. Particularly, we found a network dealing with detoxification of aminoacetone, which we propose as a target for therapeutic intervention.

The papers described above, with the exception of **Paper I**, are primarily resource and methodology papers. The sampling algorithm, the *P. chrysogenum* GEM, the RAVEN Toolbox, the INIT algorithm, and the cell type-specific GEMs are arguably larger contributions to the scientific community than the biological interpretations drawn from applying them. The last two papers have a stronger biological component, and show how the methods and resources developed in the first set of papers can be applied.

In **Paper V** we reconstructed a GEM for adipocytes based on proteomics data generated together with our collaborators in the HPA project. The model represents a significant step forward since it is the first human GEM with extensive lipid metabolism incorporated. The

model was built on the generic database developed in **Paper IV** and validated in the RAVEN Toolbox (**Paper III**). The model could correctly capture the reduced dynamics of lipid droplet formation in obese subjects, and we could see that this was associated to mitochondrial dysfunction. The model was then used together with the algorithm developed in **Paper II** to identify metabolic differences between healthy and obese siblings. This led us to hypothesize that obesity could be treated with interventions aiming at reactivating the mitochondria by increasing the availability of acetyl-CoA. An alternative approach was to target the degradation of heparan sulfate proteoglycans.

An important lesson from this project was how important it is to have collaborators in the medical field. The reconstruction of such a high-quality model would not have been possible without the large-scale proteomics data generated specifically for this purpose. The study represents an excellent example of the usability of GEMs as scaffolds for omics integration, as is shown in the study on metabolic differences between healthy and obese siblings.

In **Paper VI** we built on the model developed in **Paper V** and adapted it to hepatocytes by using INIT (**Paper IV** and Agren *et al.* (2013b)). During this step the algorithm included 61 proteins in the model even though the proteomics data suggested that they did not exist in hepatocytes. At least 51 (83%) of those proteins were indeed misidentified when the IHC stainings were re-evaluated and/or when tested for with different antibodies. This is an excellent example of how a network-centric analysis can pick up on targets which would not be possible to identify by other means. The model was then applied to study metabolism in patients with non-alcoholic fatty liver disease. Our main findings pointed to a central role of serine, and we proposed that enhancement of endogenous serine levels may correct the underlying etiology of the disease.

A number of methods have previously been developed in order to deal with the three issues set forward in the introduction (as described in detail in section 3). Despite that, I hope to have shown that the work put forward in this thesis has contributed in some small amount to solving them, and that in doing so it has also resulted in novel biological insights.

5.2 Future perspectives

During the last couple of decades the constraint-based approach to modelling has proven to be very well-suited for metabolic engineering purposes. More recently, it has also started to prove its applicability to human health and disease. Extensive method development has been carried out in order to improve on the quality of GEMs and reduce the efforts involved in reconstructing them, part of which has been performed within this Ph.D. project. This has lowered the bar for reconstructing high-quality models, and GEMs for prokaryotes can now be routinely reconstructed with little manual input. Much effort has also gone into the development of methods for guiding metabolic engineering and strain design. Although by no means solved problems, there now exists so many algorithms for these purposes that I would suggest that the field turns its attention to some other remaining challenges while survival of the fittest sorts out the most applicable algorithms.

One issue that I think is of paramount importance is that of transport reactions. The fraction of transport reactions is often >20% for eukaryotic models and the evidence level for them is significantly worse than for the enzymatically catalysed reactions. These reactions are routinely included “for connectivity reasons”. This has large implications for modelling of eukaryotic organisms, since much of the complexity of metabolism comes from the

compartmentalization of redox, charge and energy balancing. A large-scale screening effort of transporter substrate specificity, membrane localization and transport direction for a eukaryotic model organism would be immensely valuable for the field.

For applications in human health and disease there are two issues that I think warrant special attention. The first is the absence of a standard operating procedure (SOP) for reconstruction of cell type-specific GEMs. The development of detailed SOPs for reconstruction of microbial models had a hugely positive effect on the field, and a similar push is needed for multicellular organisms as well. The second issue deals with data availability. Since constraint-based modelling has evolved in close interaction with metabolic engineering, which allows for carefully set up fermentations and quantification of internal and/or external fluxes, many of the methods are based on fluxomics data. However, this is not a common experimental setup for cultivation of mammalian cells, which are most often grown in complex media with no quantification of exchange fluxes. It would be highly relevant for the field with a medically oriented project tailored to supply the data best fitted for constraint-based modelling. This would serve as an example, both to the medical field and to the metabolic engineering field, of the capabilities of constraint-based modelling of human cells.

A third area which I'm confident will be a future focus is that of interactions between models with different objectives. Possible applications include interactions between organs, bacteria in mixed fermentations such as in the gut, cells with the same genotype but with different sets of expressed genes due to stochastic noise, or between pathogens and their hosts. The first steps towards this goal have already been taken, but there is a long way left.

Lastly, I think the field is approaching maturity when it comes to microbial systems, but that it has yet to prove itself when it comes to medical applications. A large proportion of the published papers are still based on some type of novel algorithm or method, and there are not all that many examples of applications of known methods to provide answers to concrete biological questions. The modelling is still mainly used for data analysis/mining, and not interactively and iteratively for testing hypotheses and generating new data. A few success stories would open up for more funding and collaborations.

To conclude, the field has developed immensely in the short few years that I have worked in it. All the signs point to that the coming years will be just as exciting.

Acknowledgements

First and foremost, I would like to extend my warmest gratitude to my supervisor Jens Nielsen. I have always felt like you believed in my abilities, and that you had my back no matter what. You gave me the freedom to follow my ideas whenever possible, and structure and tight reins whenever needed. I have learnt immensely from you, and your breadth of knowledge has always been an inspiration.

The following people have been fundamental for my development from a young student to something vaguely resembling a scientist. Jose Manuel Otero, without your enthusiasm and love of science I would not be where I am today. Sergio Bordel Velasco, the discussions with you during the first year of my studies were the most intellectually stimulating and rewarding ones, and really made me feel like I was actually part of academia now. Intawat Nookaew, you have been somewhat of an older brother all throughout my studies; guiding me and teaching me about the ins and outs of working in academia. My sincerest thank to all of you.

