RWTH Aachen University
Institute of Applied Microbiology
MSc. Biology

Generation of a genome-scale metabolic model of *Ustilago maydis*, causative agent of corn smut

Master's Thesis

Author: **Christian Lieven**
Examiner: **Prof. Dr. Lars M. Blank**
Second Examiner: **Prof. Dr. Björn Usadel**
Supervisor: **Thiemo Zambanini**

Januar 2015

Selbstständigkeitserklärung

Die vorliegende Masterarbeit wurde im Institut für Angewandte Mikrobiologie der RWTH Aachen in der Zeit vom 01. April 2014 bis 12. Januar 2015 angefertigt.

Gutachter der Arbeit : Prof. Dr. Lars Blank
iAMB

Prof. Dr. Björn Usadel
IBG-2

Ich erkläre hiermit ehrenwörtlich, dass:

1. ich die vorliegende Masterarbeit selbständig unter Anleitung verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe;
2. die Übernahme wörtlicher Zitate aus der Literatur/Internet sowie die Verwendung der Gedanken anderer Autoren an den entsprechenden Stellen innerhalb der Arbeit gekennzeichnet wurde;
3. ich meine Masterarbeit bei keiner anderen Prüfung vorgelegt habe.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Aachen,den10.01.2015 Unterschrift

Hinweis:
Ohne Unterschrift, Datums- und Ortsangabe ist diese Erklärung ungültig und damit auch die Masterarbeit.

Sperrvermerk für studentische Arbeiten

Die vorliegende Masterarbeit beinhaltet interne vertrauliche Informationen der Firma B.R.A.I.N. AG. Die Weitergabe des Inhalts der Arbeit im Gesamten oder in Teilen ist grundsätzlich untersagt. Es dürfen keinerlei Kopien oder Abschriften, auch nicht in digitaler Form, angefertigt werden. Ausnahmen bedürfen der schriftlichen Genehmigung des Unternehmens.

Acknowledgements

I would like to thank Henrik Cordes, Suresh Sudarsan, Lars Küpfer and Birgitta Ebert for always giving me new perspectives on solving computational problems and for sharing their experience with metabolic modeling. Credit especially goes out to Henrik and Suresh who surely guided me through the initial steps of programming in MatLab, as well as operating the COBRA Toolbox and were always able to supply me with insightful literature as well as nifty software tools.

I am grateful for the friendly, patient and helpful correspondences I have had with several people over the course of my research: Jörg Büscher and Malte Herold at B.R.A.I.N. AG for their insight on operating the Model Borgifier and Pathway Tools' data export function, Markus Krummenacker and Peter Karp at SRI International for their excellent support of fixing bugs and tailoring Pathway Tools to my needs, Ulrich Güldner and Martin Münsterkötter from the IBIS at the Helmholtz Centre Munich for providing the most recent MUMDB annotated genome of *Ustilago maydis* exported via the sequence analysis suite Pedant-Pro by Biomax.

For giving me the opportunity to work on this interesting subject, for inciting my interest in stoichiometric modeling and for being the first examiner of this thesis, I would like to thank Prof. Dr. Lars M. Blank. I would like to thank Prof. Dr. Björn Usadel, for agreeing to become the second examiner. I would like to pay a very special tribute to Thiemo Zambanini, who has not only been a very helpful, motivating and genial supervisor, colleague and, ultimately friend, yet also a truly competent source of knowledge, providing me with just the right ideas at the right times.

To all my friends and colleagues at the IAMB: Isabella Albert, Judith Alferink, Hélène Aschmann, Arno Baues, Gisela Beissel, Carola Berger, Melanie Beudels, Monika Brehm, Eik Czarnotta, Promi Das, Dr. Birgitta Ebert, Lara Eisenbach, Jan Förster, Elena Geiser, Andrea Grego, Marco Grull, Christoph Halbfeld, Malte Heyer, Jana Kampmeier, Wiebke Kleineberg, Thomas Kirchner, Dr. Lars Küpfer, Jannis Küpper, Matthias Lehnen, Christoph Lenzen, Bern Leuchtle, Wing-Jin Li, Sarah Maurer, Christian Müller, Dr. Bastian Molitor, Salome Nies, Maike Otto, Tim Runge, Eda Sarikaya, Friedrich Sauer, Ivan Schlembach, Andreas Schmitz, Simone Schmitz, Annette Schreer, Sandra Schulte, Christiane Sonntag, Suresh Sudarsan,

Acknowledgements

Hamed Tehrani Christoph Thiel, Till Tiso, Kerstin Walter, Dr. Nick Wierckx Benedikt Wynands, Dr. Martin Zimmermann, Rabea Zauter.
Thank you for making my time in the office and lab as much fun as it has been!

Lastly I would like to thank my parents for providing ideas and opinions on my research as well as the hard cash necessary to stay alive. You rock!

Abstract

A genome-scale network reconstruction (GENRE) for *Ustilago maydis*, a plant pathogen of maize, was compiled by combining genomic, physiological and experimental information available for this fungus. The network was built using the automatic reconstruction software Pathway Tools, followed by extensive manual curation on the initial draft. The reconstruction serves as an easy-to-navigate, offline knowledgebase for biochemical processes as well as a means of accessing and maintaining genome annotations. Using the GENRE as the basis, a genome-scale stoichiometric model (GEM) was generated, which includes 1987 metabolites across 2218 reactions. 1658 of these reactions can be classified as conversion reactions, whereas 560 reactions describe transport processes between the 4 included compartments (Extracellular, Cytosol, Mitochondrial Lumen, Mitochondrial Intermembrane Space). Since the model contains exactly 1079 genes that are specifically assigned to about 86% of the reactions, it was termed iCL1079. Using the stoichiometry of all reactions and assuming a steady-state equilibrium, cellular metabolic fluxes could be determined using Flux Balance Analysis (FBA). After in-silico characterization, iCL1079 predicted a growth rate of 0.397 h^{-1} on glucose and 0.227 h^{-1} on glycerol. Moreover, product yields of malate on glucose (1.276 $g_{malate}\ g^{-1}_{glucose}$) and on glycerol (1.456 $g_{malate}\ g^{-1}_{glycerol}$) could be calculated. The presented GENRE and GEM are comprehensive, extendable, up-to-date resources and thus valuable tools for system-wide discoveries regarding the biochemistry, metabolism and genome of *U. maydis*.

Zusammenfassung

Eine Genom-skalige Netzwerk Rekonstruktion (GENRE, engl. genome-scale network reconstruction) wurde aus verfügbaren experimentellen Informationen sowie Informationen über das Genom und die Physiologie für den pflanzenpathogenen Pilz, *Ustilago maydis* zusammen gestellt. Ein Entwurf des Netzwerks wurde mit der automatischen Rekonstruktionssoftware, Pathway Tools erstellt, und daraufhin umfassend manuell kuratiert. Die Rekonstruktion eignet sich als einfach zu navigierende, offline Datenbank für biochemische Prozesse sowie als Mittel um auf Informationen zum Genom zuzugreifen und zu ergänzen. Mit der GENRE als Grundlage, wurde ein Genom-skaliges stöchiometrisches Modell entwickelt, welches aus 1987 Metaboliten verteilt über 2218 Reaktionen bestand. Transformierende Reaktionen machen mit einer Menge von 1658 den größten Teil aus, während Transportprozesse bei 560 Reaktionen liegen. Transportprozesse beschreiben den Metabolittransport zwischen den 4 vorhandenen Kompartimenten (Extracellular, Cytosol, Mitochondrial Lumen, Mitochondrial Intermembrane Space). Weil das Model exakt 1079 Gene enthält, welche spezifisch zu 86% der Reaktionen zugewiesen sind, wurde es mit iCL1079 bezeichnet. Unter Verwendung der Stöchiometrie aller Reaktionen und der Annahme, dass diese sich dauerhaft im Equilibrium befinden, können mithilfe der Fließgleichgewichtsanalyse die zellulären Flussverteilungen bestimmt werden. Nach einer Charakterisierung in-silico sagte iCL1079 eine Wachstumsrate von 0,397 h^{-1} auf Glukose und 0,227 h^{-1} auf Glycerol voraus. Darüberhinaus konnten auch Produktausbeuten von Malat auf Glukose (1,276 $g_{Malat}\ g^{-1}_{Glukose}$) und auf Glycerol (1,456 $g_{Malat}\ g^{-1}_{Glycerol}$) berechnet werden. Die hier gezeigten GENRE und GEM sind erweiterbare, aktuelle, offline Ressourcen und aufgrund dessen wertvolle Werkzeuge zur Schaffung von systemischen Erkenntnissen bezüglich der Biochemie, des Metabolismus und des Genoms von *U. maydis*.

Table of contents

Table of contents

List of Figures

List of Tables

List of Supplements

Please refer to chapter 8. Appendix for further information on the content of these files.

s

Abbreviations

AA	Amino acid
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Tool
CDS	Coding sequences
CDW	Cell dry weight
COBRA Toolbox	Constraint-Based Reconstruction and Analysis Toolbox
DDBJ	DNA Database of Japan
DNA	Desoxyribonucleic acid
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory
E-Value	Expect-value
FAD/FADH2	Flavinadenin dinucleotide
FBA	Flux Balance Analysis
FTP	File Transfer Protocol
FunCat	Functional Catalogue
g	Gram
g L-1	Gram per litre
g (100 g_{CDW})$^{-1}$	Gram per 100 gram cell dry weight
g g^{-1} polymer	Gram per gram polymer
GBK	GenBank
GEM	Genome-scale model
GENRE	Genome-scale network reconstruction
GLPK	GNU Linear Programming Kit
GO	Gene ontology
GPR	Gene-Protein-Reaction
h^{-1}	Per hour
IBIS	Institute of Bioinformatics and Systems Biology
KEGG	Kyoto Encyclopedia of Genes and Genomes
L	Litre
MIPS	Munich Information Center for Protein Sequences
ml	Mililitre
mmol g_{CDW}	Milimole per gram cell dry weight
MUMDB	MIPS Ustilago maydis DataBase
NAD/NADH	Nicotinamide adenine dinucleotide
NADP/NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
PEDANT	Protein Extraction, Description and ANalysis Tool

PGDB	Pathway/genome database
p-value	Probability-value
PPi	Pyrophosphate
RefSeq	Reference sequence
RNA	Ribonucleic acid
SBML	Systems Biologie Markup Language
SRI	Stanford Research Institute
TCA	Tricarboxylic acid

1. Introduction

1.1. *Ustilago maydis*

Ustilago maydis is a basidiomycete plant pathogen that has adapted to grow depending on the living tissue of maize, severely affecting the growth of the plant and thus its cereal yield in the agricultural industry (Kämper et al. 2006). Over the course of its dimorphic lifecycle, the dikariotic filamentous form invades the host cells by developing a specialized hypha, called appressorium, thus establishing a mutual interaction surface between the membrane of the host cell and that of the fungus (Figure 1a). The cell wall of the protoplast remains intact and forms a sheath around the invading appressorium. While the fungus proliferates and differentiates into branches of hyphae (Figure 1b), and ultimately forms spore-filled tumors (Figure 1c), the host cells remain viable until very late during the time of infection (Kämper et al. 2006). After the spores undergo meiosis during their germination, haploid sporidiae are formed in the tumors (Figure 1d). These haploid cells no longer depend on living tissue for growth. Instead they are saprotrophic (Steinberg & Perez-Martin 2008). Two haploid cells fuse to form a dikaryotic filament via the formation of a conjugation tube, which triggers another lifecycle (Figure 1e).

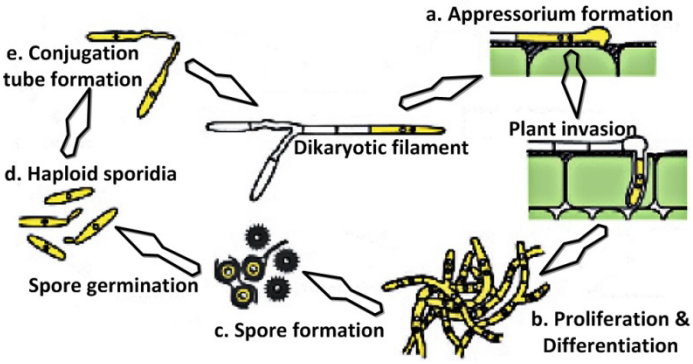


Figure 1: Growth cycle of *Ustilago maydis*

The major developmental stages of *U. maydis* are shown, commencing with the infection of a host cell. By forming an appressorium (a), the fungus pierces the cell wall and establishes a surface for interaction ensheathed by the cell membrane. Proliferation and differentiation (b) lead to the formation of spores (c) in visible tumors. Haploid sporidiae (d) are produced after the germination of the spores. The fusion of two haploid cells triggers sexual development. Figure adapted from (Kämper et al. 2006)

Due to its stealthy biotrophy and its strong impact on one of the world's most important crops used as food and animal feed, the mechanism of pathogenicity of *U. maydis* had been the primary target of research in the past (Bölker et al. 1995; Weber et al. 2003; Brefort et al. 2009). Despite the fact that the molecular mechanisms of infection and the circumvention of the host's immune response facilitated by *U. maydis* still need to be fully deciphered (Wahl et al. 2010; Hemetsberger et al. 2012), the organism has become a well-established model for

microbe-plant-interaction (Brefort et al. 2009). In addition to that, *U. maydis* has been proposed as a model organism for cell biology (Steinberg & Perez-Martin 2008). Evidence suggests that the molecular processes involved in long-distance transport (Schuchardt et al. 2005), DNA repair (Kojic et al. 2002) and mitosis (García-Muse et al. 2004) are conserved between the fungus and humans (Steinberg & Perez-Martin 2008). Furthermore, *U. maydis* shares more protein sequence similarity with humans than *S. cerevisiae*, which became the most prominent fungal model system in the past decades (Münsterkötter & Steinberg 2007).

