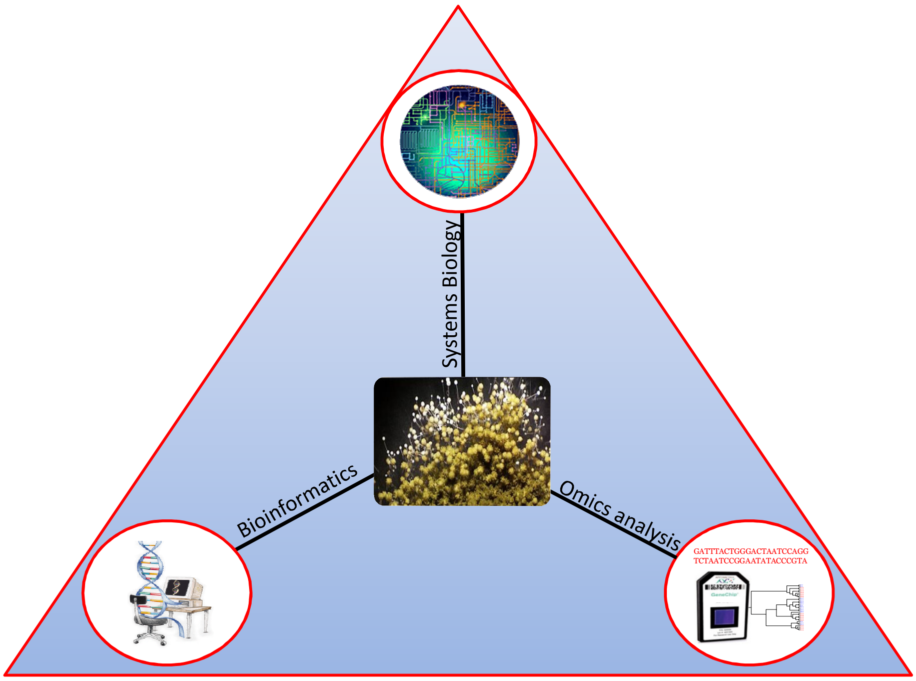


SYSBIOMICS of Aspergilli:

SYStems Biology, **BI**oinformatics and **OMICS** analysis **of Aspergilli** cell factories

Wanwipa Vongsangnak

Department of Chemical and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2009

Thesis for the Degree of Doctor of Philosophy

SYSBIOMICS of Aspergilli:

SYStems Biology, **BI**oinformatics and **OMICS** analysis **of Aspergilli** cell factories

Wanwipa Vongsangnak

Systems Biology
Department of Chemical and Biological Engineering
Chalmers University of Technology

Göteborg, Sweden 2009

SYSBIOMICS of Aspergilli:
SYStems Biology, **BI**oinformatics and **OMICS** analysis **of Aspergilli** cell factories

WANWIPA VONGSANGNAK

ISBN 978-91-7385-330-9

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3011
ISSN 0346-718X

Systems Biology
Department of Chemical and Biological Engineering
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Overview of SYSBIOMICS for *Aspergillus* research. The center picture is sporulating *A. oryzae* donated by Mikako Sasa from Novozymes A/S.

Printed by Chalmers Reproservice
Göteborg, Sweden 2009

To my family and friends

SYSBIOMICS of Aspergilli: **SYS**tems Biology, **BI**oinformatics and **OMICS** analysis **of Aspergilli** cell factories

Wanwipa Vongsangnak
Systems Biology, Department of Chemical and Biological Engineering, Chalmers University of Technology

Abstract

Aspergilli represent a group of filamentous fungi that plays a key role in industrial biotechnology, and as human, animal and plant pathogens. Here three *Aspergillus* species, namely *Aspergillus oryzae*, *Aspergillus niger* and *Aspergillus nidulans* are considered. These three species serve as working horses in industrial production of enzymes and chemicals and as key models for basic scientific work. Due to their wide applications, it is valuable to gain understanding of their metabolism, regulation and evolution with respect to genotypes and phenotypes, as this may lead to improved industrial fermentation processes for desired product formation (e.g. enzymes). We therefore applied three approaches for this investigation, namely SYStems biology, BIoinformatics and OMICS analysis (SYSBIOMICS). Firstly, we developed BIoinformatics methods to improve the genome annotation of *A. oryzae* and this improved annotation was used to reconstruct a high quality genome-scale metabolic network that could be used for mathematical modeling of the physiology and for OMICS data integration, which are the core of SYStems biology. Secondly, we designed a tri-*Aspergillus* DNA microarray chip to monitor the global regulation response at the transcriptional level. This DNA chip has been exploited to reveal conserved regulatory responses through evolution in the three aspergilli in response to change in carbon source. This resulted in mapping of key regulatory points of metabolism in these fungi, and it showed that SYSBIOMICS analysis of transcriptional data can lead to reconstruction of how carbon metabolism is regulated. Lastly, we also applied the SYSBIOMICS concept to identify possible key players/targets associated with protein production in a high producing strain of *A. oryzae*. This analysis may enable diagnosis and improvement of industrial process of protein production. In conclusion, through a number of studies, it has been demonstrated in this thesis that SYSBIOMICS can find wide applications in industrial biotechnology and assist in improving industrial process required for sustainable production of enzymes and chemicals in the future.

Keywords: Aspergilli, Bioinformatics, Omics analysis, Systems Biology

List of publications and manuscripts

The outlines presented in this thesis have formed the basis of the following papers, manuscripts, and book chapters:

Paper publications included in this thesis

Paper 1. **Vongsangnak. W**., Olsen. P., Hansen. K., Krogsgaard. S., Nielsen. J: Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. BMC Genomics 2008, 9: 245.

Paper 2. Andersen. M., **Vongsangnak. W**., Panagiotou. G., Margarita. S., Lehmann. L., Nielsen. J: A tri-species *Aspergillus* microarray - advancing comparative transcriptomics. Proceedings of the National Academy of Sciences 2008, 105:4387-4392.

Paper 3. **Vongsangnak. W***., Salazar. M*., Hansen. K., Nielsen. J: Genome-wide analysis of maltose utilization and regulation in aspergilli, Microbiology 2009.

Paper 4. Salazar. M*., **Vongsangnak. W***., Andersen. M., Panagiotou. G., Nielsen. J: Uncovering transcriptional regulation of glycerol metabolism in aspergilli through genome-wide gene expression data analysis, Molecular Genetics and Genomics, 2009.

Paper 5. **Vongsangnak. W**., Nookaew I., Salazar. M., Nielsen. J: Analysis of genome-wide co-expression and co-evolution of *Aspergillus oryzae* and *Aspergillus niger*, Manuscript in submission.

Paper 6. **Vongsangnak. W**., Hansen. K., Nielsen. J: Protein production by *Aspergillus oryzae*, Manuscript in preparation.

Book chapter publication included in this thesis

Paper 7. **Vongsangnak. W**., Nielsen. J: Systems Biology of Aspergilli: Molecular Biology and Genomics, Horizon Scientific Press, 2009.

Other publications that are not included in this thesis.

Paper 8. Andersen. M., Kjeldsen. K., Nookaew. I., **Vongsangnak. W**., Nielsen, J: ReMapper: A genome-scale model-based tool for analysis of systemic data in industrially relevant microorganisms, Manuscript in preparation.

Paper 9. Otero. J., **Vongsangnak. W**., Asadollahi. M., Olivares. R., Maury. J., Farinelli. L., Barlocher. L., Schalk. M., Clark. A., Nielsen. J: Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications, Manuscript in submission.

Paper 10. Ruenwai. R., Laoteng. K., Cheevadhanarak. S., Petranovic. D., **Vongsangnak. W**., Nielsen. J: Genome-wide regulation of Gamma- Linolenic Acid (GLA) biosynthesis in *Saccharomyces cerevisiae*, Manuscript in preparation.

Paper 11. Geijer. C., **Vongsangnak. W**., Nielsen. J., Hohmann. S: Transcriptional analysis of spore germination in *Saccharomyces cerevisiae,* Manuscript in preparation.

Paper 12. Piddocke. M., Fazio. A., **Vongsangnak. W**., Wong. M., Heldt-Hansen. H., Workman. C., Nielsen. J., Olsson. L: Revealing the beneficial effect of protease supplementation to high gravity beer fermentations using "–omics" techniques, Manuscript in submission.

Paper 13. Piddocke. M., Fazio. A., **Vongsangnak. W**., Wong. M., Nielsen. J., Olsson. L: Transcriptional study of high gravity beer fermentations- the effect of glucose repression and nitrogen limitation, Manuscript in submission.

Paper 14. Cvijovic. M*., Olivares. R*., Ågren. R*., Dahr. N., **Vongsangnak. W**., Nookaew. I., Nielsen. J: BioMet Toolbox: A toolbox for systematic analysis of metabolism, Manuscript in submission.

(*) Authors contributed equally.

My contributions to papers included in this thesis are followed:

Paper 1. I carried out the improved annotation, performed the genome-scale metabolic modeling and wrote the manuscript.

Paper 2. I designed probes for the *A. oryzae* EST collection and performed the *A. oryzae* experiments.

Paper 3. I carried out the fermentation experiments with *A. oryzae*, conducted the microarray experiments, performed the data analysis and wrote the manuscript.

Paper 4. I performed the fermentation experiments with *A. oryzae*, conducted the microarray experiments, performed the transcriptome data analysis and helped for preparation of the manuscript.

Paper 5. I performed all integrated data analysis and wrote the manuscript.

Paper 6. I participated in the experimental design, carried out the fermentation experiments with *A. oryzae*, and performed the transcriptome data analysis and wrote the manuscript.

Abbreviations

AFE - Animal Feed Enzyme
AIST - National Institute of Advanced Industrial Science of Technology (Japan)
ANOVA - ANalysis Of VAriance
BLAST - Basic Local Alignment Search Tool
bp - Base pairs
C_{12} - Twelve-carbon (compound)
C_3 - Three-carbon (compound)
C_5 - Five-carbon (compound)
C_6 - Six-carbon (compound)
CADRE - Central *Aspergillus* data repository
cDNA - complementary DNA
COG - Clusters of Orthologous Group
DOE - Department of Energy
DOGAN - Database Of the Genomes Analyzed at NITE
EC - Enzyme Commission
ER - Endoplasmic Reticulum
EST - Expressed Sequence Tag
FBA - Flux Balance Analysis
FDA - Food and Drug Administration
FE - Food Enzyme
FGI - Fungal Genome Initiative
FGSC Fungal Genetics Stock Center
GFAOP - Gap Filler for *Aspergillus oryzae* Pathway
GO - Gene Ontology
GRAS - Generally Regarded As Safe
HOG - High Osmolarity Glycerol
JGI - Joint Genome Institute
kbp - Kilo base pairs
Mb - Megabases
NCBI - National Center for Biotechnology Information
NHGRI - National Human Genome Research Institute
NIAID - National Institute of Allergy and Infectious Diseases
NITE - National Institute of Technology and Evaluation
NR - Non-Redundant
ORF - Open Reading Frame
Pfam - Protein family
PPP - Pentose Phosphate Pathway
PSI-BLAST- Position Specific Iterative-Basic Local Alignment Search Tool
SBML - Systems Biology Markup Language
SVG -Scalable Vector Graphics
SYSBIOMICS - SYStems biology, BIoinformatics, and OMICS analysis
TA - Transcriptional Analysis
TCA - Tri-Carboxylic Acid (cycle)
TE - Technical Enzyme
TIGR - The Institute for Genomic Research
UPR - Unfolded Protein Response
USDA/ARS - United States Department of Agriculture/Agricultural Research Service
WHO - World Health Organization

Table of contents

Chapter 1

1. Introduction

SYSBIOMICS is a novel term defined from this Ph.D. study. It is an abbreviation of three terms: **SYS**tems biology, **BI**oinformatics, and **OMICS** analysis. My definition of SYSBIOMICS is a systematic study of biological systems by integration of bioinformatics and omics analysis. Today the availability of bioinformatics techniques and tools as well as multi-level omics data of different aspergilli has allowed for systems studies of these fungi.

Aspergillus belongs to a group of filamentous ascomycete fungi that plays a key role as being of biotechnological importance. Particularly, it is illustrated by the use of *A. oryzae* in fermentation industries. *A. oryzae* has been applied for hundreds of years for the production of soy sauce, miso and sake with safe use. It can be used for large-scale production of enzymes and other proteins. Nowadays the use of *A. oryzae* has been facilitated in modern biotechnology.

Currently, there is much interest in SYSBIOMICS of *A. oryzae* and comparative analysis to other important *Aspergillus* species (e.g. *A. niger* as a citric acid producer and *A. nidulans* as a gene regulation model organism) as it is expected that this may lead to improve industrial fermentation processes, but also to enhance our understanding of the context of cellular metabolism, regulation and evolution and hereby assist us to further improve strains and production processes. Today the genome sequencing and bioinformatics in aspergilli have been established. It is followed by a more-gradual process of genome annotation in order to identify all the genes and their functions in biological processes as done for several other organisms. As we know, genome sequencing and bioinformatics in different organisms have resulted in a revolution in biology including development of many new experimental techniques that enables analysis at the genome-scale. Moreover, mathematical analysis, computational tools and information contents obtained in experimental biology are combined. These new approaches can be applied for understanding biological processes in the field of basic biology. However, these tools have been difficult to apply in industrial processes. A main reason is that most systems biology and bioinformatics algorithms have not been developed for industrial biotechnology. Therefore, the objective of this study is to develop systems biology tools and bioinformatics methods aiming at the construction of a genome-scale metabolic model of *A. oryzae* and further use the model for high-throughput omics data analysis from industrial fermentations for improved strains and processes. The knowledge accumulated in SYSBIOMICS approach can be further applied in metabolic engineering field to rationally enhance product formations and cellular properties of the producing organism. Therefore, the SYSBIOMICS approach applied throughout this study has a broad impact on the field of systems biology and metabolic engineering of filamentous fungi, and it will lead to a substantial improvement in our understanding of the important cell factory *A. oryzae* and related fungi. An overview of the SYSBIOMICS approach applied throughout this Ph.D. study is presented in Figure 1.1.

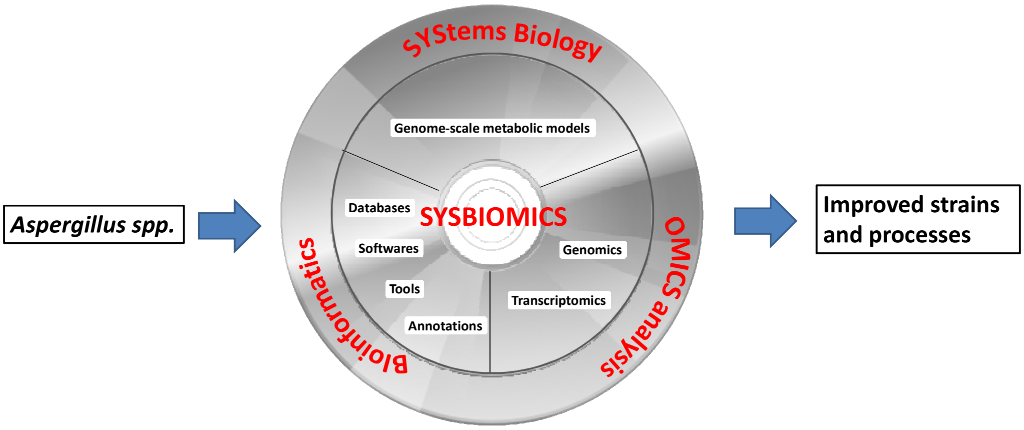


Figure 1.1 An applied SYSBIOMICS approach for *Aspergillus* research

1.1 Ph.D. thesis structure

My Ph.D. study has resulted in one book chapter and several published research articles and manuscripts submitted for publications. The thesis is divided into two parts. The first part is an extended summary describing the overall work during this study: introduction of the thesis, literature review, description of applied methods, illustration of obtained results and discussion, and conclusions drawn. The second part contains six research articles that are based on the results of this study, which are published or submitted for publication to international journals. The extended summary starts with an overview and broad application of aspergilli as well as an introduction to *Aspergillus* as efficient microbial cell factory. **Chapter 2** gives an illustration of the central carbon metabolism and linking it to protein production. **Chapter 3** gives an overview of SYSBIOMICS approach applied in this study. Moreover, we describe application of SYSBIOMICS for three case studies. The first case study presented in **Chapter 4** shows an improved annotation through development of genome-scale metabolic model for *A. oryzae*. This chapter describes how bioinformatics, systems biology and metabolic engineering tool can be combined and how it is provided genotypic and phenotypic relationship. In addition, it discusses the results of simulation and validation of the *A. oryzae* metabolic model at various growth conditions (see full details in Paper 1). **Chapter 5** represents the second case study which summarizes the results and discussion from Paper 2, 3, 4 and 5. In this chapter, the first part is a description of growth physiology for three *Aspergillus* species (i.e. *A. oryzae*, *A. niger* and *A. nidulans*) on four different carbon sources. The second part describes comparative transcriptomics of three *Aspergillus* species across four different carbon sources. The third part describes analysis of genome-wide co-expression and co-evolution of aspergilli. **Chapter 6** provides the third case study which describes how to identify possible key players/targets for protein production (see full details in Paper 6). This chapter discusses growth physiology and comparative transcriptome results of two different *A. oryzae* strains (i.e. α-amylase high producer and wild type).

Chapter 2

2. Aspergilli
2.1 An overview and broad application

The genus *Aspergillus* belongs to a group of filamentous ascomycete fungi that plays an important role in the ecosystem, involved in decomposing natural organic matter [1]. This genus comprises approximately 180 species [2], of which several species are human, animal or plant pathogens and several other species play an important role in the biotechnological industry for the production of chemicals, enzymes and drugs [3]. The biotechnological importance is illustrated by the use of *A. oryzae* for production of fermented food products. It has been attributed with Generally Regarded As Safe (GRAS) status by the U.S. Food and Drug Administration (FDA) [4] and its safety has also been supported by the World Health Organization (WHO) [5]. Today, *A. oryzae* is also used as a microbial cell factory for large scale production of industrial enzymes (e.g. amylases, proteases, and other hydrolytic enzymes) and heterologous protein productions [6, 7]. Of other relevant aspergilli can be mentioned, *A. niger* that is used for organic acid production (e.g. citric acid) [8] and large scale production of industrial enzymes (e.g. glucoamylase) [9], *A. nidulans* that is an important model organism for studying gene regulation and cell biology [10], *Aspergillus sojae* that has been used throughout Asia for fermented foods (e.g. koji foods) and beverage processes [11], *Aspergillus terreus* that is used for production of the cholesterol lowering agent lovastatin [12] and for itaconic acid production [13], and *Aspergillus awamori* that is used for protein production [14]. Clearly, these different *Aspergillus* species play a key role as microbial cell factories for different industries such as the food industry, the alcohol beverage industry, and the pharmaceutical industry.

Some *Aspergillus* species are causative agents of opportunistic infections in man, e.g. *Aspergillus fumigatus* that is the most common pathogenic mould in humans and animals, which may cause diseases that range from allergies to life-threatening diseases [15]. Furthermore, *Aspergillus clavatus* and *Aspergillus fischeri* (*Neosartorya fischeri*) which are close relatives to *A. fumigatus*, are also common fungal pathogen in the US. *A. clavatus* is rarely pathogenic, although it is potently allergenic. *A. clavatus* can produce patulin [16], a substance which may be associated with disease development in humans and animals. *A. fischeri* is an extremely rare pathogen which can cause keratitis and possibly pulmonary aspergillosis in transplant patients [17]. Besides human and animal pathogens, additional *Aspergillus* species are plant pathogens (e.g. *Aspergillus aculeatus*) and toxin producers (e.g. *Aspergillus flavus, Aspergillus parasiticus*, and *Aspergillus carbonarius*). *A. flavus* and *A. parasiticus* are aflatoxin producers, which are among the most toxic natural compounds that exist [18, 19]. *A. carbonarius* is an ochratoxin producer resulting in food spoilage. Table 2.1 summarizes the impact of different *Aspergillus* species on society.

Table 2.1 The biotechnological importance by different *Aspergillus*

Aspergillus species	Impact	Biotechnological importance
A. oryzae	Positive	Fermented food and industrial enzyme
A. niger	Positive	Citric acid, glucoamylase and heterologous enzyme
A. nidulans	Positive	Gene regulation model and cell biology
A. terreus	Positive	Cholesterol lowering agent and itaconic acid
A. awamori	Positive	Protein production
A. fumigatus	Negative	Pathogenic mould in human and animal
A. clavatus	Negative	Pathogenic mould in human and animal
A. fischeri	Negative	Pathogenic mould in plant and human
A. aculeatus	Negative	Pathogenic mould in plant
A. flavus	Negative	Pathogenic mould in plant, human and animal and aflatoxin producer
A. parasiticus	Negative	Aflatoxin producer
A. carbonarius	Negative	Ochratoxin producer caused in food spoilage

2.2 *Aspergillus* as efficient microbial cell factory

2.2.1 Enzyme production and global market

Currently, it is estimated that there are more than 1000 enzymes [20] that have industrial relevance and a fraction of these are currently being produced as mono-component enzymes (http://www.report2008.novozymes.com). These enzymes play an important role in the manufacture of many different products, including detergents, foods, animal feeds and cleaning agents. Recently, according to an updated technical market research report, Enzymes for Industrial Applications (BIO030E) from BCC Research (www.bccresearch.com), the global market for industrial enzymes increased from $2.2 billion in 2006 to an estimated $2.3 billion by the end of 2007. The market is expected to increase to over $2.7 billion by 2012 based on three main application sectors: technical, food and animal feed enzymes as shown in Figure 2.1. Enzyme production is dominated by a few leading companies, with Novozymes A/S as an undisputable market leader and having an overall global market share that developed slightly positively in 2008, ending at approximately DK 16 billion with increased market sizes on technical and food enzymes while there was a slight decline in the market share of animal feed enzymes (http://www.report2008.novozymes.com). Table 2.2 summarizes some important industrial enzymes produced by *Aspergillus* species across these three sectors.

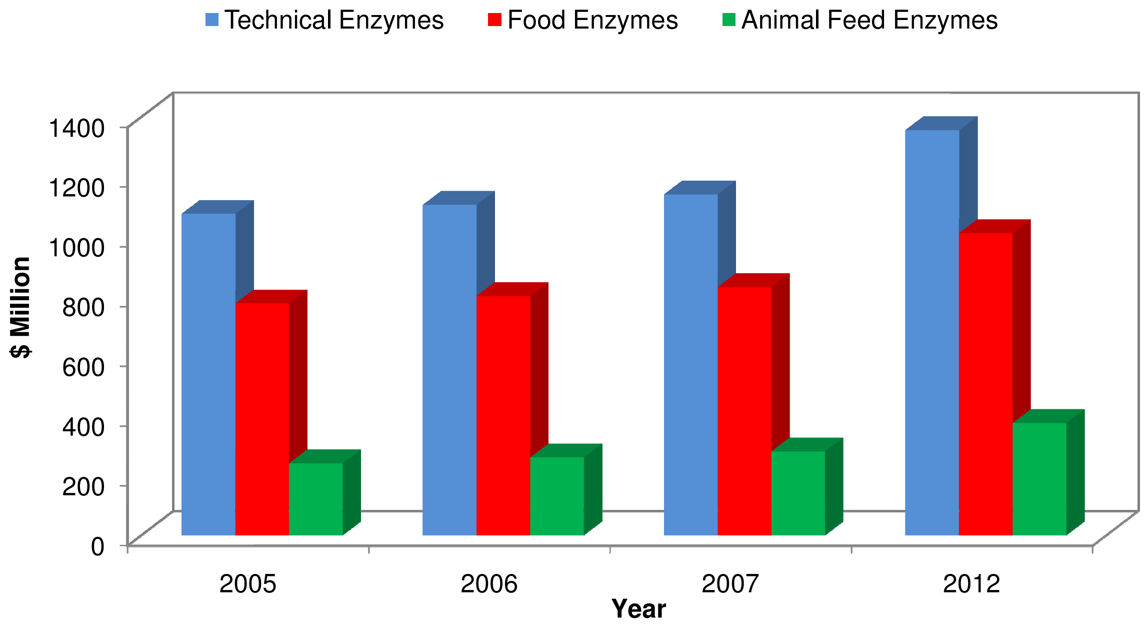


Figure 2.1 Global market of industrial enzymes based on three application sectors ($Millions)

Table 2.2 Some important industrial enzymes produced by *Aspergillus* species

Enzyme name	Application Sector			Producer
	TE	FE	AFE	
α-amylase	×	×	×	*A. oryzae, A. niger*
γ-amylase	×	×	×	*A. oryzae, A. niger*
Glucoamylase	×	×	×	*A. oryzae, A. niger*
Catalase	×	×		*A. niger*
Cellulase	×	×	×	*A. niger, A. sojae*
α-galactosidase	×	×		*A. niger*
Lactase	×	×		*A. oryzae, A. niger*
Inulinase	×	×		*A. niger*
β-glucanase	×	×	×	*A. oryzae, A. niger*
Glucose oxidase	×	×		*A. niger*
Hemicellulase	×	×		*A. oryzae, A. niger*
Invertase	×	×		*A. oryzae, A. niger*
Lipase	×	×		*A. oryzae, A. niger*
Peroxidase	×	×		*Aspergillus spp.*
Pectinase	×	×		*A. niger*
Pectin esterase	×	×		*A. niger*
Protease	×	×		*A. flavus, A. niger*
Proteinase	×	×	×	*A. oryzae*
Tannase		×		*A. oryzae, A. niger*
Xylanase	×	×	×	*A. niger*

Three application sectors: Technical Enzyme (TE), Food Enzyme (FE), Animal Feed Enzyme (AFE)

2.2.2 Utilization of different carbon sources

Aspergillus basically can utilize a wide variety of carbon sources. This advantage provides increased flexibility in the design and improves the economic feasibility of industrial fermentation process. The major source of carbon and energy for *Aspergillus* is derived from carbohydrates such as disaccharide (C_{12}), pentose (C_5), and hexose (C_6). Usually hexose, such as glucose, is the favorable carbon source due to the high efficiency in the uptake and metabolism of this sugar. Pentose such as xylose is less favorable, but still filamentous fungi can metabolize it quite efficiency. The disaccharide maltose is also a preferred carbon source in *Aspergillus*, however it has to be hydrolyzed before it can be metabolized, either intracellularly or extracellularly. This hydrolysis is catalyzed by enzyme such as maltase (α-glucosidase). Besides of the carbon sources as mentioned earlier, *Aspergillus* can also use glycerol (C_3) as a carbon source. In this study, these four different carbon sources have been used to investigate growth physiology and comparative transcriptome analysis in three *Aspergillus* species (i.e. *A. oryzae*, *A. niger* and *A. nidulans*) as discussed later in the thesis (Chapter 5).

2.2.3 Central carbon metabolism through amino acid synthesis

The central carbon metabolism primarily involves the glycolysis pathway, the pentose phosphate pathway (PPP) and the tricarboxylic acid (TCA) cycle that serve the purpose of converting sugars to Gibbs free energy and precursor metabolites that are required for biomass synthesis, including all amino acids needed for protein synthesis.