None of the work presented in this thesis has been carried out by me on my own. I would like to thank the people who have worked with me to make it possible: Wanwipa Vongsangnak, Liming Liu, Saeed Shoaie, Marija Cvijovic, Natapol Pornputtpong, Stephan Pabinger, and Roberto Olivares-Hernandez. A special thanks to Adil Mardinoglu. We have worked closely together for several years now, and much of the work in this thesis is thanks to you. You have also grown to become a valued friend. I want to give a warm thank you to Erica Dahlin, Martina Butorac, Malin Nordvall, and Marie Nordquist for helping me when I was looking lost in the lab or needed help with all kinds of practicalities.

I would also like to extend my gratitude to my external collaborators at the Human Protein Atlas and at Sandoz. The data and knowledge that you have provided have enabled several of the studies presented here. A special thanks to Timo Hardiman and Rudolf Mitterbauer at Sandoz for taking care of me when in Austria.

In addition, I would like to thank all the people that I haven't collaborated with professionally, but whose minds and wits and laughs have made this whole thing the fantastic experience that it has been: Fredrik Karlsson, Leif Väre, Siavash Partow, Amir Feizi, Verena Sievers, Kuk-ki Hong, Tobias Österlund, Christoph Knuf, Luis Caspeta-Guadarrama, Keith Tyo, Marta Papini, Goutham Vemuri, Dina Petranovic, and many more.

My deepest gratitude goes to my parents. Your unwavering love and support has made me who I am.

Evelina, you own my heart. I have no clue where life will take us now, but I know we're going to have so much fun!

References

- Acencio, M.L. and Lemke, N. (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information, *BMC Bioinformatics*, 10, p. 290.
- Adrio, J.L. and Demain, A.L. (2003) Fungal biotechnology, *Int Microbiol*, 6(3), pp. 191-199.
- Afratis, N., et al. (2012) Glycosaminoglycans: key players in cancer cell biology and treatment, *Febs Journal*, 279(7), pp. 1177-1197.
- Agren, R., et al. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, *PLoS Comput Biol*, 8(5), p. e1002518.
- Agren, R., et al. (2013a) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*, *PLoS Comput Biol*, 9(3), p. e1002980.
- Agren, R., et al. (2013b) Drug discovery through the use of personalized genome-scale metabolic models for liver cancer, (Submitted).
- Agren, R., et al. (2013c) Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production, *J Ind Microbiol Biot*, (In press).
- Aho, T., et al. (2010) Reconstruction and validation of RefRec: a global model for the yeast molecular interaction network, *PLoS One*, 5(5), p. e10662.
- Aiba, S. and Matsuoka, M. (1979) Identification of Metabolic Model - Citrate Production from Glucose by *Candida Lipolytica*, *Biotechnol. Bioeng.*, 21(8), pp. 1373-1386.
- Aiba, S., et al. (1980) Enhancement of Tryptophan Production by *Escherichia-Coli* as an Application of Genetic-Engineering, *Biotechnol Lett*, 2(12), pp. 525-530.
- Akesson, M., et al. (2004) Integration of gene expression data into genome-scale metabolic models, *Metab Eng*, 6(4), pp. 285-293.
- Alcantara, R., et al. (2012) Rhea--a manually curated resource of biochemical reactions, *Nucleic Acids Res*, 40(Database issue), pp. D754-760.
- Altschul, S.F., et al. (1990) Basic local alignment search tool, *J Mol Biol*, 215(3), pp. 403-410.
- Andersen, M.R., et al. (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*, *Mol Syst Biol*, 4, p. 178.
- Andersen, M.R., et al. (2009) Systemic analysis of the response of *Aspergillus niger* to ambient pH, *Genome Biol*, 10(5), p. R47.
- Andersen, M.R., et al. (2011) Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88, *Genome Res*, 21(6), pp. 885-897.
- Anstee, Q.M. and Day, C.P. (2012) S-adenosylmethionine (SAME) therapy in liver disease: a review of current evidence and clinical utility, *J Hepatol*, 57(5), pp. 1097-1109.
- Arakawa, K., et al. (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes, *BMC Bioinformatics*, 7, p. 168.
- Asadollahi, M.A., et al. (2009) Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through in silico driven metabolic engineering, *Metab Eng*, 11(6), pp. 328-334.
- Ascha, M.S., et al. (2010) The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis, *Hepatology*, 51(6), pp. 1972-1978.
- Auffray, C., et al. (2009) Systems medicine: the future of medical genomics and healthcare, *Genome Med*, 1(1), p. 2.
- Baffy, G., et al. (2012) Hepatocellular carcinoma in non-alcoholic fatty liver disease: An emerging menace, *J Hepatol*, 56(6), pp. 1384-1391.
- Bailey, J.E. (1991) Toward a science of metabolic engineering, *Science*, 252(5013), pp. 1668-1675.
- Balagurunathan, B., et al. (2012) Reconstruction and analysis of a genome-scale metabolic model for *Scheffersomyces stipitis*, *Microb Cell Fact*, 11, p. 27.
- Baranano, D.E., et al. (2002) Biliverdin reductase: a major physiologic cytoprotectant, *Proc Natl Acad Sci U S A*, 99(25), pp. 16093-16098.
- Beard, D.A., et al. (2002) Energy balance for analysis of complex metabolic networks, *Biophys J*, 83(1), pp. 79-86.
- Becker, S.A. and Palsson, B.O. (2008) Context-specific metabolic networks are consistent with experiments, *PLoS Comput Biol*, 4(5), p. e1000082.
- Beg, Q.K., et al. (2007) Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity, *Proc Natl Acad Sci U S A*, 104(31), pp. 12663-12668.
- Bengtsson, G. and Olivecrona, T. (1980) Lipoprotein lipase. Mechanism of product inhibition, *Eur J Biochem*, 106(2), pp. 557-562.
- Bennett, J.W. (1998) Mycotechnology: the role of fungi in biotechnology, *J Biotechnol*, 66(2-3), pp. 101-107.
- Berglund, L., et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies, *Mol Cell Proteomics*, 7(10), pp. 2019-2027.
- Bonarius, H.P.J., et al. (1997) Flux analysis of underdetermined metabolic networks: The quest for the missing constraints, *Trends Biotechnol.*, 15(8), pp. 308-314.
- Bordbar, A., et al. (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions, *Mol Syst Biol*, 6, p. 422.
- Bordbar, A., et al. (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology, *BMC Syst Biol*, 5, p. 180.