In its haploid form, *U. maydis* is of special interest as a potential fungal producer organism. This is due to the fact that *U.* maydis not only exhibits a number of traits that are desirable in an industrial fermentation setting, but also natively produces a wide range of interesting compounds. Growing as single cells without the development of filaments; the ability to withstand hydromechanical stress present in stirred tank bioreactors; and the ability to utilize plant sugars like xylose, as well as being tolerant towards impurities in crude feedstock, qualify *U. maydis* for growth in up-scale, industrial fermenters as a facilitator of value-added bioconversions (Klement et al. 2012). Native products include glycolipids such as ustilagic acid and mannosylerythrytol lipids; iron-chelating siderophores, such as ferrichrome and dicarbonic acids, such as itaconate, and malate (Bölker et al. 2008; Geiser et al. 2014). Glycolipids are non-toxic biological surfactants, which can be employed in an array of pharmaceutical, cosmetic and food applications (Kitamoto et al. 2002), whereas ferrichrome has potential uses as a chelating agent for extracting heavy metals from contaminated soil (Renshaw et al. 2002). Itaconate and malate, however, represent biochemical building blocks, that can be used as the basic monomers in the production of different types of high value chemicals including fuels, plastics, resins, synthetic fibers and paints (Voll et al. 2012).

In order to rationally enhance the production capabilities of *Ustilago maydis* especially when grown on plant biomass as substrate, a thorough understanding of the genome, the physiology and the biochemistry of the fungus is a powerful asset. Knowledge about this can be used in targeted approaches to improve the yields of certain products or help identify novel functions. In 2006, progress has been made in that direction, as an updated genome sequence of the haploid *Ustilago maydis strain 521* was published by Kämper et al. (2006). Since then, the relatively small genome of about 20.5 million bases and 6,902 predicted protein-encoding genes provided an important source of knowledge for more in-depth research about *U. maydis*. Among discoveries on clusters of genes that affected virulence (Kämper et al. 2006),

the primary metabolic pathway namely the glycolysis (Saavedra et al. 2008), and the amino acid, sulfur and nitrogen metabolism (McCann & Snetselaar 2008) of the organism could be described in greater detail. However, research remained more focused on the microbe-plant-interaction and cell biology aspects of *U. maydis*, which means in turn that fundamental physiological and metabolic connections remain unknown.

1.2. Online Databases

For researchers that are interested to investigate unknown metabolic networks and functions of *U. maydis* there is the opportunity of consulting online databases, which depending on the type of database, not only store information on the genome but also possible functions that can be inferred from it. The following section aims at giving a short introduction to most of the well-established databases which can be a valuable source of information to researchers investigating *U. maydis.*

Before being able to publish a paper describing a sequenced genome, the sequence itself has to be uploaded to one of the three principal international sequence databases: The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database at the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK (Kanz et al. 2005), the GenBank (GBK) at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA (Benson et al. 2013), and the DNA Database of Japan (DDBJ) at the National Institute of Genetics in Mishima, Japan (Tateno et al. 2002). These three databases are synchronized daily to ensure that the sequence information stays the same between them. Those researchers who investigate *U. maydis* can access the genome and the wealth of information that can be gathered from it, through the online graphical interface of these databases or download the corresponding sequence in the FASTA-Format. The NCBI, for example, offers raw complete genome sequences for download from their FTP site (ftp://ncbi.nlm.nih.gov/Entrez/Genomes, GenBank), while updated genome sequence entries, which have been checked by NCBI members of staff, are available from their Reference Sequence project website (RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq/) (Pruitt et al. 2012).

Another important primary source of genetic information about *U. maydis* used to be the Munich Information Center for Protein Sequences (MIPS) *Ustilago maydis* DataBase (MUMDB), which has been decommissioned in June 2014 (http://mips.helmholtz-

muenchen.de/genre/proj/ustilago) (Mewes et al. 1999). At the time of this thesis, the entire dataset was accessible either via the MUMDB or through its successive replacement: the Protein Extraction, Description and ANalysis Tool (PEDANT) interface (http://pedant.gsf.de/), which is now maintained by the Institute of Bioinformatics and Systems Biology (IBIS) replacing the MIPS at the Helmholtz Centre Munich. Initially based on a draft genome sequence published to GBK by the Eli and Edythe L. Broad Institute of MIT and Harvard (Cambridge, Massachusetts, USA) in 2004 but then updated with the improved sequence by Kämper et al. in 2006, the MUMDB presented continually updated, manually processed information on the statistics and location of genes and the putative function, localization and structure of their protein products in *U. maydis* for comparative genome analysis. While permanently accessible online, these pieces of information can be downloaded as annotations to the genome sequence. Moreover, it is important to note that updates from ongoing genome analysis at the IBIS are pushed to NCBI's RefSeq in irregular intervals.

Unlike the MUMDB, the Kyoto Encyclopedia of Genes and Genomes (KEGG) as part of the GenomeNet website run by the Kyoto University Institute for Chemical Research focuses solely on the connection of genomic sequence information to higher metabolism and regulatory functions, thus providing a complete visual catalogue of general metabolic pathway maps (Tanabe & Kanehisa 2012). Sub-selections of these general maps can be accessed for several organisms including *U. maydis*. In these sub-selections, highlighted reactions are present in the organism of interest, whereas reactions that are greyed out have not been found. The data for specific organisms in KEGG is obtained from NCBI's RefSeq and updated on a regular basis (Koonin & Galperin 2003).

Another database with a similar focus is MetaCyc (http://metacyc.com), which is maintained by the Stanford Research Institute (SRI) in Menlo Park, California, USA. MetaCyc represents a combined knowledgebase of more than 2255 non-redundant metabolic pathways from the primary and secondary metabolism of over 2579 distinct microbial organisms (Caspi et al. 2014). The underlying data is obtained primarily from published literature, but MetaCyc also draws on several other databases to get additional information on a reaction or a compound (http://metacyc.org/MetaCycUserGuide.shtml).

In summary, online databases are easily accessible sources that at first glance can provide a detailed overview over the genome or metabolism of an organism of interest. Their strength and weakness lies in the vast amount of information they contain. For each item within a database there are often several entries showing the results of mathematical or bioinformatics methods, whose acronyms may appear cryptic to an uninvolved researcher, and whose informational value as well as reliability, may not be easily discernible. Since the data on one particular organism is frequently derived from cross-comparison to experimental results or the genome of another organism and the fact that databases generally try to present all data in one place, following links can quickly lead off track away from the target organism. Although some databases frequently synchronize or update their contents, it is not immediately apparent how data flows through them and which databases draw from where. Hence, for the purpose of obtaining reliable results quickly, it is worthwhile to consider unifying the available data into a genome-scale network reconstruction (GENRE).

The Pathway Tools software , which is fundamental to the MetaCyc database by providing its data retrieval and visualization algorithms, is available as a download for free (Karp et al. 2010). It offers a wide range of features that aid not only the analysis but also the construction of genome-scale biological networks. On the basis of an annotated genome, for example from MUMDB or RefSeq, and with access to MetaCyc, Pathway Tools can infer the reactome of an entire organism, all metabolic pathways and even propose genes, whose products catalyze previously unidentified reactions (Karp et al. 2011). In order to elucidate the physiology and metabolism of *U. maydis* in greater detail than online databases, this tool can be used to produce such a GENRE (Hamilton & Reed 2014).

1.3. GENREs and GEMs

A genome-scale network reconstruction (GENRE) represents an organized collection of information about the metabolism of a specific organism based on the annotated genome and experimental results from primary literature (Hamilton & Reed 2014). It is made up of a network of metabolic chemical conversions in the form of enzymatic reactions and their underlying genes. The bare network of reactions is extended by including further information about genes, proteins and compounds in the form of database identifiers and notes on experimental observations (Thiele & Palsson 2010). Recently, the number of published

GENREs has increased drastically as system-level approaches to analyzing microbial metabolism have become more powerful (Patil et al. 2004; Liu et al. 2010) and thus more popular. Besides being an all-encompassing knowledgebase, specifically for the organism of interest, a GENRE makes up the basis of a genome-scale model (GEM), which is simply a mathematical representation of this network of reactions represented in a GENRE (Price et al. 2004). Hence a GEM allows for in-silico manipulation of the phenotype and simulation of metabolic behavior in certain conditions (Price et al. 2004).

The creation of a GEM is an iterative process that can be divided into four distinct stages with more than 90 substeps in total (Thiele & Palsson 2010). In stage 1, the draft reconstruction is carried out using an annotated genome, primary literature and online databases such as KEGG or MetaCyc. With the aid of a reconstruction software such as Pathway Tools, the annotations to the genome can be used to identify metabolic genes and link them to their corresponding functions by parsing the mentioned databases. This process is automatic and only requires limited input from a user (Karp et al. 2010).

In the next stage, the draft network is curated manually according to general physical principles and experimental physicochemical knowledge about the organism of interest. In this step, network gaps are filled, and all reactions are scrutinized for their localization within the cell, the charge and mass balance of their substrates and their gene-protein-reaction (GPR) relationships. These GPR relationships are the underlying factors in generating Boolean statements in the model that determine how the genes of an organism relate to the corresponding gene products. For an enzyme that is composed of multiple subunits which are equally important to the functioning of said enzyme, the genes encoding for these subunits relate to each other with a Boolean "AND", meaning that if either of them were to be knocked-out the enzyme won't function. Several genes relate to each other with a Boolean "OR", when they individually code for isoenzymes, which function even if the other genes, respectively, are not expressed.

Stage 3 is an automatic process where the metabolic network is converted into a GEM, which involves compiling each substrate in each reaction of a GENRE into one large stoichiometric matrix. This can be facilitated using the Constraint-Based Reconstruction and Analysis (COBRA) Toolbox, which is a free of charge collection of scripts in the MATLAB™ programming language (Schellenberger et al. 2011). The mathematical concept of this

conversion is further elucidated in Chapter 3.1 of this thesis. In stage 4, on the basis of the GEM created in stage 3, experimental results and the results of simulations are compared to evaluate the model. This can also be achieved by employing the algorithms and functions delivered with the COBRA Toolbox. If they agree well, the model can be considered for publication, however it is more likely that further curation is necessary, which is why stage 2, 3 and 4 are usually repeated until a satisfactory level of agreement is reached.

The greatest challenge in the construction of a GEM are erroneous or lacking annotations of metabolic genes, which can persist until late in the reconstruction process and skewer the results of modeling, for example the false introduction of a non-existent reaction could lead to an in-silico knockout not resulting in lethality after simulation. Another issue that has similar outcomes are outdated, generic, duplicate or faulty entries in the online databases, which the software refers to during stage 1 of building a GEM. Furthermore, the existing software tools of today differ strongly in the level of support during certain substeps, which can lead to stage 2 becoming even more time intensive.

A carefully curated GEM, however, is an excellent basis for a number of tasks. It can be used to build hypotheses about the natural physiology of an organism (Schellenberger et al. 2011), predict the outcome of genetic modifications (Latendresse et al. 2012), analyze synthetic lethality and allow the depiction of intracellular flux distributions in order to identify metabolic bottlenecks (Karp et al. 2010). It can be extended to include information about the regulation of genes by providing the scaffold for the display of high-throughput transcriptional data and thus help researchers to identify possible correlations (Kim & Reed 2014).

1.4. Aim of Thesis

For the Master's thesis, a genome-scale metabolic model of *U. maydis* is to be constructed based on an up-to-date annotated genome. Using Pathway Tools as the software of choice, the draft reconstruction is to be carried out automatically, followed by an iterative cycle of curation and amendment until the generated model can predict in-silico growth. In addition to that, high-throughput growth-experiments are to be carried out to characterize *U. maydis* for

the growth on various carbon sources., which can be compared to results from simulations with the model.

2. Materials

2.1. Software

- COBRA Toolbox 2.0.5(Schellenberger et al. 2011)available from http://opencobra.sourceforge.net/openCOBRA/Welcome.html

- Excel 2007 (Microsoft) commercially available from http://microsoftstore.com

- MATLAB R2014a (Mathworks) commercially available from http://mathworks.co.uk

- Pathway Tools 18.0 (Karp et al. 2010) available from http://bioinformatics.ai.sri.com/ptools/

- SBML Toolbox 4.1.0 (Keating et al. 2006) available fromhttp://sbml.org/Software/SBMLToolbox

- Sublime Text 2 available at http://www.sublimetext.com/

2.2. Databases

- MetaCyc (Caspi et al. 2014) at http://metacyc.com (last date of access: 05.12.2014)

- MUMDB(Mewes et al. 1999)via the PEDANT Interface (Frishman et al. 2001) at http://pedant.helmholtz-muenchen.de/pedant3htmlview/pedant3view?Method=start_method&Db=p3_t237631_Ust_maydi_v2 (last date of access: 15.12.2014)

- KEGG (Kanehisa et al. 2014; Ogata et al. 1999) at http://kegg.jp/ (last date of access: 23.10.2014)

2.3. Used Strain

For the growth experiments *Ustilago maydis* FB1 mating type a1b1 was used.

3. Methods

3.1. Draft reconstruction

The genomic DNA sequence of *Ustilago maydis (Strain 521 FGSC 9021)* (Kämper et al. 2006) was obtained from NCBI's RefSeq project (Tatusova et al. 2014) in the FASTA format. A corresponding annotation file (Mannhaupt et al. 2013) was then exported from the MIPS Ustilago Maydis Database via the PEDANT Interface (Frishman et al. 2001). The exported file in the GBK format contained Gene Ontology (GO) terms (Ashburner et al. 2000), Functional Catalogue (FunCat) (Ruepp et al. 2004) and Enzyme Commission (EC) Number (Barrett 1995) annotations for the sequenced genome.