An illustration of how the 20 amino acids are derived from precursor metabolites that are intermediates in the glycolysis, the PPP and the TCA cycle is given in Figure 2.2. Amino acids derived from intermediates of the glycolysis pathway are glycine, serine and cysteine (using 3-phosphoglycerate as a precursor), alanine, valine and leucine (using pyruvate as a precursor). Amino acids formed from intermediates of the PPP are phenylalanine, tyrosine, tryptophan (using erythrose-4-phosphate and phosphoenolpyruvate from the glycolysis as precursors) and histidine (using ribose-5-phosphate as a precursor). The amino acids that are derived from intermediates of the TCA cycle are aspartate, asparagine, methionine, threonine, isoleucine (using oxaloacetate as a precursor). Furthermore, glutamate, glutamine, proline, arginine, and lysine are formed using α-ketoglutarate as a precursor. Limitation in any of these precursors will obviously influence the formation of the corresponding amino acids and hereby protein synthesis. Oxaloacetate and α-ketoglutarate are formed in the TCA cycle and are serving as building blocks for 10 amino acids. Therefore the limitation of these precursors will have a serious impact on protein synthesis.

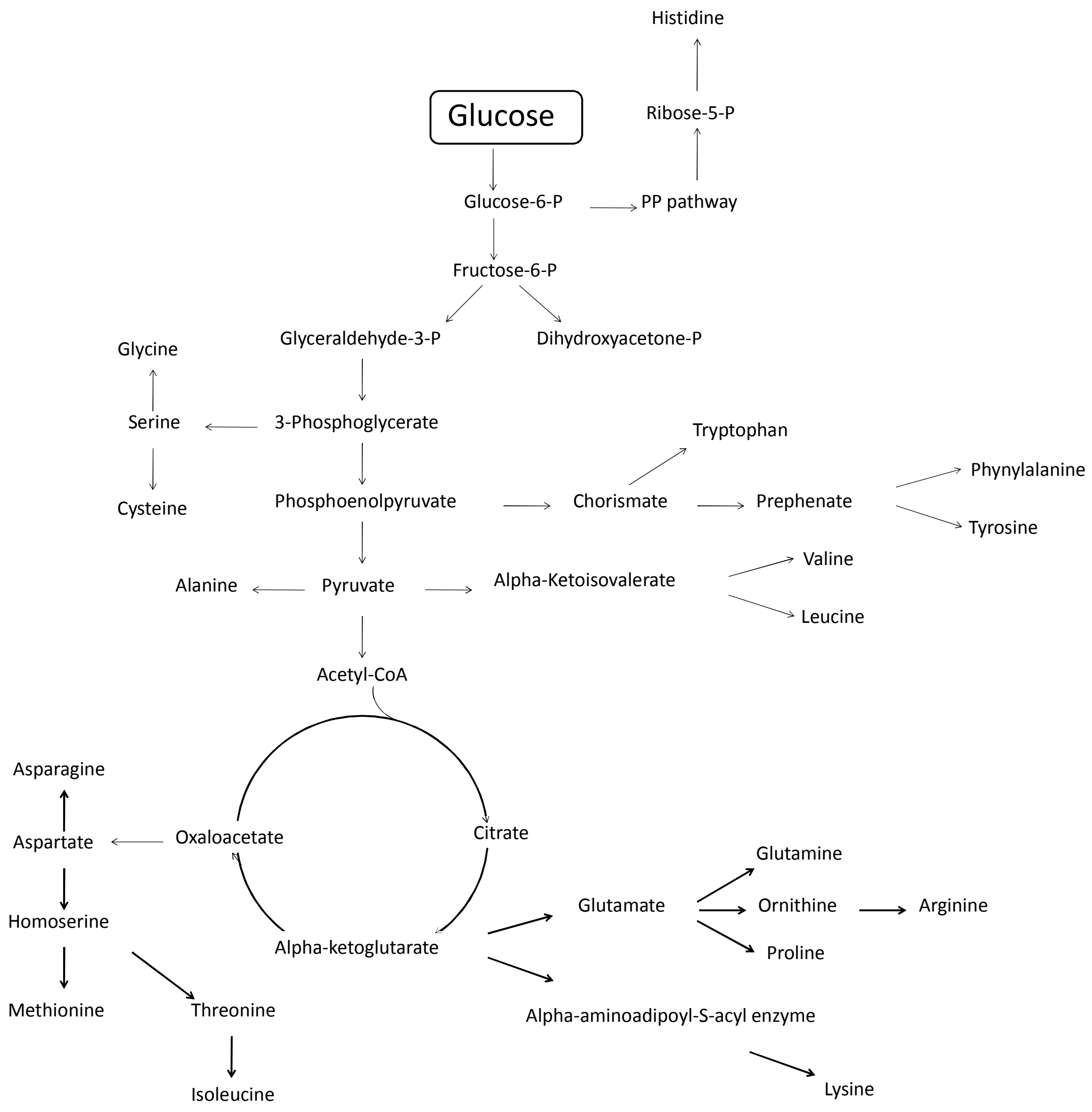


Figure 2.2 Central carbon metabolism and amino acid synthesis from using glucose as a carbon source.

2.2.4 Protein production by genetic engineering

In order to improve protein production from microbial cell factories, it is important to consider molecular mechanisms. In the last decade, one successful method for improving protein production was genetic engineering techniques. Based on these techniques, new strains of microorganisms are constructed that can produce desired products (e.g. protein, metabolite or antibiotic) in high amounts and with few by-products is the core of biotechnology. Such strains can be obtained by the use of recombinant DNA technology for targeted introduction of genetic changes, as exemplified by the construction of different *Aspergillus* strains that produce industrial enzymes. One way to increase the productivity of a given enzyme is to introduce extra copies of the gene encoding the product (a protein). An example is to increase the number of genes encoding an endogenous gene, e.g. the gene encoding for α-amylase in *A. oryzae*. In chapter 6 is shown a case study of the growth physiology and gene expression study of an *A. oryzae* recombinant strain that contains additional copies of the homologous α-amylase gene.

Chapter 3

3. SYStems biology, BIoinformatics and OMICS analysis (SYSBIOMICS) of Aspergilli

This chapter describes SYSBIOMICS approach applied throughout this study (See Figure 3.1). The chapter gives an overview of SYSBIOMICS in aspergilli. In the following, three chapters are given three case studies of SYSBIOMICS. The details presented in this chapter have formed the basis of a recently published book chapter [21]. The general idea of combined omics analysis (i.e. genomics, transcriptomics, proteomics and metabolomics), bioinformatics and systems biology (i.e. genome-scale metabolic models) of aspergilli is presented. As mentioned this study involved several applications of the SYSBIOMICS approach, and these applications are divided into three case studies, which are described in the following three chapters (chapter 4, 5 and 6). The first case study (chapter 4) shows strategies to perform improved genome annotation (e.g. gene prediction and functional assignment). It also illustrates how to use a systems biology approach to reconstruct a metabolic network and how to apply metabolic engineering tool (i.e. Flux Balance Analysis) to perform metabolic modeling. The second case study (chapter 5) shows how SYSBIOMICS can be used to analyze data from growth of aspergilli on four different carbon sources, and hereby gain information on how the different carbon sources effect the cellular metabolism and based on this reveals new insight into transcriptional regulation in *Aspergillus* species. In this part, the focus is on *A. oryzae* but with comparative transcriptomics analysis with two other important species (i.e. *A. niger* and *A. nidulans*). The last case study (chapter 6) demonstrates the use of SYSBIOMICS for analyzing protein production by *A. oryzae*.

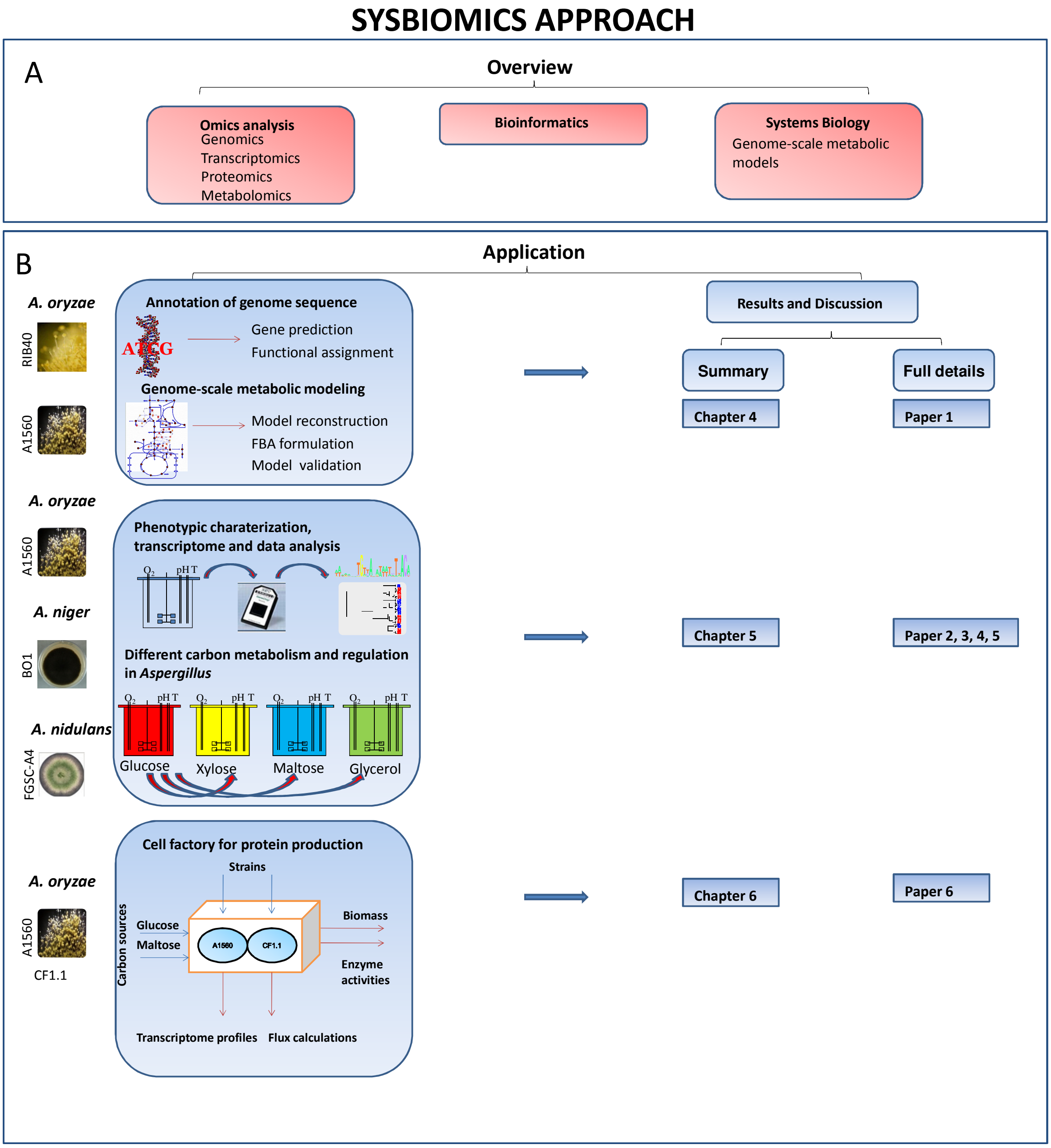


Figure 3.1 A flow diagram showing an overview (A) and application (B) of SYSBIOMICS approach applied throughout this study and how different elements are connected to the different chapters and to the corresponding papers.

3.1 Omics analysis

The availability of multi-level omics data of aspergilli has allowed for systems studies of these fungi, and in recent years there have been several studies on the use of omics technologies for studies of different *Aspergillus* species. In the following, we present a brief history of high-throughput omics technologies, namely genomics, transcriptomics, proteomics, and metabolomics as well as review studies where these technologies have been applied to *Aspergillus*.

3.1.1 Genomics

In the last decade of the 20th century there was a paradigm shift in the biological sciences. The focus of research shifted away from the study of individual genes to genomes. This scientific shift was driven by the availability of complete genomic sequences. Today, deciphering the complete sequence of smaller genomes such as bacteria is fairly straight forward. However, determining and structuring the genome sequences of more complex microorganisms with large genome sizes and complexities are still a considerable challenge [22], and genomics research on fungi has therefore obtained much focus [23]. The first sequenced fungus was the baker's yeast, *Saccharomyces cerevisiae* [24, 25]. Shortly after the *S. cerevisiae* genome was released, fungal workshops were held that discussed initiating genome projects on filamentous fungi [23]. Herein, the details for genome sequencing projects of several different *Aspergillus* species are discussed.

Genome sequencing projects

The history of *Aspergillus* sequencing projects started in 1998, where Cereon Genomics (Monsanto) sequenced the genome of *A. nidulans* [23] strain FGSC-A4 using the whole-genome shotgun approach with a three fold (3x) coverage. The sequence was, however, firstly released for public use several years later. A few years afterward, the *A. nidulans* genome was further sequenced with a thirteen fold (13x) coverage and here the Cereon data was incorporated. This complete genomic sequence was released in 2003 by the Whitehead Institute/MIT Center for Genome Research, now the Broad Institute. The size of the *A. nidulans* genome is approximately 30.1 Megabases (Mb), and it is organized in 8 chromosomes [26]. Parallel to this work, a group of scientists decided to accelerate the developments in fungal genomics. The Fungal Genome Initiative (FGI) agenda was established with the overall aim of sequencing several key fungal species [27]. The first White paper was submitted to the National Human Genome Research Institute (NHGRI), and it was here proposed to sequence many fungal species. Three *Aspergillus* species, namely *A. nidulans* FGSC-A4, *A. flavus* NRRL 3357, and *A. terreus* NIH 2624 were in the initial sequencing list [28], but several other genome sequencing projects in *Aspergillus* species have been initiated in recent years. For instances, the sequencing project of the genome of *A. fumigatus* strain Af293 was begun by both the Wellcome Trust Sanger Institute and TIGR, and 29.4 Mb of sequence data distributed on 8 chromosomes was obtained [29]. Three different strains of *A. niger* were completely sequenced by three institutes, namely Gene Alliance/DSM, Integrated Genomics/Genencor and the Joint Genome Institute (JGI). First, the availability of a completed genome sequence for *A. niger* strain CBS 513.88 was announced in a press release by Gene Alliance/DSM and published some years later. The genome size of *A. niger* strain CBS 513.88 is 33.9 Mb and consists of 8 chromosomes [30]. Thereafter the genome of *A. niger* strain ATCC 9029 was sequenced [31] by Integrated Genomics, however, only with a 4x coverage and approximately 32 Mb of genome size. In the mean-time, the Joint Genome Institute (JGI), with funding support from the U.S. Department of Energy (DOE), undertook to complete the genome sequence of *A. niger* strain ATCC 1015 with 34.9 Mb and a final 8.9x coverage, and these data were publicly released [31]. Sequencing of the genome of *A. oryzae* strain RIB 40 was initiated at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. The project was extended from 1998 until 2001 and was carried out by the collaboration with the Japanese National Institute of Technology and Evaluation (NITE) and other members of the *A. oryzae* Genome Analysis Consortium [32]. The complete genome sequence was publicly released in 2005 with a 9x coverage. The total *A. oryzae* genome size is 37.2 Mb, and it consists of 8 chromosomes [33]. *A. terreus* clinical strain NIH 2624, which has been reported to be a potential human pathogen causing aspergillosis, was chosen by the Steering Committee of the Fungal Genome Initiatives (FGI) for genome sequencing. Currently, the completed genome sequence of the *A. terreus* clinical strain has been finished. It was publicly released in 2005 with a 11x coverage by the Broad Institute. The genome size of *A. terreus* strain NIH 2624 has been estimated to be about 29.3 Mb distributed over 8 chromosomes [34]. In addition to sequencing of the *A. terreus* clinical strain, the genome sequence of the *A. terreus* industrial strain ATCC 20542 is in the process of being sequenced by the Microbia company [35]. For *A. flavus* strain NRRL 3357, a sequencing project was carried out through a collaboration project between TIGR, University of Oklahoma and United States Department Of Agriculture/Agricultural Research Service (USDA/ARS) [36]. The USDA/NRI Microbial Genome Sequencing Project provided funding for whole genome sequencing of *A. flavus*. The size of *A. flavus*'s genome is approximately 36.3 Mb with a 10x coverage [36, 37]. Last but not least, two additional genome projects of *A. clavatus* strain NRRL 1 and *A. fischeri* strain NRRL 181 have been started [38] with the goals of better elucidating the *A. fumigatus* genome and improving its genome annotations. These projects are being sequenced by TIGR with funding provided by the National Institute of Allergy and Infectious Diseases (NIAID) [39]. An estimated genome size of *A. clavatus* is 27.9 Mb and of *A. fischeri* is 32.5 Mb. A brief summary of available genomic information for the *Aspergillus* species along with some status of their genomes is presented in Table 3.1. Several databases are now available for *Aspergillus* genomes, which provide access to the sequence data and include bioinformatics tools for biological sequence analysis. For examples, one is the CADRE (Central Aspergillus Data REpository) which is a major public resource for *Aspergillus* species hosting genome data. In addition, various genome projects of *Aspergillus* can be found in the Fungal Genetic Stock Center (FGSC). Table 3.2 presents *Aspergillus* genome sequence websites that are available for public access.

Table 3.1 Summary of information on *Aspergillus* genomes

Aspergillus species	Strain	Genome size (Mb)	Institution/Company	Status
A. nidulans	FGSC-A4	30.1	Broad Institute	Complete (13x)
		30.1	Cereon Genomics (Monsanto)	Complete (3x)
A. niger	CBS 513.88	33.9	Gene Alliance/DSM	Complete (7.5x)
	ATCC 9029	32	Integrated Genomics/Genencor	Complete (4x)
	ATCC 1015	34.9	DOE Joint Genome Institute	Complete (8.9x)
A. fumigatus	Af293	29.4	Sanger Institute, TIGR	Complete (10.5x)
	A1163	29.2	TIGR, Celera Genomics, Merck & Co., USA	Complete
A. oryzae	RIB 40	37.2	Japanese National Institute of Technology and Evaluation	Complete (9x)
A. terreus	ATCC 20542	N/A*	Microbia	Incomplete
	NIH 2624	29.3	Broad Institute	Complete (11x)
A. flavus	NRRL 3357	36.3	TIGR, University of Oklahoma, USDA/ARS	Complete (10x)
A. clavatus	NRRL 1	27.9	TIGR	Complete (11.4x)
A. fischeri	NRRL 181	32.5	TIGR	Complete
A. parasiticus		N/A*	University of Oklahoma	Incomplete
A. aculeatus	ATCC 16872	N/A*	DOE Joint Genome Institute	Incomplete
A. carbonarius	IMI 388653	N/A*	DOE Joint Genome Institute	Incomplete

Data in the list was gathered from the reviews [40, 41] and the Broad Institute database at (http://www.broad.mit.edu/annotation/genome/ aspergillus_group/MultiHome.html)
*Not Available

Table 3.2 Public genome websites for different *Aspergillus* species

Species	**Public genome websites**
A. oryzae RIB 40	http://www.aist.go.jp/RIODB/ffdb/welcome.html
A. nidulans FGSC-A4	http://www.broad.mit.edu/annotation/fungi/aspergillus/
A. fumigatus Af293	http://www.tigr.org/tdb/e2k1/afu1/ http://www.sanger.ac.uk/Projects/A_fumigatus/
A. niger CBS 513.88	http://www.dsm.com/en_US/html/dfs/genomics_aniger.htm
A. niger ATCC 1015	http://genome.jgi-psf.org/Aspni1/Aspni1.home.html
A. clavatus NRRL 1	http://msc.tigr.org/aspergillus/aspergillus_clavatus_nrrl_1/index.shtml
A. fischeri NRRL 181	http://msc.tigr.org/aspergillus/neosartorya_fischeri_nrrl_181/index.shtml
A. terreus NIH 2624	http://www.broad.mit.edu/annotation/fungi/aspergillus/
A. terreus ATCC 20542	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=AABT00000000
A. flavus NRRL 3557	http://www.aspergillusflavus.org/genomics/
Other *Aspergillus* sites	http://www.aspergillus.man.ac.uk/ http://www.fgsc.net/aspergenome.htm http://www.aspergillus-genomics.org/ http://www.genome.ou.edu/fungal.html http://www.cadre.man.ac.uk/Hosts http://www.cs.man.ac.uk/~cornell/eFungi/database.html

3.1.2 Transcriptomics

The study of transcriptomics examines the expression level of mRNA by using high-throughput techniques based on DNA microarray. Over the last decade, DNA microarray has become an increasingly important tool to analyze the expression level at the genome level as it allows for identification of which genes are active and to what extent. The principle of DNA microarray is based on the hybridization of nucleic acids extracted from samples to high-density arrays of immobilized nucleotide sequences, each corresponding to a specific gene or EST from the organism in question. The biochemical potential of DNA microarray is based on the specificity and affinity of complementary base pairing [42]. The precise high-density positioning of the probes allows them to act as molecular detectors [43]. In applications, the transcriptional analysis by DNA microarray is often used to study gene expression on a genomic scale, to reveal regulatory patterns and systematic features [42] or to explore the function of genes by determining their patterns of expression [44]. An advantage of the DNA microarray technology is that it allows parallel and automated expression monitoring, and it is therefore suitable for functional genomics [45]. By comparing DNA microarray to more traditional methods such as Northern Blotting, a larger number of mRNA can be measured under different conditions in a quantitative way, while using very small amounts of biological sample [46]. The array technology is flexible, quite cheap and fast [42].

DNA microarray was introduced several years ago, and the first fungal microarray studies was reported in *S. cerevisiae* [47, 48]. Afterward DNA microarray has been widely used for genome-wide transcription for different microorganisms, including *Aspergillus* species. At present DNA microarray studies have been used for global gene expression in six different *Aspergillus* species encompassing a broad variety of research areas such as metabolism, pathogenesis, and industrial applications [49]. A list of some updated DNA microarray studies in different array format is listed in Table 3.3. Nowadays, the field of transcriptomics for aspergilli is being progressed in many research areas. Comparing expression profiles of different species of *Aspergillus* may provide insight into how this diverse group of fungi has evolved. With the possibility for comparative transcriptomics between *Aspergillus* species, it might dramatically increase the potential for global gene expression studies in aspergilli. Furthermore, it can raise the number of studies available for comparative functional genomics in filamentous fungi.

Table 3.3 List of some DNA microarray studies in aspergilli

***Aspergillus* spp.**	**DNA microarray studies**	**Array format**
*A. oryzae**	Expression analysis of energy catabolism and hydrolytic enzymes	cDNA
*A. nidulans**	Identification of biosynthetic terrequinone A gene cluster	Nimblegen
*A. nidulans**	Expression analysis of previously characterized metabolic genes	cDNA
*A. nidulans**	Identification of aflatoxin biosynthesis genes	cDNA
*A. nidulans**	Identification of unfolded protein response genes	cDNA
*A. parasiticus**	Identification of aflatoxin biosynthesis genes	cDNA
*A. flavus**	Identification of aflatoxin biosynthesis genes	cDNA
*A. fumigatus**	Identification of voriconazole adaptation genes	cDNA
*A. fumigatus**	Identification of temperature-regulated genes	cDNA
*A. niger**	UPR-independent dithiothreitol stress-induced genes	Affymetrix
A. nidulans	Identification of regulatory genes in different carbon sources	Nimblegen

*Data were referred to the review [49]

3.1.3 Proteomics

Proteomics aims at identifying and quantifying every protein at the same time in the cell. Proteome analysis is important for increasing our biological understanding of aspergilli and identifying enzyme with biotechnological potential as well as for new antifungal drug target identification [50]. The aim of proteomics is to obtain quantitative data of differential protein expression in response to environmental alteration. Several researches have shown that mRNA levels do not correlate well with protein expression levels, and the study of the whole dynamic proteome has therefore gained more significance [51, 52]. To date, the proteomics studies have developed a large diversity in terms of methods used, such as Two-Dimensional PolyAcrylamide Gel Electrophoresis combined with mass spectrometry [53], protein arrays [54], Isotope-Coded Affinity Tagging [55] and techniques for investigation of protein interactions such as a yeast two-hybrid system [56]. The availability of genome sequences together with proteomics technologies are beginning to reveal the complex and dynamic nature of *Aspergillus* species.

3.1.4 Metabolomics

Metabolomics is the study of all, or a large group of, intracellular metabolites in a biological organism with a single method and metabolite profiling gives a snapshot of the physiology of the cell. Nowadays, many high-throughput methods for quantitative metabolome analysis are used such as Nuclear Magnetic Resonance, Gas Chromatography Mass Spectrometry, Gas Chromatography Time-Off Flight, and Liquid Chromatography-Mass Spectrometry [57]. However, only a few publications about aspergilli have appeared.

3.2 Bioinformatics

The term “bioinformatics” was introduced only a few years ago, and it represents a growing area of science that uses computational approaches together with modern molecular biology techniques to answer biological questions [58]. In answering these queries, bioinformatics tools are utilized to analyze biological data sets in a rigorous fashion. Basically, the meaning of bioinformatics encompasses the generation, collection, storage in digital form, and efficient exploitation of data and information from high-throughput technologies such as genomics, transcriptomics, proteomics, metabolomics, statistical data from various experimental trials, and also scientific literature. At present the major uses of bioinformatics are in the area of completed genome analysis, genome annotation and comparative genomics [58]. With the availability of complete genome sequences of different *Aspergillus* species, the use of sequence information to annotate genes and to identify gene products has been extensively studied. In principle, the goal of annotation is to deduce from the genome sequence to its corresponding biological features, exploring and describing all intermediate levels such as molecular and cellular processes. Annotation is thus a complex process that requires, besides of analysis of the raw sequencing data, the integration of much additional information [22]. For examples, information is gathered from using different bioinformatics tools, extracting data from generic or specific databases, collecting biological knowledge accumulated in the literature over the years and obtaining data from genome-wide experiments. Thus, several different types of information needs to be integrated and it is therefore important in the annotation processes to use suitable bioinformatics tools [22]. Because of the complexity of the genome structure of different aspergilli, a multilayer annotation analysis pipeline as illustrated in Figure 3.2 is often used. This shows the individual step and workflow of individual module involved in the annotation of an *Aspergillus* genome. The major players in this workflow are bioinformatics algorithms (i.e. nucleotide level, protein level and bioprocess level annotation) and integration of ome-data. Several gene finding programs and functional assignment programs have been developed for use in *Aspergillus*. They differ in the underlying parameters used for gene and function detection. Combinatorial approaches are often used for performance accuracy of gene finding and function prediction. In Figure 3.2, an overall annotation process is applied to predict the genes and to assign protein functions. For each species, the annotation algorithm depends on which method is used for gene prediction (e.g. using Expressed Sequence Tag (EST) library or comparative genomics) and functional assignment (e.g. homology detection, domain detection, or etc). Also, which software or database is selected for suitable species.