- Bordbar, A. and Palsson, B.O. (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology, *J Intern Med*, 271(2), pp. 131-141.
- Bordel, S., et al. (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes, *PLoS Comput Biol*, 6(7), p. e1000859.
- Bro, C., et al. (2006) In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production, *Metab Eng*, 8(2), pp. 102-111.
- Brochado, A.R., et al. (2010) Improved vanillin production in baker's yeast through in silico design, *Microb Cell Fact*, 9, p. 84.
- Brooks, J.P., et al. (2012) Gap detection for genome-scale constraint-based models, *Advances in bioinformatics*, 2012, p. 323472.
- Bundy, J.G., et al. (2007) Evaluation of predicted network modules in yeast metabolism using NMR-based metabolite profiling, *Genome Res*, 17(4), pp. 510-519.
- Burgard, A.P. and Maranas, C.D. (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions, *Biotechnol. Bioeng.*, 82(6), pp. 670-677.
- Burgard, A.P., et al. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnol Bioeng*, 84(6), pp. 647-657.
- Burgard, A.P., et al. (2004) Flux coupling analysis of genome-scale metabolic network reconstructions, *Genome Res*, 14(2), pp. 301-312.
- Cakir, T., et al. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions, *Mol Syst Biol*, 2, p. 50.
- Camarasa, C., et al. (2003) Investigation by ¹³C-NMR and tricarboxylic acid (TCA) deletion mutant analysis of pathways for succinate formation in *Saccharomyces cerevisiae* during anaerobic fermentation, *Microbiology*, 149(Pt 9), pp. 2669-2678.
- Camarasa, C., et al. (2007) Role in anaerobiosis of the isoenzymes for *Saccharomyces cerevisiae* fumarate reductase encoded by OSM1 and FRDS1, *Yeast*, 24(5), pp. 391-401.
- Cao, H., et al. (2008) Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism, *Cell*, 134(6), pp. 933-944.
- Caspeta, L., et al. (2012) Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials, *BMC Syst Biol*, 6, p. 24.
- Caveney, E., et al. (2011) Pharmaceutical interventions for obesity: a public health perspective, *Diabetes Obes Metab*, 13(6), pp. 490-497.
- Chan, C., et al. (2003) Metabolic flux analysis of cultured hepatocytes exposed to plasma, *Biotechnology and Bioengineering*, 81(1), pp. 33-49.
- Chaneton, B., et al. (2012) Serine is a natural ligand and allosteric activator of pyruvate kinase M2, *Nature*, 491(7424), pp. 458-462.
- Chang, R.L., et al. (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model, *PLoS Comput Biol*, 6(9), p. e1000938.
- Chechik, G., et al. (2008) Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network, *Nat Biotechnol*, 26(11), pp. 1251-1259.
- Chen, L. and Vitkup, D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles, *Genome Biol*, 7(2), p. R17.
- Cherry, J.M., et al. (1998) SGD: *Saccharomyces* Genome Database, *Nucleic Acids Res*, 26(1), pp. 73-79.
- Choi, B.K., et al. (2003) Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast *Pichia pastoris*, *Proc Natl Acad Sci U S A*, 100(9), pp. 5022-5027.
- Chung, B.K., et al. (2010) Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement, *Microb Cell Fact*, 9, p. 50.
- Cimini, D., et al. (2009) Global transcriptional response of *Saccharomyces cerevisiae* to the deletion of SDH3, *BMC Syst Biol*, 3, p. 17.
- Corning, P.A. (2012) The re-emergence of emergence, and the causal role of synergy in emergent evolution, *Synthese*, 185(2), pp. 295-317.
- Costenoble, R., et al. (2011) Comprehensive quantitative analysis of central carbon and amino-acid metabolism in *Saccharomyces cerevisiae* under multiple conditions by targeted proteomics, *Mol Syst Biol*, 7, p. 464.
- Covert, M.W., et al. (2001) Regulation of gene expression in flux balance models of metabolism, *J Theor Biol*, 213(1), pp. 73-88.
- Croft, D., et al. (2011) Reactome: a database of reactions, pathways and biological processes, *Nucleic Acids Res*, 39(Database issue), pp. D691-697.
- Cvijovic, M., et al. (2010) BioMet Toolbox: genome-wide analysis of metabolism, *Nucleic Acids Res*, 38(Web Server issue), pp. W144-149.
- Daran-Lapujade, P., et al. (2007) The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels, *Proc Natl Acad Sci U S A*, 104(40), pp. 15753-15758.
- David, H., et al. (2006) Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans*, *Genome Biol*, 7(11), p. R108.
- David, H., et al. (2008) Analysis of *Aspergillus nidulans* metabolism at the genome-scale, *BMC Genomics*, 9, p. 163.
- de Koning, T.J., et al. (2004) Prenatal and early postnatal treatment in 3-phosphoglycerate-dehydrogenase deficiency, *Lancet*, 364(9452), pp. 2221-2222.