Pathway Tools 18.0 was installed and launched. Using the inherent PathoLogic Tool (Karp et al. 2011), the sequence and annotation files were parsed and a new Pathway/Genome Database (PGDB) was created. The PathoLogic algorithm starts by creating gene entries in the PGDB based on the input files. Entries for genes, their base-pair position on a chromosome in the genome and the type of product they code for, are inferred that way. Depending on a gene's product, either rRNA, tRNA or polypeptide, the corresponding product entry is created. Then, PathoLogic checks the annotation file for either EC numbers or GO terms, in order to connect those polypeptides that are enzymes to the corresponding reactions which they catalyze. The corresponding reactions are taken from the MetaCyc database which is included in Pathway Tools. In case EC numbers or GO terms are not available, PathoLogic can compare the annotated names with a range of keywords in order to infer the catalyzed reaction. Finally all of the inferred reactions are matched against the pathways stored in the MetaCyc database.

The pathways that have a single reaction matching with the ones from the PGDB are imported entirely, including all associated compounds and reactions that could not be inferred on the basis of the annotation file. Subsequently, those pathways are subjected to an automatic pruning step, which removes some of them again based on two parameters: A pathway's evidence score and its taxonomic range. The former is calculated based on the fraction of reactions in MetaCyc that can be catalyzed by enzymes from the PGDB and on how many of these reactions are unique to that pathway only. The latter is based on a tagging system within MetaCyc: By default each pathway is tagged for the taxa it was observed in experimentally.

Hence pathways from other taxa than the PGDB are pruned, unless there are enzymes matching to all of the reactions.

The polypeptides which could not be matched to reactions or those that resulted in having several ambiguously matching reactions in MetaCyc are listed for further manual curation. Within PathoLogic, the command "Assign Probable Enzymes" was then carried out to access this list in order to manually assign appropriate reactions or to discard polypeptides entirely, which cannot be assigned, due to lack of information. Each polypeptide of this list was queried on PEDANT, MUMDB, MetaCyc or KEGG until they were all assigned unambiguously. Inconclusive polypeptides as well as those that are involved in signaling and other non-metabolic pathways were discarded.

Up to this point, Pathway Tools has used the annotated genome to find matching metabolic reactions in the MetaCyc database. Additionally, reactions for polypeptides that could not be matched automatically were determined via manual curation. On the basis of automatic and manual matches Pathway Tools has imported entire pathways from the MetaCyc database into the PGDB and pruned them according to an internal evidence score and taxonomic information on the reconstructed organism. Hence, according to PathoLogic's algorithm and the quality of the underlying annotation file, the organism of interest usually may lack enzymes that are necessary to complete a pathway. For that reason the "Pathway Hole Filler" command within PathoLogic was executed.

Since PathoLogic has already inferred the entire pathway at this point, the missing reaction can be looked up. A set of amino acid sequences of isozymes that encode the missing reaction in other genomes is automatically retrieved from the most recent version of Swiss-Prot and Protein Information Ressource (PIR). This set is then used to carry out a pBLAST on the proteome of *U. maydis*, in order to find a potential polypeptide candidate to fill the pathway hole. PathoLogic then uses a Bayes classifier as a means of scoring potential candidates. The focus here is put on predicting a polypeptide that likely has the same function, rather than merely focusing on sequence similarities. Hence, several aspects are considered by the classifier: E-Values, alignment lenghts, a candidate's rank in the pBLAST results as well as whether a candidate's gene is next to a gene that catalyzes a reaction immediately up- or downstream in the same pathway. Furthermore the classifier also takes into account whether a candidate's gene is in the same operon as another gene from the same pathway.

Before the Bayes classifier can make reliable predictions it has to learn how likely it is for two genes that catalyze sequential reactions in a pathway to also be on the same operon, or even be sequential neighbors. Moreover it learns how high the average BLAST ranks are for those genes that have been assigned to reactions in the PGDB. This training of the Bayes classifier was carried out with all the reactions that PathoLogic was able to match unambiguously.

After the actual prediction has run, it is left up to the user to pick or reject potential polypeptide candidates from a list, based on the shown Bayes classifier as well as the raw BLAST parameters. Candidates with a P-Value ranging from 0.90 to 0.99, an average rank of 1 and the highest E-value were accepted without further investigation, whereas candidates that deviated from these values were manually checked for further evidence in MUMDB.

3.2. Flux Balance Analysis

Quantification of metabolic fluxes via Flux Balance Analysis (FBA) (Orth & Palsson 2011) was used as a means to curate the model manually and check for consistency. To do this, the PGDB from Pathway Tools was exported in the SBML format version 1 level 2 (Hucka et al. 2007), and subsequently imported into MatLab using the COBRA Toolbox (Schellenberger et al. 2011). Pathway Tools summarizes enzymatic reactions that act non-specifically on functional groups or multiple substrates into generic reactions that react with generic substrates. With the following command it is possible to export the specific instances of these reactions as well as the generic ones. The command was run from the Pathway Tools Lisp console:

(1) > *(select-organism :org-id 'umay)(2cobra \"FILEPATH" (get-model-instantiations (all-rxns :all)))* <

As part of the import process into MatLab, all reactions are converted into a stoichiometric matrix $\boldsymbol{S}$ of the dimensions $m \times n$. The m rows represent the number of metabolites and the n columns the number of reactions. During the exponential growth phase or during constant phases like dormancy, it can be assumed that for the entire network of reactions the metabolic fluxes are in a steady-state. This means that the sum of influx equals the sum of

efflux across all reactions. Hence, since there is no change in metabolite concentration over time, the model can be expressed as:

(2) $\boldsymbol{S} * \boldsymbol{v} = 0$

The vector $\boldsymbol{v}$ lists the fluxes $v_1 - v_n$ which are inherent variables to the metabolic model. A distinct set of solutions in the vector is called flux distribution. The entire system is solved by minimizing/maximizing a specific objective function obj with a linear programming solver. The linear programming problem can be represented as:

(3) $min/\max Z = obj * \boldsymbol{v}^T$

(4) Subject to $\boldsymbol{S} * \boldsymbol{v} = 0$

(5) $lb_i \leq v_i \leq ub_i$

(6) $lb_i, ub_i = |25|$

The scalar value$Z$ is obtained by multiplying the transposed flux distribution vector$\boldsymbol{v}^T$with the objective function obj. Individual fluxes v_i are constrained to lower bounds lb_i and upper bounds ub_i. Depending on the direction and reversibility of certain reactions either lb_i or ub_i are set to zero. Fluxes for all other reactions, including transport reactions (efflux and influx) are limited to 25.The objective functionobj represents which cellular action the calculation is to be focused on. Common focus points are the maximization of the biomass production, minimization of substrate uptake or maximization of product secretion. In this case, the chosen optimization parameter was to maximize the flux of the biomass reaction. All calculations were carried out employing the standard GNU Linear Programming Kit (GLPK) linear solver, utilized by the COBRA Toolbox in MatLab.

By executing a custom MatLab script (Supplement 3) prior to each FBA the modeling conditions were determined, by adding missing reactions that are important to the simulations. These reactions included reactions that allow the uptake of certain vitamins and sink reactions that qualitatively correspond to the medium components of a modified Tabuchi minimal

medium (Table 1) (Guevarra & Tabuchi 2014), which *U. maydis* is known to be able to grow on (Geiser 2014). Upper and lower bounds of the sink reactions are not limited, which means that these components can be taken up or secreted freely during the calculations. In addition, the script adds demand reactions which allow a free secretion, but no uptake of various carbonic acids, mannose and all amino acids.

Table 1: Composition of a modified Tabuchi medium (Guevarra & Tabuchi 2014)
The components of the modified Tabuchi medium can only be incorporated qualitatively into the model. Due to the steady state and mass-balancing requirement compound pools cannot be modeled. Hence, the exact amount of the components is neglected in the script (Supplement 3) for the simulations.

Mineral medium	
Ammonium chloride	5 g
Potassium dihydrogenphosphate	3 g
Magnesium sulphate 7 H2O	0.5 g
Trace element solution (see below)	1 ml
vitamin solution (see below)	1 ml
Demineralised water	1000 ml
Trace element solution (1 L)	
EDTA (Titriplex III®)	15.00 g
Zinc sulfate heptahydrate	4.50 g
Manganese chloride dihydrate	0.84 g
Cobalt(II)chloride hexahydrate (**toxic**)	0.30 g
Copper(II)sulfate pentahydrate	0.30 g
Di-sodium molybdenum dihydrate	0.40 g
Calcium chloride dihydrate	4.50 g
Iron sulfate heptahydrate	3.00 g
Boric acid	1.00 g
Potassium iodide	0.10 g
Vitamin Solution (1 L)	
Biotin (D-)	0.05 g
Calcium D(+) panthotenate	1.00 g
Nicotinic acid	1.00 g
Myo-inositol (for microbiology)	25.0 g
Thiamine hydrochloride	1.00 g
Pyridoxol hydrochloride	1.00 g
Para-aminobenzoic acid	0.20 g

3.3. Growth Experiments

In order to determine the scope of potential substrates for growth, *Ustilago maydis* FB1 mating type a1b1 was cultivated for 162 hours, at 30°C, 300 rpm, shaking diameter 50 mm and at 80% Humidity in PM1 MicroPlate™ and PM2A MicroPlate™ (Biolog Inc, USA) on 190 different potential carbon sources (Table 2: 96-well plate tables for carbon sources used in high-throughput growth experiments (Biolog Inc., USA).Table 2). The wells were inoculated according to the manufacturer's standard protocol. A preculture of *Ustilago maydis*

was grown in YEP medium (10 g L^{-1} Yeast extract, 10 g L^{-1} Bacto peptone, 5 g L^{-1} NaCl) over night.

After all data had been obtained by using the Synergy Mx (BioTek) well plate reader to measure the optical density at 600 nm and exporting the raw data using the included Gen5 analysis software, the data was automatically processed and evaluated with a custom MatLab script (Supplement 4). This script used the excel files exported from the reader to compile result vectors containing the values of OD_{600} at the measured time points for each well and subsequently normalized the individual values with those of the negative sample. The normalized data was smoothed using an exponentially weighted moving filter with an alpha value of 0.5. The signal smoothing was carried out to eliminate noisy peaks in the data. An alpha value of 1 corresponds to less smoothing, while an alpha value of 0 corresponds to strong smoothing (http://de.mathworks.com/help/signal/examples/signal-smoothing.html). Then, the natural logarithm $ln(x)$ of the result vectors was calculated and a threshold was introduced that only accepted those carbon sources for further calculations which reached a maximum OD_{600} of above 0.368.

A line was fitted to 3 consecutive values in the result vector paired with the corresponding measurement times, and thus the slope was calculated for all measured results. 2 consecutive slopes had to agree with each other to at least 10% and one of them had to be the largest measured slope of all slopes. Only then the mean of both slopes was considered the growth rate of *U. maydis* on that particular carbon source. In short, this script was written to automatically approximate the slope of the exponential growth phase which can be described as the steepest, linear section of a line, if the data is converted logarithmically. Lastly the script plotted all growth curves individually(Supplement 9, Supplement 10) and arranged them according to the grid positions of the 96-well plates from Biolog (Table 2).

Table 2: 96-well plate tables for carbon sources used in high-throughput growth experiments (Biolog Inc., USA).

PM1	1	2	3	4	5	6	7	8	9	10	11	12
A	Negative control	L-arabinose	N-acetyl-D-glucosamine	D-saccharic acid	Succinic acid	D-galactose	L-aspartic acid	L-proline	D-alanine	D-trehalose	D-mannose	Dulcitol
B	D-serine	D-sorbitol	Glycerol	L-fucose	D-glucuronic acid	D-gluconic acid	D,L-α-glycerol-phosphate	D-xylose	L-lactic acid	Formic acid	D-mannitol	L-glutamic acid
C	D-glucose-6-phosphate	D-galactonic acid-gamma-lactone	D,L-malic acid	D-ribose	Tween 20	L-rhamnose	D-fructose	Acetic acid	α-D-glucose	Maltose	D-melibiose	Thymidine
D	L-asparagine	D-aspartic acid	D-glucosaminic acid	1,2-Propanediol	Tween 40	α-Keto-glutaric acid	α-Keto-butyric acid	α-methyl-D-galactoside	α-D-lactose	Lactulose	Sucrose	Uridine
E	L-glutamine	M-tartaric acid	D-glucose-1-phosphate	D-fructose-6-phosphate	Tween 80	α-Hydroxy glutaric acid-gamma-lactone	α-Hydroxy butyric acid	β-methyl-D-glucoside	Adonitol	Matlotriose	2-Deoxy adenosine	Adenosine
F	Glycyl-L-aspartic acid	Citric acid	M-inositol	D-threonine	Fumaric acid	Bromo succinic acid	Propionic acid	Mucic acid	Glycolic acid	Glyoxylic acid	D-cellobiose	Inosine
G	Glycyl-L-glutamic acid	Tricarballylic acid	L-serine	L-threonine	L-alanine	L-alanyl glycine	Acetoacetic acid	N-acetyl-β-D-monosamine	Mono methyl succinate	Methyl pyruvate	D-malic acid	L-malic acid
H	Glycyl-L-proline	P-hydroxy phenyl acetic acid	M-hydroxy phenyl acetic acid	Tyramine	D-psicose	L-lyxose	Glucuronamide	Pyruvic acid	L-galactonic acid-γ-lactone	D-galacturonic acid	Phenylethylamine	2-Aminoethanol
PM2A	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A	Negative control	Chodroitin sulfate c	α-Cyclodextrin	β-Cyclodextrin	γ-Cyclodextrin	Dextrin	Gelatin	Glycogen	Inulin	Laminarin	Mannan	Pectin
B	N-acetyl-D-galactosamine	N-acetyl-neuraminic acid	β-D-allose	Amygdalin	D-arabinose	D-arabitol	L-arabitol	Arbutin	2-deoxy_dribose	L-erythritol	D-fucose	3-0-β-D-galactopydanosyl-D-arabionose
C	Gentobiose	L-glucose	Lactitol	D-melezitose	Maltitol	α-Methyl-D-glucoside	β-Methyl-D-galactoside	3-Methyl glucose	β-Metyl-D-glucuronic acid	α-Methyl-D-mannoside	β-Metyl-D-xyloside	Palatinose
D	D-raffinose	Salicin	Sedoheptulosan	L-sorbose	Stachyose	D-tagatose	Turanose	Xylitol	N-acetyl-D-glucosaminitol	γ-Amino butyric acid	δ-Keto valeric acid	Butyric acid
E	Capric acid	Caproic acid	Citraconic acid	Citramalic acid	D-glucosamine	2-Hydroxy benzoic acid	4-Hydroxy benzoic acid	β-Hydroxy butyric acid	γ-Hydroxy butyric acid	α-Keto valeric acid	Itaconic acid	5-Keto-D-gluconic acid
F	D-lactic acid methyl ester	Malonic acid	Melibionic acid	Oxalic acid	Oxalomalic acid	Quinic acid	D-ribono-1,4-lactone	Sebacic acid	Sorbic acid	Succinamic acid	D-tartaric acid	L-tartaric acid
G	Acetamide	L-alaninamide	N-acetyl-L-glutamic acid	L-arginine	Glycine	L-histidine	L-homoserine	Hydroxy-L-proline	L-isoleucine	L-leucine	L-lysine	L-methionine
H	L-ornithine	L-phenylalanine	L-pyroglutamic acid	L-valine	D,L-carnitine	Sec-butylamine	D,L-octopamine	Putrescine	Dihydroxy Acetone	2,3-Butanediol	2,3-Butanone	3-Hydroxy 2-butanone