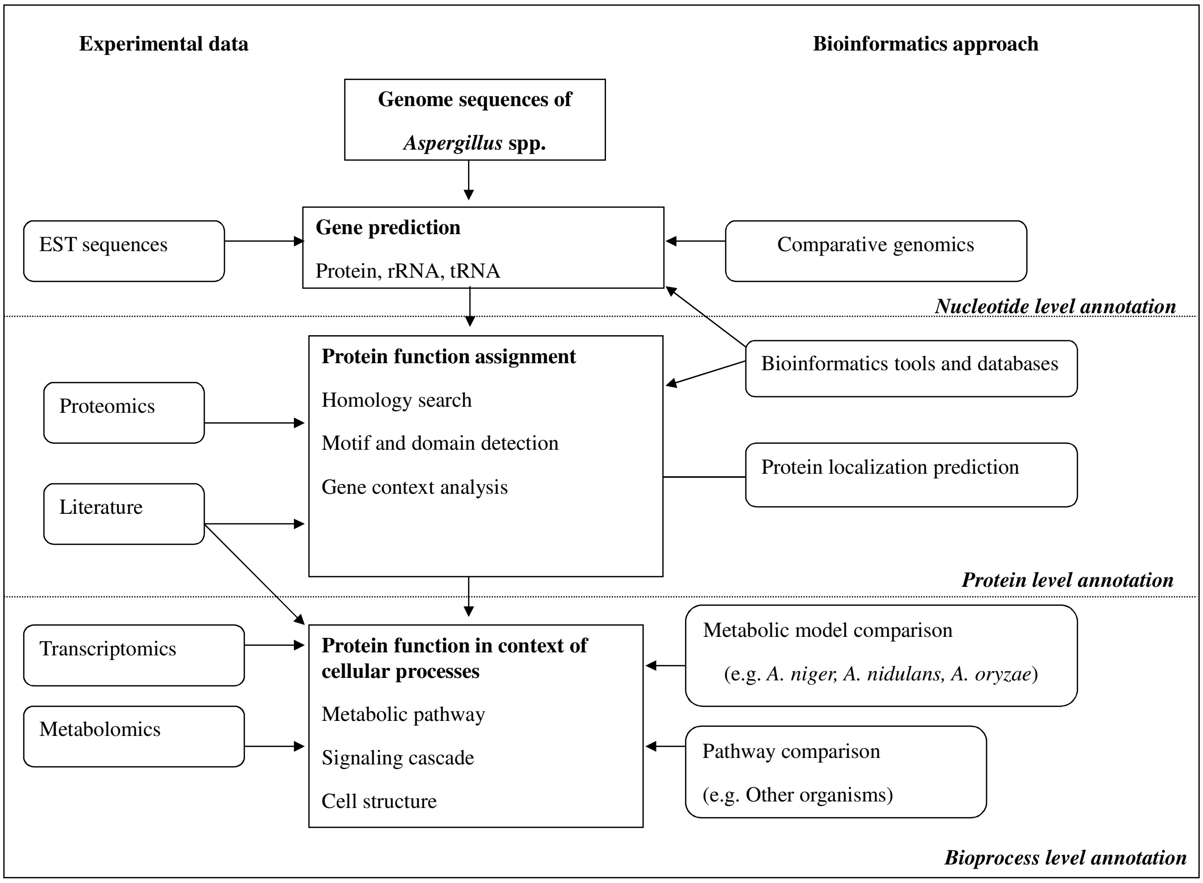

Experimental data
Bioinformatics approach
Genome sequences of *Aspergillus* spp.
EST sequences
Gene prediction
Protein, rRNA, tRNA
Comparative genomics
Nucleotide level annotation
Protein function assignment
Homology search
Motif and domain detection
Gene context analysis
Bioinformatics tools and databases
Proteomics
Literature
Protein localization prediction
Protein level annotation
Transcriptomics
Metabolomics
Protein function in context of cellular processes
Metabolic pathway
Signaling cascade
Cell structure
Metabolic model comparison
(e.g. *A. niger, A. nidulans, A. oryzae*)
Pathway comparison
(e.g. Other organisms)
Bioprocess level annotation

Figure 3.2 Schematic analysis of the pipeline for gene prediction and annotation in aspergilli.

3.3 Systems biology

There are many definitions of systems biology, but most of these contain elements such as mathematical modeling, global analysis (or ome analysis), mapping of interactions between cellular components, and quantification of dynamic responses in living cells. In most cases the objective of systems biology is to obtain a quantitative description of the biological system under study, and this quantitative description may be in the form of a mathematical model. In some cases, the model may be the final result of the study, i.e. the model captures key features of the biological system and can hence be used to predict the behaviour of the system at conditions different from those used to derive the model. In other cases, mathematical modeling rather serves as a tool to extract information of the biological system, i.e. to enrich the information content in the data. There is not necessarily a conflict between the two, and generally, mathematical modeling goes hand in hand with experimental work [59]. Systems biology do not necessarily involve the use of omics data, but still the use of global studies are often particularly rewarding for biological systems where there are many unknown factors and components, as is the case for aspergilli. However, the use of omics studies without thorough analysis, where the data are put into the context of the whole system, is often not providing much information about how the system operates. Here genome-scale metabolic models offer an attractive scaffold for analysis of omics data, as well as discussed here they offer good opportunities for simulation of cellular behaviour. In the following, we describe introduction of genome-scale metabolic models.

Genome-scale metabolic models

Briefly, there are several steps for building a genome-scale metabolic model. Normally, the modeling process starts with reconstruction of the metabolic network from the collected information, such as annotated genomic data, biochemistry textbooks, literature and bioinformatics databases. Thereafter follows the development of a stoichiometric metabolic model, where the stoichiometry of the individual reactions is checked and the reactions are connected to form a metabolic model. This metabolic model can then be used for simulations using Flux Balance Analysis, and hereby it is possible to gain quantitative data for cellular phenotypic behavior and these can be compared with experimental data. In this process, the model may have to be refined such that its predictions are in closer agreement with experimental observations. Furthermore, data from different levels of cellular processes may be incorporated into the genome-scale metabolic model for improvement of the predictions [60].

To date, genome-scale metabolic models have been developed for several microorganisms and used for many different applications. The first fungal metabolic model, namely for the yeast *S. cerevisiae*, was reconstructed by Forster *et al.* [61] and in 2008 this model was updated with additions of many new reactions, in particular reactions involved in lipid metabolism [62]. In these models, the metabolic reactions were compartmentalized between the cytosol and the mitochondria, and transport steps between the compartments and the environment were included. The Forster *et al*. model was the first comprehensive reconstructed metabolic network for an eukaryotic organism, and it has been used as the basis for *in silico* analysis of phenotypic functions. In recent years, comprehensive knowledge regarding *S. cerevisiae* based modeling has accumulated, and today this yeast model has been used for metabolic engineering for biotechnological application [63]. In addition to yeast, metabolic models for *Aspergillus* have been developed. For instance, a metabolic model of the central carbon metabolism of *A. niger* was reconstructed [64, 65]. Additionally, recently several genome-scale models for whole metabolism of *Aspergillus* species have been reconstructed: one for *A. nidulans* [66] and one for *A. niger* [67]. There are several applications of these metabolic models, e.g. the *A. niger* model [64] was applied to search for strategies to improve succinic acid production. Another example is the application of a developed genome-scale model of *A. niger* for simulating the operation of the oxidative pathways during production of citrate at high yields [67]. The *A. nidulans* model [66] was used to integrate information on transcriptome data in order to identify subnetworks structure, and this showed that subnetworks structure can point to coordinate regulation of genes that are involved in many different parts of the metabolism [68].

Chapter 4

4. Improved annotation through genome-scale metabolic model of *A. oryzae*

This chapter describes the first case study that shows how SYSBIOMICS approach can be used to to link genome sequence information to a functional metabolic model, and how it is possible to improve genome annotation through development of a genome-scale metabolic model for *A. oryzae*. Furthermore, we discuss the results of simulation and validation of the *A. oryzae* metabolic model at various growth conditions (see full details in Paper 1).

4.1 SYSBIOMICS-based improved genome annotation and metabolic model reconstruction

To improve the annotation through reconstructing a metabolic model of *A. oryzae*, we followed four main steps as illustrated in Figure 4.1, i.e. the SYSBIOMICS paradigm was used to link from genome sequence (genotypic level), to gene, to protein function, to metabolic reaction, and further to metabolic model (phenotypic level).

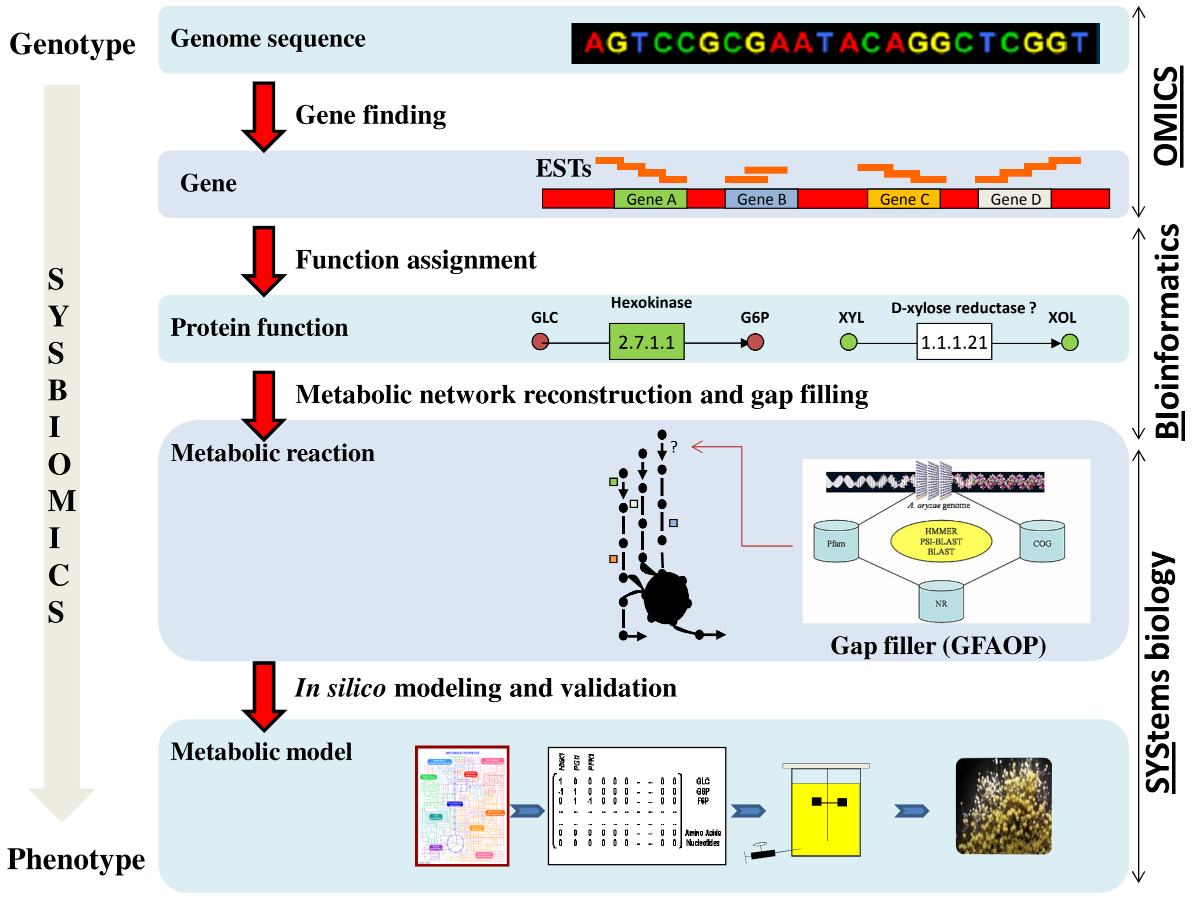


Figure 4.1 Genotype-Phenotype relationship based SYSBIOMICS paradigm.

4.1.1 Gene finding

First improved gene finding was carried out based on our assembled library of 9,038 EST sequences of *A. oryzae* (GenBank accession number: "EY424375–EY433412") together with public EST data from *A. flavus* [69]. *A. oryzae*'s EST data and *A. flavus*'s EST data were compared to the genes previously identified [33] in the genome of *A. oryzae* strain RIB 40 by BLASTN [70]. The purpose of this comparison was to validate genes that were already annotated and to discover new genes that had not been annotated from previous work presented by Machida *et al.* [33].

According to our strategy implemented for gene finding, the 9,038 EST sequences were searched against the 12,074 previously identified genes obtained from Machida *et al.* [33] in the sequenced genome using various search parameters to create lists of predicted genes with different match stringencies. Using the criteria, many dissimilar sequences between the EST sequences and previously identified gene sequences of *A. oryzae* [33] were found. This suggests the presence of many newly predicted genes. Interestingly, approximately 12% (1,046 out of the 9,038 EST sequences) were categorized as newly predicted genes in the genome. Many homolog sequences were also found strongly validating previously identified genes [33], with approximately 75% of the total EST sequences (6,773 out of the 9,038 EST sequences) matching earlier identified genes. To confirm that all the EST sequences do existed in the *A. oryzae* genome, the 9,038 EST sequences were searched against the complete genome, and the results showed that only 20 EST sequences could not be found to be present in the genome. Therefore, this suggests that the assembled EST data of *A. oryzae* had very high quality and showed an excellent success rate for gene discovery and validation, even though approximately 13% (1,219 out of the 9,038 EST sequences) could not be used to predict genes, because 6% (582 out of the 9,038 EST sequences) were too short and about 7% (637 out of the 9,038 EST sequences) were too weakly validated in the original gene list using a conservative cut-off. In another attempt to predict new genes in *A. oryzae* genome, *A. flavus* EST data stored in the TIGR public database [69] were also used because *A. flavus* and *A. oryzae* are very closely related. However, using these *A. flavus* EST sequences to search against the genes in our new gene list for the *A. oryzae* genome, no new genes were predicted. Based on all the results of the gene finding a total of 13,120 protein-encoding genes were identified in the *A. oryzae* genome. This total number of genes deviates from the 12,074 previously annotated genes by Machida *et al.* and 1,046 newly predicted genes were identified from our assembled EST library.

4.1.2 Function assignment

In order to assign protein functions to the 13,120 predicted genes, sequence alignment analysis based homology searching was performed. The alignment was done through pairwise comparison of protein sequences by BLASTP [70] between *A. oryzae* RIB40 (version 1) [71] and other related fungi (i.e. *A. nidulans* FGSC-A4 (version 3) [72], *A. fumigatus* Af293 (version 1) [73], *S. cerevisiae* S288C [74, 75]. For newly predicted genes, sequence alignment was done to assign putative function by BLASTX [70] against the non-reduntdant protein database [76] and protall_e database (Novozymes's database). The annotation process used here (see Methods in Paper 1) resulted in the improved annotated data shown in Table 4.1 where the data are compared with values in the *A. oryzae* genome database by Machida *et al.* [33]. The results show that the number of improved annotated genes is 13,120 which are higher than the number of genes in the database [33]. In total, the annotated genome of *A. oryzae* contains 7,258 putative protein functions of which 3,894 proteins have metabolic functions. Even though the genome still contains 5,862 hypothetical proteins this is less than the 6,683 hypothetical proteins currently reported in the database [33], and the work therefore resulted in a substantial improvement of the genome annotation. The enhanced annotated data were mapped on the *A. oryzae* genome by using the Perl Scalable Vector Graphics (SVG) Module V2.33. Figure 4.2 shows an example of gene and EST mapping on the contig of AP007151 which is a part of chromosome 1 of the *A. oryzae* genome.

Table 4.1 Statistical values of our improved genome annotation compared with previous publication[*]

Characteristics of annotation	**This study**	**Machida *et al.*[*]**
Protein-encoding genes	13,120	12,074
Putative protein functions	7,258	5,391
- Metabolic genes	3,894	3,178
- Other functional groups of protein-encoding genes	3,364	2,213
Hypothetical protein encoding genes	5,862	6,683

[*]Machida *et al.* [33]

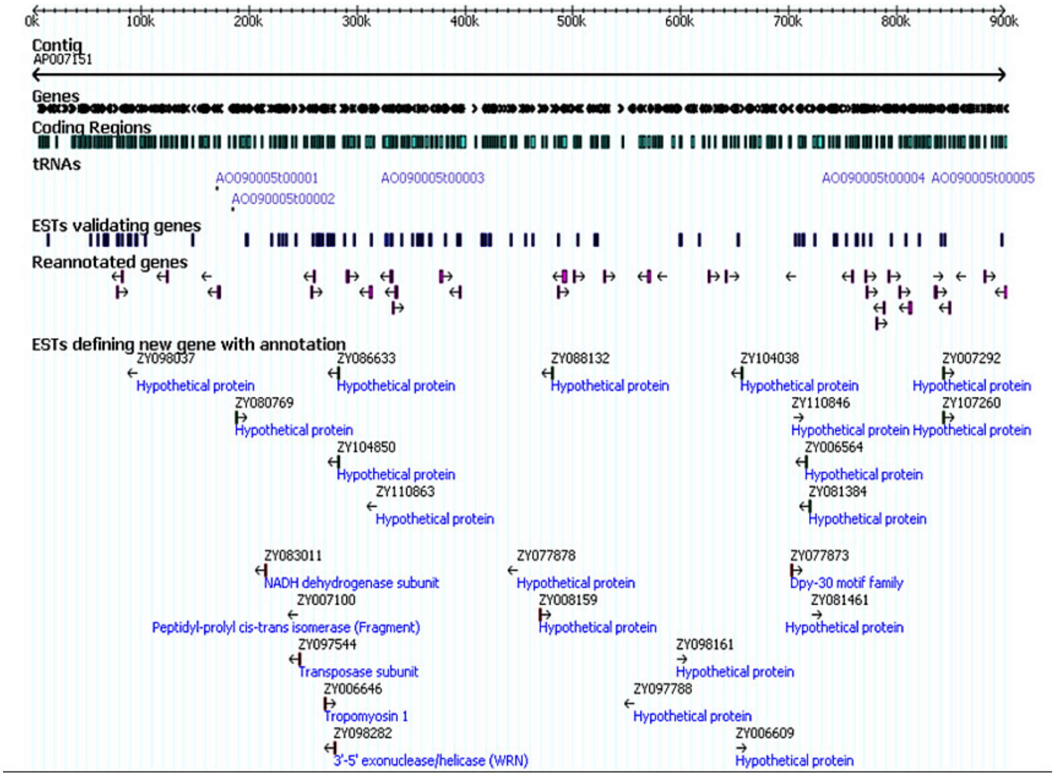


Figure 4.2 Example of gene and EST mapping on the *A. oryzae* genome (In contig of AP007151).

4.1.3 Metabolic network reconstruction and gap filling

After finishing the improved annotation process, a metabolic network for *A. oryzae* was reconstructed. The metabolic network reconstruction aimed at representing the whole metabolism of *A. oryzae*, which consists of primary metabolism of carbohydrates, amino acids, nucleotides, lipids, cofactors and energy, as well as of secondary metabolism. Combination of different types of information was essential to carry out a solid reconstruction. Information was collected from the improved annotated data of *A. oryzae*, an initial metabolic reaction list from other available metabolic models (e.g. *S. cerevisiae* [61], *A. nidulans* [68], and *A. niger* [64, 67]), biochemical pathways (e.g. KEGG database [77] and BioCyc database [78]), publications on specific enzymes, online protein databases (e.g. Swiss-Prot database [79]) and also literature. In addition, physiological evidence for the presence of a reaction or a pathway in *A. oryzae* was used to add reactions, e.g. when there was information of presence of a specific enzyme activity or presence of a pathway involved in consumption of a given substrate or formation of a given metabolic product, then the underlying reaction was added to the model, even if there was no annotated gene supporting the presence of the reaction. In the processes of stoichiometry for cofactors as well as the information on reversibility or irreversibility for each reaction, these were considered and added as information into the reconstructed network. Different cellular compartments were considered and consequently biochemical reactions were distributed into four different compartments: the extracellular space, the cytosol, the mitochondria, and the peroxisome. The localization of each biochemical reaction was analyzed according to enzyme localization, which was performed by applying protein localization predictors. Herein, pTARGET [80] and CELLO [81] were selected to predict sub-cellular protein localization because they contain databases of known eukaryotic protein localizations. If there no information on localization of a biochemical reaction or its corresponding enzyme could not be found, then by default this reaction was considered to occur in the cytosol. In addition, the reconstructed metabolic network included transport steps between the different intracellular compartments and between the cell and the environment.

The reconstructed metabolic network contained many gaps. In order to identify genes encoding enzymes with specific functions and hereby reduce the number of gaps in the metabolic network, an integrated bioinformatics tool was developed and it was used to identify these missing enzymes. This tool called "Gap Filler for *Aspergillus oryzae* Pathway (GFAOP)" was developed in- house by combining different bioinformatics tools (i.e. BLAST [70], HMMER [82], and PSI-BLAST [83]) and databases (i.e. *A. oryzae* genome [33], Pfam [84], COG [85], and NR [76]). An overview of GFAOP is shown in Figure 4.3 with an example and the implementation is illustrated in Paper 1. As a result of using GFAOP, the missing enzyme of D-xylose reductase (Figure 4.3) was entered into the pathway. Our method resulted in an improved annotation of the genome using the context of the metabolic network. An iterative process was done for filling all the gaps in the whole metabolic network and ultimately, 210 gaps in the metabolic network were closed using GFAOP.

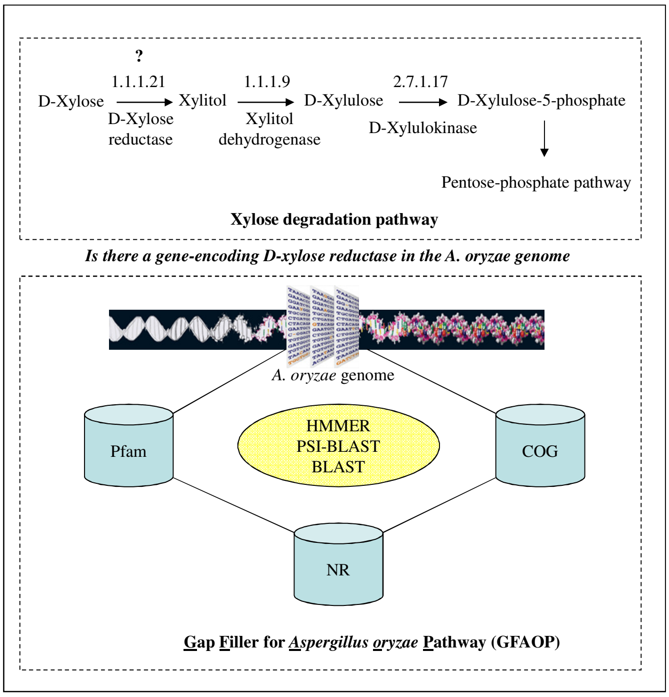


Figure 4.3 Filling gap by integrated bioinformatics approach

The improved annotation process resulted in a final reconstructed metabolic network that contains 1,314 genes, 729 enzymes, 1,846 (1,053 unique) biochemical reactions and 1,073 metabolites. The large number of isoenzymes (indicated by the difference between the total number of biochemical reactions and unique biochemical reactions) points to a very high degree of flexibility in the metabolic network of *A. oryzae*. The 1,053 unique biochemical reactions are distributed into 832 cytosolic, 172 mitochondrial, 19 peroxisomal, and 30 extracellular reactions. There are 281 (161 unique) reactions that function as transport processes, and of these 173 (53 unique) are included on the basis of gene assignments whereas there are no annotated genes for 108 of the transport reactions. All the genes and functions involved in metabolism were inspected manually.

4.1.4 *In silico* modeling and validation

After the metabolic network was reconstructed, the network was transformed into a mathematical framework to allow for Flux Balance Analysis (FBA). This approach is based on conservation of mass at steady-state conditions. This transformation requires information about the stoichiometry of the metabolic reactions, metabolic demands and a few specific parameters. Using the FBA model growth can be simulated using linear programming by introducing an appropriate objective function, e.g. one reaction is selected as an objective function that is to be maximized or minimized. For physiologically meaningful results, the objective functions must be defined as the ability to produce the required components of cellular biomass for a specified uptake rate of a selected carbon source. By maximizing the flux towards biomass formation, a flux is obtained for each reaction in the metabolic network. In this process it is important to validate the model. In this study, the model was validated by our experimental data simulating the rate of biomass formation on different carbon sources in batch experiments and also it was validated by literature data [86, 87] simulating biomass yield in chemostat experiments. Here the uptake rate of the carbon source was given as input to the simulations. Different carbon sources namely glucose (C_6), maltose (C_{12}), glycerol (C_3) and xylose (C_5), were selected as they result in widely different physiological responses. The strain used for generating these data was *A. oryzae* wild type strain A1560, which was obtained from Novozymes A/S, Denmark. The validation results (see Paper 1) indicate that the model can accurately predict the maximum specific growth rate (h^{-1}) during batch cultivations and the biomass yield (gDW/mmol substrate) during chemostat cultivations on different carbon sources. In this case study, we conclude that the model serves as an important resource for gaining further insight into our understanding of *A. oryzae* physiology.

4.2 Comparison of *Aspergillus* metabolic models

The key statistics of the *A. oryzae* metabolic model were compared to that of other available *Aspergillus* genome-scale metabolic models. In order to evaluate the models in a systematic way and to allow for comparison across the three models, we have updated all metabolite names across three *Aspergillus* metabolic models using the ChEBI ontology database, the KEGG database and manual curation at the end [88]. Due to this update the number of metabolites of the three models is slightly changed when compared to the original references and these changes led to slightly difference in the unique reactions. Number of genes, metabolites and reactions have been updated as shown in Table 4.2 and the three models of aspergilli have been converted to use the Systems Biology Markup Language (SBML) and are available at www.sysbio.se [88].

To evaluate the overlap of the metabolic models of the three aspergilli, the reactions in *A. oryzae i*WV 1314 were compared to the genome-scale metabolic models of *A. niger i*MA871 and *A. nidulans i*HD666. The results are shown in Figure 4.4. We found 426 reactions are overlapped for the three models. In *A. oryzae i*WV1314 there are 536 unique reactions that are not present in the other models, and besides this the main difference is due to divergences in compartmentalization used. Additionally, a number of pathways are found in *A. oryzae*, but not in *A. niger* and *A. nidulans* (e.g. aminoacyl-tRNA biosynthesis, metabolism of other amino acids, cofactors and vitamins metabolism). *A. niger i*MA871 contains 600 unique reactions in the network. These are large extent reactions relating to lipid biosynthesis and xenobiotic catabolism. The *A. nidulans i*HD666 model contains 156 unique reactions, which is a lower number than for the *A. oryzae* and *A. niger* models because the *A. nidulans* model is less complex and in this model many reactions are lumped, especially in lipid metabolism.

Table 4.2 Comparison of metabolic models for three *Aspergillus* species

Organism	Model name	Metabolic genes	Metabolites[a]	Unique metabolic reactions[b]	References
A. nidulans	*iHD666*	666	543 (740)	789	[68]
A. niger	*iMA871*	871	782 (1047)	1194	[67]
A. oryzae	*iWV1314*	1314	808 (1102)	1243	This study

[a] Number of chemically distinct metabolites, not counting presence in multiple compartments. Metabolite number in parenthesis counting presence in multiple compartments

[b] Unique reactions are defined as reactions being biochemically unique in their own compartment or transport reactions. Isoenzymes are thus not included in this number. Directionality is not taken into account.