References

- Dean, J.T., et al. (2009) Resistance to diet-induced obesity in mice with synthetic glyoxylate shunt, *Cell Metab*, 9(6), pp. 525-536.
- DeJongh, M., et al. (2007) Toward the automated generation of genome-scale metabolic networks in the SEED, *BMC Bioinformatics*, 8, p. 139.
- del Rio, G., et al. (2009) How to identify essential genes from molecular networks?, *BMC Syst Biol*, 3, p. 102.
- Delcher, A.L., et al. (1999) Improved microbial gene identification with GLIMMER, *Nucleic Acids Res*, 27(23), pp. 4636-4641.
- Derrien, T., et al. (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps, *Bioinformatics*, 23(4), pp. 498-499.
- Deutscher, D., et al. (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network, *Nat Genet*, 38(9), pp. 993-998.
- Diamant, I., et al. (2009) A network-based method for predicting gene-nutrient interactions and its application to yeast amino-acid metabolism, *Mol Biosyst*, 5(12), pp. 1732-1739.
- Dikicioglu, D., et al. (2008) Integration of metabolic modeling and phenotypic data in evaluation and improvement of ethanol production using respiration-deficient mutants of *Saccharomyces cerevisiae*, *Appl Environ Microbiol*, 74(18), pp. 5809-5816.
- Dobson, P.D., et al. (2010) Further developments towards a genome-scale metabolic model of yeast, *BMC Syst Biol*, 4, p. 145.
- Draths, K.M., et al. (1992) Biocatalytic Synthesis of Aromatics from D-Glucose - the Role of Transketolase, *Journal of the American Chemical Society*, 114(10), pp. 3956-3962.
- Duarte, N.C., et al. (2004a) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model, *Genome Res*, 14(7), pp. 1298-1309.
- Duarte, N.C., et al. (2004b) Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*, *BMC Genomics*, 5, p. 63.
- Duarte, N.C., et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data, *Proc Natl Acad Sci U S A*, 104(6), pp. 1777-1782.
- Dudakovic, A., et al. (2011) Geranylgeranyl diphosphate depletion inhibits breast cancer cell migration, *Invest New Drugs*, 29(5), pp. 912-920.
- Durot, M., et al. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications, *FEMS Microbiol Rev*, 33(1), pp. 164-190.
- Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, 14(9), pp. 755-763.
- Edwards, J.S., et al. (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data, *Nat Biotechnol*, 19(2), pp. 125-130.
- Eisenberg, T., et al. (2009) Induction of autophagy by spermidine promotes longevity, *Nat Cell Biol*, 11(11), pp. 1305-1314.
- Elander, R.P. (2003) Industrial production of beta-lactam antibiotics, *Appl Microbiol Biotechnol*, 61(5-6), pp. 385-392.
- Eruslanov, E., et al. (2009) Altered expression of 15-hydroxyprostaglandin dehydrogenase in tumor-infiltrated CD11b myeloid cells: a mechanism for immune evasion in cancer, *J Immunol*, 182(12), pp. 7548-7557.
- Fabry, M.E., et al. (1981) Some aspects of the pathophysiology of homozygous Hb CC erythrocytes, *J Clin Invest*, 67(5), pp. 1284-1291.
- Famili, I., et al. (2003) *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network, *Proc Natl Acad Sci U S A*, 100(23), pp. 13134-13139.
- Farese, R.V., Jr. and Walther, T.C. (2009) Lipid droplets finally get a little R-E-S-P-E-C-T, *Cell*, 139(5), pp. 855-860.
- Fell, D.A. and Small, J.R. (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints, *Biochem J*, 238(3), pp. 781-786.
- Fisher, C.D., et al. (2009) Hepatic cytochrome P450 enzyme alterations in humans with progressive stages of nonalcoholic fatty liver disease, *Drug Metab Dispos*, 37(10), pp. 2087-2094.
- Fletcher, D.A. and Theriot, J.A. (2004) An introduction to cell motility for the physical scientist, *Physical biology*, 1(1-2), pp. T1-10.
- Folger, O., et al. (2011) Predicting selective drug targets in cancer through metabolic networks, *Mol Syst Biol*, 7, p. 501.
- Fong, S.S., et al. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid, *Biotechnol Bioeng*, 91(5), pp. 643-648.
- Forberg, C. and Haggstrom, L. (1987) Effects of Cultural Conditions on the Production of Phenylalanine from a Plasmid-Harboring *Escherichia-Coli* Strain, *Appl. Microbiol. Biotechnol.*, 26(2), pp. 136-140.
- Forberg, C., et al. (1988) Correlation of Theoretical and Experimental Yields of Phenylalanine from Non-Growing Cells of a Rec *Escherichia-Coli* Strain, *Journal of Biotechnology*, 7(4), pp. 319-332.
- Forster, J., et al. (2003a) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network, *Genome Res*, 13(2), pp. 244-253.
- Forster, J., et al. (2003b) Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*, *OMICS*, 7(2), pp. 193-202.
- Fu, P.C. (2009) Gene expression study of *Saccharomyces cerevisiae* under changing growth conditions, *J Chem Technol Biot*, 84(8), pp. 1163-1171.
- Gevorgyan, A., et al. (2008) Detection of stoichiometric inconsistencies in biomolecular models, *Bioinformatics*, 24(19), pp. 2245-2251.
- Ghosh, A., et al. (2011) Genome-scale consequences of cofactor balancing in engineered pentose utilization pathways in *Saccharomyces cerevisiae*, *PLoS One*, 6(11), p. e27316.