4. Results

4.1. Construction of iCL1079

4.1.1. Overview and statistics of the construction process

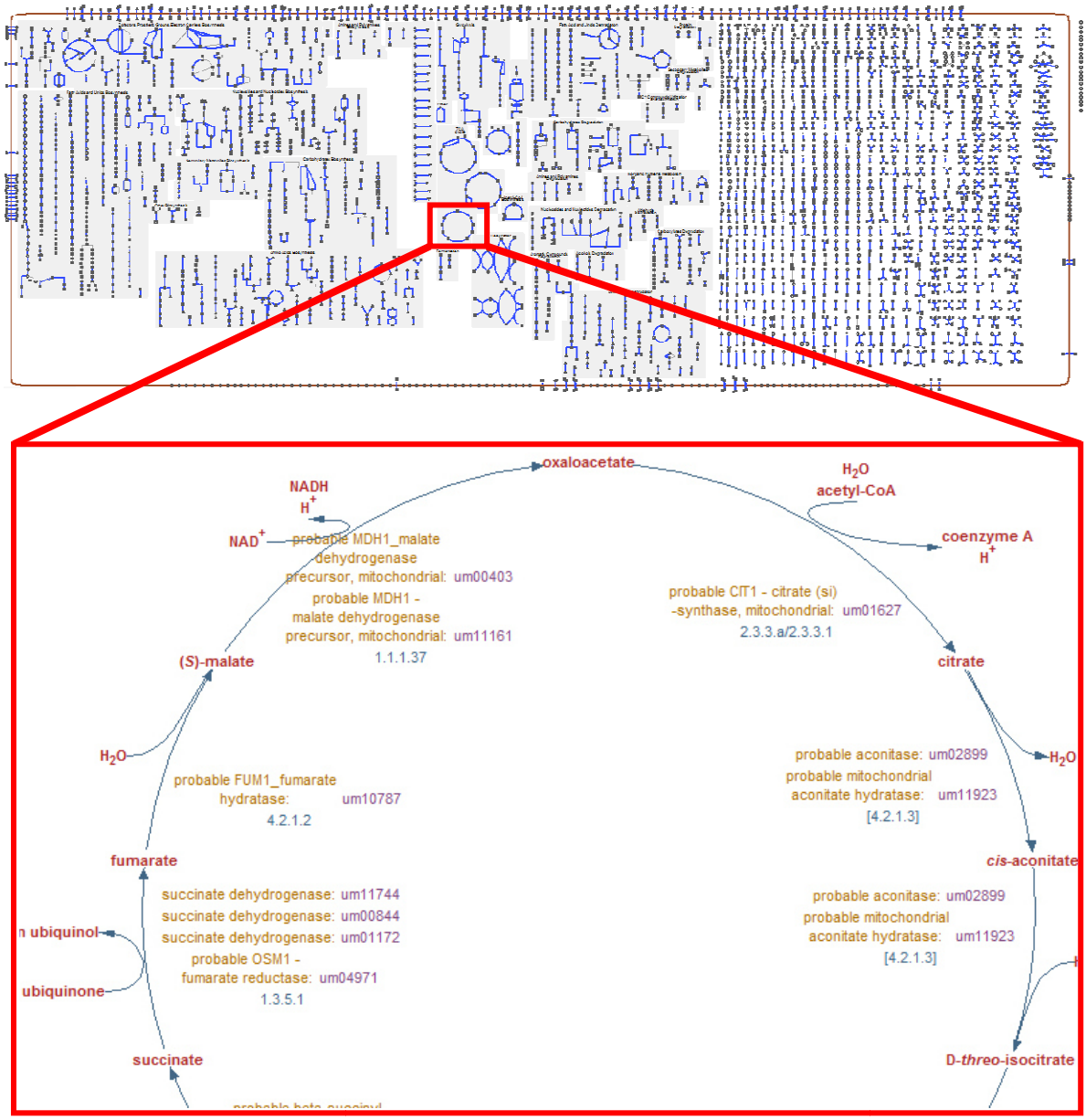
The *Ustilago maydis* metabolic model was constructed automatically on the basis of the annotated genome (Mannhaupt et al. 2013) using Pathway Tools (Karp et al. 2010) and subsequently curated manually by gathering information from literature, including both biochemistry textbooks as well as scientific articles, and from online databases.

The genome-scale metabolic model aims to provide an all-encompassing insight into the metabolic capacity of *U. maydis*. By providing an interconnection between genes, gene products and mass-balanced chemical reactions, the model describes the basic aspects of the cell's metabolism. The very basis of the metabolic reconstruction was an annotated genome. The most recent export of an annotated genome provided by the Helmholtz Centre Munich contained the 19,666,356 bp genome sequence subdivided into 6784 coding sequences (CDS) including corresponding functional annotations (EC Numbers, GO Terms & FunCat). By connecting *U. maydis* specific identifiers (e.g. um02332) to certain loci on the genome and providing the corresponding functions of the gene products, the genome annotation supplied information about the GPR associations.

Using Pathway Tools' semiautomatic draft reconstruction and pathway hole filling algorithms iteratively, along with manual curation of putative proteins, 1710 unique enzymatic reactions and 106 unique transport reactions could be inferred from the annotated genome. Then, on the basis of these enzymatic reactions 277 pathways were constructed automatically from MetaCyc (Figure 2). The pathways depicted do not only include the central carbon metabolism composed of glycolysis, pentose-phosphate pathway and TCA cycle, but also biosynthesis and degradation pathways of amino acids, fatty acids, lipids, nucleotides, cofactors, prosthetic groups and electron carriers among others, as well as the oxidative phosphorylation in the electron transport chain.

Figure 2: Visual depiction of all pathways and reactions in Pathway Tools

Blue lines correspond to reactions that could be linked to genes in *U.* maydis. Faint grey lines denote reactions without a corresponding gene. Grey dots correspond to metabolites. Pathways are shown as separate units without interconnected metabolites. Inferred Pathways include among others: The Glycolysis, the pentose phosphate pathway, the TCA cycle, amino acid biosynthetis and degradation,, fatty acid and lipid biosynthesis and degradation, Inorganic nutrient metabolism, and oxidative phosphorylation. The zoomed section shows the TCA cycle in greater detail.



In order to refine the model manually, it was exported from Pathway Tools, including all metabolic transport and conversion reactions and the corresponding instances thereof. Manual changes involved the implementation of a reaction that represented all the components of biomass in *U. maydis* (BIOMASS-Formation) and fixing those pathway holes that lead to the production of these components. Furthermore, in order to represent the respiratory chain as detailed as possible, two mitochondrial compartments were added (Mitochondrial Lumen &

Mitochondrial Intermembrane Space). The reactions and metabolites of the pyruvate dehydrogenase complex, the TCA cycle, the electron transport chain and all necessary transport systems were moved or added by hand. In addition to that, a putative itaconate synthesis pathway was adapted from literature to reflect the behavior of *U. maydis*, producing this organic acid (Geiser 2014). Lastly, artificial transhydrogenase cycles, generic duplicate reactions and bypasses of the respiratory reactions were fixed, by either removing or changing the reversibility of the corresponding reactions.

After all manual changes had been completed the model measured 2218 reactions and 1987 metabolites, in total, with 560 of the reactions being transport reactions. The model contained 1079 genes ,which is why it was coined iCL1079. The included genes were involved in GPR with about 86% (1906) of the total reactions, while only about 14% (314) of all reactions lacked a corresponding gene entirely. Furthermore, each reaction was assigned a confidence score based on a simplified version of Pathway Tool's Evidence Ontology (Karp et al. 2004). Table 3 explains the scoring system and shows the distribution of confidence scores across the total amount of reactions.

Table 3: Confidence scoring system and distribution of reactions.
All values displayed have the unit [absolute amount (relative amount)].

Type of Confidence	Score Value	Meaning	Reactions
Experimental	3	Manually curated reaction that has underlying experimental evidence.	15(0.6)
PathoLogic	2	Automatically inferred by PathoLogic on the basis of the annotated genome. Has a GPR association.	1875(85)
Manual	1	Manually added reaction in order to fill a metabolic gap for modeling. Candenote artificial, biomass or lumped reactions. If a GPR association is present it is manually curated.	42(1.9)
Not evaluated	0	Automatically inferred by PathoLogic as part of an entire pathway, without a GPR association.	286 (13)
Total			**2218(100)**

Reactions and metabolites are distributed across four distinct compartments: Extracellular Space, Cytoplasm, Mitochondrial Lumen, and Mitochondrial Intermembrane Space (Table 4). Unless specific information on their localization within intracellular compartments was present in MUMDB, metabolites were assigned to the Cytoplasm.

Table 4: Amount of metabolites and reactions in the exported model and their localization across compartments.

Reactions are classified into conversion and transport processes. Conversion processes are all chemical reactions whose substrates, cofactors and products are in the same compartment, whereas transport processes are reactions where at least one reactant is in a different compartment. Thetransport reactions that are listed as being localized in the Cytoplasm catalyze reactions that involve metabolites from the Cytoplasm and the Extracellular Space. Reactions in the Mitochondrial Lumen involve metabolites from the Mitochondrial Lumen and the Cytoplasm, whereas reactions localized in the Mitochondrial Intermembrane Space contain metabolites from the intermembrane space and the Mitochondrial Lumen. All values displayed have the unit [absolute amount (relative amount)].

Entity	Cytoplasm	Mitochondrial Lumen	Mitochondrial Intermembrane Space	Extracellular Space	Total
Metabolites	**1649(83)**	**48(2.4)**	**1(0.05)**	**286(14.4)**	**1987 (100)**
Reactions					
Conversion	1630	28	0	0	1658(75.8)
Transport	512	42	6	0	560(25.2)
Total	**2142(96.6)**	**70(3.2)**	**6(0.3)**	**0**	**2218(100)**

4.1.2. Calculation of the biomass composition

Biological growth can be interpreted as the synthesis of highly reduced non-polymeric macromolecules; lipids and steroids, as well as the construction of the typical biopolymers; nucleic acids, proteins and carbohydrates from their respective subunits. The exact cellular composition, meaning the percentage of each of the building blocks measured during growth, changes drastically between species. The biomass composition of *U. maydis* also differs depending on environmental conditions and varies over the course of its lifespan.

In iCL1079, growth is represented by a single reaction which transforms all required monomers and subunits of essential macromolecules into an artificial metabolite called "BIOMASS". Due to a lack of sufficient research in that regard, the stoichiometric coefficients, which indicate how much of each individual monomer is involved in the reaction relative to "BIOMASS", had to be calculated based on information gathered from various sources. The composition of proteins, RNA and DNA was estimated based on the protein and genome sequence respectively, whereas the composition of lipids and the cell wall were results mined from scientific articles(Hernandez et al. 1997; Ruiz-Herrera et al. 1996). There was no information available on the exact biomass composition of *U. maydis* in terms of how much each of the principal macromolecules (proteins, DNA, RNA, lipids, cell wall components) contributed to the biomass in percent. However, after further literature queries information on the specific elemental composition of *U. maydis* (Klement et al. 2012) and the biomass composition for fungi in general (Griffin 1994) could be obtained. Using linear programming it was now possible to approximate the total biomass composition of *U. maydis* (Table 5). The composition values of each monomer were then converted into stoichiometric values according to Thiele's and Palsson's "Protocol for generating a high-quality genome-scale metabolic reconstruction" (Thiele & Palsson 2010). In the following section, the individual steps will be explained using the calculation for amino acids (AA) as an example.