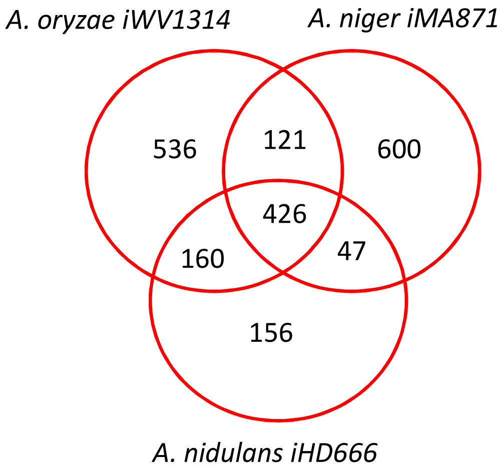


Figure 4.4 Venn diagram of reaction statistics for three genome-scale models of *Aspergillus*. The diagram shows the number of unique reactions shared and reactions that are specific for the three models.

4.3 Graphical representation of *A. oryzae* metabolic network

To illustrate the whole metabolic network, an overall metabolic map of *A. oryzae* was drawn. This map allows for graphical presentation of modeling results and of omics data. Figure 4.5 presents this network that link genes, enzymes, metabolic reactions and metabolites. This map was further used as graphical representation of genome-scale data in chapter 5 and 6 to map gene expression data or flux data.

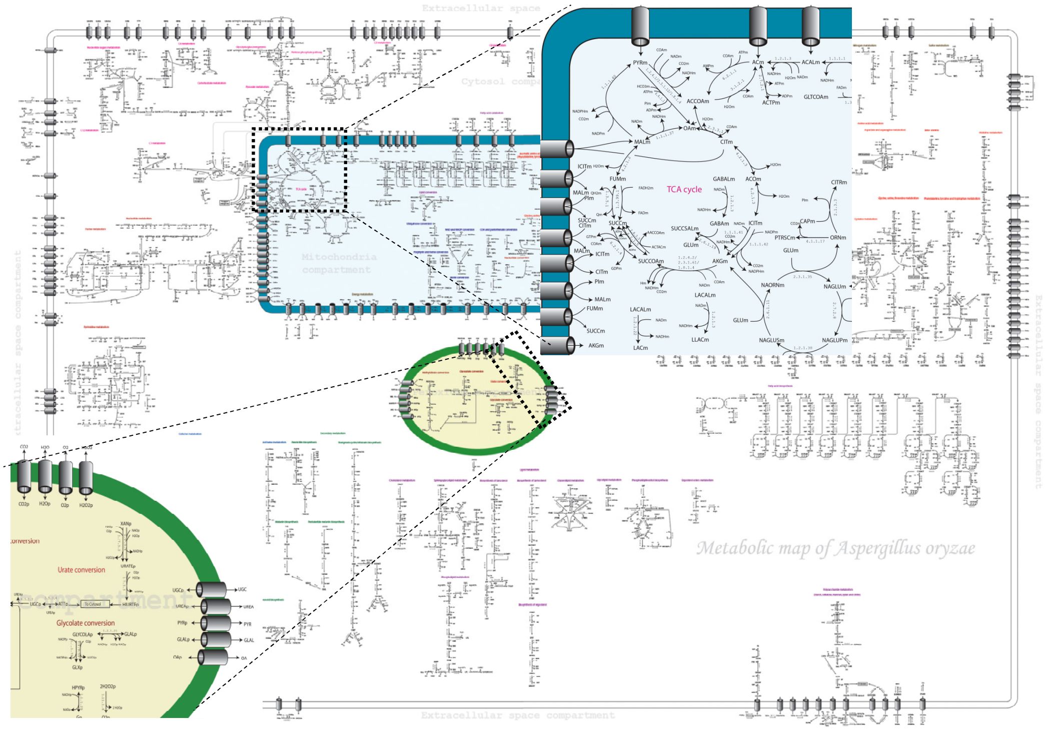


Figure 4.5 Overall metabolic map of *A. oryzae.*

Chapter 5

5. Different carbon sources on metabolism and transcriptional regulation across three *Aspergillus* species

Carbon metabolism and regulation in *Aspergillus* species is intensively studied. In many cases, the carbon source effects transcription of genes and the effect can sometimes be foreseen from the functions of their gene products. To further investigate this principle, this chapter represents the second case study that describes the obtained results from applying the SYSBIOMICS approach. The chapter is divided into three main parts. Firstly, we summarize the results and discussion of a study on the growth physiology on four different carbon sources. The different carbon sources discussed are glucose (see full details in Paper 2, 3 and 4), xylose (see full details in Paper 2), maltose (see full details in Paper 3) and glycerol (see full details in Paper 4). In the second part, we present the results and discuss transcriptome studies on these four carbon sources as well as present results of comparative analysis of three aspergilli. In the last part, we present analysis of genome-wide co-expression and co-evolution of aspergilli (see full details in Paper 5).

5.1 Growth physiology

In order to study carbon metabolism and transcriptional regulation in aspergilli, batch cultivations on four different carbon sources (i.e. glucose, maltose, xylose or glycerol) were performed and further comparative transcriptome analysis was done for *A. oryzae*, *A. niger* and *A. nidulans*. Table 5.1 gives an overview of all these investigations.

Table 5.1 List of different carbon sources for study of growth physiology and transcriptome profiling in three aspergilli

	Different carbon sources			
***Aspergillus* species**	**Glucose**	**Maltose**	**Xylose**	**Glycerol**
A. oryzae A1560 (Wild type)	+	+	+	+
A. niger BO1 (Wild type)	+	+	+	+
A. nidulans FGSC-A4 (Wild type)	+		+	+

+ indicates that the carbon source was used to study growth physiology and transcriptome profiles in the *Aspergillus* species

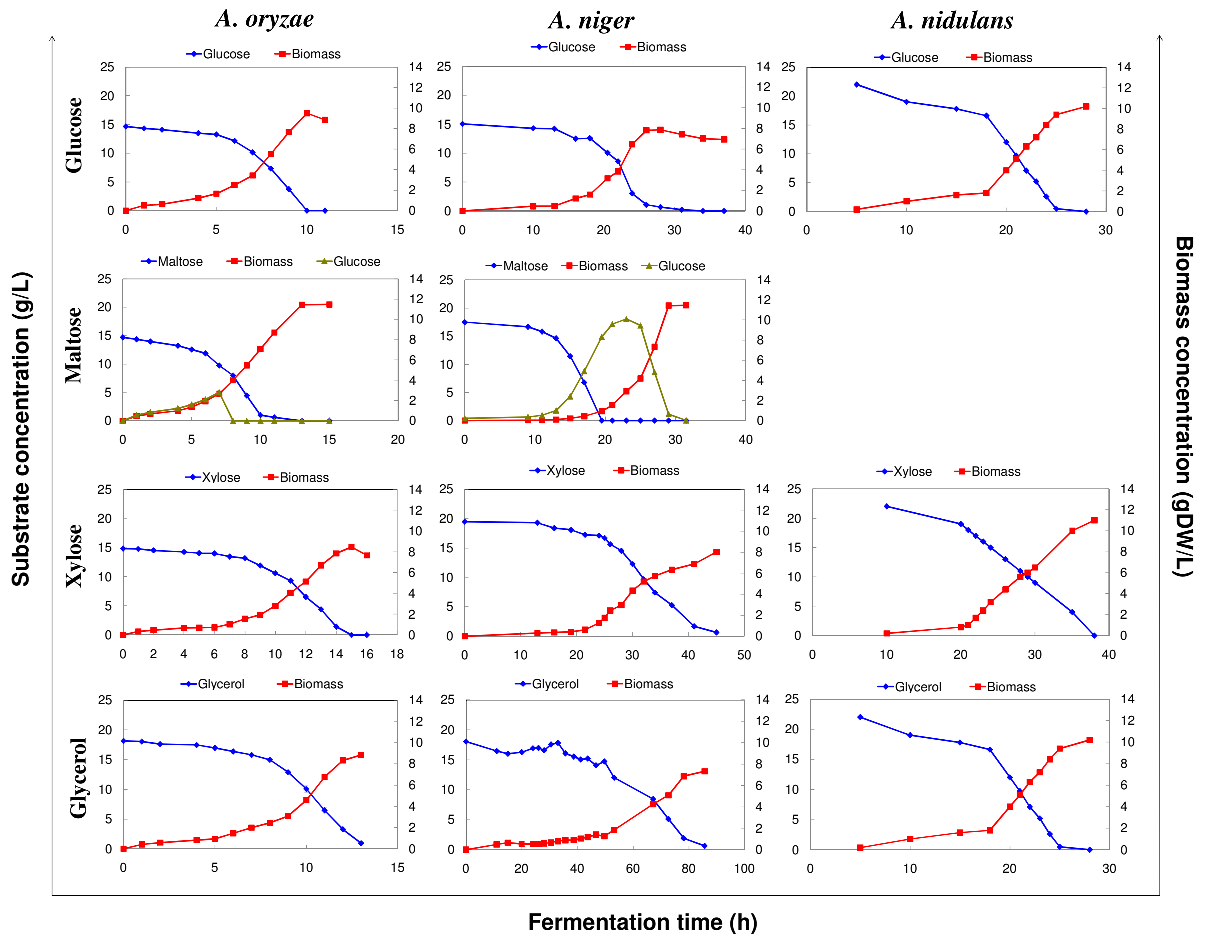


Figure 5.1 Summary of fermentation profiles of batch cultivations on four carbon sources carried out with *A. oryzae* A1560, *A. niger* BO1 and *A. nidulans* FGSC-A4

The cultures of these three *Aspergillus* species were all carried out in well-controlled bioreactors. All cultivations were batch cultures grown on defined salt medium with glucose, xylose, maltose or glycerol as the carbon source. For each of the three species, three biological replicates cultivations were performed on each carbon source. Figure 5.1 presents fermentation profiles of the four different carbon sources and Table 5.2 summarizes the physiological characterization data. For *A. oryzae*, glucose was exhausted in 10 h, maltose in 12 h, glycerol in 14 h and xylose in 15 h. The maximum specific growth rate of *A. oryzae* on glucose was 0.38±0.004 h^{-1}, which is due to the high efficiency in the uptake and metabolism of this sugar. Slower growth was achieved on maltose, where the maximum specific growth rate was 0.32±0.160 h^{-1}. For glycerol as carbon source, the maximum specific growth rate on of *A. oryzae* was 0.30±0.004 h^{-1}. Xylose is a less favorable carbon source for *A. oryzae*, but it could still be efficiently metabolized allowing for growth at a maximum specific growth rate of 0.27±0.010 h^{-1}. For *A. niger,* glucose was exhausted in 32 h, whereas maltose, xylose and glycerol were consumed after 19 h, 45 h and 86 h, respectively. *A. niger* growth on maltose was the fastest. Slower growth was achieved on glucose, with a maximum specific growth rate of 0.22±0.015 h^{-1}. In the case of maltose consumption, there was an accumulation of glucose due to a very high extracellular glucosidase activity expressed by *A. niger*, which allowed the fungus to grow very fast on this carbon source at a maximum specific growth rate of 0.31±0.020 h^{-1}. Xylose was a less favorable carbon source for *A. niger*, but still *A. niger* is able to efficiently metabolize it and grow at a maximum specific growth rate of 0.19±0.030 h^{-}

[1]. For glycerol, the growth of *A. niger* was 4 times slower when compared to glucose. In the case of *A. nidulans*, the maximum specific growth rate on glucose (0.23±0.020 h^{-1}) was faster than on xylose (0.16±0.010 h^{-1}) and the growth rate was double than that on glycerol (0.11±0.010 h^{-1}). Besides of growth rates for the three aspergilli, transcriptional analysis (TA) sampling times and biomass concentrations at the specific TA sampling time, and biomass yields were recorded for the three species on the four different carbon sources (see Table 5.2).

Table 5.2 Data for batch cultivations of *A. nidulans* FGSC-A4, *A. oryzae* A1560 and *A. niger* BO1. Fermentations were performed in three biological replicates. Values are shown as average ± standard deviations

***Aspergillus* Strains**	**Carbon source**	**μ_{max}[1]**	**Y_{sx}**	**Time of sampling[2]**	**Biomass concentration[3]**
		(h^{-1})	**(g DW/g substrate)**	**(h)**	**(g DW/L)**
	Glucose	0.38±0.004	0.54±0.013	~6	2.50±0.09
A. oryzae	Xylose	0.27±0.010	0.53±0.130	~10	2.80±0.11
	Maltose	0.32±0.160	0.49±0.150*	~7	2.27±0.09
	Glycerol	0.30±0.004	0.52±0.008	~8	2.44±0.05
	Glucose	0.22±0.015	0.57±0.053	~23	4.37±0.42
A. niger	Xylose	0.19±0.030	0.45±0.010	~31	3.73±0.55
	Maltose	0.31±0.020	0.62±0.020*	~24	3.55±0.51
	Glycerol	0.05±0.007	0.40±0.022	~36	0.88±0.29
	Glucose	0.23±0.020	0.47±nd	~22	6.33±0.40
A. nidulans	Xylose	0.16±0.010	0.45±nd	~33	6.43±0.23
	Glycerol	0.11±0.010	0.42±nd	nd	6.50±0.50

[1]μ_{max}: maximum specific growth rate.
[2]Time of sampling: average time of sampling for transcriptome analysis.
[3]Biomass concentration: biomass concentration at the sampling time
*Biomass yield was calculated based on glucose (g DW/g glucose)
nd: not detectable

5.2 Transcriptome analysis

To perform initial analysis of microarray data, we designed an *Aspergillus* Affymetrix GeneChip (see details in Paper 2) that consists of probes for the genes from the genomes of *A. oryzae*, *A. niger* and *A. nidulans*. Then we applied our own deigned chip to detect the effect of different carbon sources on genome-wide expression data and performed comparative analysis of the three species. The results and discussion are described as follows.

5.2.1 Xylose metabolism and regulation

For all three sample sets from glucose or xylose fermentations, transcriptome data analysis was performed for all three aspergilli (see Paper 2 for details). In summary, in order to study xylose metabolism and regulation in aspergilli, we analyzed gene expression that was conserved across three species *A. oryzae*, *A. niger* and *A. nidulans*. First homologues genes for the three species were identified by using pairwise BLASTP comparisons [70]. Using this approach, based on bi-directional best hits with an E-value cut-off of 1E-30, 5,561 predicted genes were found to be conserved in all three species (1:1:1 orthologues). The three sets of 5,561 conserved genes were initially used for further analysis.

Pairwise xylose versus glucose comparisons

From statistical analysis of gene expression data, the significantly regulated genes in all three species were compared to the list of the 5,561 conserved genes as well as with each other. This resulted in the identification of 23 conserved genes (Figure 5.2 and Table 5.3) that are differentially regulated in all three species as well as 365 genes that are differentially expressed in only two of the aspergilli. The 23 genes that are significant in all three species can be seen as a conserved response across the *Aspergillus* genus. A further inspection of the expression values of the 23 common genes revealed that the homologues are regulated in the same direction, with 22 of the genes being up-regulated on the xylose medium and only one gene being down-regulated. With the annotation of 23 conserved genes, the protein-encoding genes are enzymes and sugar transporters. As mentioned, most of the genes were induced in all three species and they were involved in D-xylose degradation pathway. Interestingly, the xylanolytic transcriptional activator XlnR was also identified as having a conserved transcriptional response. However, we found one down-regulated gene, which encodes a monosugar transporter (mstC), and this gene may encode a conserved transporter that has a higher affinity for glucose.

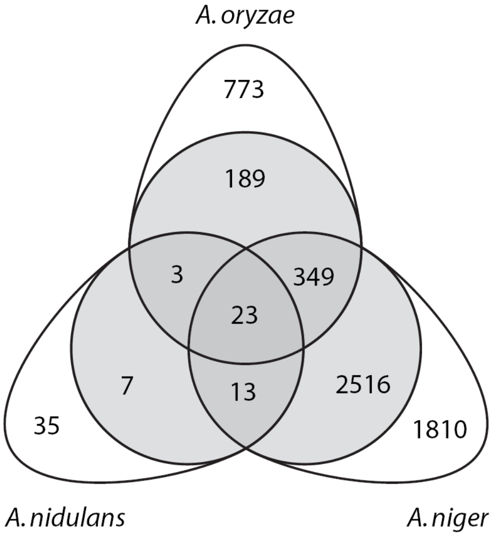


Figure 5.2 Venn diagram of significantly differentially expressed genes from t-test pairwise comparisons in the three aspergilli. The colored circles contain the genes that are significantly differentially expressed and conserved in the three *Aspergillus* species. The numbers on a white background are not conserved in all three species, but differentially expressed in each species.

Table 5.3 The 23 genes with conserved transcriptional responses in all three *Aspergillus* species

A. nidulans	*A. oryzae*	*A. niger*	*A. niger* annotation
AN0250	AO090001000069	55668	Sugar transporter
AN0280	AO090005000767	55419	Glucosyl hydrolase
AN0423	AO090003000859	51997	D-xylose reductase (xyrA)
AN0942	AO090005001078	46405	L-arabitol dehydrogenase
AN10124	AO090003000497	213437	β-glycosidase
AN10169	AO090038000426	177736	Short-chain dehydrogenase
AN1677	AO090023000688	54541	Short-chain dehydrogenase
AN2359	AO090005000986	205670	β-xylosidase (xlnD/xylA)
AN3184	AO090012000809	55604	Aldose 1-epimerase
AN3368	AO090010000208	212893	Glycoside hydrolase
AN3432	AO090020000042	56084	Aldose 1-epimerase
AN4148	AO090009000275	205766	Sugar transporter
AN4590	AO090011000483	180923	Sugar transporter
AN5860	AO090026000494	197162	Monosugar-transporter (mstC)
AN7193	AO090023000264	55928	Aldo/keto reductase
AN7610	AO090012000267	48811	XlnR transcriptional activator
AN8138	AO090010000684	212736	α-galactosidase
AN8400	AO090020000324	199510	Sugar transporter
AN8790	AO090020000603	209771	D-xylulokinase (xkiA)
AN9064	AO090038000631	203198	Xylitol dehydrogenase (xdhA)
AN9173	AO090010000063	194438	Sugar transporter
AN9286	AO090026000127	56619	α-glucuronidase (aguA)
AN9287	AO090701000345	54859	Lipolytic enzyme

Identification of conserved motif for XlnR regulator

Realizing that one or more conserved transcriptional regulators might be active in all three species and be responsible for the conserved response of the 23 genes, statistical promoter analysis was performed for all three species sets of 22 genes up-regulated on xylose (see Methods for detection of conserved motifs in Paper 2). A motif of "GGNTAAA" was found to be significant in the promoter sequences of 46 of the 3*22 genes. Based on this analysis, we propose that the "GGNTAAA" motif is indeed the XlnR motif that is conserved in *A. nidulans*, *A. niger*, and *A. oryzae*.

5.2.2 Maltose utilization and regulation

For all three sample sets from glucose and maltose fermentations, transcriptome data analysis was performed in two different *Aspergillus* species, namely *A. oryzae* and *A. niger* (see Paper 3 for details). In brief, in order to study maltose utilization and regulation in aspergilli, yeast *S. cerevisiae* was used as model organism since regulation of maltose transport and metabolism is well studied in this organism [89, 90]. The presence of maltose in the environment is necessary for induction of synthesis of the maltase and the maltose transporter. The metabolism and regulation of maltose requires the presence of a *MAL* loci, of which there are several identified in different strains of *S. cerevisiae*, but the *MAL6* locus is the most well studied [89]. The *MAL6* locus is composed of a cluster of three genes: *MAL61* (*MALT*) encoding maltose permease, *MAL62* (*MALS*) encoding maltase and *MAL63* (*MALR*), encoding transcriptional activator specifically activating expression of the *MALT* and the *MALS* genes [91]. The aim of this study was to identify the *MAL* gene cluster in different sequenced *Aspergillus* genomes using the gene cluster of the *MAL*6 locus of *S. cerevisiae* as a model. We further validated the presence or absence of the *MAL* gene cluster in *A. oryzae* and *A. niger* by using our custom designed Affymetrix GeneChip for transcriptome analysis (Paper 2) [92].

Comparative analysis of MAL gene clusters in S. cerevisiae and Aspergillus species

First we searched for the presence of the *MAL* gene cluster in 10 different sequenced *Aspergillus* genomes. For this purpose, the gene cluster of the *MAL*6 locus of *S. cerevisiae* was used as a model and BLASTP was applied (see Methods in Paper 3). The results showed that six different *Aspergillus* strains (i.e. *A. oryzae*, two strains of *A. fumigatus*, *A. flavus*, *A. clavatus*, and *A. fischeri*) contain at least one *MAL* gene cluster as illustrated in Figure 5.3. In contrast, we could not find any *MAL* cluster in *A. nidulans*, *A. terreus* and two strains of *A. niger* with the statistical constraints imposed. These results could suggest that these four *Aspergillus* strains most likely do not have the *MAL* regulon for maltose utilization. Notably, in all the sequenced *Aspergillus* genomes, we could identify multiple orthologue genes encoding maltase enzymes and maltose transporters as shown in Figure 5.3. To prove the presence or absence of the *MAL* regulon at the transcriptional level, we further evaluated our results obtained from comparative genomics through transcriptomics analysis. In the following, we show an example of using our previously designed *Aspergillus* GeneChip (Paper 2) [92] to validate the presence of the *MAL* regulon in the *A. oryzae* genome and the absence of the *MAL* regulon in the *A. niger* genome.

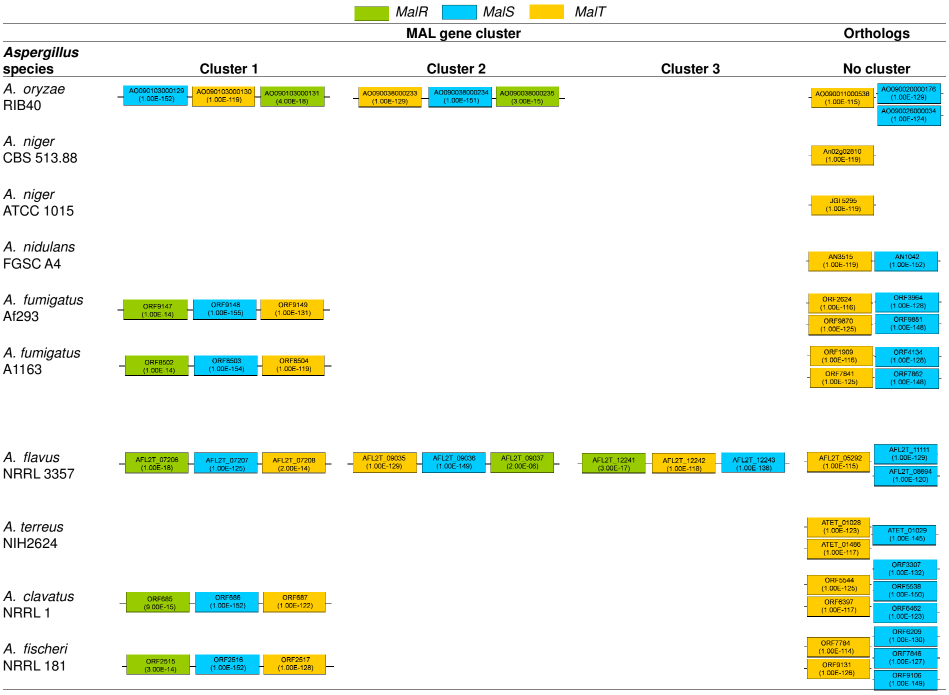


Figure 5.3 Diagram shows comparative sequence analysis of *MAL* gene cluster between *S. cerevisiae* and 10 different *Aspergillus* species. Values in each rectangle represent the ORF name.

Pairwise maltose versus glucose comparisons

To further test our assumption obtained from comparative genomics for *MAL* regulon presence in *A. oryzae* or not present in *A. niger*, the genome-wide gene expression data obtained from glucose or maltose cultivations were pairwise compared for each species. To detect transcriptional changes in response to a change in the carbon source, Student's t-test statistics were used to identify significantly different gene expression levels with p-value cut-off of 0.05. List of genes (see results in Paper 3) were significantly differentially expressed in *A. oryzae* between glucose and maltose (16 gene expression changes). In contrast, for *A. niger*, no genes were statistically differentially expressed.

As presented in Paper 3, 16 genes showed higher expression level on maltose compared to glucose in *A. oryzae*. Interestingly, among the 16 genes with significantly higher expression on maltose, we found genes encoding maltase in *A. oryzae* (AO090103000129 and AO090038000234), which correspond to orthologous genes of *MALS* in *S. cerevisiae* (see Figure 5.3). Moreover, we also found up-regulated genes encoding maltose permease, AO090103000130 and AO090038000233, which are the functionally related orthologous genes of *MALT* in *S. cerevisiae* (see Figure 5.3). The two *A. oryzae* genes orthologous to the *S. cerevisiae MALR* transcription factor, AO090103000131 and AO090038000235, were also up-regulated, but not statistically significant. This suggests that the mechanism behind the *MALR* regulon in *A. oryzae* is similar to *MALR* function in *S. cerevisiae*, where it is activated by maltose and repressed by glucose.

According to the results of protein sequence analysis and synteny gene analysis [93] of the *MAL* gene cluster, we could conclude that *A. oryzae* has two *MAL* regulons and each regulon contains gene that is likely to be *MALR* transcriptional activator (see Figure 5.3). From this significant evidence combined with the physiological response of *A. oryzae* growth on maltose (see Figure 5.1), where *A. oryzae* continuously consumed maltose having almost no accumulated glucose over time, we propose that *A. oryzae* has global regulation of maltose utilization by these two *MAL* regulons, where the *MALR* transcription factor induces maltose permeases (*MALT*) to transport extracellular maltose into the cell and *MALR* also induces maltase (*MALS*) that hydrolyzes intracellular maltose into glucose, which is then channelled through glycolysis. In contrast, we could not identify any *MAL* gene cluster in *A. niger* that is closely homologous to the one existing in *S. cerevisiae* (see Figure 5.3). Furthermore, transcription data analysis of the pairwise comparison between maltose and glucose in *A. niger*, did not show any significant gene expression change that can point out the presence of *MAL* cluster either. We therefore propose that maltose utilization in *A. niger* most likely do not involve a *MAL* regulon, but occurs through another regulatory system and the recent publication of genome-wide expression analysis in *A. niger* by Yuan and coworkers [94] supports these results. Yuan *et al.* suggested that AmyR is an important transcription factor that is found in *A. niger* and that *amyR* itself is induced by the presence of maltose. In addition, their studies indicated that *amyR* gene transcription regulation takes place in *A. niger* and showed that a disruption of the AmyR transcription factor resulted in low levels of extracellular enzymes i.e. glucoamylase (GlaA) converting maltose to glucose and consequently activating a stress response due to low glucose levels. The low availability of glucose transferred a signal to down-regulate glucose transporters [94]. This is in accordance with our transcriptome results and the physiological response obtained in *A. niger* maltose cultivations (see Figure 5.1), where maltose degradation occurred faster than glucose uptake and metabolism leading to a high extracellular accumulation of glucose over time. We therefore support the conclusions from previous studies where it is stated that *A. niger* has global regulation of maltose utilization by the AmyR transcriptional activator. It activates genes encoding known extracellular starch degrading enzymes, such as *glaA*. The *glaA* gene product, an extracellular glucoamylase, can convert extracellular maltose to extracellular glucose and then glucose can be taken up by glucose transporters. Figure 5.4 illustrates the proposed different mechanisms for global regulation of maltose utilization between *A. oryzae* (Panel A) and *A. niger* (Panel B).