- Gille, C., et al. (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology, *Mol Syst Biol*, 6, p. 411.
- Girard-Globa, A., et al. (1972) Long-term adaptation of weanling rats to high dietary levels of methionine and serine, *J Nutr*, 102(2), pp. 209-217.
- Gorden, D.L., et al. (2011) Increased Diacylglycerols Characterize Hepatic Lipid Changes in Progression of Human Nonalcoholic Fatty Liver Disease; Comparison to a Murine Model, *Plos One*, 6(8), p. e22775.
- Greco, D., et al. (2008) Gene expression in human NAFLD, *Am J Physiol Gastrointest Liver Physiol*, 294(5), pp. G1281-1287.
- Gulsen, M., et al. (2005) Elevated plasma homocysteine concentrations as a predictor of steatohepatitis in patients with non-alcoholic fatty liver disease, *J Gastroenterol Hepatol*, 20(9), pp. 1448-1455.
- Hao, T., et al. (2010) Compartmentalization of the Edinburgh Human Metabolic Network, *BMC Bioinformatics*, 11, p. 393.
- Harris, D.M., et al. (2006) Enzymic analysis of NADPH metabolism in beta-lactam-producing *Penicillium chrysogenum*: presence of a mitochondrial NADPH dehydrogenase, *Metab Eng*, 8(2), pp. 91-101.
- Heavner, B.D., et al. (2012) Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network, *BMC Syst Biol*, 6, p. 55.
- Heinken, A., et al. (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut, *Gut microbes*, 4(1), pp. 28-40.
- Heinrich, R., et al. (1977) Metabolic regulation and mathematical models, *Progress in biophysics and molecular biology*, 32(1), pp. 1-82.
- Henry, C.S., et al. (2006) Genome-scale thermodynamic analysis of *Escherichia coli* metabolism, *Biophys J*, 90(4), pp. 1453-1461.
- Henry, C.S., et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models, *Nat Biotechnol*, 28(9), pp. 977-982.
- Herrgard, M.J., et al. (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*, *Genome Res*, 16(5), pp. 627-635.
- Hjersted, J.L. and Henson, M.A. (2009) Steady-state and dynamic flux balance analysis of ethanol production by *Saccharomyces cerevisiae*, *IET systems biology*, 3(3), pp. 167-179.
- Hoffman, J., et al. (2006) Effect of creatine and beta-alanine supplementation on performance and endocrine responses in strength/power athletes, *Int J Sport Nutr Exerc Metab*, 16(4), pp. 430-446.
- Jerby, L., et al. (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism, *Mol Syst Biol*, 6, p. 401.
- Jorgensen, H., et al. (1995a) Metabolic flux distributions in *Penicillium chrysogenum* during fed-batch cultivations, *Biotechnol Bioeng*, 46(2), pp. 117-131.
- Jorgensen, H., et al. (1995b) Analysis of penicillin V biosynthesis during fed-batch cultivations with a high-yielding strain of *Penicillium chrysogenum*, *Appl Microbiol Biotechnol*, 43(1), pp. 123-130.
- Jouhten, P., et al. (2012) Dynamic flux balance analysis of the metabolism of *Saccharomyces cerevisiae* during the shift from fully respirative or respirofermentative metabolic states to anaerobiosis, *FEBS J*, 279(18), pp. 3338-3354.
- Joyce, A.R. and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets, *Nat Rev Mol Cell Bio*, 7(3), pp. 198-210.
- Kalapos, M.P. (1994) Methylglyoxal toxicity in mammals, *Toxicol Lett*, 73(1), pp. 3-24.
- Kaleta, C., et al. (2009) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns, *Genome Res*, 19(10), pp. 1872-1883.
- Kalhan, S.C., et al. (2011) Plasma metabolomic profile in nonalcoholic fatty liver disease, *Metabolism*, 60(3), pp. 404-413.
- Kanehisa, M., et al. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res*, 38(Database issue), pp. D355-360.
- Kang, Y., et al. (1996) Effect of methylglyoxal on human leukaemia 60 cell growth: modification of DNA G1 growth arrest and induction of apoptosis, *Leuk Res*, 20(5), pp. 397-405.
- Karp, P.D., et al. (2002) The Pathway Tools software, *Bioinformatics*, 18 Suppl 1, pp. S225-232.
- Karp, R.M. (2008) George Dantzig's impact on the theory of computation, *Discrete Optim*, 5(2), pp. 174-185.
- Kharchenko, P., et al. (2005) Expression dynamics of a cellular metabolic network, *Mol Syst Biol*, 1, p. 2005 0016.
- Kitano, H. (2002a) Computational systems biology, *Nature*, 420(6912), pp. 206-210.
- Kitano, H. (2002b) Systems biology: a brief overview, *Science*, 295(5560), pp. 1662-1664.
- Koppenol, W.H., et al. (2011) Otto Warburg's contributions to current concepts of cancer metabolism, *Nat Rev Cancer*, 11(5), pp. 325-337.
- Kuepfer, L., et al. (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*, *Genome Res*, 15(10), pp. 1421-1430.
- Kumar, V.S. and Maranas, C.D. (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions, *PLoS Comput Biol*, 5(3), p. e1000308.
- Kummel, A., et al. (2006a) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data, *Mol Syst Biol*, 2, p. 2006 0034.
- Kummel, A., et al. (2006b) Systematic assignment of thermodynamic constraints in metabolic network models, *BMC Bioinformatics*, 7, p. 512.
- Kusminski, C.M. and Scherer, P.E. (2012) Mitochondrial dysfunction in white adipose tissue, *Trends Endocrinol Metab*, 23(9), pp. 435-443.
- Lafontan, M. (2008) Advances in adipose tissue metabolism, *Int J Obes (Lond)*, 32 Suppl 7, pp. S39-51.