Table 5: Total biomass composition of *U. maydis*

Biomass Component	Molecular mass [g mol^{-1} of monomer in polymer]	Composition [g g^{-1} polymer]	Stoichiometric coefficient [mmol g_{CDW}^{-1}]
Alanine	7.51	0.02	0.29
Arginine	9.97	0.063	0.18
Asparagine	3.83	0.01	0.09
Aspartic Acid	6.68	0.09	0.16
Cystein	1.10	0.003	0.03
Glutamine	5.74	0.016	0.12
Glutamic Acid	6.86	0.09	0.15
Glycine	3.68	0.01	0.18
Histidine	3.61	0.01	0.07
Isoleucine	4.70	0.01	0.11
Lysine	10.14	0.03	0.25
Methionine	5.65	0.015	0.12
Phenylalanine	2.52	0.007	0.05
Proline	4.75	0.01	0.09
Serine	8.80	0.02	0.28
Threonine	6.08	0.01	0.17
Tryptophan	2.21	0.006	0.03
Tyrosine	3.67	0.01	0.06
Valine	5.81	0.016	0.16
Total Protein	**108.9**	**30.18 [g (100 g_{CDW})$^{-1}$]**	**2.76**
dATP	72.91	0.00070	0.00220
dTTP	70.84	0.00068	0.00220
dGTP	90.03	0.00086	0.00259
dCTP	79.21	0.00076	0.00259
Total DNA	**312.9**	**0.3 [g (100 g_{CDW})$^{-1}$]**	**0.01**
ATP	76.59	0.071	0.024
UTP	71.30	0.071	0.022
GTP	94.35	0.083	0.029
CTP	83.54	0.083	0.026
Total RNA	**325.7**	**10.00 [g (100 g_{CDW})$^{-1}$]**	**0.31**
Phosphatidylethanolamine	241.05	0.166	0.206
Phosphatidylcholine	160.15	0.110	0.130
Ergosterol	139.45	0.096	0.241
Palmitate	11.00	0.019	0.0297
Stearic Acid	13.72	0.023	0.0331
Oleate	3.70	0.006	0.0090
Linoleic Acid	14.88	0.025	0.0364
Linolenic Acid	0.19	0.00033	0.0005
Total Lipids	**584.13**	**40.11 [g (100 g_{CDW})$^{-1}$]**	**0.68**
Chitin	32.71	0.216	0.174
Mannose	42.28	0.280	0.317
Galactose	20.22	0.134	0.152
Glucose	47.79	0.316	0.358
Ribose	7.28	0.048	0.069
Total Cell Wall	**151.2**	**16.34 [g (100 g_{CDW})$^{-1}$]**	**1.08**
Ash		**3.07**	

4. Results

The relative average content of the principal components of proteins, DNA and RNA was calculated straight from the proteome and genome sequence available as FASTA file. In the case of the protein composition, the absolute amount of each AA codon was counted and subsequently the molar percentage MP was calculated.

(7) $$Molar\ percentage\ MP\ [\frac{mol}{mol}] = \frac{Absolute\ amount\ of\ codons\ for\ single\ AA}{Sum\ of\ all\ codons}$$

Desoxy- and ribonucleotides were calculated in the same way. For the components of the lipid pool and the cell wall relative compositions were taken from literature. The relative amounts were adapted to only contain those metabolites that existed in the model (Supplement 6).

In order to calculate the weight of an AA per mole of an average protein, the molecular weight of one water molecule had to be subtracted from the molecular weight of the AA, before multiplying it with the molar percentage. This was because one molecule of water is formed during the polymerization of two AA. For the same reasons, when calculating the weight of a nucleotide per mole of average DNA and RNA, the molecular weight of one molecule of pyrophosphate (PPi) was subtracted.

(8) $$Weight\ of\ AA_i\ per\ mole\ protein\ [\frac{g\ AA_i}{mol\ Protein}] = MP * (\frac{g}{mol} AA_i - \frac{g}{mol} H_2O)$$

(9) $$Weight\ of\ AA_i\ per\ gram\ protein\ [\frac{g\ AA_i}{g\ Protein}] = \frac{\frac{g\ AA_i}{mol\ Protein}}{\Sigma(\frac{g\ AA_n}{mol\ Protein})}$$

The following formula was used to calculate the stoichiometric fraction of each AA per gram dry weight of biomass [g_{CDW}]. The variable$X$ denotes the relative portion of proteins to biomass in percent.

(10) $$Stoichiometric\ factor\ of\ AA_i\ \left[\frac{mmol\ AA_i}{g\ DW}\right] = \frac{(\frac{g\ AA_i}{mol\ Protein} * X * 1000)}{(\frac{g}{mol} AA_i - \frac{g}{mol} H_2O)}$$

Due to lack of information in literature, $\boldsymbol{X}$ was calculated with linear programming.First of all, the average elemental composition of each macromolecule (Protein, DNA, RNA, Lipid,

Cell Wall) was determined by summing up the products of the absolute amount of each element (Carbon c, Hydrogen h, Nitrogen n, Oxygen o, Phosphorous p and Sulphur s) multiplied with the molar percentageMP of each monomer respectively. As an example:

(11) $$Amount\ of\ carbon\ atoms\ per\ mole\ of\ average\ protein\ = \sum c * MP_{AA_i}$$

With regard to the specific elemental composition of *U. maydis* in Klement et al. (2012), the amount of each element was normalized to the amount of carbon, yielding the C-mole content of each macromolecule. Before the elemental composition outlined in literature could be used in linear programming, values for Phosphorous and Sulfur had to be added. They were added from the elemental composition of *S. cerevisiae*, which was assumed to have a similar Phosphorous and Sulfur content in biomass as *U. maydis*. During the linear programming process their influence on the final result was tested, by varying them by several orders of magnitude respectively. Ultimately, the values for Phosphorous and Sulfur adapted from *S. cerevisiae* proved to have no significant influence. Hence, the linear programming problem was outlined as follows:

(12) $$\boldsymbol{A} * \boldsymbol{X} = \boldsymbol{b}$$

(13) Subject to: $$xlb_i \leq x_i \leq xub_i$$

The rows of matrix $\boldsymbol{A}$ correspond to the six individual elements c, h, o, n, p, s whereas the columns correspond to the five types of macromolecules (Protein, DNA, RNA, Lipids, Cell Wall) that make up the biomass. The vector $\boldsymbol{b}$ depicts the measured elemental biomass composition of *U. maydis* for c, h, o, n supplemented by the elemental content of p and s of *S. cerevisiae*. The equation was solved for vector $\boldsymbol{X}$, that containes the unknown fractional values with which each of the macromolecules contributed to the biomass composition. The standard LP solver was allowed to find a solution in between the bounds xlb and xub outlined in Table 6. The results of equation 13 are displayed in Table 5. Furthermore the raw data from literature is available as Supplement 5, the excel spreadsheet used for the calculation as Supplement 6 and the MatLab workspace used during the linear programming as Supplement 7.

Table 6: Upper and lower bounds used in linear programming to define the total biomass composition
The values for the bounds were adapted from (Griffin 1994, S.24, Table 1)

Principal Biomass components	Lower Bounds (xlb)	Upper Bounds (xub)
DNA	0.15	0.3
RNA	1	10
Protein	14	44
Lipids	0.2	87
Cell Wall	16	85

In addition to the biomass reaction itself, constraints in the form of ATP were implemented to account for the growth-associated maintenance costs (GAM costs), which is the amount of ATP associated with the synthesis of macromolecules relevant to biomass. Furthermore a simple ATP degrading reaction with a fixed rate was added to account for the non-growth-associated maintenance costs (NGAM costs).The values amounted to 46.3 $mol_{ATP}\ g_{CDW}^{-1}$ for GAM and 1.9 $mol_{ATP}\ g_{CDW}^{-1}\ h^{-1}$ for NGAM.

After adding the biomass reaction, each pathway leading up to the principal biomass components was checked for the ability to carry flux. This led to the discovery of metabolic gaps which made it impossible to simulate in-silico growth by maximizing the biomass reaction. The gaps existed in the biosynthesis of chorismate, Coenzyme A, linolenic acid, and ubiquinol. A list of all reactions that were added manually to fix the production of biomass components can be found in Supplement 8. This list includes reactions that were part of the addition of compartments, as well as reactions that were added during the manual curation of cycles, bypasses and duplicates.

4.1.3. Fixing erroneous cycles and reactions

After the model had successfully been exported from Pathway Tools, and the biomass reaction had been implemented and tested for functionality, an array of manual curation operations ensued on the basis of the SBML-formatted export file. Aided by Flux Balance Analysis (FBA), most effort of the manual curation process was focused on the removal of artificial transhydrogenase cycles, on fixing respiration reactions, and on the removal of

duplicate reactions. A full changelog is available as Supplement 8. From the initial export until the version presented herein, a total of 119 existing reactions had to be altered to eradicate such cycles, bypasses and duplicates.

Transhydrogenase cycles erroneously generated flux through the conversion of NADH into NADHP without the actual conversion of any metabolites. For many enzymes the exact redox cofactor usage was often unknown, which is why these reactions were exported into two instances, that contained either NADH or NADPH as a cofactor, but otherwise involved the same metabolites. By default, all of the reactions in the model were set to be reversible, unless there was explicit information that they were not. Hence in the simulation, flux could be channeled through one instance of the reaction pair consuming NADH or NAD and then through the respective other instance producing NADP or NADPH, without actually influencing the net amount of the involved metabolites. However, these cycles are most likely not a natural physiological process, which is why either one of the reactions was constrained to be irreversible or deleted entirely.

An example that details how these instanced reactions contribute to transhydrogenase cycles was the reversible interconversion of glycerate to hydroxypyruvate catalyzed by the hydroxypyruvatereductase (EC 1.1.1.81). In Pathway Tools, there is the general reaction which can use either cofactor, NADH or NADPH. During the export to an SBML file from Pathway Tools, the reaction was instanced into two reactions,"RXN0-300" and "GLYCERATE-DEHYDROGENASE-RXN", that were dependent on one type of cofactor respectively, while both remained reversible. This led to the following cycle: The GLYCERATE-DEHYDROGENASE-RXN catalyzed the conversion of Hydroxypyruvate to Glycerate oxidizing NADH to NAD, after which the RXN0-300 converted Hydroxypyruvate back to Glycerate reducing NADP to NADPH. The cycle was avoided by setting both reaction instances to be irreversible, which remained consistent with the initial reaction catalyzed by the hydroxypyruvatereductase (EC 1.1.1.81), meaning that Hydroxypyruvate could be reduced to Glycerate by either NADH or NADPH.

Since *U. maydis* is grown aerobically during fermentations, it was very important to be able to model the aerobic respiration as detailed as possible. Hence, reactions depicting the enzyme-complexes of the respiratory chain Complex I-V in the inner membrane of the mitochondria were manually added from the KEGG database. All uses of the generic

metabolites, "an electron-transport-related-quinol" or "ubiquinols" and "electron-transport-related-quinone" or "ubiquinones" were replaced with the already existing, distinct metabolites "ubiquinol-6" and "ubiquinone-6", respectively. Thus, pathways that involved such generic metabolites were reconnected to the reactions catalyzing aerobic respiration. Moreover, in order to ensure that no flux was able to bypass the reduction of cytochrome-C via the oxidation of ubiquinol-6 as catalyzed by the ubiquinol-cytochrome-C reductase (complex III), all reactions involving these coenzymes were constrained to be irreversible. Flavinadenin dinucleotide (FAD) is an electron-transport related coenzyme which is primarily involved in oxidative phosphorylation and the β-oxidation of fatty acids. Like with ubiquinol, all reactions that use FAD or FADH2 as a coenzyme have been constrained to be irreversible. An artificial lumped reaction termed "Oxidization-FADH2" has been added as a simplified workaround to connect all FAD-related red-ox conversions to the respiratory chain. As a means to prevent the circumvention of the Adenosine triphosphate synthase reaction "ATPSYN-RXN", all cytosolic reactions that use ATP as a substrate, were also fixed to be irreversible.

Lastly, in an effort to streamline the model, duplicate and partial reactions were removed. These had also appeared after exporting an SBML file from Pathway Tools. Furthermore, the PEDANT database entries of all genes associated with a transport reaction were double-checked for further information regarding the localization of the gene product. For 27 transport reactions the database indicated a compartment that did not exist in the model, so both upper and lower boundaries were set to zero and a note was left in the SBML file accordingly. The reactions were not deleted, to allow for an easy, more accurate reimplementation in the future.

4.2. Screening of carbon sources

A high-throughput growth experiment was carried out to determine which carbon sources can be utilized by *U. maydis*. In total 190 unique carbon sources were tested to uncover the possible range of feasible substrates and thus provide qualitative insight into potentially interesting pathways. Employing a custom MatLab script (Supplement 4), the raw data was automatically imported, smoothed, the growth rates were approximated by a custom linear fitting algorithm and ultimately the data was plotted (Supplement 9, Supplement 10). Table 7 displays the entire range of tested carbon sources and a qualitative assessment of growth judged from a measured change in OD_{600} over time. An approximated growth rate is also provided, wherever the raw data was in agreement with the fitting parameters.

Based on the approximated growth rates, five C-sources that triggered the fastest growth and five that *U. maydis* grew on slowest were examined further. The carbon source for fastest growth was L-Phenylalanine with an approximated growth rate of $\mu=0.778\ h^{-1}$. D,L-Malic Acid closely followed with $\mu=0.588\ h^{-1}$. With L-Rhamnose ($\mu=0.437\ h^{-1}$) and L-Leucine ($\mu=0.325\ h^{-1}$) in front, L-Isoleucine ($\mu=0.324\ h^{-1}$) was the substrate that allowed the fifth fastest growth. The five slowest substrates, in decreasing order were N-Acetyl-D-Glucosamine ($\mu=0.103\ h^{-1}$), L-Glutamine ($\mu=0.099\ h^{-1}$), L-Lyxose ($\mu=0.099\ h^{-1}$), 5-Keto-D-Gluconic Acid ($\mu=0.095\ h^{-1}$) and lastly D-Glucosamine at $\mu=0{,}064\ h^{-1}$.