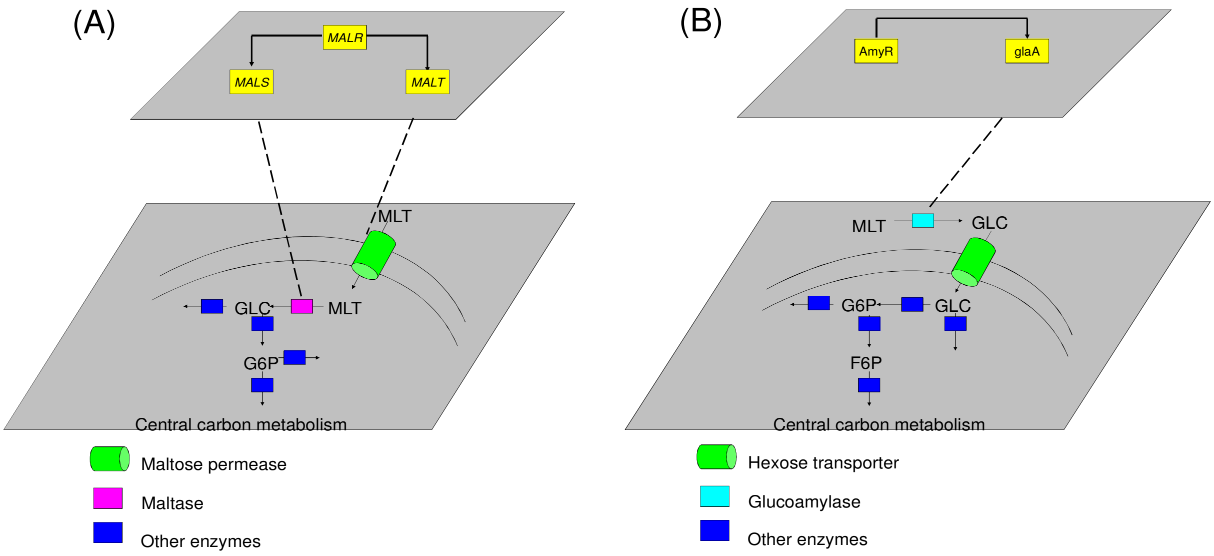


Figure 5.4 Diagram shows comparative maltose utilization and global regulation in *A. oryzae* (A) and in *A. niger* (B).

Key metabolites identification and metabolic subnetworks analysis

In order to analyze the overall metabolic responses to changes in the carbon source, i.e. using glucose or maltose, we applied the reporter metabolites and subnetworks algorithm to identify key metabolites and to search for highly correlated metabolic subnetworks for the pairwise comparison [95]. This analysis relied on the reconstructed genome-scale metabolic network of *A. oryzae* (see Paper 1), and therefore we demonstrated how these metabolic networks can be used to map global regulatory responses in this *Aspergillus* spp. Figure 5.5 shows the list of key genes-encoding enzymes and transporters comprising the subnetwork of *A. oryzae* investigated upon a change of carbon source, from glucose to maltose.

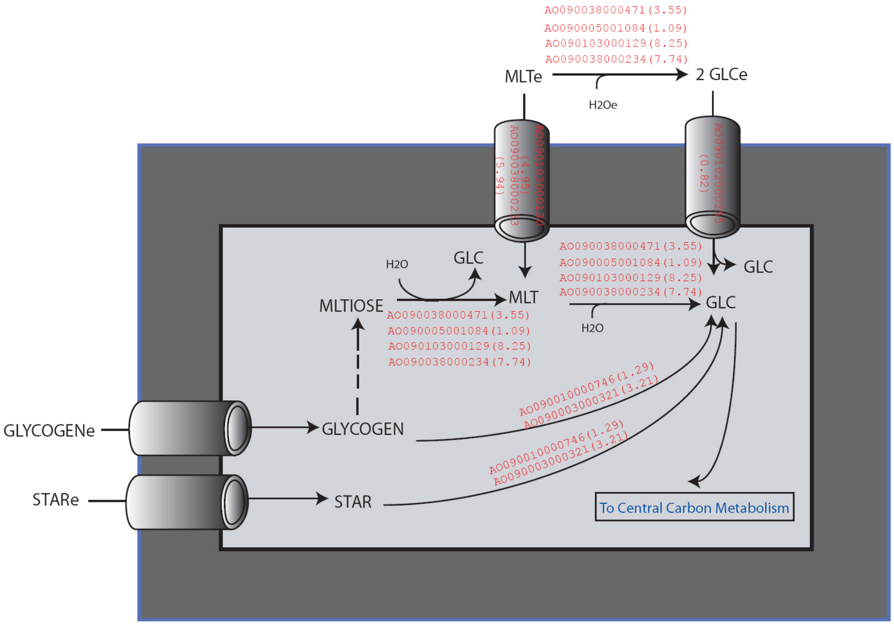


Figure 5.5 Schema of key up-regulated enzymes and transporters comprising the subnetwork of *A. oryzae* in maltose condition.

5.2.3 Glycerol metabolism and regulation

In this part (see Paper 4 for full details), we aimed at the identification of global regulatory patterns of gene expression during a metabolic transition from repressed (glucose) to derepressed (glycerol) condition. We identified a conserved regulatory response among the three *Aspergillus* species, which was found to be consistent with the response reported previously in *S. cerevisiae* [96]. At the beginning, to identify conserved regulatory systems as well as to exploit both similarities and differences at the protein level, genes having orthologues in the three species were identified by using a BLASTP based comparison and the data sets obtained from homology search were compiled for further analysis.

Pairwise glycerol versus glucose comparisons

Genome-wide gene expression data was analyzed for all three sets of glucose or glycerol batch fermentations for the three *Aspergillus* species. A t-test pairwise comparison for each *Aspergillus* species on glycerol versus glucose revealed 904, 1,145 and 3,058 significantly differentially expressed genes for *A. nidulans*, *A. oryzae* and *A. niger*, respectively. Subsequently, these three subsets of significant genes in all three species were compared to the list of conserved orthologous genes in the three aspergilli as well as with each other. This resulted in the identification of 88 conserved genes that were differentially expressed in all three species. Among them, 81 genes were up-regulated during growth on glycerol, 5 genes were down-regulated and 2 genes did not show a clear trend. The obtained response of 88 differentially expressed genes in the three species suggests a conserved regulatory response across the *Aspergillus* genus. From our transcriptome data through analysis of pathway utilization of glycerol in *Aspergillus* species, interestingly, the data obtained suggests that the catabolic pathway via glycerol 3-phosphate is the major route for glycerol catabolism in *A. nidulans* and *A. niger*. The genes encoding glycerol kinase as well as the genes encoding the FAD$^+$ dependent glycerol 3-phosphate dehydrogenase were significantly up-regulated on glycerol media compared to glucose media. In contrast, in *A. oryzae*, the most statistically significant up-regulated ones were the genes encoding the enzyme glycerone kinase and glycerol dehydrogenase. Therefore the most active pathway in *A. oryzae* is probably the one using glycerol dehydrogenase and glycerone kinase to produce glycerone phosphate. Besides of a major transcriptional change in genes involved in glycerol utilization pathways, other metabolic changes occurred in response to polyol metabolism mainly due to a stress response. The polyols can be produced in high concentrations, and therefore make a significant contribution to the osmotic pressure in the cell. In yeast and fungi, glycerol has a role in regulating the osmotic pressure in the cells [97]. Here we show the presence of the High Osmolarity Glycerol (HOG) pathway in *S. cerevisiae* with comparison to the proposed HOG pathway for three *Aspergillus* speices (see full information in Paper 4).

Regulation of gene expression by the Adr1 transcriptional activator

One or more conserved transcriptional regulators were suspected to be up-regulating the subset of 81 genes or down-regulating the subset of 5 genes within the group of 88 genes having a conserved transcriptional response. Statistical promoter analysis was conducted for all three data sets of 81 up-regulated genes on glycerol medium. By inspecting the upstream sequences of each up-regulated orthologues dataset, giving a subset of 243 promoters (3*81 promoters), we found the most over-represented pattern to be "TGCGGGGA". A logo plot was constructed (Figure 5.6). This result suggests that these genes are up-regulated by a common cross species conserved transcription factor. Based on a literature search, it was proposed that "TGCGGGGA" is the consensus binding sequence of the transcriptional activator Adr1, which has been found to regulate several pathways in *S. cerevisiae* [96] and in humans [98]. The same kind of analysis was conducted with the subset of down-regulated genes, but no motif was found.

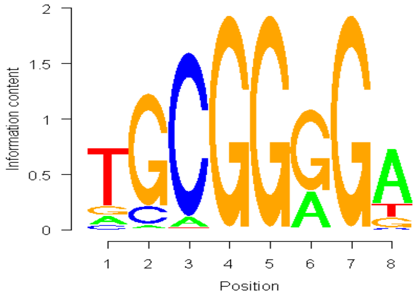


Figure 5.6 Logo plot of the over-represented motif as Adr1 binding site for three *Aspergillus* species

Key metabolites identification and metabolic subnetworks analysis of A. oryzae

We applied the reporter metabolites algorithm and subnetworks identification [95] in order to examine overall metabolic responses in *A. oryzae* studied upon a change of a repressing (glucose) to a derepressing (glycerol) carbon source. Figure 5.7 illustrates the list of key enzymes and transporters comprising the subnetworks of *A. oryzae* investigated upon a shift of carbon source. With our expectation, we found key enzymes mainly involved in glycerol metabolism and fatty acid metabolism as well as linked to TCA cycle that were up-regulated when using glycerol compared to glucose as carbon source. For reporter metabolites analysis, it was interestingly found that all these key metabolites such as monoacylglycerol, diacylglycerol, triacylglycerol, TCA cycle intermediates in the mitochondria, i.e. succinate, and some others like lactaldehyde were identified as significantly changed in *A. oryzae* when glycerol was used as a carbon source.

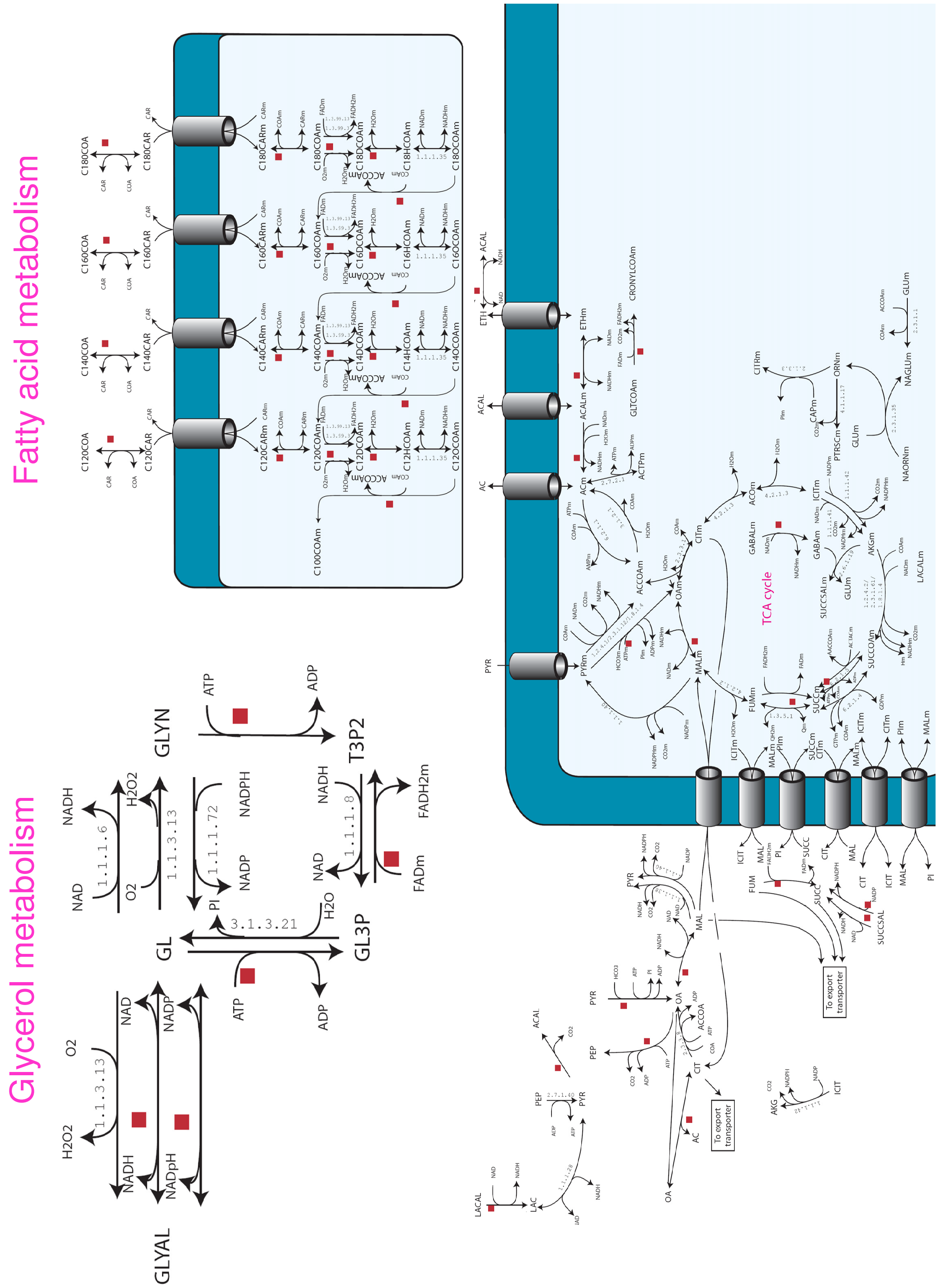


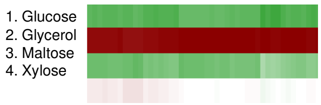
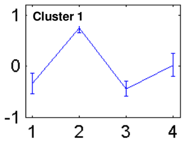
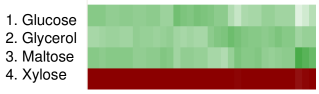
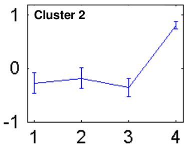
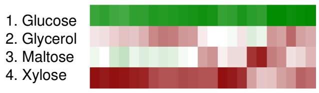
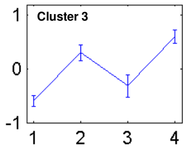
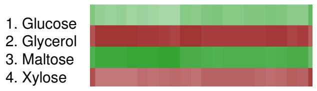
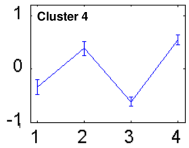
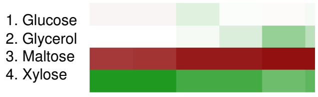
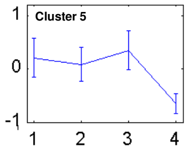
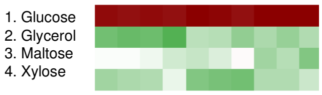
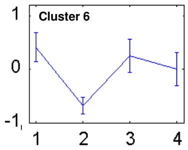
Figure 5.7 Reporter enzymes comprising the subnetworks of *A. oryzae* investigated upon a shift of carbon source from glucose to glycerol.

5.3 Analysis of genome-wide co-expression and co-evolution in Aspergilli

One important goal of analyzing gene expression data is to identify co-expressed genes. Transcriptomics can be used for identification of co-expressed genes which provide hints toward inferring gene function based on the concept of guilt-by-association, i.e. co-expressed genes are likely to serve similar purposes and to be regulated by similar mechanisms [99, 100]. In this section, we summarize genome-wide transcriptome analysis that leads to underline the core biological processes on both genotype and expression in different aspergilli and hereby identify the possible coexistence of DNA regulatory motifs and transcription factors. Full details of the work described in this section are given in Paper 5.

In this study, we carried out cross-species analysis of genome-wide transcriptional co-expression patterns under different growth conditions, specifically to identify concerted transcriptional changes of genes that clearly reflect cellular adaptations. Our analysis first focus on comparative analysis between *A. oryzae* and *A. niger* at the genome and the transcriptome levels. The results showed that most of the genes in these two species have similar environmental responses in terms of gene expression patterns as presented in Table 5.4. In particular, there are many genes in cluster 1 and cluster 2 that are orthologues and co-expressed in the two fungi. Considering cluster 1, it is clear that genes were up-regulated in response to growth on glycerol, which is also found by the identification of genes associated with the metabolic pathway of glycerol metabolism. The cluster contains a lot of co-evolved and co-expressed genes that has been annotated with similar functions, i.e. genes encoding enzymes involved in fatty acid catabolism by the beta-oxidation pathway, fatty acid transport (e.g. mitochondria carnitine-acylcarnitine carrier protein and peroxisomal long-chain acyl-CoA transporter), the glyoxylate bypass, peroxisomal biogenesis and function. These were also found in reporter Gene Ontology (GO), such as peroxisome (GO:0005777), glyoxylate cycle (GO:0006097), lipid metabolic process (GO:0006629), and glycerol metabolic process (GO:0006071) (see Paper 5). The results indicate that there is a co-regulation of metabolic pathways involved in glycerol and fatty acid catabolism, probably due to the co-existence of these compounds as triacylglycerides and phospholipids in nature. In fungi, the beta-oxidation pathway has been studied for localization by Shen *et al.* [101], based on a large scale *in silico* screening of localization prediction for all relevant enzymes in more than 50 fungal species, the results showed that this pathway mainly takes place in the mitochondria and the peroxisome. To evaluate whether the two aspergilli contain co-evolved pathways, we mapped the conserved co-expressed genes identified in cluster 1 onto the genome-scale metabolic networks of *A. oryzae* (Paper 1) [102] and *A. niger* [67]. Once gene-metabolic pathway mapping was performed, the results showed that the conserved co-expressed genes are involved in fatty acid catabolism by beta-oxidation. Besides, we also found enzymes/protein functions involved in the glyoxylate bypass and peroxisomal protein functions (peroxins) to be conserved. The common pathways and core protein functions for the two *Aspergillus* species are illustrated in Figure 5.8. For the genes in cluster 2, the pattern indicates up-regulation of genes in response to growth on xylose, and not surprisingly it is found that many of these genes are associated with the arabinose and xylose metabolism (Table 5.4) and reporter GO terms like xylan catabolic process (GO:0045493), transaldolase activity (GO:0004801), and D-xylulose reductase activity (GO:0046526). We also performed mapping of the conserved genes onto the metabolic network. As expected, we found the co-evolved pathways which are mainly involved in pentose metabolism, especially the xylose degradation pathway and nucleotide sugar metabolism. This result showed a good agreement with our previous dedicated study of the conserved response in three *Aspergillus* species to growth on xylose (see Paper 2).

Table 5.4 Gene expression profiles and cluster patterns of *A. oryzae* and *A. niger*

Gene expression profiles	Cluster patterns*	Number of co-expressed genes in cluster	Over-represented metabolic pathway
		Total gene number: 777 (238)# *A. oryzae* : 290 genes *A. niger* : 487 genes	Glycerol metabolism Glycolysis and gluconeogenesis Propanoate and butanoate metabolism Polysaccharide metabolism Valine/leucine/isoleucine metabolism Phenylalanine/tyrosine/tryptophan biosynthesis
		Total gene number: 172 (23)# *A. oryzae :* 120 genes *A. niger* : 52 genes	Arabinose and xylose metabolism Pentose phosphate pathway Polysaccharide metabolism
		Total gene number: 152 (2)# *A. oryzae :* 138 genes *A. niger* : 14 genes	Polysaccharide metabolism Pyruvate metabolism
		Total gene number: 160 (7)# *A. oryzae :* 100 genes *A. niger* : 60 genes	Phenylalanine/tyrosine/tryptophan biosynthesis
		Total gene number: 30 (1)# *A. oryzae :* 7 genes *A. niger* : 23 genes	ND
		Total gene number: 83 (14)# *A. oryzae :* 32 genes *A. niger* : 51 genes	Polysaccharide metabolism Carbohydrates transport Other compounds transport

*Cluster patterns: The *x* axis represents the four different carbon sources investigated: 1 - Glucose; 2 - Glycerol, 3 - Maltose; 4 - Xylose; the *y* axis represents normalized gene expression intensities.

#Number of orthologous genes across two aspergilli

ND: Not Detectable

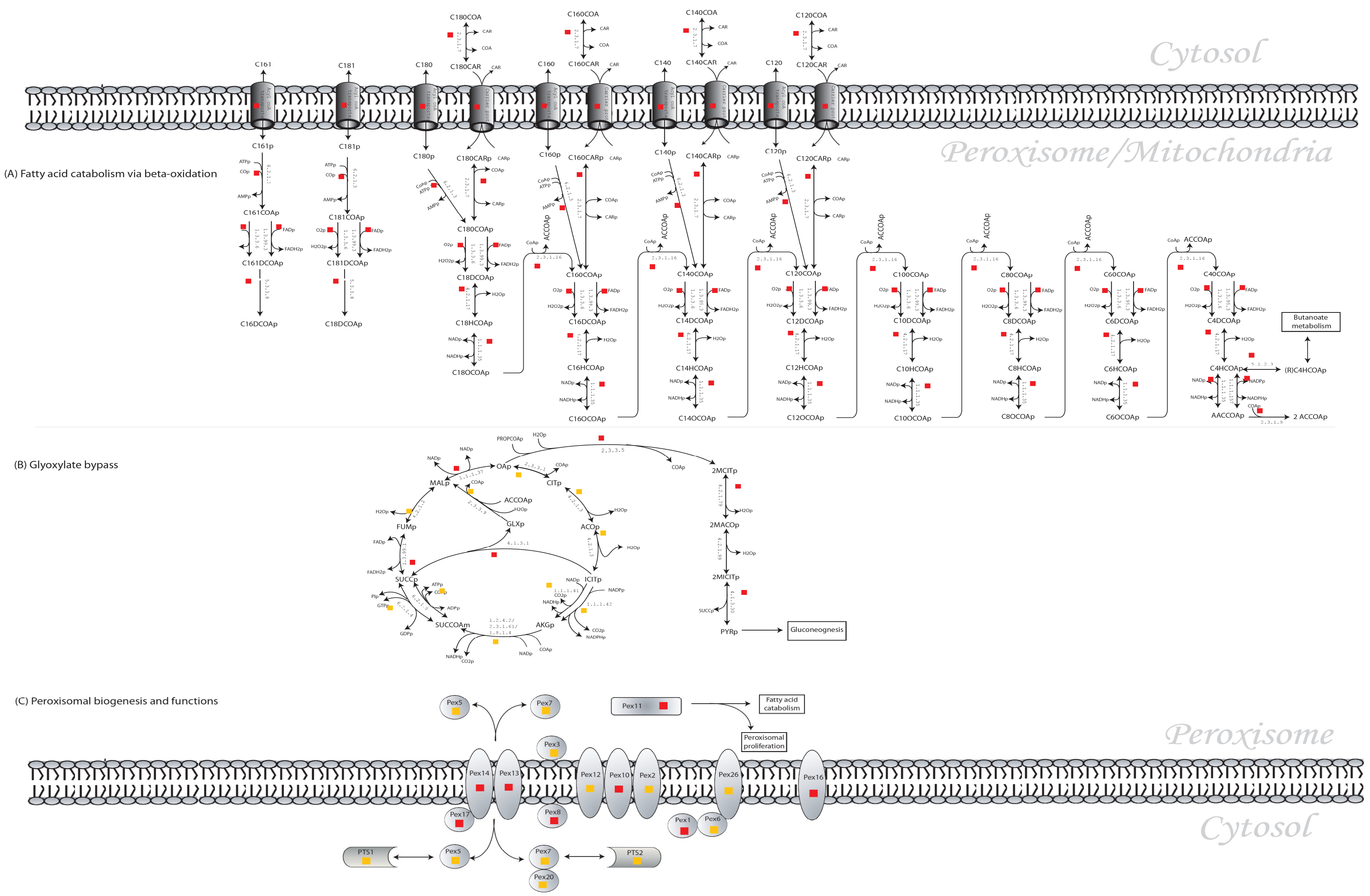


Figure 5.8 Co-evolved pathways and functions identified between *A. oryzae* and *A. niger*
(A) Fatty acid catabolism via beta-oxidation, (B) Glyoxylate bypass, (C) Peroxisomal biogenesis and functions. Red boxes refer to conserved orthologous genes with co-expression profiles for cross-species of *A. oryzae* and *A. niger*. Yellow boxes indicate that genes with conserved orthologues for cross-species of *A. oryzae* and *A. niger* but not co-expression profiles.

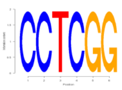
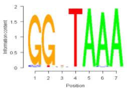
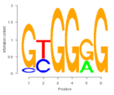
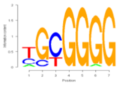
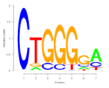
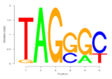
5.3.1 Analysis of DNA regulatory motif and transcription factor underlying co-expression

Genes with similar expression profiles will often have their promoter regions bound by common transcription factors at specific motifs and potentially regulated through common regulatory mechanisms. By promoter sequence analysis, we sought potential regulatory motifs in the upstream DNA sequences and further searched for the corresponding transcription factors that underlie the transcriptional co-expression patterns. The 1,000 base pairs (bp) of the upstream regions from the start codon of relevant genes in each cluster were analyzed to find the most over-represented common motif (see Methods in Paper 5). For example, for the genes belonging to cluster 1, a 1 kilo base pairs (kbp) upstream sequence from the start codon was scanned to identify an over-represented pattern. This was done for all the genes (see Table 5.4) from the *A. oryzae* genome (290 genes) and the *A. niger* genome (487 genes). Hereby we identified the motif, “CCTCGG” (reverse complement, “CCGAGG”) for this cluster. Several other common motifs of other genes in the other co-expression clusters were also found and the corresponding logo plots of all detected motifs are presented in Table 5.5.

In order to analyze a transcription factor that potentially bind to the identified DNA motifs, we compared the motifs with known or predicted *Aspergillus* consensus motifs from the public databases or the literature. We found that four out of six of the over-represented motifs are highly conserved to fungal species which allowed for identification of putative transcription factors as summarized in Table 5.5. The identified motifs were consistent with known binding sites of known transcription factors in aspergilli, such as FarA [103], FarB [103], XlnR (see Paper 2), CreA [104] and Adr1 (see Paper 4). As described above, the core cellular processes with co-evolution and co-expression found in cluster 1, are likely regulated by the FarA and the FarB proteins that can potentially bind to the over-represented motif pattern “CCTCGG”. There are few studies on these proteins that are existing in both the genome of *A. oryzae* and *A. niger*. However, a number of studies have been reported that the Far protein family governs transcriptional activator controlling the utilization of fatty acids in several fungi [103].