References

- Lago, F., et al. (2007) Adipokines as emerging mediators of immune response and inflammation, *Nature clinical practice. Rheumatology*, 3(12), pp. 716-724.
- Lake, A.D., et al. (2011) Analysis of global and absorption, distribution, metabolism, and elimination gene expression in the progressive stages of human nonalcoholic fatty liver disease, *Drug Metab Dispos*, 39(10), pp. 1954-1960.
- Lee, D., et al. (2012) Improving metabolic flux predictions using absolute gene expression data, *BMC Syst Biol*, 6, p. 73.
- Lee, J.M., et al. (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks, *PLoS Comput Biol*, 4(5), p. e1000086.
- Lee, S.J., et al. (2005) Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation, *Appl Environ Microbiol*, 71(12), pp. 7880-7887.
- Lewis, N.E., et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain, *Nat Biotechnol*, 28(12), pp. 1279-1285.
- Lewis, N.E., et al. (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods, *Nat Rev Microbiol*, 10(4), pp. 291-305.
- Liu, L., et al. (2010) Use of genome-scale metabolic models for understanding microbial physiology, *FEBS Lett*, 584(12), pp. 2556-2564.
- Liu, T., et al. (2012) A constraint-based model of *Scheffersomyces stipitis* for improved ethanol production, *Biotechnology for biofuels*, 5(1), p. 72.
- Loira, N., et al. (2012) A genome-scale metabolic model of the lipid-accumulating yeast *Yarrowia lipolytica*, *BMC Syst Biol*, 6, p. 35.
- Lookene, A., et al. (1996) Interaction of lipoprotein lipase with heparin fragments and with heparan sulfate: stoichiometry, stabilization, and kinetics, *Biochemistry*, 35(37), pp. 12155-12163.
- Lund, K., et al. (1985) The reactions of the phosphorylated pathway of L-serine biosynthesis: thermodynamic relationships in rabbit liver in vivo, *Arch Biochem Biophys*, 237(1), pp. 186-196.
- Ma, H., et al. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis, *Mol Syst Biol*, 3, p. 135.
- Machado, M.V. and Cortez-Pinto, H. (2012) Non-invasive diagnosis of non-alcoholic fatty liver disease. A critical appraisal, *J Hepatol*.
- Mahadevan, R. and Schilling, C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, *Metab Eng*, 5(4), pp. 264-276.
- Mahadevan, R. and Lovley, D.R. (2008) The degree of redundancy in metabolic genes is linked to mode of metabolism, *Biophys J*, 94(4), pp. 1216-1220.
- Mardinoglu, A. and Nielsen, J. (2012) Systems medicine and metabolic modelling, *J Intern Med*, 271(2), pp. 142-154.
- Mardinoglu, A., et al. (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte, *Mol Syst Biol*, 9, p. 649.
- Matsuda, F., et al. (2011) Engineering strategy of yeast metabolism for higher alcohol production, *Microb Cell Fact*, 10, p. 70.
- Matthews, L., et al. (2009) Reactome knowledgebase of human biological pathways and processes, *Nucleic Acids Res*, 37(Database issue), pp. D619-622.
- McKinlay, J.B., et al. (2007) Prospects for a bio-based succinate industry, *Appl Microbiol Biotechnol*, 76(4), pp. 727-740.
- McQuaid, S.E., et al. (2011) Downregulation of adipose tissue fatty acid trafficking in obesity: a driver for ectopic fat deposition?, *Diabetes*, 60(1), pp. 47-55.
- Michaelis, L., et al. (2011) The original Michaelis constant: translation of the 1913 Michaelis-Menten paper, *Biochemistry*, 50(39), pp. 8264-8269.
- Mintz-Oron, S., et al. (2009) Network-based prediction of metabolic enzymes' subcellular localization, *Bioinformatics*, 25(12), pp. i247-252.
- Mo, M.L., et al. (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast, *BMC Syst Biol*, 3, p. 37.
- Nacher, J.C., et al. (2006) Identification of metabolic units induced by environmental signals, *Bioinformatics*, 22(14), pp. e375-383.
- Neuschwander-Tetri, B.A. and Caldwell, S.H. (2003) Nonalcoholic steatohepatitis: summary of an AASLD Single Topic Conference, *Hepatology*, 37(5), pp. 1202-1219.
- Newgard, C.B. (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance, *Cell Metab*, 15(5), pp. 606-614.
- Ng, C.Y., et al. (2012) Production of 2,3-butanediol in *Saccharomyces cerevisiae* by in silico aided metabolic engineering, *Microb Cell Fact*, 11, p. 68.
- Nielsen, J. and Jorgensen, H.S. (1995) Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of *Penicillium chrysogenum*, *Biotechnol Prog*, 11(3), pp. 299-305.
- Nielsen, J.H. (1995) Physiological engineering aspects of penicillium chrysogenum, Denmark, Polyteknisk forlag.
- Nookaew, I., et al. (2008) The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism, *BMC Syst Biol*, 2, p. 71.
- Notebaart, R.A., et al. (2006) Accelerating the reconstruction of genome-scale metabolic networks, *BMC Bioinformatics*, 7, p. 296.
- Ogata, H., et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res*, 27(1), pp. 29-34.
- Oh, Y.K., et al. (2007) Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data, *Journal of Biological Chemistry*, 282(39), pp. 28791-28799.