The growth rate on Alpha-D-Glucose amounted to $\mu=0.138\ h^{-1}$, while that on Glycerol was slightly lower at $\mu=0.126\ h^{-1}$. The rate of Xylitol was higher than both at $\mu=0.172\ h^{-1}$. In the next section, the modeling capabilities of iCL1079 are analyzed with respect to the results of this screening

Table 7: Results of high-throughput growth experiment

Growth for all 190 carbon sources was assessed visually from observing the plotted data and the growth rates were approximated using a custom algorithm. The five highest (green) and lowest (red) growth rates and respective carbon sources are marked.

Substrate	Obs. Growth	Est. Growth Rate (h^{-1})	Substrate	Obs. Growth	Est. Growth Rate (h^{-1})
Negative Control			Capric Acid		
1,2-Propanediol			Caproic Acid		
2,3-Butanediol			Chondroitin Sulfate C		
2,3-Butanone			Citraconic Acid		
2-Aminoethanol	Yes		Citramalic Acid		
2-Deoxy Adenosine			Citric Acid		
2-Deoxy-D-Ribose			D,L-α-Glycerol-Phosphate		
2-Hydroxy Benzoic Acid	Yes	0.156	D,L-Carnitine		
3-0-β-D-Galactopyranosyl-D-Arabinose			D,L-Malic Acid	Yes	0.588
3-Hydroxy 2-Butanone			D,L-Octopamine		
3-Methyl-D-Glucoronic Acid			D-Alanine	Yes	0.206
4-Hydroxy Benzoic Acid	Yes	0.117	D-Arabinose	Yes	0.107
5-Keto-D-Gluconic Acid	Yes	0.095	D-Arabitol	Yes	
Acetamide			D-Aspartic Acid		
Acetic Acid	Yes		D-Cellobiose		
Acetoacetic Acid			δ-Amino Valeric Acid		
Adenosine			Dextrin		
Adonitol			D-Fructose	Yes	0.168
α-Cyclodextrin			D-Fructose-6-Phosphate		
α-D-Glucose	Yes	0.138	D-Fucose	Yes	0.205
α-D-Lactose			D-Galactonic Acid-γ-Lactone		
α-Hydroxy Butyric Acid	Yes		D-Galactose	Yes	0.182
α-Hydroxy Glutaric Acid-γ-Lactone			D-Galacturonic Acid	Yes	
α-Keto Valeric Acid	Yes		D-Gluconic Acid	Yes	0.150
α-Keto-Butyric Acid	Yes		D-Glucosamine	Yes	0.064
α-Keto-Glutaric Acid	Yes	0.220	D-Glucosaminic Acid		
α-Methy-D-Glucoside	Yes	0.142	D-Glucose-1-Phosphate		
α-Methyl-D-Galactoside			D-Glucose-6-Phosphate		
α-Methyl-D-Mannoside			D-Glucuronic Acid		
Amygdalin			Dihydroxy Acetone	Yes	0.148
Arbutin			D-Lactic Acid Methyl Ester	Yes	
β-Cyclodextrin			D-Malic Acid		
β-D-Allose			D-Mannitol	Yes	
β-Hydroxy Butyric Acid	Yes	0.112	D-Mannose	Yes	0.303
β-Methyl-D-Galactoside			D-Melezitose	Yes	0.136
β-Methyl-D-Glucoronic Acid			D-Mellibiose		
β-Methyl-D-Glucoside			D-Psicose		
β-Methyl-D-Xyloside			D-Raffinose	Yes	0.129
Bromo Sucinic Acid			D-Ribono-1,4-Lactone		
Butyric Acid	Yes	0.138	D-Ribose	Yes	0.113

Substrate	Obs. Growth	Est. Growth Rate (h^{-1})	Substrate	Obs Growth	Est. Growth Rate (h^{-1})
D-Saccharic Acid			L-Erythritol	Yes	0.244
D-Serine			L-Fucose	Yes	0.158
D-Sorbitol			L-Galactonic Acid-γ-Lactone	Yes	
D-Tagatose	Yes		L-Glucose		
D-Tartaric Acid			L-Glutamic Acid	Yes	
D-Threonine			L-Glutamine	Yes	0.099
D-Trehalose	Yes	0.301	L-Histidine	Yes	
Dulcitol			L-Homoserine	Yes	
D-Xylose	Yes	0.134	L-Isoleucine	Yes	0.324
Formic Acid			L-Lactic Acid	Yes	0.106
Fumaric Acid	Yes	0.136	L-Leucine	Yes	0.325
γ-Amino Butyric Acid	Yes	0.248	L-Lysine	Yes	
γ-Cyclodextrin			L-Lyxose	Yes	0.099
γ-Hydroxy Butyric Acid	Yes		L-Malic Acid	Yes	0.320
Gelatin			L-Methionine		
Gentiobiose	Yes		L-Ornithine	Yes	0.146
Glucoronamide			L-Phenylalanine	Yes	0.778
Glycerol	Yes	0.126	L-Proline	Yes	0.228
Glycine	Yes		L-Pyroglutamic Acid		
Glycogen			L-Rhamnose	Yes	0.437
Glycolic Acid			L-Serine	Yes	0.158
Glycyl-L-Aspartic Acid			L-Sorbose	Yes	
Glycyl-L-Glutamic Acid			L-Tartaric Acid		
Glycyl-L-Proline	Yes		L-Threonine	Yes	0.138
Glyoxylic Acid			L-Valine	Yes	0.234
Hydroxy-L-Proline			Malonic Acid		
Inosine			Maltitol	Yes	0.120
Insulin			Maltose	Yes	0.108
Itaconic Acid			Maltotriose	Yes	
Lacitol			Mannan		
Lactulose			Melibionic Acid		
L-Alaninamide	Yes	0.144	Methyl Pyruvate	Yes	0.109
L-Alanine	Yes	0.213	m-Hydroxy Phenyl Acetic Acid		
L-Alanyl-Glycine	Yes		M-Inositol	Yes	0.129
Laminarin	Yes	0.242	Mono Methyl Succinate	Yes	0.170
L-Arabinose	Yes	0.134	M-Tartaric Acid		
L-Arabitol	Yes	0.123	Mucic Acid		
L-Arginine		0.234	N-Acetyl-β-D-Mannosamine		
L-Asparagine	Yes	0.249	N-Acetyl-D-Galactosamine		
L-Aspartic Acid	No		N-Acetyl-D-Glucosamine	Yes	0.103

Substrate	Obs. Growth	Est. Growth Rate (h^{-1})	Substrate	Obs. Growth	Est. Growth Rate (h^{-1})
N-Acetyl-D-Glucosaminitol			Sedoheptulosan		
N-Acetyl-L-Glutamic Acid			Sorbic Acid	Yes	
N-Acetyl-Neuraminic Acid			Stachyose	Yes	0.158
Oxalic Acid			Succinamic Acid	Yes	
Oxalomalic Acid			Succinic Acid	Yes	0.177
Palatinose	Yes	0.171	Sucrose	Yes	0.148
Pectin			Thymidine		
Phenylethylamine			Tricarballylic Acid		
p-Hydroxy Phenyl Acetic Acid			Turanose	Yes	
Propionic Acid			Tween 20		
Putrescine			Tween 40		
Pyruvic Acid	Yes	0.132	Tween 80		
Quinic Acid	Yes		Tyramine		
Salicin	Yes		Uridine		
Sebacic Acid	Yes	0.133	Xylitol	Yes	0,172
Sec-Butylamine					

4.3. Modeling capacity of iCL1079

Flux Balance Analysis was carried out to gain insight into the predictive capabilities of iCL1079. Based on the results of the high-throughput growth screening, the five highest and lowest growth rates as well as the rates on glucose, glycerol and xylitol were chosen, and the respective carbon sources were provided as the sole carbon sources for in-silico simulations with the model (Table 8).

The Flux Vectors generated by FBA on each of the substrates were compared (Supplement 11). When the simulations were run with α-D-glucose, D,L-malic Acid or L-glutamine as the sole carbon sources, the model predicted the production of succinate: 2.93 mmol g_{CDW}^{-1} h^{-1} on α-D-glucose, 2.12 mmol g_{CDW}^{-1} h^{-1} on D,L-malic acid and 2.48 mmol g_{CDW}^{-1} h^{-1} on L-glutamine. Malate was produced only when the model was grown using α-D-glucose as the carbon source. The production rate amounted to 5.15 mmol g_{CDW}^{-1} h^{-1}. When grown on L-phenylalanine, L-leucine or L-glutamate, excess ammonium was secreted. While the rates were similar for L-phenylalanine (1.08 mmol g_{CDW}^{-1} h^{-1}) and L-leucine (1.44 mmol g_{CDW}^{-1} h^{-1}), the rate for L-glutamine was much higher in comparison (14.81 mmol g_{CDW}^{-1} h^{-1}). Regardless of the substrate used in simulations, the model predicted an average secretion of nicotinic acid of about 0.041 mmol g_{CDW}^{-1} h^{-1}.

Table 8: Comparison of approximated growth rates with growth rates predicted by iCL1079
The respective carbon sources are listed from fastest approximated growth to slowest approximated growth. Alpha-D-Glucose, glycerol and xylitol were included, as they are of potential interest for an industrial application using *U. maydis*.

C-Source Common Name	Est. Growth Rate based on Experimental Results (h^{-1})	Predicted Growth Rate based on FBA (h^{-1})
L-Phenylalanine	0.778	0.276
D,L-Malic Acid	0.588	0.221
L-Rhamnose	0.437	Not included in the model
L-Leucine	0.325	0.201
L-Isoleucine	0.324	0.000
N-Acetyl-D-Glucosamine	0.103	0.000
L-Glutamine	0.099	0.268
L-Lyxose	0.099	Not included in the model
5-Keto-D-Gluconic Acid	0.095	0.000
D-Glucosamine	0.064	Not included in the model
α-D-Glucose	0.138	0.397
Glycerol	0.126	0.227
Xylitol	0.172	0.263

The simulated yield of malate for *U. maydis* was calculated using three industrially relevant or for *Ustilago* interesting carbon sources (α-D-glucose, glycerol and xylitol). This was achieved by maximizing the excretion flux of malate in FBA instead of the biomass production rate. Considering the production of malate from glucose, the yield amounted to 1.276 $g_{malate}\ g^{-1}_{glucose}$, whereas the production from glycerol was at 1.456 $g_{malate}\ g^{-1}_{glycerol}$. Using xylitol as the substrate for the production resulted in a yield of 1.322 $g_{malate}\ g^{-1}_{xylitol}$.

5. Discussion

The genome-scale metabolic model for *U. maydis* presented as a result of this work, has been constructed bottom-up from an annotated genome using the semiautomatic draft reconstruction software Pathway Tools, followed by an extensive bibliome and database survey and manual curation.

Since the annotated genome marked the very foundation of all the processes which the GEM was constructed from, errors in the genes or annotations would have been transported all the way through, possibly leading to false-positive or false-negative single reactions and even entire pathways. While the initial high-quality of the exported GBK file (Mannhaupt et al. 2013) provided a trustworthy starting point, further manual curation during each step of the creation process ensured that no profound errors persisted. However, concrete experimental evidence on the physiology of *Ustilago maydis* has only been found for a few primary metabolic pathways (Saavedra et al. 2008; McCann & Snetselaar 2008), and for even fewer secondary metabolic pathways (Geiser 2014). While these pathways have been included into the model, as a part of this work, the inclusion of most other metabolic pathways in iCL1079 is at best hypothetical.

The Pathway Tools software offered reliable algorithms that helped to build the GENRE in the form of a PGDB, which had to be exported to an SBML format in order to generate a GEM using the COBRA toolbox. This export process unfortunately introduced some errors attributed to the inherent data structures within Pathway Tools, which became apparent as artificial cycles and duplicate reactions. Most of these cycles that affected the modeling conditions used during this study were purged from the model either by removing the corresponding duplicate reactions or by fixing the direction of conversion. The removal of these cycles is a common practice in literature (David et al. 2003), and was carried out according to available protocols (Thiele & Palsson 2010). However, other cycles may become apparent should different modeling conditions be applied.

Another reconstruction of the metabolic pathways of *U. maydis* was previously automatically created as a part of the Path2Models project by Büchel et al. (2013) using the SuBliMinaL Toolbox in a top-down approach, meaning that instead of starting from an annotated genome, reconstruction started with extracted data from KEGG and MetaCyc.

5. Discussion

The Path2Models project resulted in a metabolic model consisting of 2199 reactions and 1109 metabolites spread across three compartments (intracellular, extracellular, biomass). The model of Büchel et al. (2013) is less precise in depicting the correct localization of reactions due to lacking the correct compartments. It seems that the Path2Model network agrees well with the model created for this thesis, iCL1079, regarding the number of reactions, although a closer look at the model of Büchel et al. reveals that 1108 of the reactions are merely transport reactions added for modeling purposes with no gene associations at all. This leaves about 1091 of metabolic conversion reactions. Comparing this to the 1658 conversion reactions in iCL1079, can equally be interpreted as the latter being a more comprehensive or a more redundant depiction of the metabolism of *U. maydis*. Given the fact that Pathway Tools introduced plenty of duplicate reactions only differing in the cofactors used, contributes to inflating the number of reactions in iCL1079, making it redundant. However, since only 14% of reactions lack GPR associations, meaning that for most reactions there is an evident foundation in the genome, errors introduced by Pathway Tools cannot be the only reason for a difference in size. Whereas the model presented in this thesis contains a biomass reaction specifically calculated for *U. maydis* based on experimental and genomic evidence, Büchel et al. only included generic components to their biomass reaction, with no effort spent on determining the stoichiometry. In addition to the differences in size and the specificity of the biomass reaction, the iCL1079 model utilizes a more up-to-date set of *U. maydis* gene identifiers ('um1xxxx') than the model by Büchel et al. which uses the initial gene call set ('um00001-um06521') as proposed by the Broad Institute before manual curation was carried out by the MIPS during their work on the MUMDB (http://mips.helmholtz-muenchen.de/genre/proj/ustilago). Due to the fact that the Path2Model project was focused on creating functional metabolic models for a mass of different organisms automatically, rather than focusing on *U. maydis*, it is only meaningful to a certain extent to compare the two models based on their statistics and not including their predictive capabilities, albeit it is already clear that iCL1079 is more up-to-date, more detailed and more specific at depicting the metabolism of *U. maydis*.