An elegant study presented by Hynes *et al.* [103], showed that these transcription factors FarA and FarB can bind to DNA sequences at 5′ region of a large number of genes involved in fatty acid catabolism and related processes in *A. nidulans.* Following their results, they concluded that FarA and FarB mainly induce genes via binding to the 6-bp core sequence “CCTCGG” in the 5’ regions. They also found that the genes involved in catabolism of fatty acid have a high enrichment of the core motif on their promoters. From their conclusions, we further performed our analysis of occurrence of “CCTCGG” sequence in the upstream regions of core genes that have transcriptional co-expression and co-evolution in cluster 1, and which have cellular processes related to those reported by Hynes and coworkers [103]. The results clearly showed that this pattern was also enriched in genes involved in fatty acid catabolism, glyoxylate bypass and peroxisome biogenesis (see Paper 5). Based on this, we postulate that the core cellular processes are all conserved at the genetic, transcriptional and regulatory level in *A. oryzae* and *A. niger.*

Table 5.5 List of identified putative DNA regulatory motifs and transcription factors

Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
DNA regulatory motifs						
Putative transcription factors	FarA FarB	XlnR	CreA	Adr1	Unknown	Unknown

Conserved FarA and FarB transcription factors in aspergilli

Both FarA and FarB proteins are classified as Zn(II)$_2$Cys$_6$ transcription factors. FarA is required for induction by both short- and long-chain fatty acids, while FarB is likely required only for short-chain fatty acid induction in *A. nidulans* [103]. As shown in Figure 5.9, we found highly conserved protein sequences of FarA and FarB across 8 species, *A. nidulans* (two strains), *A. oryzae* and *A. niger* (two strains), *A. flavus*, *A. clavatus*, *A. terreus*, *A. fischeri* and *A. fumigatus* (two strains). As shown in the phylogenic tree, considering known conserved domains analysis, the FarA protein contains both of a Zn$_2$-Cys$_6$ binuclear cluster domain (PF00172) and a fungal specific transcription factor domain (PF04082) that are conserved for all 8 species. For the FarB protein, we could not identify a conserved domain of the Zn$_2$-Cys$_6$ binuclear cluster (PF00172) in the *A. niger* strain ATCC 1015 and in *A. terreus*, while we found this domain in the other aspergilli. According to the conserved domains analysis, both proteins FarA and FarB have the similar architecture of the two conserved domains, and from these results, we can conclude that these transcription factors are evolutionary conserved among aspergilli. We further evaluated whether the FarA and the FarB proteins are conserved among other fungi, and here it is found to be highly homologue to genes identified in *Penicillium* spp, *Fusarium* spp, *Neurospora* spp, *Sclerotinia* spp, *Ajellomyces* spp, *Paracoccidioides* spp, *Coccidioides* spp, *Talaromyces* spp and *Microsporum* spp, whereas there is no conservation to yeast genes (see full details in Paper 5). Through this study, one can obtain a better understanding of the complex relationships between co-expression of genes. We found that FarA and FarB are conserved regulators of aspergilli that govern regulation of co-evolved and co-expressed genes related with core biological processes. Our work therefore improved functional annotation and the reconstruction of gene regulatory network in aspergilli.

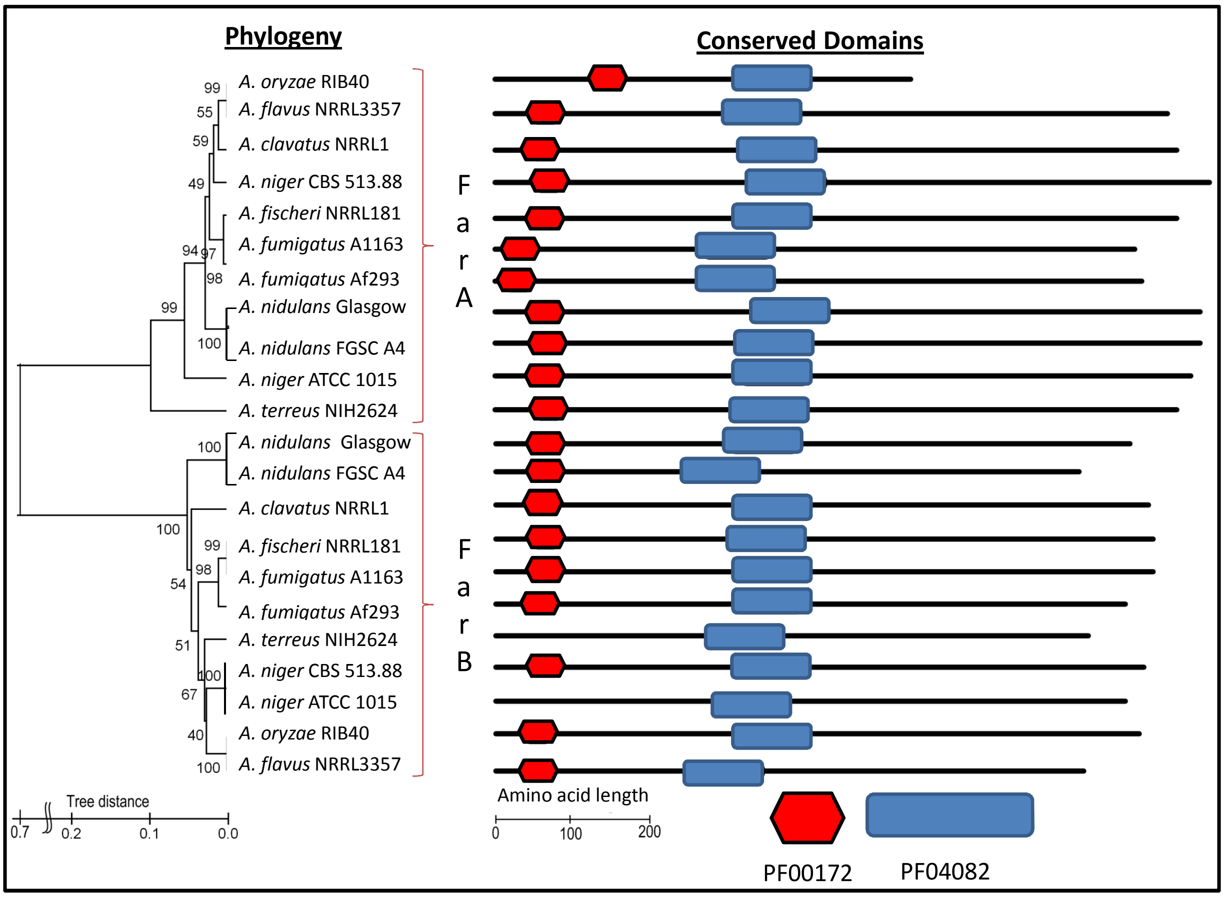


Figure 5.9 Phylogeny and conserved domains analysis of FarA and FarB proteins in 8 *Aspergillus* species

Chapter 6
6. Protein production by *A. oryzae*

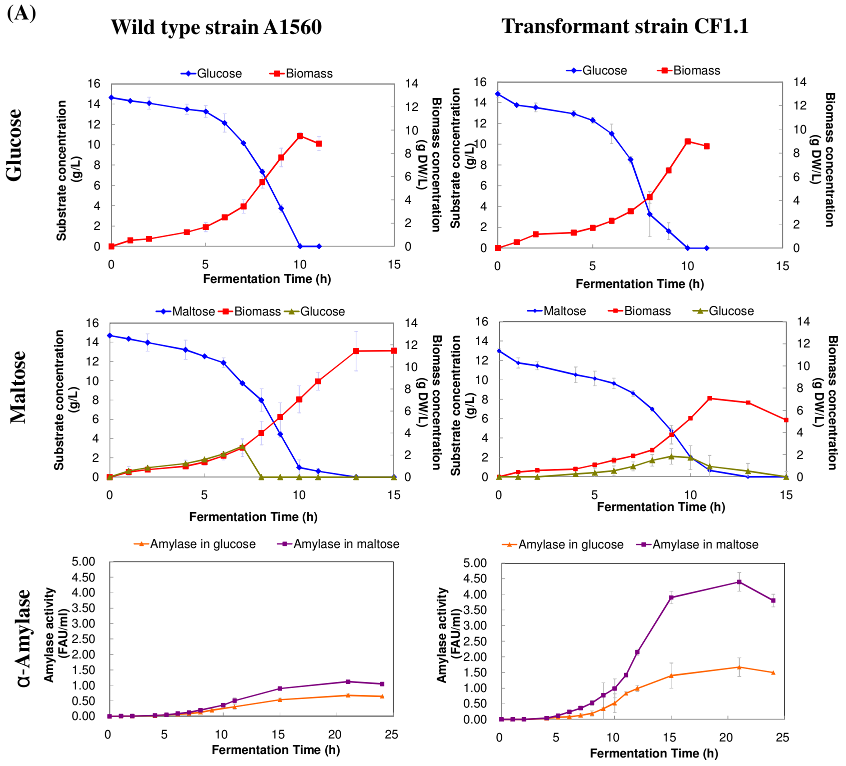
A. oryzae has a very high natural protein secretion capacity, which enables high level production of many fungal enzymes that find applications in the technical, feed and food industries. In connection with sustainable development of chemicals and biofuel productions, raw materials such as starches, celluloses, and hemicelluloses are widely used in production of e.g. ethanol and lactic acid. In order to degrade starch to glucose, there is a need for efficient enzyme biocatalysts. *A. oryzae* seems to be a producing organism that secretes significant amounts of α-amylases that break polysaccharides into sugars which are further fermented by yeast and lactic acid bacteria [105]. Therefore a demand for α-amylase production is growing at a fast pace. Despite the industrial importance of *A. oryzae* as mentioned, there is relatively little known information about its fundamental process of protein production. Such knowledge is quite important for optimization of an industrial enzyme fermentation process. For example, information about which genes/pathways are key players/targets for high level protein production.

In this study, the aim was to perform integrative data analysis (i.e. genomes, transcriptomes, metabolic networks, interactomes and fluxes) for diagnosis of industrial enzyme production in fermentation process. Our integrative approach involves comparative transcriptome analysis of a high-producing strain of α-amylase with a reference strain. We compared the *A. oryzae* transformant strain CF1.1 that contains multiple additional gene copies of the α-amylase production strain with the wild type strain A1560. To identify fundamental metabolic process of protein production, we further combined the genome-scale metabolic network of *A. oryzae* (Paper 1) [102] with the transcriptome data. Moreover, analysis of the global regulatory structure underlying protein synthesis and secretion was also analyzed. We further reconstructed an interaction network of *A. oryzae* based on identification of putative components through comparative genomics and interactomics (e.g. protein-protein interaction) between *A. oryzae* and *S. cerevisiae*. The reconstructed interaction network was used for identification of key proteins in the transcriptional response to high-level protein production. Additionally, flux calculation was performed for analysis of amino acid consumption for protein production.

Here we summarize the integration of these multiple data dimensions for diagnosis of the protein production process. We present the key players/targets (i.e. genes, enzymes, proteins, metabolites and pathways) in response to protein production that may lead us to further improve industrial enzyme fermentation processes. Full details of this study are given in Paper 6.

6.1 Growth physiology

The growth physiology of the two *A. oryzae* strains was examined in well-controlled bioreactors. Batch cultures were carried out using the same defined salt medium with glucose or maltose as the carbon source (see Paper 6). Three biological replicates cultivations were performed for each strain and each carbon source. The results illustrated in Figure 6.1 (panel A) present profiles of the growth and α-amylase enzyme activity of the two strains (wild type strain A1560 and transformant strain CF1.1) for each carbon source (glucose or maltose). Panel B summarizes the physiology data for the batch cultures. Comparison of the fermentation profiles of the two strains showed that the wild type strain A1560 has a slightly higher maximum specific growth rate than the transformant strain CF1.1, but the maximum activity of extracellular α-amylase enzyme produced was higher for the transformant strain CF1.1 than for the wild type strain A1560. For growth on glucose, the increase was approximately 2.3-fold higher for transformant strain CF1.1, whereas for the maltose medium the enzyme production was approximately 4-fold higher for the transformant strain CF1.1 than for the wild type strain A1560 (see Figure 6.1). These results raised an important question: what are the key players/targets that cause increased level of protein production of transformant strain CF1.1 (compared to wild type strain A1560)? To find out the key players/targets and their functional roles, comparative transcriptome analysis was performed as described in the following section.



(B)

A. oryzae strain	Carbon sources	Sampling time (h)	μ_{max} (h^{-1})	Biomass (gDW/L)	Maximum α-Amylase activity (FAU/ml)
A1560	Glucose	6	0.38±0.01	2.50±0.09	0.68±0.21
	Maltose	7	0.33±0.02	2.24±0.09	1.12±0.11
CF1.1	Glucose	6	0.34±0.01	2.35±0.24	1.53±0.25
	Maltose	8	0.30±0.03	2.42±0.12	4.40±0.18

Figure 6.1 Fermentation profiles of *A. oryzae*
(A) Growth and enzyme activity profiles of two strains (i.e. wild type strain A1560 and transformant strain CF1.1) during growth on glucose and maltose as carbon sources.
(B) Overview of time for sampling for transcriptome analysis, the maximum specific growth rate, biomass concentration at the sampling time and the maximum enzyme activity for the two strains grown on the two different carbon sources. For all values, average values and standard deviations for the three replicates are shown.

6.2 Comparative transcriptome analysis

To identify the key players/targets for α-amylase production in *A. oryzae*, the genome-wide gene expression data obtained from wild type A1560 and transformant CF1.1 cultivations were pairwise compared for each of the two carbon sources (either glucose or maltose). To detect transcriptional changes in response to the strain background, Student's t-test statistics was used to identify significantly different gene expression levels with a p-value cut-off of 0.05. To avoid influence of carbon source, the list of conserved genes that respond to the strain background for each carbon source was overlapped. Figure 6.2 shows 2,560 overlapped genes that were significantly differentially expressed in the two *A. oryzae* strains. Among these 2,560 genes, 1,916 (~75%) were up-regulated genes in the transformant strain CF1.1. Based on gene classification of *A. oryzae* from the Database Of the Genomes Analyzed at NITE (DOGAN) database [71], we found that 474 out of the 2,560 genes were involved in the functional category of protein synthesis and secretion. These results are highly reasonable as α-amylase is a protein and therefore not surprisingly the process of synthesis and secretion of protein is significantly changed in the transformant strain CF1.1. In addition, we also classified the protein synthesis and secretion into sub-functional categories (see Paper 6). Hereby we found many genes involved in RNA processing and translation as well as post-translation modification and secretion process. A number of genes with relative difference in gene expressions between the transformant CF1.1 and the wild type A1560 were also found (e.g genes encoding protein functions involved in energy metabolism, amino acid metabolism, DNA processing and transcription, and cellular development process). The results are shown in Figure 6.2.

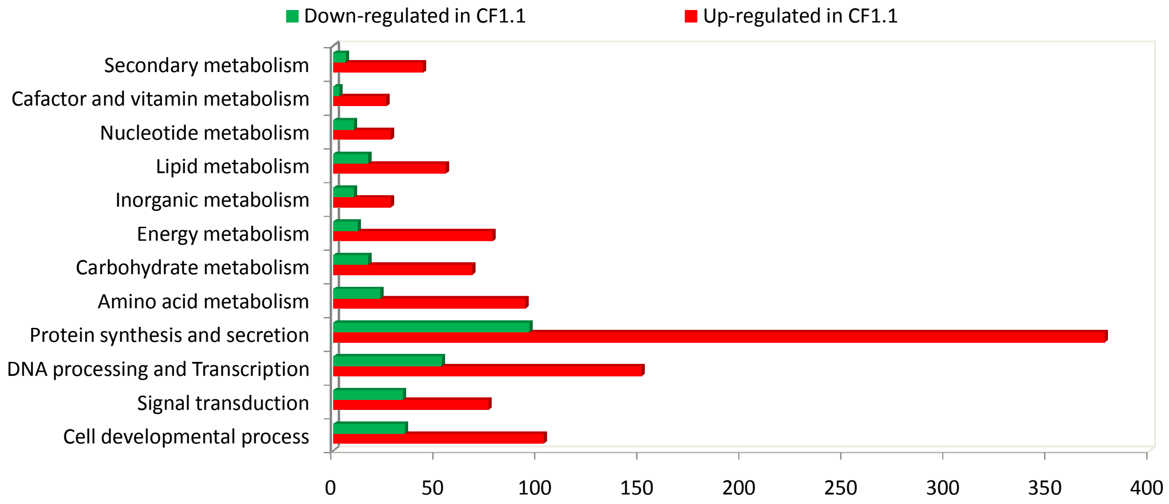


Figure 6.2 Bar graph representing significantly differentially expressed genes between the transformant strain (CF1.1) and the wild type strain (A1560), distributed into different functional categories.

6.2.1 Key metabolites and metabolic subnetworks analysis in response to protein production

The reporter metabolites and subnetworks algorithm was applied to identify key metabolites involved in protein production and to search for highly correlated metabolic subnetworks for the pairwise comparison [95]. This analysis relies on the reconstructed genome-scale metabolic network of *A. oryzae* (Paper 1) [102] and gene expression data. Here, we demonstrated how a metabolic network can be used to map global regulatory responses for protein production in *A. oryzae*. The top 25 high-scoring key metabolites for *A. oryzae* are listed in Table 6.1.

The fact that tRNA (both cytosol and mitochondria tRNA) was identified as a key metabolite is biologically reasonable, as charged tRNAs are precursors for protein synthesis. Of the 25 metabolites, 20 key metabolites are involved in purine and pyrimidine nucleotide biosynthesis, namely 5-phospho-α-ribosyl-1-pyrophosphate (PRPP), mitochondria and cytosol pyrophosphate (PPI), inosine monophosphate (IMP), xanthosine monophosphate (XMP), guanosine monophosphate (GMP), mitochondria and cytosol adenosine monophosphate (AMP), adenosine triphosphate (ATP), deoxy-uridine monophosphate (dUMP), deoxy-guanosine monophosphate (dGMP), guanine, adenine, cytosine, ADP-ribose, 3',5'-cyclic deoxy-adenosine monophosphate (cdAMP), 3',5'-cyclic inosine monophosphate (cIMP), 3',5'-cyclic adenosine monophosphate (cAMP), 3',5'-cyclic guanosine monophosphate (cGMP), and 3',5'-cyclic cytosine monophosphate (cCMP). The results are in agreement with classical molecular biology, where formation of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) is very important in protein synthesis. Besides, we found that ferricytochrome C and ferrocytochrome C are involved in energy metabolism and also nicotinate-D-ribonucleotide is involved in nicotinate and nicotinamide metabolism. In addition to key metabolites, we also identified key enzymes or transporters in response to increased protein production. We performed metabolic subnetworks analysis using the whole reaction set from the reconstructed metabolic network of *A. oryzae* (Paper 1) [102]. Figure 6.3 captures key genes encoding enzymes in nucleotide metabolism (purine and pyrimidine biosynthesis) and key genes encoding enzymes involved in amino acid metabolism that are significantly changed in the metabolic subnetworks identified from pairwise strains comparison in *A. oryzae* (wild type strain A1560 vs. transformant strain CF1.1).

Table 6.1 Reporter metabolites analysis

High α-amylase producer strain (CF1.1) versus reference strain (A1560)	
Key metabolite	**P-value**
Pyrophosphate (PPI)	4.62E-04
Guanosine monophosphate (GMP)	8.81E-04
tRNA	1.02E-03
Pyrophosphate (PPI), mitochondria	2.68E-03
Inosine monophosphate (IMP)	3.98E-03
Adenosine monophosphate (AMP), mitochondria	4.04E-03
Adenosine monophosphate (AMP)	5.02E-03
tRNA, mitochondria	6.85E-03
xanthosine monophosphate (XMP)	9.43E-03
Ferricytochrome C, mitochondria	1.11E-02
Ferrocytochrome C, mitochondria	1.11E-02
Adenosine triphosphate (ATP)	1.22E-02
3',5'-cyclic adenosine monophosphate (cAMP)	1.23E-02
Adenine	1.30E-02
5-phospho-α-ribosyl-1-pyrophosphate (PRPP)	1.52E-02
deoxy-uridine monophosphate (dUMP)	1.63E-02
Nicotinate-D-ribonucleotide	1.67E-02
Cytosine	1.75E-02
deoxy-guanosine monophosphate (dGMP)	2.02E-02
Guanine	2.27E-02
ADP-ribose	2.89E-02
3',5'-cyclic deoxy-adenosine monophosphate (cdAMP)	2.99E-02
3',5'-cyclic inosine monophosphate (cIMP)	2.99E-02
3',5'-cyclic guanosine monophosphate (cGMP)	2.99E-02
3',5'-cyclic cytosine monophosphate (cCMP)	2.99E-02

Reporter metabolite analysis identified metabolites around which the most significant transcriptional changes occur. The algorithm used the pairwise t-test analysis referring to strain background effect as an input. The *P*-value gives a measure of significance and all results with *P*< 0.03 are reported.

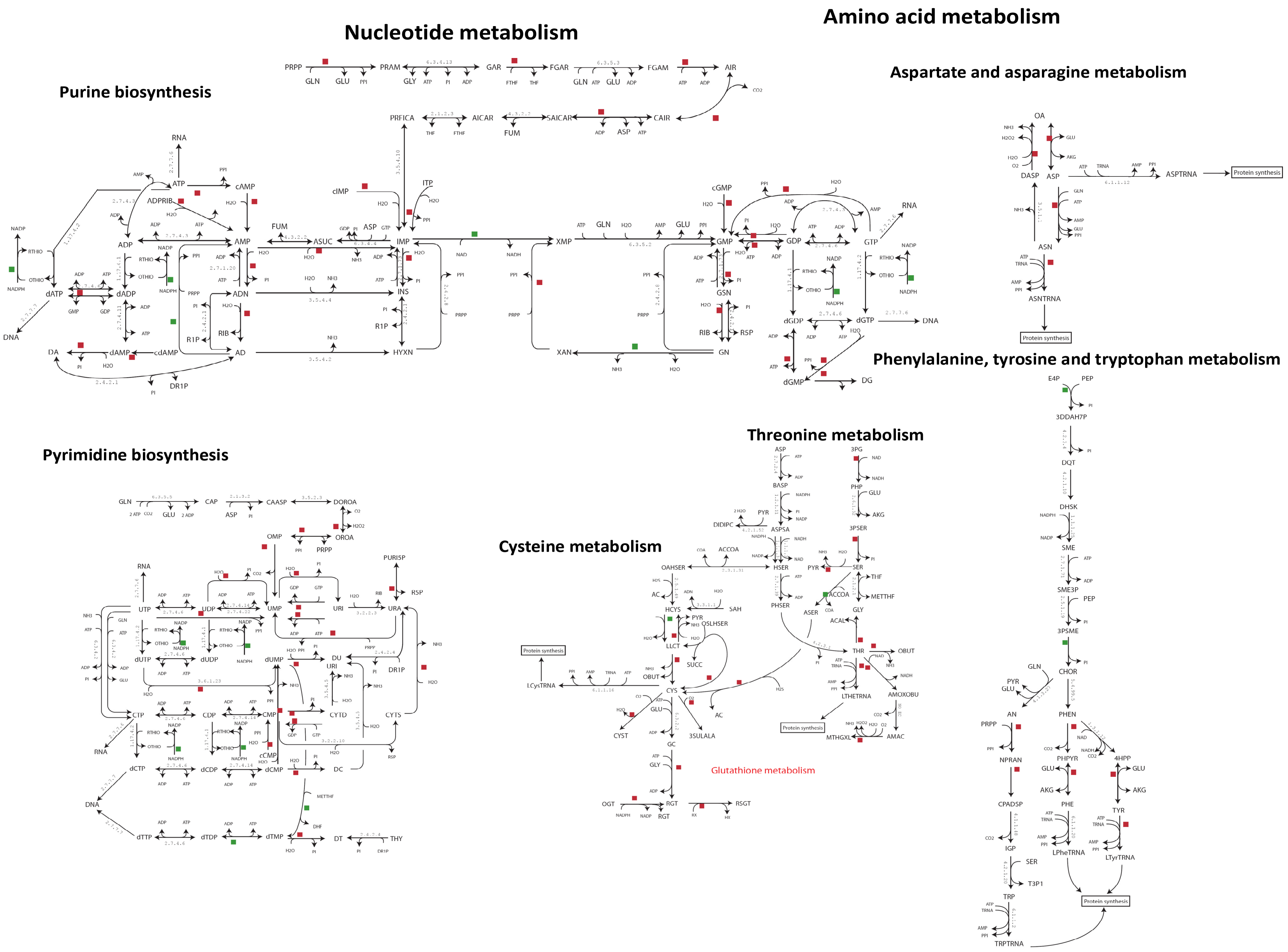


Figure 6.3 Parts of *A. oryzae* metabolism in which genes are significantly changed in the metabolic subnetworks identified from pairwise strains comparison. Red boxes represent up-regulated genes. Green boxes represent down-regulated genes.

6.2.2 Identification of key proteins in response to protein production

In order to find key proteins regulating the global transcriptional response to the increased protein production level, the reporter features algorithm [106] was applied. In this study, the reporter features for key protein identification is based on reconstructed networks covering each protein-protein interaction combined with gene expression data. In order to apply the algorithm, we first needed to reconstruct a protein-protein interaction network of *A. oryzae* (see Methods in Paper 6). Each protein pairs from yeast *S. cerevisiae* BIOGRID database [107] was used as a query interaction and searched against *A. oryzae* genes obtained from pairwise protein sequences comparison (see Methods in Paper 6). Considering only 1:1 orthologous genes of *A. oryzae* and *S. cerevisiae*, we identified 3,514 genes. We searched 140,849 protein pairs of *S. cerevisiae* used as a query interaction database against the 3,514 genes of *A. oryzae* to reconstruct a putative protein-protein interaction network (hypothesis presented by Jonsson *et al*. [108] and as described in Methods (see section Reconstruction of protein-protein interaction network in Paper 6). The reconstructed interaction network of *A. oryzae* contains 2,704 individual proteins with 48,483 putative interactions of protein pairs. In order to identify key proteins, the gene expression data set from comparative transcriptome analysis were combined with the reconstructed network of *A. oryzae.*

Applying the reporter features algorithm with specific thresholds (see Methods in Paper 6), we could identify 33 proteins (see Figure 6.4) that are possible to be the key targets in gene expression regulation in response to increased protein production. These 33 proteins can be divided into five functional groups, namely 4 proteins involved in transcription, 8 proteins involved in RNA processing and translation, 6 proteins involved in the proteasome, 7 proteins involved in post-translation modification and protein secretion, and 8 proteins involved in cell cycle and structure. In the following, we discuss two interesting cases of how increased enzyme production causes a global response in *A. oryzae* and hereby may impact the overall physiology of the organism.

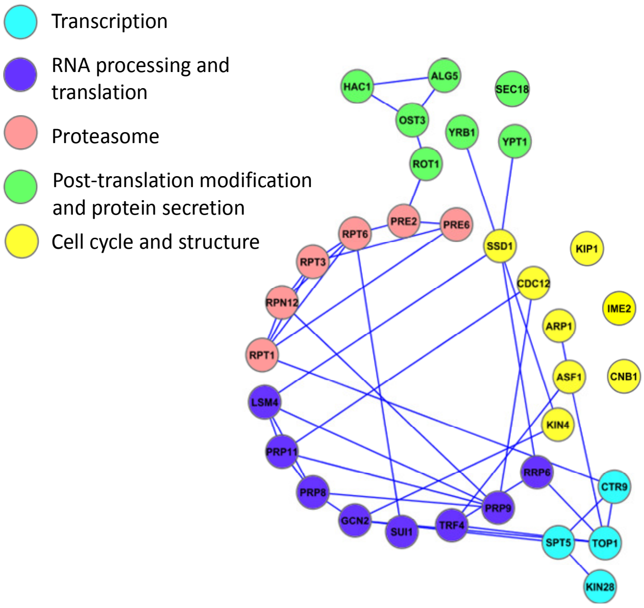


Figure 6.4 An interaction network of 33 key proteins identified as a global response to increased protein production. The connectivity among the proteins (the nodes) is based on the interactions stored at the BIOGRID database [107] of the yeast *S. cerevisiae*. The network was drawn by using Cytoscape [109].