- Olivecrona, G. and Beisiegel, U. (1997) Lipid binding of apolipoprotein CII is required for stimulation of lipoprotein lipase activity against apolipoprotein CII-deficient chylomicrons, *Arterioscl Throm Vas*, 17(8), pp. 1545-1549.
- Olivecrona, T. and Olivecrona, G. (2009) The ins and outs of adipose tissue, in Ehnholm, C. (ed), *Cellular lipid metabolism*, New York, Springer.
- Ostergaard, S., et al. (1998) Identification and purification of O-acetyl-L-serine sulphhydrylase in *Penicillium chrysogenum*, *Appl. Microbiol. Biotechnol.*, 50(6), pp. 663-668.
- Ostergaard, S., et al. (2001) The impact of GAL6, GAL80, and MIG1 on glucose control of the GAL system in *Saccharomyces cerevisiae*, *FEMS Yeast Res*, 1(1), pp. 47-55.
- Osterlund, T., et al. (2012) Fifteen years of large scale metabolic modeling of yeast: developments and impacts, *Biotechnol Adv*, 30(5), pp. 979-988.
- Otero, J.M., et al. (2013) Industrial systems biology of *Saccharomyces cerevisiae* enables novel succinic acid cell factory, *PLoS One*, 8(1), p. e54144.
- Othmer, H.G. (1976) The qualitative dynamics of a class of biochemical control circuits, *Journal of mathematical biology*, 3(1), pp. 53-78.
- Pabinger, S., et al. (2011) MEMOSys: Bioinformatics platform for genome-scale metabolic models, *BMC Syst Biol*, 5, p. 20.
- Paley, S.M. and Karp, P.D. (2006) The Pathway Tools cellular overview diagram and Omics Viewer, *Nucleic Acids Res*, 34(13), pp. 3771-3778.
- Papini, M., et al. (2010) Phosphoglycerate mutase knock-out mutant *Saccharomyces cerevisiae*: physiological investigation and transcriptome analysis, *Biotechnology journal*, 5(10), pp. 1016-1027.
- Papp, B., et al. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast, *Nature*, 429(6992), pp. 661-664.
- Patil, K.R. and Nielsen, J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology, *Proc Natl Acad Sci U S A*, 102(8), pp. 2685-2689.
- Patil, K.R., et al. (2005) Evolutionary programming as a platform for in silico metabolic engineering, *BMC Bioinformatics*, 6, p. 308.
- Patnaik, R. and Liao, J.C. (1994) Engineering of *Escherichia coli* central metabolism for aromatic metabolite production with near theoretical yield, *Appl Environ Microbiol*, 60(11), pp. 3903-3908.
- Patterson, B.W., et al. (2002) Regional muscle and adipose tissue amino acid metabolism in lean and obese women, *American journal of physiology. Endocrinology and metabolism*, 282(4), pp. E931-936.
- Philips, M.R. and Cox, A.D. (2007) Geranylgeranyltransferase I as a target for anti-cancer drugs, *J Clin Invest*, 117(5), pp. 1223-1225.
- Pinney, J.W., et al. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*, *Nucleic Acids Res*, 33(4), pp. 1399-1409.
- Popescu, L. and Yona, G. (2005) Automation of gene assignments to metabolic pathways using high-throughput expression data, *BMC Bioinformatics*, 6, p. 217.
- Possemato, R., et al. (2011) Functional genomics reveal that the serine synthesis pathway is essential in breast cancer, *Nature*, 476(7360), pp. 346-350.
- Prabha, A.N.L., et al. (1988) Similar Effects of Beta-Alanine and Taurine in Cholesterol-Metabolism, *J Bioscience*, 13(3), pp. 263-268.
- Price, N.D., et al. (2003) Genome-scale microbial in silico models: the constraints-based approach, *Trends Biotechnol*, 21(4), pp. 162-169.
- Price, N.D., et al. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints, *Nat Rev Microbiol*, 2(11), pp. 886-897.
- Raab, A.M., et al. (2010) Metabolic engineering of *Saccharomyces cerevisiae* for the biotechnological production of succinic acid, *Metab Eng*, 12(6), pp. 518-525.
- Rector, R.S., et al. (2008) Non-alcoholic fatty liver disease and the metabolic syndrome: an update, *World journal of gastroenterology : WJG*, 14(2), pp. 185-192.
- Reed, J.L., et al. (2006) Systems approach to refining genome annotation, *Proc Natl Acad Sci U S A*, 103(46), pp. 17480-17484.
- Rokhlenko, O., et al. (2007) Constraint-based functional similarity of metabolic genes: going beyond network topology, *Bioinformatics*, 23(16), pp. 2139-2146.
- Rokholm, B., et al. (2010) The levelling off of the obesity epidemic since the year 1999--a review of evidence and perspectives, *Obes Rev*, 11(12), pp. 835-846.
- Romero, P., et al. (2005) Computational prediction of human metabolic pathways from the complete human genome, *Genome Biol*, 6(1), p. R2.
- Rossouw, D., et al. (2009) Comparative transcriptomic approach to investigate differences in wine yeast physiology and metabolism during fermentation, *Appl Environ Microbiol*, 75(20), pp. 6600-6612.
- Satish Kumar, V., et al. (2007) Optimization based automated curation of metabolic reconstructions, *BMC Bioinformatics*, 8, p. 212.
- Sauer, M., et al. (2008) Microbial production of organic acids: expanding the markets, *Trends Biotechnol*, 26(2), pp. 100-108.
- Sauer, U., et al. (2007) Genetics. Getting closer to the whole picture, *Science*, 316(5824), pp. 550-551.
- Schellenberger, J. and Palsson, B.O. (2009) Use of randomized sampling for analysis of metabolic networks, *J Biol Chem*, 284(9), pp. 5457-5461.
- Schilling, C.H., et al. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era, *Biotechnol Prog*, 15(3), pp. 296-303.

References

- Schilling, C.H. and Palsson, B.O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis, *J. Theor. Biol.*, 203(3), pp. 249-283.
- Schneider, C. and Pozzi, A. (2011) Cyclooxygenases and lipoxygenases in cancer, *Cancer Metastasis Rev*, 30(3-4), pp. 277-294.
- Schomburg, I., et al. (2002) BRENDA, enzyme data and metabolic information, *Nucleic Acids Res*, 30(1), pp. 47-49.
- Schuetz, R., et al. (2012) Multidimensional optimality of microbial metabolism, *Science*, 336(6081), pp. 601-604.
- Schuster, S., et al. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, *Trends Biotechnol*, 17(2), pp. 53-60.
- Schwartz, J.M., et al. (2007) Observing metabolic functions at the genome scale, *Genome Biol.*, 8(6).
- Sebti, S.M. and Hamilton, A.D. (2000) Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies, *Oncogene*, 19(56), pp. 6584-6593.
- Segre, D., et al. (2002) Analysis of optimality in natural and perturbed metabolic networks, *Proc Natl Acad Sci U S A*, 99(23), pp. 15112-15117.
- Segre, D., et al. (2005) Modular epistasis in yeast metabolism, *Nat Genet*, 37(1), pp. 77-83.
- Seiler, N. (2003a) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 2. Structural analogues and derivatives, *Curr Drug Targets*, 4(7), pp. 565-585.
- Seiler, N. (2003b) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 1. Selective enzyme inhibitors, *Curr Drug Targets*, 4(7), pp. 537-564.
- Seo, S. and Lewin, H.A. (2009) Reconstruction of metabolic pathways for the cattle genome, *BMC Syst Biol*, 3, p. 33.
- Sheikh, K., et al. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*, *Biotechnol Prog*, 21(1), pp. 112-121.
- Shlomi, T., et al. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations, *Proc Natl Acad Sci U S A*, 102(21), pp. 7695-7700.
- Shlomi, T., et al. (2007) Systematic condition-dependent annotation of metabolic genes, *Genome Res*, 17(11), pp. 1626-1633.
- Shlomi, T., et al. (2008) Network-based prediction of human tissue-specific metabolism, *Nat Biotechnol*, 26(9), pp. 1003-1010.
- Shlomi, T., et al. (2009) Predicting metabolic biomarkers of human inborn errors of metabolism, *Mol Syst Biol*, 5, p. 263.
- Shlomi, T., et al. (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect, *PLoS Comput Biol*, 7(3), p. e1002018.
- Simeonidis, E., et al. (2010) Why does yeast ferment? A flux balance analysis study, *Biochem Soc Trans*, 38(5), pp. 1225-1229.
- Smallbone, K., et al. (2010) Towards a genome-scale kinetic model of cellular metabolism, *BMC Syst Biol*, 4, p. 6.
- Snitkin, E.S., et al. (2008) Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions, *Genome Biol*, 9(9), p. R140.
- Sohn, S.B., et al. (2010) Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for in silico analysis of heterologous protein production, *Biotechnology journal*, 5(7), pp. 705-715.
- Sohn, S.B., et al. (2012) Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of in silico/in vivo mutant growth, *BMC Syst Biol*, 6, p. 49.
- Song, H. and Lee, S.Y. (2006) Production of succinic acid by bacterial fermentation, *Enzyme and Microbial Technology*, 39(3), pp. 352-361.
- Srivastava, A., et al. (2012) Reconstruction and visualization of carbohydrate, N-glycosylation pathways in *Pichia pastoris* CBS7435 using computational and system biology approaches, *Systems and Synthetic Biology*.
- Stein, L. (2001) Genome annotation: from sequence to biology, *Nat Rev Genet*, 2(7), pp. 493-503.
- Stephanopoulos, G., et al. (1998) *Metabolic engineering: principles and methodologies*, San Diego, Academic Press.
- Su, A.I., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*, 101(16), pp. 6062-6067.
- Sun, J. and Zeng, A.P. (2004) IdentiCS--identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence, *BMC Bioinformatics*, 5, p. 112.
- Sun, J., et al. (2007) Metabolic peculiarities of *Aspergillus niger* disclosed by comparative metabolic genomics, *Genome Biol*, 8(9), p. R182.
- Suthers, P.F., et al. (2007) Metabolic flux elucidation for large-scale models using C-13 labeled isotopes, *Metab. Eng.*, 9(5-6), pp. 387-405.
- Tanabe, A., et al. (2009) Obesity causes a shift in metabolic flow of gangliosides in adipose tissues, *Biochem Biophys Res Commun*, 379(2), pp. 547-552.
- Teh, K.Y. and Lutz, A.E. (2010) Thermodynamic analysis of fermentation and anaerobic growth of baker's yeast for ethanol production, *J Biotechnol*, 147(2), pp. 80-87.
- Theilgaard, H., et al. (2001) Quantitative analysis of *Penicillium chrysogenum* Wis54-1255 transformants overexpressing the penicillin biosynthetic genes, *Biotechnol Bioeng*, 72(4), pp. 379-388.
- Thiele, I. and Palsson, B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction, *Nat Protoc*, 5(1), pp. 93-121.
- Thiele, I., et al. (2013) A community-driven global reconstruction of human metabolism, *Nat Biotechnol*.
- Thykaer, J. and Nielsen, J. (2003) Metabolic engineering of beta-lactam production, *Metab Eng*, 5(1), pp. 56-69.
- Titov, V.N., et al. (2010) [Methylglyoxal--test for biological dysfunctions of homeostasis and endoecology, low cytosolic glucose level, and gluconeogenesis from fatty acids], *Ter Arkh*, 82(10), pp. 71-77.