Since iCL1079 was drafted in Pathway Tools, it contains MetaCyc specific identifiers and names for reactions and metabolites, which are less cryptic to understand, than for example the MetaNetX Ids (Bernard et al. 2014) used by Büchel et al. However, due to Pathway Tools' inherent structure of classes and instances, some names, especially the ones of transport reactions are rather long and convoluted. Citations and some cross-database annotations that

are included in the PGDB version of the model presented in this work were lost during the export from Pathway Tools to SBML. Using a supplied MatLab script (Supplement 12), however, these missing annotations can be parsed from a tab-delimited Pathway Tools export or SBML file if needed. EC Number annotations have been included in the SBML for about 67% (1496) of the total reactions, which represent an excellent means to specifically identifying and comparing reactions.

The stoichiometry of the components belonging to the biomass reaction was calculated according to the protocol by Thiele & Palsson (2010). Because there is little experimental evidence available, the compositions of DNA, RNA and proteins had to be estimated from the genome sequence, and those of lipids and cell wall had to be estimated from primary literature, with the issue of experimental conditions supporting the compositions differing wildly (Ruiz-Herrera et al. 1996; Hernandez et al. 1997). It is important to note that in order to simplify the process of defining the biomass from the scarce experimental results, certain omissions and assumptions had to be made, especially regarding the exact composition of lipids and cell wall components. Despite of experimental evidence predicting their existence (Hernandez et al. 1997), a number of free sterols such as ergosta-5,7-dienol, ergosta-7-ol, 14-methylfecosterol, eburicol, ergosta-5.8,22-trienol, ergosta-8,22-dienol and ergosta-8-enol and free fatty acids like arachidic acid, 11-eicosenoic acid, behenic acid and erucic acid were purposely left out. The reason why these metabolites were not present in the model is likely that the pathways leading up to them were either not predicted by Pathway Tools or not existent in MetaCyc. Likewise the cell wall of *U. maydis* was simplified to only contain a few basic neutral polysaccharides and chitin, omitting xylose, arabinose and fucose (Ruiz-Herrera et al. 1996), of which mostly degradation but no biosynthesis pathways could be found in MetaCyc. Hence, these omissions should definitely be included in a later update, although all aspects of the biomass composition should be determined in an experiment using unified parameters first. Such an experiment would help to verify the assumptions that had to be made due to missing knowledge on *U. maydis*. This specifically applies to the generic boundaries used in calculation and the fact that the elemental values for phosphorous and sulphur were adapted from *S. cerevisiae*. Furthermore, also the values for the growth and non-growth associated maintenance costs (GAM and NGAM) should be reconsidered in the future, as they were initially reported for *Aspergillus nidulans* by David et al. (2008). Unlike previously done in literature (Voll et al. 2012; Büchel et al. 2013), and despite of the

omissions and assumptions mentioned above, iCL1079 is the first GEM to carry out calculations with a biomass composition specifically tailored to *U. maydis*.

When looking at the results of the high-throughput carbon source screening it is clear that *U. maydis* is quite versatile in taking up and utilizing common and rare carbon sources alike, by growing on almost half of the 190 tested substrates (about 90). The range of amino acids, mono- and disaccharides, carboxylic acids and other organic compounds, that *U. maydis* is capable of growing on, can be attributed to the diversity of compounds available during the degradation of organic plant matter, for example maltose, α-D-glucose and sucrose play the key role in the starch metabolism of plants (Zeeman et al. 2010); arabinose, xylose, mannose, and galactose are monomers found in hemicellulose; and free amino acids are ubiquitous in living cells.

Nevertheless, there is also one rather surprising carbon source that *U. maydis* can grow on, lyxose. It is an extremely rare epimer of xylose, only found as a component of the glycolipids of two distinct species of bacteria *Mycobacterium phlei* and *Mycobacterium smegmatis* (Khoo et al. 1996), yet it can be catabolized by the fungus. In *Escherichia coli K-12 substr. MG1655* a possible pathway for lyxose degradation was described by Badia et al. (1991). L-lyxose is taken up through a rhamnose:proton-symporter (RhaT) with dual specificity for L-rhamnose and L-lyxose. A subsequent unspecific isomerisation step facilitated by a pentose isomerase that can also accept L-xylose, yields L-xylulose which enters the metabolism through the pentose-phosphate pathway. A similar mechanism has already been elucidated for *Aerobacter aerogenes* (Anderson & Wood 1962). Since *U. maydis* growths well on L-rhamnose, uptake and conversion could be facilitated by similar enzymatic processes to that found in *E. coli K-12 substr. MG1655* and *A. aerogenes*.

Using a custom algorithm in order to facilitate a bulk evaluation of the high-throughput growth data, specific growth rates were calculated for 61 of the substrates. As calculated by the algorithm, L-phenylalanine is the carbon source that exhibits the fastest approximated growth rate, and D-glucosamine is the carbon source exhibiting the slowest approximated growth rate. When these calculated results are compared with the plotted graphs (Supplement 9, Supplement 10) for the two compounds, it becomes apparent that there must have been an error in the algorithm for the approximation of growth rates. The slope of the graph just below the threshold line is evidently steeper for D-glucosamine than it is for L-phenylalanine. This drastic difference becomes even more apparent when taking the graph of D,L-malic acid into

account, which does not resemble a typical growth curve at all. It seems as if the algorithm was not sensitive enough to correctly detect the initial steep linear increases that the graphs of D-glucosamine and L-phenylalanine show, but instead calculated the growth rate from the later parts of the curve depicting the stationary phase. A solution for this could be to decrease the step-size for the linear fit from 3 to 2 consecutive values, making the algorithm follow the shape of the curve more closely, but also making it more sensitive to remaining noise. Alternatively, the amount of agreement that is required of two consecutive slopes could be raised from 10% to at least 50%, forcing the algorithm to only consider long segments of the same slope. A third option would be to restrict the algorithm to calculating the slope only for the first X points of measurement that display values greater than zero, which would ensure that the algorithm is limited to the exponential growth phase. Despite of the employed signal smoothing the results remained quite distorted. This can be improved by conducting growth experiments in triplicates, which also enables the application of statistical methods for further analysis.

Flux Balance Analysis was carried out as a means to determine the predictive capabilities of iCL1079. The growth rates predicted by the model averaged to about 0.25 for the carbon sources included in the model, with the exception of the growth rate on α-D-glucose which was the highest at 0.4. Given, the complexity of the metabolic network included in the model, it is hard to discern the exact reason for this difference. Further investigations using a visual representation capable of displaying the flux distributions across the different pathways of the model would be necessary. L-rhamnose, L-lyxose and D-glucosamine are not included in the model, since no degradation pathways could be inferred on the basis of the annotated genome from MetaCyc, nor any experimental evidence existed prior to this study, that outlined that these substrates could be metabolized by *Ustilago maydis*. The reason why the model did not show any growth on L-isoleucine, N-acetyl-D-glucosamine and 5-keto-D-gluconic acid, can very likely be attributed to missing reactions in the corresponding degradation pathways. In fact, no reactions can be found in the model for the S-2-methylbutryl-CoA:FAD oxidoreductase (EC 1.3.8.5) and the N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25), which are respectively listed as initial steps in "Isoleucine degradation" and the "N-acetylglucosamine degradation" pathways in MetaCyc. While no information could be found on the S-2-methylbutryl-CoA:FAD oxidoreductase in the GBK file either, surprisingly, an "uncharacterized protein" ("um11364") was annotated as coding for the corresponding N-

acetylglucosamine-6-phosphate deacetylase. In future, these pathways ought to be revisited and the corresponding gaps filled, by including the correct reactions.

Regardless of remaining gaps, iCL1079 is capable of correctly predicting the secretion of secondary metabolites produced by *U. maydis.* When the model is grown on α-D-glucose as a carbon source, with the object function of maximizing the biomass reaction, it also produced 5.15 mmol g_{CDW}^{-1} h^{-1} of malate and 2.93 mmol g_{CDW}^{-1} h^{-1} of succinate, which have been shown experimentally to be natural products of *Ustilaginaceae* (Guevarra & Tabuchi 2014). succinate was the only secondary metabolite that was secreted when the simulations were run with D,L-malic acid (2.12 mmol g_{CDW}^{-1} h^{-1}) and L-glutamine (2.48 mmol g_{CDW}^{-1} h^{-1}) as carbon source instead of α-D-glucose. The values had the same magnitude, although they were slightly lower than those of glucose. Since the calculations in FBA adhere to a steady-state condition, all influxes are balanced by equal effluxes, which can link the occurrence of succinate and malate as the predicted products to an excess of carbon, oxygen or hydrogen, which is equaled through excretion. Alternatively, the products could also be excreted in order to close the balance of cofactors (NAD, NADP and CoA) that take part in the specific degradation pathways of the underlying carbon sources. This means, that in order to maximize the biomass function and given a fixed set of reactions the model has no other mathematical choice than to excrete the products to fulfill the steady-state and mass-balance criteria. This principle is most likely the reason why simulated growth on all carbon sources leads to the production of nicotinic acid. Moreover, the steady-state and mass-balance conditions can be considered the cause the excretion of ammonium when the simulations are executed with L-phenylalanine (1.08 mmol g_{CDW}^{-1} h^{-1}), L-leucine (1.44 mmol g_{CDW}^{-1} $h^{-1)}$, and L-glutamine (14.81 mmol g_{CDW}^{-1} h^{-1}) as carbon source respectively. Still, the rate of excretion for L-glutamine would be expected to be only about twice as high as that of leucine or Phenylalanine, and not increased about 9-fold. A closer look at a visual representation of the flux distributions would simplify the process of determining the exact reason for this over-excretion and the production of succinate, malate, nicotinic acid and ammonium in the model.

The simulated yields of malate on three principal carbon sources of industrial relevance or interest for *Ustilago*, glucose, glycerol and xylitol, agree very well with the theoretical maximal yields, which can be calculated for each carbon source. Considering a carbon dioxide co-feed, the theoretical maximal yield on glucose amounts to 2 mol malate per 1 mol glucose (Brown et al. 2013), equaling to 1.489 g_{malate} $g^{-1}_{glucose}$, for which the model predicts a

simulated yield of 1.276 $g_{malate}\ g^{-1}_{Glucose}$. The theoretical maximal yield of malate on glycerol can be estimated to equal 1 mol malate per 1 mol glycerol i.e. 1.456 $g_{malate}\ g^{-1}_{glycerol}$, which corresponds exactly to the prediction of the model at 1.456 $g_{malate}\ g^{-1}_{glycerol}$. Xylitol is converted to xylulose-5-phosphate after oxidation to D-xylulose, and then further degraded through the pentose-phosphate pathway. From 3 mol of xylulose-5-phosphate, 2 mol of fructose-6-phosphate and 1 mol glyceraldehyde-3-phosphate can be generated. If 2 mol of fructose-6-phosphate are converted to 4 mol of pyruvate along with the single mol of glyceraldehyde-3-phosphate, the theoretical maximal yield of malate would amount to 5 mol malate per 3 mol xylitol i.e. 1.469 $g_{malate}\ g^{-1}_{xylitol}$. iCL1079 predicted a yield of 1.322 $g_{malate}\ g^{-1}_{xylitol}$, which equals to a ratio of 3 mol malate per 2 mol xylitol.

Despite the many shortcomings outlined in this discussions, the model presented in this thesis (iCL1079), is the first extensively-curated and most up-to-date genome-scale metabolic model of *Ustilago maydis*, fit for FBA and more advanced stoichiometric modeling techniques. In combination with the PGDB-version accessible via Pathway Tools it presents a comprehensive knowledgebase of the status-quo of research on the plant parasitic fungus. Although iCL1079 is already an extensive network, boasting a size of 2218 reactions, it is by no means a complete depiction of the metabolism of *U. maydis*, but rather a tool developed for further genomic and metabolic research.

6. Prospects

The generation of a GEM is an iterative process , which requires many cycles of curation and validation in order to accurately depict all aspects of a cell's metabolism. Validation is one aspect that could only partly be completed in the limited time available for this thesis, but which is very desirable to improve the predictive qualities of iCL1079. Experiments should be conducted to measure the uptake and product secretion flux rates of *U. maydis* on the three chosen carbon sources, if not on all 190 carbon sources screened in this thesis. Until experimental results on the biomass composition are available, the approximated composition will hold as it was carefully constructed on combined evidence available at the time. Other aspects that experimental efforts should be directed towards are accurately determining the GAM and NGAM requirements, as these strongly factor into the outcome of FBA calculations. Moreover, the addition of the cellobiose and mannosylerythritol lipid pathways will make it possible to simulate the production of ustilagic acid and mannosylerythritol lipids respectively. Using FBA with iCL1079 can aid in metabolic engineering processes for itaconate or malate production, as well as provide a valuable tool for the testing of other metabolism-related hypotheses. If transcriptome data was included in the future, combined with a visual representation of the model presented herein, deeper insights into the metabolic regulation of *U- maydis* can be gained.