Analysis of general amino acid control

Not surprisingly, the key proteins are involved in process of protein synthesis and secretion since the pairwise t-test analysis from two strains comparison was used as input for the network analysis. The applied reporter features algorithm can identify regulatory hot-spots in bio-molecular interaction networks that are significantly affected in response to different conditions. An important regulatory hot-spot is the *S. cerevisiae* protein kinase GCN2 or the *Aspergillus* homologue CpcC involved in RNA processing and translation process (see Figure 6.4). GCN2/CpcC is known as a sensor for amino acid abundance. It usually enhances the sensitivity of translation of the transcriptional activator GCN4 (as named in *S. cerevisiae*) or CpcA (as named in *Aspergillus* species), leading to transcriptional induction of multiple genes encoding amino acid biosynthetic enzymes upon amino acid starvation [110]. In *S. cerevisiae*, this phenomenon is called "general control of amino acids", whereas in *Aspergillus* species, this event is named "cross pathway control of amino acid biosynthesis". Evidently, if the presence of the general control/the cross pathway control of amino acid biosynthesis occurs via the GCN2/CpcC protein kinase, then increased expression of multiple enzymes in different amino acid biosynthetic pathways is found [110]. Since the reporter features analysis identified GCN2/CpcC as one of the regulatory hot-spots, we hypothesize that the cross pathway control of amino acid biosynthesis in *A. oryzae* is likely to occur in connection with the increased protein production.

To test our hypothesis, we used the amino acid biosynthetic enzymes that are known to be under general amino acid control in yeast, fungi and bacteria [110, 111] as a query list and searched against our comparative transcriptome data between the wild type strain A1560 and the transformant strain CF1.1 in order to see if we could find these enzymes. As expected, we found several enzymes that are targets for cross pathway control of amino acid biosynthesis in *A. oryzae*, and this indicates that amino acid starvation is likely to occur in the transformant CF1.1 as most of the genes encoding amino acid enzymes were up-regulated, such as multiple enzymes in tyrosine, tryptophan, ariginine, histidine, lysine, isoleucine, valine, and general aromatic amino acids biosynthesis. A list of the enzymes subject to the cross pathway control is given in Table 6.2.

Analysis of occurrence of Unfolded Protein Response (UPR)

HAC1 was also identified as one of the key proteins in the protein-protein interaction network. This protein is a key regulatory component of UPR pathway that is activated in response to poor protein folding that leads to block in the protein secretion pathway, which is obviously an important step for protein production. In eukaryotic cells, the synthesized proteins are folded and assembled in the Endoplasmic Reticulum (ER). The ER provides an oxidising environment in which protein folding is assisted by a number of molecular chaperones and folding enzymes. Protein folding in the ER can be compromised by several endogenous and exogenous factors such as changing environmental conditions or genetic perturbations. This event leads to the accumulation of unfolded proteins within the ER and this lead to ER stress conditions. To maintain homeostasis of ER functions, the cell reacts to the accumulation of unfolded proteins in the ER by inducing a pathway known as the UPR. The UPR pathway has been studied in *A. oryzae* [112] and four key components of this pathway are: (1) The HAC1 protein is a transcriptional activator that up-regulates the transcription of various target-genes of the UPR pathway; (2) The Bip protein is a chaperone of the HSP70 class that plays an important role in the UPR; (3) The Pdi is a luminal ER enzyme that catalysts the mechanism of disulfide bond formation; (4) The Ppi is an enzyme of catalyzing the cis-trans isomerisation of a peptide bond on the N-terminal side of proline residues in polypeptides.

Since the reporter features algorithm revealed that HAC1 is an important protein in response to the protein secretion, we reasoned that the UPR pathway is activated in the α-amylase over-producing strain CF1.1. To test our assumption, three UPR-relevant genes known to be controlled by HAC1 in *A. oryzae* were selected, and our transcriptome results showed that the UPR pathway is very likely to be active, since the three target genes in the UPR pathway were up-regulated in the transformant strain CF1.1, namely AO090003000257 gene-encoding Bip protein, AO090001000733 gene-encoding Pdi protein, AO090023000811 gene-encoding Ppi protein.

6.3 Analysis of amino acid consumptions

We performed comparative analysis of amino acid consumptions in terms of flux calculations for synthesis of protein content in biomass and α-amylase between the two strains (wild type A1560 and transformant CF1.1). Based on the amino acid composition in biomass protein and α-amylase, the specific consumption rate of each of the 20 amino acids (as mmol/gDW/h) was calculated for the two strains. In terms of flux calculation (see more information in Paper 6), we found that in particularly four amino acids are drained substantially more in the CF1.1 strain due to the over-production of α-amylase, and the biosynthesis of these amino acids could be possible targets for further increasing the protein production by the transformant strain CF1.1. These four candidate amino acids are tyrosine, aspartate, cysteine and threonine.

6.4 Integrated data analysis as a scaffold for diagnosis of industrial enzyme fermentation process

We demonstrated that by performing integrated data analysis, i.e., genomes, transcriptomes, interactomes (protein–protein interaction), metabolic networks and flux calculations, it is possible to identify key metabolic/regulatory pathways involved in protein production. From comparative transcriptome analysis of two strains (wild type A1560 and transformant CF1.1), we identified that several key processes involved in protein synthesis and secretion are affected at the transcriptional level in response to high-level protein production. We found key metabolites and key enzymes in nucleotide metabolism (purine and pyrimidine biosynthesis) used for synthesis of DNA and RNA (i.e. mRNA, tRNA and rRNA). In addition, we found several amino acid biosynthetic enzymes whose genes are significantly changed in the metabolic subnetworks with respect to protein production, such as tyrosine, aspartate, cysteine and threonine (see Figure 6.3). The results obtained from flux calculations support that these four amino acids play the key roles for increased α-amylase production. To find out key regulatory steps, an interaction network (protein-protein interaction) of *A. oryzae* was reconstructed. The reporter features algorithm was applied to this network and hereby 33 proteins were identified that are possible key targets regulating gene expression in response to increased level of protein production. 2 proteins out of the 33 proteins, namely GCN2/CpcC and HAC1, suggest that the limiting step for production of α-amylase is the control of general amino acids upon starvation and the lack of folding capacity in the ER resulting in an UPR. In addition to these 2 proteins, the other key proteins also imply that other steps in protein production are limited, such as transcription, RNA processing and translation, post-translational modification, and proteasome degradation.

From this study, one can obtain a better understanding of the complex relationship of biological processes in response to high level protein production. Our work therefore revealed the key players/targets (i.e. genes, enzymes, proteins, metabolites and pathways) in response to α-amylase production that may lead us to further improvements of industrial enzyme fermentation processes. We believe that the integrated data analysis can be a scaffold for identifying possible limiting steps for protein production and hereby strategies for strain improvement or process optimization of *A. oryzae* in relation to industrial enzyme production.

Table 6.2 List of enzymes subject to general amino acid control/cross pathway control targets (p-value<0.05)

Gene name	Enzyme name	Common name	Up/down
Tryptophan biosynthesis			
AO090012000581	Anthranilate synthase (Multifunctional protein)	TRP2	Up
AO090012000581	indole-3-glycerol-phosphate synthase	TRP3	Up
AO090012000581	phosphoribosylanthranilate isomerase	TRP1	Up
AO090003001011	Anthranilate phosphoribosyltransferase	TRP4	Up
AO090005001315	Tryptophan synthase beta chain	TRP5	Up
Arginine biosynthesis			
AO090026000498	Acetylglutamate kinase	ARG2	Up
AO090026000498	Acetylglutamate synthase	ARG6	Up
AO090020000418	Argininosuccinate lyase	ARG4	Up
AO090701000214	Multifunctional pyrimidine synthesis protein CAD (includes carbamoyl-phophate synthetase, aspartate transcarbamylase, and glutamine amidotransferase)	CPA1	Down
Histidine biosynthesis			
AO090206000105	Histidinol-phosphatase	HIS2	Up
AO090012000450	Histidinol phosphate aminotransferase	HIS5	Up
Lysine biosynthesis			
AO090026000245	Transaminases (Aromatic aminotransferases)		Up
AO090001000516	Alpha-aminoadipate reductase and related enzymes	LYS2	Up
Isoleucine and valine biosynthesis			
AO090166000076	Acetolactate synthase, large subunit	ILV2	Up
Leucine biosynthesis			
AO090010000218	Isoleucyl-tRNA synthetase	ILS1	Up
Tyrosine biosynthesis			
AO090003000301	Tyrosine decarboxylase		Up
AO090001000383	Tyrosinase		Up

7. Conclusions and Future perspectives

The work presented in this thesis deals with development of SYStems biology tools and BIoinformatics methods to analyze multi-level OMICS data. SYSBIOMICS is a novel term defined from this Ph.D. study. We applied SYSBIOMICS approach aiming at the construction of a genome-scale metabolic model of *A. oryzae*. This model was further used for high-throughput omics data analysis from industrial fermentations for investigating cellular metabolism and regulation in *A. oryzae* and mainly comparative analysis with *A. niger* and *A. nidulans*. Beyond regulation and metabolism, the work also contributed towards integrated data analysis for studying genotype-phenotype relationships and further identifying possible key players/targets for improvements of industrial enzyme production. Based on the SYSBIOMICS approach, we believe that it will find wide applications in industrial biotechnology and life science in the future.

The main conclusions of this work can be summarized in three points as follows:

- **Improved annotation through development of metabolic model of *A. oryzae* at a genome-scale**
  With a complete genome sequence of *A. oryzae* becoming available and top-down reconstruction of a metabolic network developing fast and now in wide use, it has opened possibilities for studying the cellular physiology of this fungus on a systematic level. To explore this, we developed bioinformatics methods for improved annotation of the genome sequence of *A. oryzae*. We performed gene discovery and validation by using an EST library as well as assignment of protein function by using advanced bioinformatics algorithms. The resulting improved annotation was used to reconstruct the metabolic network leading to a genome scale metabolic modeling of *A. oryzae*. With the case study (**Chapter 4**), we present a framework for the systems biology paradigm to link a component (gene), to its function, further to a biological network, and finally to a functional model that can describe the physiology of the organism. The model serves as an important resource for gaining further insight into our understanding of the *A. oryzae* physiology and it can also be applied as a scaffold for integrated data analysis.

- **Investigation of cellular metabolism and regulation in aspergilli**

  Carbon metabolism and its regulation is one of the most intensively studied in many different organisms and it represents a complex metabolic system. Here a case study (**Chapter 5**), on how the carbon source affects transcription of genes, is presented. To investigate this principle, an Affymetrix GeneChip was designed for transcriptome analysis of three *Aspergillus* species, and the DNA microarray was used for transcriptome analysis of mainly *A. oryzae* but with cross-species analysis of expression data from *A. niger* and *A. nidulans* on four different carbon sources (i.e. glucose, xylose, maltose and glycerol). With this case study, we present the developed tool for transcriptome analysis of three *Aspergillus* species and our methodology for conducting cross-species evolutionary studies within a genus using comparative genomics and transcriptomics.

- **Protein production by *A. oryzae***
  With integrated data analysis (i.e. genomes, transcriptomes, metabolic networks, interactomes and fluxes), we performed comparative analysis of high and low level α-amylase production in *A. oryzae* in order to identify key players/targets involved in

high-level protein production. From this case study (**Chapter 6**), some possible key players/targets in protein production were identified and these can be used for diagnosis of industrial enzyme production in fermentation process and for further improvement of industrial strains used for protein production.

Acknowledgements

This dissertation serves as a record for the work carried out during my Ph.D. study under Department of Systems Biology at the Technical University of Denmark (DTU), Novozymes A/S, Denmark, and under Department of Biological and Chemical Engineering at Chalmers University of Technology (Chalmers), Sweden in the period between 2006 and 2009. This Ph.D. work was financial supported by first DTU and Novozymes Bioprocess Academy (NBA) and later by the Chalmers Foundation.

First of all I am very grateful to Professor Jens Nielsen, who strongly inspired me and allowed for pleasant scientific discussions. There are many good opportunities which I have been given during my four years at DTU and Chalmers. He encouraged me to write this dissertation and he has made a number of wonderful suggestions. Since I decided to start my Ph.D. study, I very much appreciated everything he has offered. I would also like to thank my external co-supervisors. First, Kim Hansen from Novozymes for his advice and many good ideas and discussions related to fermentation processes. His fungal physiology department at Novozymes was a very nice place for me to work and carried out my experiments. I always enjoyed it and learned a lot. I would like to thank Dr. Dina Petranovic for my internal co-supervisor. I am very indebted to Dr. Peter Olsen and Steen Krogsgaard from Novozymes for their excellent assistance with bioinformatics and DNA microarray data analysis, respectively. I very much benefited from their guidance and support, especially when I worked at the department of bioinformatics at Novozymes which was a very cool place to practice my computational programming skills.

I would like to give a special thank to all of my wonderful close collaborators. Mikael Andersen at DTU and Margarita Salazar at Chalmers, who worked with me to study *Aspergillus* systems biology since I started my Ph.D. I am very happy to work with many of my close collaborators, who directly or indirectly contributed to my work, such as Manny, Roberto, Sergio, Intawat, Rawisara, Marija, Subir, Goutham and Verena . I would also like to thank the rest of the colleagues from DTU and Chalmers, who I worked with some time of my study. I would also like to particular thank all the technicians at DTU and Novozymes, who assisted me since starting the experiment and introduction of equipments, namely Lone Vuholm, Anne Jensen, Lene Christiansen, and Pia Friis for their experience. I also thank the administrative staff Trine Bro, Birgitte Karsbol, and Jytte Laursen at DTU and Erica Dahlin at Chalmers for taking care of many practical things.

I also deeply appreciate my dearest friends for sharing life with me in Denmark and Sweden, such as Kanchana, Arwit and Panaram’s family, Swee, Xiaole, Maya, Songsak, Rawisara, Pramote, Kanokarn, and Intawat. My close foreign friends share experience and enjoyable time together in Göteborg during the last stage of my PhD, especially Marta, Gionata**,** Martina, Andrea, Rasmus, Tobias, Jie, Siavash, and Christoph. Lastly, a very big thanks to my parents and families as well as friends in Thailand for their love, support, encouragement and understanding at all times.

Wanwipa Vongsangnak

November 2009

References

1. Fincham JRS: **The Genus *Aspergillus* - from Taxonomy and Genetics to Industrial Applications**. *Nature* 1994, **371**:116-117.
2. Gugnani HC: **Ecology and taxonomy of pathogenic Aspergilli**. *Front Biosci* 2003, **8**:S346-S357.
3. Baker SE, Bennett JW: **in The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods, eds Osmani SA, Goldman GH (CRC Press, Boca Raton, FL)**. 2008:3-13.
4. Tailor MJ, Richardson T: **Applications of microbial enzymes in food systems and in biotechnology**. *Adv Appl Microbiol* 1979, **25**:7-35.
5. FAO/WHO: **Committee on food additives** *World Health Organization Technical Report Series, Geneva* 1987, **31**.
6. Abe K, Gomi K, Hasegawa F, Machida M: **Impact of *Aspergillus oryzae* genomics on industrial production of metabolites**. *Mycopathologia* 2006, **162**:143-153.
7. Archer DB: **Filamentous fungi as microbial cell factories for food use**. *Curr Opin Biotechnol* 2000, **11**:478-483.
8. Kubicek CP, Rohr M: **Citric Acid Fermentation**. *Crc Cr Rev Biotechn* 1986, **3**:331-373.
9. Pedersen H, Beyer M, Nielsen J: **Glucoamylase production in batch, chemostat and fed-batch cultivations by an industrial strain of *Aspergillus niger***. *Appl Microbiol Biotechnol* 2000, **53**(3):272-277.
10. Martinelli SD: ***Aspergillus nidulans* as an experimental organism**. *In Aspergillus: 50 years* 1994:33-58.
11. Chang PK, Matsushima K, Takahashi T, Yu JJ, Abe K, Bhatnagar D, Yuan GF, Koyama Y, Cleveland TE: **Understanding nonaflatoxigenicity of *Aspergillus sojae*: a windfall of aflatoxin biosynthesis research**. *Appl Microbiol Biotechnol* 2007, **76**(5):977-984.
12. Manzoni M, Rollini N: **Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs**. *Appl Microbiol Biotechnol* 2002, **58**(5):555-564.
13. Bonnarme P, Gillet B, Sepulchre AM, Role C, Beloeil JC, Ducrocq C: **Itaconate Biosynthesis in *Aspergillus terreus***. *J Bacteriol* 1995, **177**(12):3573-3578.
14. Johansen CL, Coolen L, Hunik JH: **Influence of morphology on product formation in *Aspergillus awamori* during submerged fermentations**. *Biotechnol Progr* 1998, **14**(2):233-240.
15. Denning DW: **Invasive aspergillosis**. *Clinical Infectious Diseases* 1998, **26**(4):781-803.
16. Sabater-Vilar M, Maas RFM, De Bosschere H, Ducatelle R, Fink-Gremmels J: **Patulin produced by an *Aspergillus clavatus* isolated from feed containing malting residues associated with a lethal neurotoxicosis in cattle**. *Mycopathologia* 2004, **158**(4):419-426.
17. Gori S, Pellegrini G, Filipponi F, Della Capanna S, Biancofiore G, Mosca F, Lofaro A: **Pulmonary aspergillosis caused by *Neosartorya fischeri* (*Aspergillus fischerianus*) in a liver transplant recipient**. *J Med Mycol* 1998, **8**(2):105-107.
18. Roze LV, Beaudry RM, Arthur AE, Calvo AM, Linz JE: ***Aspergillus* volatiles regulate aflatoxin synthesis and asexual sporulation in *Aspergillus parasiticus***. *Appl Environ Microbiol* 2007, **73**(22):7268-7276.

19. Hedayati MT, Pasqualotto AC, Warn PA, Bowyer P, Denning DW: ***Aspergillus flavus:* human pathogen, allergen and mycotoxin producer**. *Microbiology-Sgm* 2007, **153**:1677-1692.

20. Prathumpai W: **Enzyme production in aspergilli on different carbon sources**. *PhD thesis* 2003:Technical Univeristy of Denmark, Department of Systems Biology.

21. Vongsangnak W, Nielsen J: ***Aspergillus*: Molecular Biology and Genomics**. *Horizontal press* 2009.

22. Lesk AM: **Database Annotation in Molecular Biology: Principles and Practice** *John Wiley & Sons, Chichester* 2005:299.

23. Hamer L: **From genes to genomes: Sequencing of filamentous fungal genomes**. *Fungal Genet Biol* 1997, **21**(1):8-10.

24. Goffeau A: **The yeast genome**. *Pathol Biol* 1998, **46**(2):96-97.

25. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-567.

26. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J *et al*: **Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae***. *Nature* 2005, **438**:1105-1115.

27. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: **Genomics of the fungal kingdom: Insights into eukaryotic biology**. *Genome Res* 2006, **16**(2):304-304.

28. Birren BW: **Fungal genome initiative—a white paper for fungal comparative genomics**. 2003.

29. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C *et al*: **Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus***. *Nature* 2006, **439**:502-502.

30. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K *et al*: **Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88**. *Nat Biotechnol* 2007, **25**(2):221-231.

31. Baker SE: ***Aspergillus niger* genomics: Past, present and into the future**. *Med Mycol* 2006, **44**:S17-S21.

32. Machida M: **Progress of *Aspergillus oryzae* genomics**. *Adv Appl Microbiol* 2002, **51**:81-106.

33. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto KI, Arima T, Akita O, Kashiwagi Y *et al*: **Genome sequencing and analysis of *Aspergillus oryzae***. *Nature* 2005, **438**:1157-1161.

34. Birren B, Denning D, Nierman B: **Comparative analysis of an emerging fungal pathogen, *Aspergillus terreus***. *White paper* 2004.

35. Askenazi M, Driggers EM, Holtzman DA, Norman TC, Iverson S, Zimmer DP, Boers ME, Blomquist PR, Martinez EJ, Monreal AW *et al*: **Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains**. *Nat Biotechnol* 2003, **21**(2):150-156.

36. Payne GA, Nierman WC, Wortman JR, Pritchard BL, Brown D, Dean RA, Bhatnagar D, Cleveland TE, Machida M, Yu J: **Whole genome comparison of *Aspergillus flavus* and *A. oryzae***. *Med Mycol* 2006, **44**:S9-S11.

37. Yu. J., Cleveland. T., Nierman. W., Bennett. J.: ***Aspergillus flavus* genomics: gateway to human and animal health, food safety, and crop resistance to diseases**. *Rev Iberoam Micol* 2005, **22**:194-202.

38. Fedorova N, Khaldi N, Joardar V, Maiti R, Amedeo P, Anderson M, Crabtree J, Silva J, Badger J, Albarraq A *et al*: **Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus***. *PLoS Genet* 2008, **4**(e1000046).
39. Wortman JR, Fedorova N, Crabtree J, Joardar V, Maiti R, Haas BJ, Amedeo P, Lee E, Angiuoli SV, Jiang B *et al*: **Whole genome comparison of the *A. fumigatus* family**. *Med Mycol* 2006, **44**:S3-S7.
40. Andersen MR, Nielsen J: **Current status of systems biology in Aspergilli**. *Fungal Genet Biol, In press* 2008.
41. Jones MG: **The first filamentous fungal genome sequences: *Aspergillus* leads the way for essential everyday resources or dusty museum specimens?** *Microbiology-Sgm* 2007, **153**:1-6.
42. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nature Genetics* 1999, **21**:33-37.
43. Holloway AJ, van Laar RK, Tothill RW, Bowtell DDL: **Options available - from start to finish - for obtaining data from DNA microarrays**. *Nature Genetics* 2002, **32**:481-489.
44. Duggan DJ, Bittner M, Chen YD, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nature Genetics* 1999, **21**:10-14.
45. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: **Microarrays: biotechnology's discovery platform for functional genomics**. *Trends in Biotechnology* 1998, **16**(7):301-306.
46. Van Hal NLW, Vorst O, Van Houwelingen A, Kok EJ, Peijnenburg A, Aharoni A, Van Tunen AJ, Keijer J: **The application of DNA microarrays in gene expression analysis**. *J Biotechnol* 2000, **78**(3):271-280.
47. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680-686.
48. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: **Yeast microarrays for genome wide parallel genetic and gene expression analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(24):13057-13062.
49. Breakspear A, Momany M: **The first fifty microarray studies in filamentous fungi**. *Microbiology-Sgm* 2007, **153**:7-15.
50. Kim YY, Nandakumar MP, Marten MR: **Proteomics of filamentous fungi**. *Trends in Biotechnology* 2007, **25**(9):395-400.
51. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R: **Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae***. *Molecular & Cellular Proteomics* 2002, **1**(4):323-333.
52. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast**. *Molecular and Cellular Biology* 1999, **19**(3):1720-1730.
53. Bader GD, Hogue CWV: **Analyzing yeast protein-protein interaction data obtained from different sources**. *Nature Biotechnology* 2002, **20**(10):991-997.
54. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**(5537):2101-2105.
55. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nature Biotechnology* 1999, **17**(10):994-999.
56. Auerbach D, Thaminy S, Hottiger MO, Stagljar I: **The post-genomic era of interactive proteomics: Facts and perspectives**. *Proteomics* 2002, **2**(6):611-623.

57. Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J: **Mass spectrometry in metabolome analysis**. *Mass Spectrometry Reviews* 2005, **24**(5):613-646.

58. Baxevanis AD, Ouellette BFF: **Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins**. *Wiley, John and Sons, Incorporated* 2004, **3**:540.

59. Jewett MC, Nielsen J: **Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae***. *FEMS Yeast Research* 2008, **8**:122-131.

60. Wiechert W: **Modeling and simulation: tools for metabolic engineering**. *J Biotechnol* 2002, **94**(1):37-63.

61. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network**. *Genome Res* 2003, **13**:244-253.

62. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, Nielsen J, Bhumiratana S: **The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism**. *Bmc Systems Biology* 2008, **2**:15.

63. Patil KR, Akesson M, Nielsen J: **Use of genome-scale microbial models for metabolic engineering**. *Curr Opin Biotechnol* 2004, **15**(1):64-69.

64. David H, Akesson M, Nielsen J: **Reconstruction of the central carbon metabolism of *Aspergillus niger***. *Eur J Biochem* 2003, **270**:4243-4253.

65. Melzer G, Dalpiaz A, Grote A, Kucklick M, Göcke Y, Jonas R, Dersch P, Francolara E, Nörtemann B, Hempel D: **Metabolic flux analysis using stoichiometric models for *Aspergillus niger*: comparison under glucoamylase-producing and non-producing conditions**. *J Biotechnol* 2007, **132**:405-417.

66. David H, Özçelik İ, Hofmann G, Nielsen J: **Analysis of *Aspergillus nidulans* metabolism at the genome-scale**. *BMC Genomics* 2008, **9**.

67. Andersen MR, Nielsen ML, Nielsen J: **Metabolic model integration of the bibliome, genome,metabolome and reactome of *Aspergillus niger***. *Mol Syst Biol* 2008, **4**:178.

68. David H, Hofmann G, Oliveira AP, Jarmer H, Nielsen J: **Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans***. *Genome Biol* 2006, **7**(11).

69. ***Aspergillus flavus* Gene Index** [http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=a_flavus].

70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *J Mol Biol* 1990, **215**:403-410.

71. ***Aspergillus oryzae* RIB40 database**.[http://www.bio.nite.go.jp/ngac/e/rib40-e.html].

72. ***Aspergillus nidulans* genome database** [www.broad.mit.edu/annotation/genome/aspergillus_nidulans].

73. ***Aspergillus fumigatus* genome database** [www.sanger.ac.uk/Projects/A_fumigatus/].

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

75. Fisk DG, Ball CA, Dolinski K, Engel SR, Hong EL, Issel-Tarver L, Schwartz K, Sethuraman A, Botstein D, Cherry JM: ***Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis**. *Yeast* 2006, **23**:857-865.

76. **Non-redundant protein database** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/].

77. **KEGG pathway database**.[www.kegg.com].

78. **BioCyc database** [http://biocyc.org/server.html].

79. **Swiss-Prot database** [www.expasy.ch/sprot/]

80. Guda C, Subramaniam S: **TARGET: a new method for predicting protein subcellular localization in eukaryotes**. *Bioinformatics* 2005, **21**:3963-3969.

81. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization**. *Proteins: Struct, Funct, Bioinf* 2006, **64**:643-651.

82. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**:755-763.

83. Altschul S, Madden T, Schaffer A, Zhang JH, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs**. *FASEB J* 1998, **12**:A1326-A1326.

84. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

85. **COG database** [www.ncbi.nih.gov/COG].