- Tomita, M., et al. (1997) E-CELL: Software Environment for Whole Cell Simulation, *Genome informatics. Workshop on Genome Informatics*, 8, pp. 147-155.
- Uhlen, M., et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics, *Mol Cell Proteomics*, 4(12), pp. 1920-1932.
- Uhlen, M., et al. (2010) Towards a knowledge-based Human Protein Atlas, *Nat Biotechnol*, 28(12), pp. 1248-1250.
- Usaite, R., et al. (2006) Global transcriptional and physiological responses of *Saccharomyces cerevisiae* to ammonium, L-alanine, or L-glutamine limitation, *Appl Environ Microbiol*, 72(9), pp. 6194-6203.
- Waldrop, M.M. (1992) *Complexity: the emerging science at the edge of order and chaos*, New York, Simon & Schuster.
- van den Berg, M.A., et al. (2008) Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*, *Nat Biotechnol*, 26(10), pp. 1161-1168.
- van Hoek, M.J. and Hogeweg, P. (2009) Metabolic adaptation after whole genome duplication, *Molecular biology and evolution*, 26(11), pp. 2441-2453.
- Wang, L. and Hatzimanikatis, V. (2006) Metabolic engineering under uncertainty--II: analysis of yeast metabolism, *Metab Eng*, 8(2), pp. 142-159.
- Wang, Y., et al. (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE, *BMC Syst Biol*, 6, p. 153.
- Varma, A. and Palsson, B.O. (1994a) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110, *Appl Environ Microbiol*, 60(10), pp. 3724-3731.
- Varma, A. and Palsson, B.O. (1994b) Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use, *Bio-Technology*, 12(10), pp. 994-998.
- Verhoef, P., et al. (2004) Dietary serine and cystine attenuate the homocysteine-raising effect of dietary methionine: a randomized crossover trial in humans, *American Journal of Clinical Nutrition*, 80(3), pp. 674-679.
- Whelan, K.E. and King, R.D. (2008) Using a logical model to predict the growth of yeast, *BMC Bioinformatics*, 9, p. 97.
- Wiechert, W. (2001) ¹³C metabolic flux analysis, *Metab Eng*, 3(3), pp. 195-206.
- Willke, T. and Vorlop, K.D. (2004) Industrial bioconversion of renewable resources as an alternative to conventional chemistry, *Appl Microbiol Biotechnol*, 66(2), pp. 131-142.
- Wintermute, E.H. and Silver, P.A. (2010) Emergent cooperation in microbial metabolism, *Mol Syst Biol*, 6, p. 407.
- Wishart, D.S., et al. (2007) HMDB: the Human Metabolome Database, *Nucleic Acids Res*, 35(Database issue), pp. D521-526.
- Vitkup, D., et al. (2006) Influence of metabolic network structure and function on enzyme evolution, *Genome Biol*, 7(5), p. R39.
- Vongsangnak, W., et al. (2008) Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*, *BMC Genomics*, 9, p. 245.
- Vongsangnak, W., et al. (2010) Integrated analysis of the global transcriptional response to alpha-amylase over-production by *Aspergillus oryzae*, *Biotechnol Bioeng*.
- Wu, C., et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, *Genome Biol*, 10(11), p. R130.
- Väremo, L., et al. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods., *Nucleic Acids Research*, p. 10.1093/nar/gkt1111.
- Xu, N., et al. (2013) Reconstruction and analysis of the genome-scale metabolic network of *Candida glabrata*, *Mol Biosyst*, 9(2), pp. 205-216.
- Yang, K.M., et al. (2012) Ethanol reduces mitochondrial membrane integrity and thereby impacts carbon metabolism of *Saccharomyces cerevisiae*, *FEMS Yeast Res*, 12(6), pp. 675-684.