7. References

Anderson, R.L. & Wood, W.A., 1962. Pathway of L-Xylose and Degradation in Aerobacter aerogenes. *The Journal of Biological Chemistry*, 237(2).

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25, pp.25–29.

Badia, J. et al., 1991. L-Lyxose Metabolism Employs the L-Rhamnose Pathway in Mutant Cells of Escherichia coli Adapted To Grow on L-Lyxose. , 173(16), pp.5144–5150.

Barrett, A.J., 1995. Enzyme Nomenclature. Recommendations 1992. *European Journal of Biochemistry*, 232, p.1. Available at: http://dx.doi.org/10.1111/j.1432-1033.1995.tb20774.x.

Benson, D.A. et al., 2013. GenBank. *Nucleic Acids Research*, 41.

Bernard, T. et al., 2014. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in bioinformatics*, 15(1), pp.123–35. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3896926&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2014].

Bölker, M. et al., 1995. Tagging pathogenicity genes in Ustilago maydis by restriction enzyme-mediated integration (REMI). *MGG Molecular & General Genetics*, 248, pp.547–552.

Bölker, M., Basse, C.W. & Schirawski, J., 2008. Ustilago maydis secondary metabolism-From genomics to biochemistry. *Fungal Genetics and Biology*, 45.

Brefort, T. et al., 2009. Ustilago maydis as a Pathogen. *Annual review of phytopathology*, 47, pp.423–445.

Brown, S.H. et al., 2013. Metabolic engineering of Aspergillus oryzae NRRL 3488 for increased production of L-malic acid. *Applied microbiology and biotechnology*, 97(20), pp.8903–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23925533 [Accessed January 8, 2015].

Büchel, F. et al., 2013. Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC systems biology*, 7, p.116. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4228421&tool=pmcentrez&rendertype=abstract [Accessed January 6, 2015].

Caspi, R. et al., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42.

David, H. et al., 2008. Analysis of Aspergillus nidulans metabolism at the genome-scale. *BMC genomics*, 9, p.163.

7. References

David, H., Kesson, M. & Nielsen, J., 2003. Reconstruction of the central carbon metabolism of Aspergillus niger. *European Journal of Biochemistry*, 270(21), pp.4243–4253. Available at: http://doi.wiley.com/10.1046/j.1432-1033.2003.03798.x [Accessed August 6, 2014].

Frishman, D. et al., 2001. Functional and structural genomics using PEDANT. *Bioinformatics (Oxford, England)*, 17, pp.44–57.

García-Muse, T., Steinberg, G. & Perez-Martín, J., 2004. Characterization of B-type cyclins in the smut fungus Ustilago maydis: roles in morphogenesis and pathogenicity. *Journal of cell science*, 117, pp.487–506.

Geiser, E., 2014. Itaconic Acid Production by Ustilago maydis Diplom-Biologin.

Geiser, E. et al., 2014. Prospecting the biodiversity of the fungal family Ustilaginaceae for the production of value-added chemicals. *Fungal Biology and Biotechnology*, 1(1), p.2. Available at: http://www.fungalbiolbiotech.com/content/1/1/2 [Accessed December 23, 2014].

Griffin, D.H., 1994. *Fungal Physiology* 2nd ed., New York: Wiley-Liss, Inc.

Guevarra, E.D. & Tabuchi, T., 2014. Agricultural and Biological Chemistry Accumulation of Itaconic , 2- Hydroxyparaconic , Itatartaric , and Malic Acids by Strains of the Genus Ustilago. *Agricultural and Biological Chemistry*, 54(9), pp.37–41.

Hamilton, J.J. & Reed, J.L., 2014. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental microbiology*, 16(1), pp.49–59. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24148076 [Accessed March 21, 2014].

Hemetsberger, C. et al., 2012. The Ustilago maydis effector Pep1 suppresses plant immunity by inhibition of host peroxidase activity. *PLoS Pathogens*, 8.

Hernandez, A. et al., 1997. Fungicides and sterol-def icient mutants of Ustiago maydis : plasma membrane physico- chemical characteristics do not explain growth inhibition. , (1 997).

Hucka, M. et al., 2007. Systems Biology Markup Language (SBML) Level 2 : Structures and Facilities for Model Definitions.

Kämper, J. et al., 2006. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature*, 444(7115), pp.97–101. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17080091 [Accessed April 15, 2014].

Kanehisa, M. et al., 2014. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, 42.

Kanz, C. et al., 2005. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 33.

Karp, P.D. et al., 2004. An Evidence Ontology for Use in Pathway/Genome Databases. *Pacific Symposium on Biocomputing*, 9, pp.190–201.

Karp, P.D. et al., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1), pp.40–79. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2810111&tool=pmcentrez&rendertype=abstract [Accessed March 25, 2014].

Karp, P.D., Latendresse, M. & Caspi, R., 2011. The pathway tools pathway prediction algorithm. *Standards in genomic sciences*, 5(3), pp.424–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3368424&tool=pmcentrez&rendertype=abstract [Accessed September 30, 2014].

Keating, S.M. et al., 2006. SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics (Oxford, England)*, 22(10), pp.1275–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16574696 [Accessed October 6, 2014].

Khoo, K. et al., 1996. Chemistry of the Lyxose-Containing Mycobacteriophage Receptors of. , 2960(970), pp.11812–11819.

Kim, J. & Reed, J.L., 2014. Refining metabolic models and accounting for regulatory effects. *Current opinion in biotechnology*, 29C, pp.34–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24632483 [Accessed March 19, 2014].

Kitamoto, D., Isoda, H. & Nakahara, T., 2002. Functions and potential applications of glycolipid biosurfactants--from energy-saving materials to gene delivery carriers. *Journal of bioscience and bioengineering*, 94, pp.187–201.

Klement, T. et al., 2012. Biomass pretreatment affects Ustilago maydis in producing itaconic acid. *Microbial cell factories*, 11(1), p.43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3364905&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014].

Kojic, M. et al., 2002. BRCA2 homolog required for proficiency in DNA repair, recombination, and genome stability in Ustilago maydis. *Molecular Cell*, 10, pp.683–691.

Koonin, E. V. & Galperin, M.Y., 2003. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*, Boston: Kluwer Academic. Available at: http://www.ncbi.nlm.nih.gov/books/NBK20256/.

Latendresse, M. et al., 2012. Construction and completion of flux balance models from pathway databases. *Bioinformatics (Oxford, England)*, 28(3), pp.388–96. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3268246&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

Liu, L. et al., 2010. Use of genome-scale metabolic models for understanding microbial physiology. *FEBS Letters*, 584, pp.2556–2564.

Mannhaupt, G. et al., 2013. Ustilago Maydis Annotated Genome.

McCann, M.P. & Snetselaar, K.M., 2008. A genome-based analysis of amino acid metabolism in the biotrophic plant pathogen Ustilago maydis. *Fungal genetics and biology : FG & B*, 45 Suppl 1, pp.S77–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18579420 [Accessed August 6, 2014].

Mewes, H.W. et al., 1999. MIPS: A database for genomes and protein sequences. *Nucleic Acids Research*, 27, pp.44–48.

Münsterkötter, M. & Steinberg, G., 2007. The fungus Ustilago maydis and humans share disease-related proteins that are not found in Saccharomyces cerevisiae. *BMC genomics*, 8, p.473.

Ogata, H. et al., 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27, pp.29–34.

Orth, J.D. & Palsson, B.Ø., 2011. What is flux balance analysis? *Nature biotechnology*, 28(3), pp.245–248.

Patil, K.R., Åkesson, M. & Nielsen, J., 2004. Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology*, 15, pp.64–69.

Price, N.D., Reed, J.L. & Palsson, B.Ø., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature reviews. Microbiology*, 2, pp.886–897.

Pruitt, K.D. et al., 2012. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Research*, 40.

Renshaw, J.C. et al., 2002. Fungal siderophores: structures, functions and applications. *Mycological Research*, 106, pp.1123–1142.

Ruepp, A. et al., 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32, pp.5539–5545.

Ruiz-Herrera, J. et al., 1996. Structure and chemical composition of the cell walls from the haploid yeast and mycelial forms of Ustilago maydis. *Fungal genetics and biology : FG & B*, 20(2), pp.133–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8810518.

Saavedra, E. et al., 2008. Glycolysis in Ustilago maydis. *FEMS yeast research*, 8(8), pp.1313–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18803552 [Accessed May 8, 2014].

Schellenberger, J. et al., 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 6(9), pp.1290–307. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319681&tool=pmcentrez&rendertype=abstract [Accessed July 15, 2014].

Schuchardt, I. et al., 2005. Myosin-V, Kinesin-1, and Kinesin-3 cooperate in hyphal growth of the fungus Ustilago maydis. *Molecular biology of the cell*, 16, pp.5191–5201.

Spoeckner, S. et al., 1999. Glycolipids of the smut fungus Ustilago maydis from cultivation on renewable resources. *Applied Microbiology and Biotechnology*, 51(1), pp.33–39. Available at: http://link.springer.com/10.1007/s002530051359.

Steinberg, G. & Perez-Martin, J., 2008. Ustilago maydis, a new fungal model system for cell biology. *Trends in cell biology*, 18(2), pp.61–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18243705 [Accessed September 20, 2014].

Tanabe, M. & Kanehisa, M., 2012. Using the KEGG database resource. *Current Protocols in Bioinformatics*.

Tateno, Y. et al., 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic acids research*, 30, pp.27–30.

Tatusova, T. et al., 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic acids research*, 42(Database issue), pp.D553–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965038&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

Thiele, I. & Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1), pp.93–121. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125167&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2014].

Voll, A. et al., 2012. Metabolic Modelling of Itaconic Acid Fermentation with Ustilago Maydis. , 27(2004), pp.367–372.

Wahl, R. et al., 2010. A novel high-affinity sucrose transporter is required for virulence of the plant pathogen Ustilago maydis. *PLoS Biology*, 8.

Weber, I., Gruber, C. & Steinberg, G., 2003. A class-V myosin required for mating, hyphal growth, and pathogenicity in the dimorphic plant pathogen Ustilago maydis. *The Plant cell*, 15, pp.2826–2842.

Zeeman, S.C., Kossmann, J. & Smith, A.M., 2010. Starch: its metabolism, evolution, and biotechnological modification in plants. *Annual review of plant biology*, 61, pp.209–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20192737 [Accessed July 9, 2014].

8. Appendix

For access to data and the source code of the scripts created over the course of this thesis, please refer to the enclosed compact disc. References to the individual files are made within the body of the thesis. The following list contains a brief summary of all the files included:

Supplement 1: iCL1079.xml

SBML-formatted, COBRA-ready genome-scale metabolic model presented in this thesis. For FBA execute script from Supplement 3 after import into MatLab.

Supplement 2: umaybase.ocelot

The genome-scale network reconstruction as a PGDB structure for use in Pathway Tools.

Supplement 3: NutrientsTraceElementFBAScript.m

Builds the necessary framework for iCL1079 to be ready for simulations via FBA. It defines the reaction bounds, adds the components of a modified Tabuchi medium as exchange reactions, adds the corresponding uptake reactions for vitamins, defines which reaction is set as the objective function and conducts an FBA calculation.

Supplement 4: TableReader.m

Imports measured values from a preformatted Excel sheet exported by the Gen5 data analysis software (Synergy Mk micro-plate reader, BioTek) according to a defined spacing scheme. The Data is then normalized, smoothed and plotted automatically. A fitting algorithm tries to identify an exponential increase in the processed data points in order to estimate the growth rate of *U. maydis* on the corresponding carbon source.

Supplement 5: Ustilago Biomass Research Compilation.vsd

Microsoft Visio 2007 file containing collected tables and graphs on the composition of the cell wall, the elemental composition and the lipid

composition of *U. maydis* (Ruiz-Herrera et al. 1996; Klement et al. 2012; Spoeckner et al. 1999; Hernandez et al. 1997)

Supplement 6: Amino Acid and GC Content.xlsx

Used calculating the exact distribution of biomass components. Contains unmarked raw and refined results. Could be used for future calculations.

Supplement 7: LinearProgrammingTotalBiomassComposition.mat

Workspace contains variables used during the calculation of the total biomass composition.

Supplement 8: Changelog of Reactions.xlsx

A list of all reactions that were manually added or changed to fix the production of biomass components as well as reactions that were part of the addition of compartments, and those that were added or changed during the manual curation of cycles, bypasses and duplicates.

Supplement 9: PM1.pdf

Collection of graphs depicting the results of the high-throughput growth experiments for PM1 (Table 2). The horizontal yellow line marks the threshold at an OD_{600} of 0.3 as a starting criterion for the fitting algorithm in Supplement 4 to approximate the growth rate. The blue line represents normalized data, whereas the green line represents normalized, smoothed data.

Supplement 10: PM2A.pdf

Collection of graphs depicting the results of the high-throughput growth experiments for PM2A (Table 2). The horizontal yellow line marks the threshold at an OD_{600} of 0.3 as a starting criterion for the fitting algorithm in Supplement 4 to approximate the growth rate. The blue

line represents normalized data, whereas the green line represents normalized, smoothed data.

Supplement 11: Carbon Sources Flux Vector Analysis.xlsx

Flux vectors obtained from FBA based on in-silico growth experiments using the carbon sources who yielded the five highest and lowest growth rates as well as glucose, glycerol and xylitol.

Supplement 12: AnnotatePthwTIsModel.m

MatLab script that parses files exported from Pathway Tools for common database ID annotations in order to complete models created with the COBRA Toolbox.