86. Carlsen M, Nielsen J: **Influence of carbon source on alpha-amylase production by *Aspergillus oryzae***. *Appl Microbiol Biotechnol* 2001, **57**:346-349.

87. Pedersen H, Carlsen M, Nielsen J: **Identification of enzymes and quantification of metabolic fluxes in the wild type and in a recombinant *Aspergillus oryzae* strain**. *Applied and Environmental Microbiology* 1999, **65**:11-19.

88. Cvijovic M, Olivares R, Agren R, Dahr N, Vongsangnak W, Nookaew I, Nielsen J: **BioMet Toolbox: A toolbox for systematic analysis of metabolism**. *Submitted*.

89. Klein CJL, Olsson L, Ronnow B, Mikkelsen JD, Nielsen J: **Alleviation of glucose repression of maltose metabolism by MIG1 disruption in *Saccharomyces cerevisiae***. *Applied and Environmental Microbiology* 1996, **62**(12):4441-4449.

90. Novak S, Zechner-Krpan V, Maric V: **Regulation of maltose transport and metabolism in *Saccharomyces cerevisiae***. *Food Technology and Biotechnology* 2004, **42**(3):213-218.

91. Needleman RB, Kaback DB, Dubin RA, Perkins EL, Rosenberg NG, Sutherland KA, Forrest DB, Michels CA: ***MAL6* of *Saccharomyces*: A complex genetic locus containing three genes required for maltose fermentation**. *Proc Natl Acad Sci USA* 1984, **81**:2811-2815.

92. Andersen MR, Vongsangnak W, Panagiotou G, Margarita PS, Lehmann L, Nielsen J: **A tri-species *Aspergillus* microarray - advancing comparative transcriptomics**. *Proc Nat Acad Sci USA* 2008, **105**:4387-4392.

93. Sinha AU, Meller J: **Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms**. *BMC Bioinformatics* 2007, **8**.

94. Yuan XL, van der Kaaij RM, van den Hondel C, Punt PJ, van der Maarel M, Dijkhuizen L, Ram AFJ: ***Aspergillus niger* genome-wide analysis reveals a large number of novel alpha-glucan acting enzymes with unexpected expression profiles**. *Mol Genet Genomics* 2008, **279**(6):545-561.

95. Patil KR, Nielsen J: **Uncovering transcriptional regulation of metabolism by using metabolic network topology**. *Proc Natl Acad Sci USA* 2005, **102**(8):2685-2689.

96. Young ET, Dombek KM, Tachibana C, Ideker T: **Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8**. *J Biol Chem* 2003, **278**(28):26146-26158.

97. Blomberg A, Adler L: **Physiology of osmotolerance in fungi**. *Adv Microb Physiol* 1992, **33**:145-212.

98. Das HK, Baez ML: **ADR1 interacts with a down-stream positive element to activate PS1 transcription**. *Front Biosci* 2008, **13**:3439-3447.

99. Schulze A, Downward J: **Navigating gene expression using microarrays - a technology review**. *Nature Cell Biology* 2001, **3**(8):E190-E195.

100. Altman RB, Raychaudhuri S: **Whole-genome expression analysis: challenges beyond clustering**. *Current Opinion in Structural Biology* 2001, **11**(3):340-347.

101. Shen YQ, Burger G: **Plasticity of a key metabolic pathway in fungi**. *Functional & Integrative Genomics* 2009, **9**(2):145-151.

102. Vongsangnak W, Olsen P, Hansen H, Krogsgaard S, Nielsen J, : **Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae***. *BMC Genomics* 2008, **9**.

103. Hynes MJ, Murray SL, Duncan A, Khew GS, Davis MA: **Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus *Aspergillus nidulans***. *Eukaryotic Cell* 2006, **5**(5):794-805.

104. Chamalaun-Hussey N: **Characterization of DNA-binding by the CreA protein of *Aspergillus nidulans***. *PhD thesis* 1996:The university of Adelaide, Department of Genetics.

105. Abe K, Gomi K: **Food products fermented by *Aspergillus oryzae***. In: *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods.* Edited by Osmani SA, Goldman GH. CRC Press; 2007: 428-438.

106. Oliveira AP, Patil KR, Nielsen J: **Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks**. *BMC Systems Biology* 2008, **2**:16.

107. **BIOGRID database**.[http://www.thebiogrid.org/].

108. Jonsson PF, Cavanna T, Zicha D, Bates PA: **Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis**. *BMC Bioinformatics* 2006, **7**:12.

109. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. In: *3rd International Conference on Systems Biology 2002: Dec 13-15 2002; Stockholm, Sweden*: Cold Spring Harbor Lab Press, Publications Dept; 2002: 2498-2504.

110. Hinnebusch AG: **Translational Regulation of GCN4 and the General Amino Acid Control of Yeast**. *Annu Rev Microbiol* 2005, **59**:407–450.

111. Niederberger P, Miozzari G, Hutter R: **Biological role of the general control of amino acid biosynthesis in *Saccharomyces cerevisiae***. *Molecular and Cellular Biology* 1981, **1**(7):584-593.

112. Ye L, Pan L: **A Comparison of the Unfolded Protein Response in Solid-State with Submerged Cultures of *Aspergillus oryzae***. *Bioscience Biotechnology and Biochemistry* 2008, **72**(11):2998-3001.

Paper 1

Improved annotation through genome-scale metabolic modeling
of *Aspergillus oryzae*

Wanwipa Vongsangnak, Peter Olsen, Kim Hansen,
Steen Krogsgaard and Jens Nielsen

BMC Genomics 2008, 9:245

BMC Genomics

Research article

Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*

Wanwipa Vongsangnak[1,3], Peter Olsen[2], Kim Hansen[2], Steen Krogsgaard[2] and Jens Nielsen*[1,3]

Address: [1]Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark, [2]Novozymes A/S, DK-2880 Bagsværd, Denmark and [3]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Email: Wanwipa Vongsangnak - wanwipa@chalmers.se; Peter Olsen - pbo@novozymes.com; Kim Hansen - kimh@novozymes.com; Steen Krogsgaard - stkh@novozymes.com; Jens Nielsen* - nielsenj@chalmers.se

* Corresponding author

Published: 23 May 2008

BMC Genomics 2008, **9**:245 doi:10.1186/1471-2164-9-245

Received: 5 April 2008
Accepted: 23 May 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/245

Abstract

Background: Since ancient times the filamentous fungus *Aspergillus oryzae* has been used in the fermentation industry for the production of fermented sauces and the production of industrial enzymes. Recently, the genome sequence of *A. oryzae* with 12,074 annotated genes was released but the number of hypothetical proteins accounted for more than 50% of the annotated genes. Considering the industrial importance of this fungus, it is therefore valuable to improve the annotation and further integrate genomic information with biochemical and physiological information available for this microorganism and other related fungi. Here we proposed the gene prediction by construction of an *A. oryzae* Expressed Sequence Tag (EST) library, sequencing and assembly. We enhanced the function assignment by our developed annotation strategy. The resulting better annotation was used to reconstruct the metabolic network leading to a genome scale metabolic model of *A. oryzae*.

Results: Our assembled EST sequences we identified 1,046 newly predicted genes in the *A. oryzae* genome. Furthermore, it was possible to assign putative protein functions to 398 of the newly predicted genes. Noteworthy, our annotation strategy resulted in assignment of new putative functions to 1,469 hypothetical proteins already present in the *A. oryzae* genome database. Using the substantially improved annotated genome we reconstructed the metabolic network of *A. oryzae*. This network contains 729 enzymes, 1,314 enzyme-encoding genes, 1,073 metabolites and 1,846 (1,053 unique) biochemical reactions. The metabolic reactions are compartmentalized into the cytosol, the mitochondria, the peroxisome and the extracellular space. Transport steps between the compartments and the extracellular space represent 281 reactions, of which 161 are unique. The metabolic model was validated and shown to correctly describe the phenotypic behavior of *A. oryzae* grown on different carbon sources.

Conclusion: A much enhanced annotation of the *A. oryzae* genome was performed and a genome-scale metabolic model of *A. oryzae* was reconstructed. The model accurately predicted the growth and biomass yield on different carbon sources. The model serves as an important resource for gaining further insight into our understanding of *A. oryzae* physiology.

Background

A. oryzae is a member of the diverse group of aspergilli that includes species that are important microbial cell factories, as well as species that are human and plant pathogens [1]. *A. oryzae* has been used safely in the fermentation industry for hundreds of years in the production of soy sauce, miso and sake. Today *A. oryzae* is also used for production of a wide range of different fungal enzymes such as α-amylase, glucoamylase, lipase and protease and it is regarded as an ideal host for the synthesis of proteins of eukaryotic origin [1]. In the post genome-sequencing era, various high-throughput technologies have been developed to characterize biological systems on the genome-scale [2]. Discovering new biological knowledge from high-throughput biological data and assigning biological functions to all the proteins encoded by the genome is, however, challenging and allowing systems level investigations of microbial cell factory. For fungi, several genome-sequencing and annotation projects have been presented, including *Saccharomyces cerevisiae* [3], *A. nidulans* [4], *A. fumigatus* [5], and *A. niger* [6,7]. Recently, genome sequence of *A. oryzae* by Machida and his coworkers has been published [8]. Based on their sequence annotation using gene-finding software tools such as ALN [9], GlimmerM [10] and GeneDecoder [11], this analysis 12,074 genes encoding proteins were predicted to be present in the genome [8]. Despite this prediction many genes had not been assigned a definite function, and of the 12,074 genes, more than 50% were annotated as hypothetical proteins. Hence, there are clearly opportunities for refining the gene prediction and improving the annotation. However, the present one dimensional data does not allow for complete annotation of all genes and it would therefore be interesting and potentially fruitful to use integrative biological tools in the process of improving the annotation of fungal genomes [12]. In this process reconstruction of a genome-scale metabolic model is a good starting point as it allows for integration of various types of data. Nowadays, there are several open sources of fungal metabolic models, such as for *S. cerevisiae* [13], *A. nidulans* [14], *A. niger* [15] and a model for the central carbon metabolism of *A. niger* [16]. These models currently are prominent as one of the most promising approaches to achieve an *in silico* prediction of cellular function in terms of physiology [17].

The aim of this study is to improve the annotation of the genome sequence of *A. oryzae* and further integrate enhanced annotated data to construct a genome-scale metabolic model of *A. oryzae*. The first *A. oryzae* EST library, sequencing and assembly were performed in order to improve gene prediction. Then functional assignment was done by our developed annotation strategy and a combination of different bioinformatics tools and databases. The bioinformatics tools used were BLAST [18], HMMER [19], and PSI-BLAST [20]. Several databases used were namely the *A. oryzae* genome database [21], the EST database of *A. flavus* [22], the *A. nidulans* genome database [23], the *A. fumigatus* genome database [24], the *S. cerevisiae* genome database [25], the Pfam protein families database [26], the COG database [27], and the Non-Redundant (NR) protein database [28]. Subsequently, manual inspection was through in order to achieve a solid annotation for enzyme functions that were needed for reconstruction of the metabolic network. Based on the improved annotated genome, the genome-scale metabolic network was reconstructed. The network was built by comparison with other related metabolic models, namely models for *S. cerevisiae* [13], *A. nidulans* [14], and *A. niger* [15,16], and biochemical pathway databases, literature, as well as experimental evidence for the presence of specific pathways. The biomass composition was taken from the literature, whereas, maintenance and growth-associated ATP consumption rates were estimated based on literature data on yields and growth rates. Finally, Flux Balance Analysis (FBA) was used to predict the flux distributions in the metabolic network, and the biomass yields as well as growth rates on different carbon sources were estimated to validate the metabolic model of *A. oryzae*.

Results and Discussion

Gene discovery and validation

The assembled EST sequences of *A. oryzae* were achieved from this study (see Additional file 1) where were deposited into Genbank database under accession numbers "EY424375–433412". Within our assembled EST data analysis of *A. oryzae*, we found 9,038 EST contig sequences with a GC content of 51.2% and an average EST length of 738 base pairs (bps). Based on analysis of sequences obtained from Machida and coworkers [8], the *A. oryzae* genome consists of eight chromosomes containing 37.2 Megabases (Mb) with a GC content of 48.2% and 12,074 annotated genes. According to the described strategy implemented for gene finding (See Methods), the 9,038 EST sequences were searched against the 12,074 previously identified genes [8] in the sequenced genome using various search parameters to create lists of predicted genes with different match stringencies. Using the criteria described in the Methods, many dissimilar sequences between the EST sequences and previously identified gene sequences of *A. oryzae* [8] were found. This suggests the presence of many newly predicted genes. Interestingly, approximately 12% (1,046 out of the 9,038 EST sequences) were categorized as newly predicted genes in the genome. Many homolog sequences were also found strongly validating previously identified genes [8], with approximately 75% of the total EST sequences (6,773 out of the 9,038 EST sequences) matching earlier identified genes (See Figure 1). To confirm that all the EST sequences do existed in the *A. oryzae* genome, the 9,038 EST

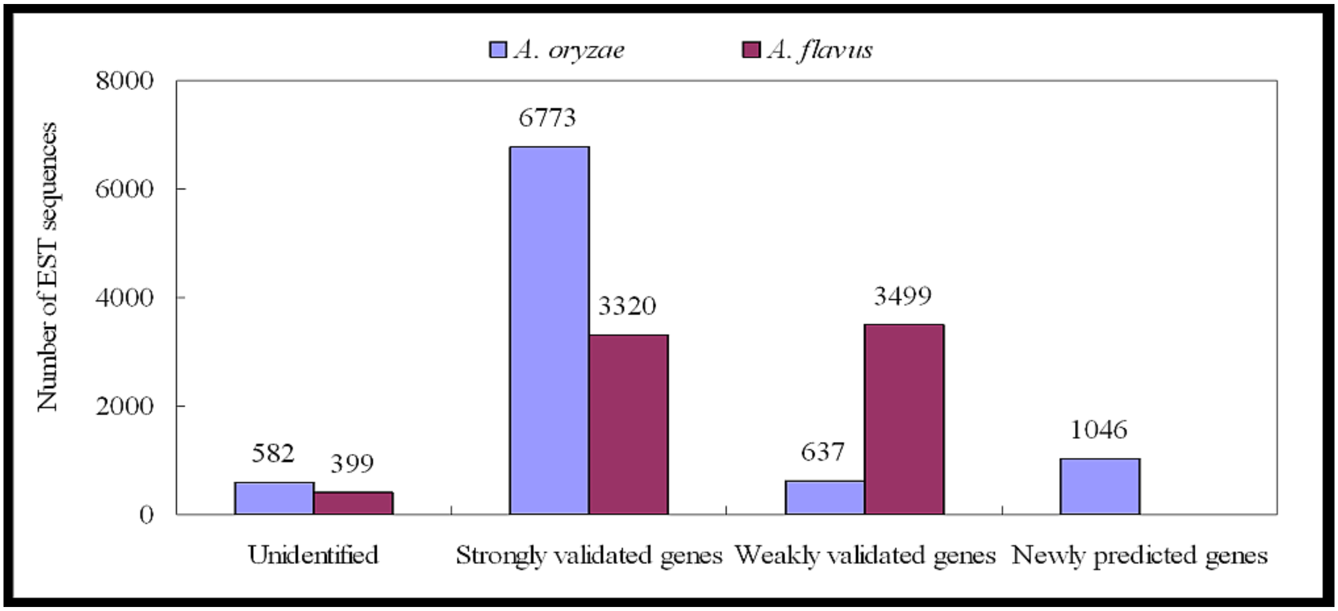


Figure 1
Gene discovery and validation of existing genes. The bars show the number of new genes discovered and the number of existing gene validated by our assembled EST sequences of *A. oryzae* and EST data of *A. flavus* [22].

sequences were searched by BLASTN [18] against the complete genome, and the results showed that only 20 EST sequences could not be found to be present in the genome. Therefore, this suggests that the assembled EST data of *A. oryzae* had very high quality and showed an excellent success rate for gene discovery and validation, even though approximately 13% (1,219 out of the 9,038 EST sequences) could not be used to predict genes, because 6% (582 out of the 9,038 EST sequences) were too short and about 7% (637 out of the 9,038 EST sequences) were too weakly validated in the original gene list using a conservative cut-off. In another attempt to predict new genes in *A. oryzae* genome, *A. flavus* EST contigs stored in the TIGR public database [22] were also used because *A. flavus* and *A. oryzae* are very closely related [29]. Also, there is a high degree of DNA homology between the two organisms (e.g. aflatoxin cluster > 96%) [29]. *A. flavus* EST library contained 7,218 sequences with a GC content of 49.7% and an average EST length of 636.2 bps. Using these *A. flavus* EST sequences to search against the genes in our new gene list for the *A. oryzae* genome, no new genes were predicted but 3,320 genes in the *A. oryzae* genome were validated by EST sequences (see Figure 1). Based on all the results of the gene finding a total of 13,120 protein-encoding genes were identified in the *A. oryzae* genome. This total number of genes derives from 12,074 previously annotated genes by Machida et al and 1,046 newly predicted genes from our assembled EST library.

Identification of protein functions by pairwise comparison

In order to assign protein functions to the 13,120 predicted genes, sequence alignment analysis by pairwise comparison between *A. oryzae* and closely related fungi was performed. These fungi included *A. nidulans*, *A. fumigatus* and *S. cerevisiae*. Table 1 shows some genome characteristics of the related fungi in comparison with *A. oryzae*. Initially pairwise comparison was done by similarity searching of the protein sequences of *A. oryzae* against the protein sequences of other related fungi as described in the Methods. With a chosen threshold of the alignment length (bps) and identity (%), a list of putative protein functions was created. The results are summarized in Table 1. Pairwise comparison shows that *A. fumigatus* has 6,274 homologs with *A. oryzae* sequences. It is the highest number of sequence homologs and this indicates the highest percentage (88%) of the homologs obtained between the three species tested. This result is consistent with the fact that *A. oryzae* and *A. fumigatus* are the phylogenetically closest species of those evaluated [4,30]. Upon completion of the similarity searching, the results suggest that 7,161 genes in *A. oryzae* could be assigned as orthologous genes from the three fungi used for comparison. Of these 7,161 protein sequences, 5,836 sequences were assigned putative protein functions for *A. oryzae*. These functions were mainly obtained from *A. fumigatus* (Table 1). The remaining 1,325 sequences that have homologs in the three other fungi could not been assigned any function yet, and they are therefore classified as hypothetical proteins. The putative functions annotated here were clas-

Table 1: Comparison of genome characteristics and function assignments between *A. oryzae* and other related fungi

Genome characteristics				
Features	***ANI***[1]	***AFU***[2]	***SC***[3]	***AO***[4]
Genome size (Mb)	30.1	29.4	12.1	37.2
Number of chromosomes	8	8	16	8
Number of total predicted genes	10,701	10,267	5,869	13,120
Function assignments				
Pairwise comparison	***ANI* and *AO***	***AFU* and*AO***	***SC* and*AO***	***AO***
Number of protein sequence homologs	6,095	6,274	1,794	7,161
Percentage of sequence homologs	85	88	25	100
Number of assigned putative functions	837	5,482	1,731	5,836
Percentage of assigned putative functions	14	93	30	100
Number of predicted genes involved in metabolism	567	1,556	837	1,924
Number of putative functions involved in metabolism	377	1,132	495	1,070

ANI[1]: Data were obtained from *A. nidulans* genome database [23]
AFU[2]: Data were obtained from *A. fumigatus* genome database [24]
SC[3]: Data were obtained from *S. cerevisiae* genome database [25]
AO[4]: Data were obtained from *A. oryzae* genome database [21]. Notably, total predicted genes were achieved from both database [21] and our EST sequence analysis.

sified using biological process (BP) type from the Gene Ontology (GO) database [31]. The genes and functions that have biological process terms involved in metabolism, including both biosynthesis and catabolism, were extracted and used for metabolic network reconstruction. The results of this process show that the *A. oryzae* genome contains 1,924 genes (15% of the 13,120 total genes) encoding 1,070 different protein functions involved in metabolism.

Metabolic pathway mapping

The metabolic models for *S. cerevisiae* [13], *A. nidulans* [14], and *A. niger* [15,16] were combined to generate an initial reaction list for the construction of the *A. oryzae* metabolic network. Duplicated reactions were removed resulting in a list of 1,924 genes and 1,070 functions involved in metabolism. For each enzyme function involved in this reaction list it was searched in the above generated list of metabolic proteins present in *A. oryzae*. If an enzyme name matched, then the enzyme-encoding genes, enzyme functions and Enzyme Commission (EC) numbers of *A. oryzae* were selected and mapped onto this reaction list. Hereafter a classification system was established to divide reactions in the whole metabolic network of *A. oryzae* into 7 main metabolic pathways: carbohydrate metabolism, energy metabolism, amino acid metabolism, nucleotide metabolism, lipid metabolism, cofactor metabolism and secondary metabolism. It is hereby found that the highest number of enzyme-encoding genes is involved in carbohydrate metabolism, which is consistent with the fact that *A. oryzae* has the ability to use a wide range of carbohydrate substrates. For amino acid and lipid metabolisms, many enzyme-encoding genes were also found. A lower number of enzyme-encoding genes were found in nucleotide, cofactor and energy metabolisms. The lowest number of enzyme-encoding genes was found in secondary metabolism. In fact, the *A. oryzae* genome contains a lot of enzyme-encoding genes involved in secondary metabolism [29], but most of these genes are without EC numbers and could therefore not be mapped onto the metabolic network. The hereby resulting metabolic network contains several gaps, which means that there are metabolic reactions without corresponding enzymes.

Filling gaps in the metabolic network using an integrated bioinformatics tool

In order to identify genes encoding more enzyme functions and hereby reduce the number of gaps in the metabolic network, an integrated bioinformatics tool was developed and used to identify these missing enzymes. This tool called "Gap Filler for *Aspergillus oryzae* Pathway (GFAOP)" was developed in- house by combining different bioinformatics tools (i.e. BLAST [18], HMMER [19], and PSI-BLAST [20]) and databases (i.e. *A. oryzae* genome [21], Pfam [26], COG [27], and NR [28]). GFAOP is similar to the McConkey searching algorithm which has been used for enzyme identification in eukaryote genomes [32]. The method is also related to Osterman's method for the identification of bacterial genes encoding metabolic functions [33]. An overview of GFAOP is shown in Figure 2. First, the tool was validated by searching for 441 known protein functions in *A. oryzae* using the information from

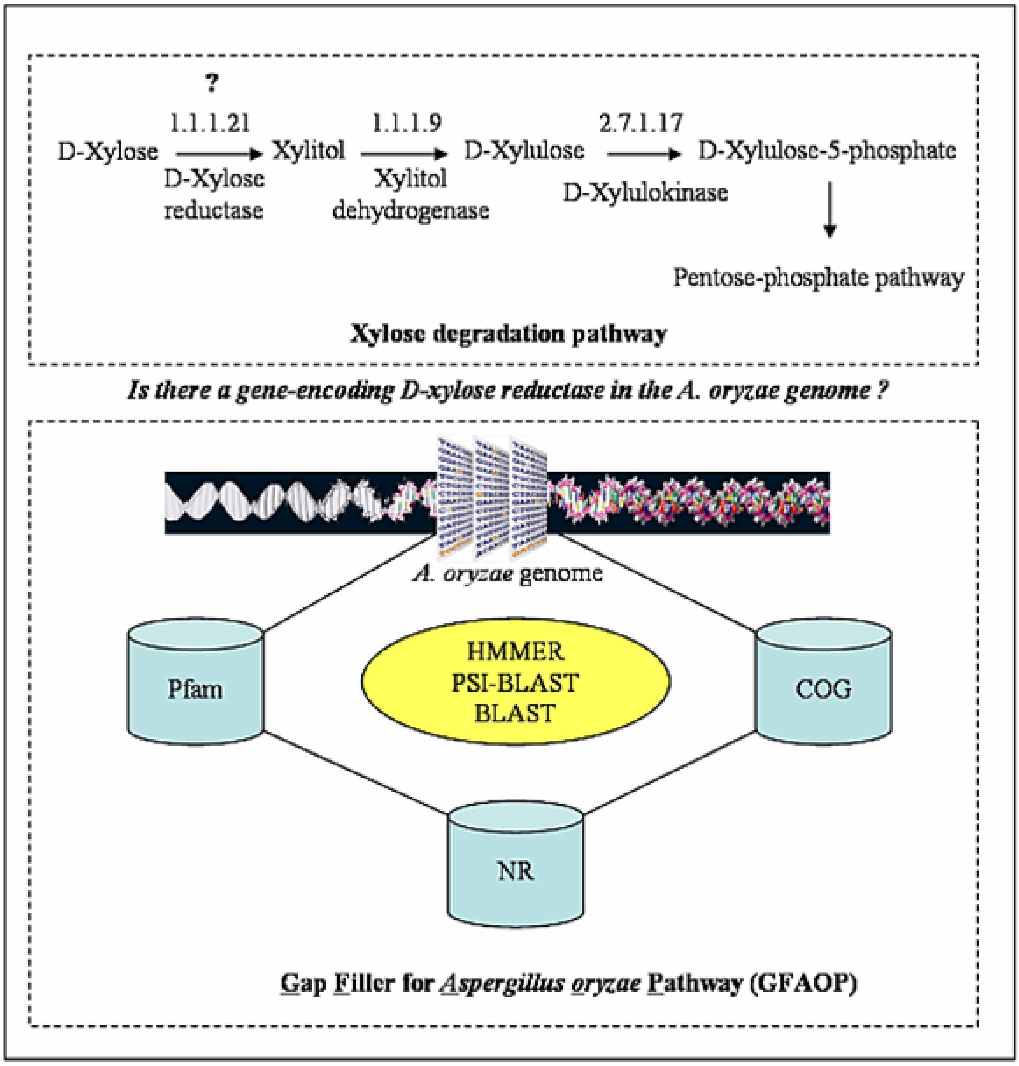

?
1.1.1.21
1.1.1.9
2.7.1.17
D-Xylose
Xylitol
D-Xylulose
D-Xylulose-5-phosphate
D-Xylose reductase
Xylitol dehydrogenase
D-Xylulokinase
Pentose-phosphate pathway
Xylose degradation pathway
Is there a gene-encoding D-xylose reductase in the A. oryzae genome ?
A. oryzae genome
Pfam
HMMER
PSI-BLAST
BLAST
COG
NR
Gap Filler for Aspergillus oryzae Pathway (GFAOP)

Figure 2

Filling gap by integrated bioinformatics approach. A diagram of the integrated bioinformatics tools used for filling the gaps in the metabolic network. A missing enzyme of D-xylose reductase in xylose degradation pathway is used as an example to illustrate the gap filling process.

the genome database [21]. The tool confirmed 100% accuracy of the prediction. This tool was then used to search for functional activity related to missing enzyme (Gap) in the metabolic reaction. To illustrate this approach, one of the missing enzymes ("D-xylose reductase" (EC: 1.1.1.21)) in the pathway of xylose degradation of *A. oryzae* is selected as an example. To answer the question of whether there is a gene encoding D-xylose reductase in *A. oryzae*. GFAOP was applied as follows. First the HMMER program generates a Hidden Markov Model (HMM) profile of this enzyme (D-xylose reductase) from the protein families databases (such as Pfam or COG). Second, a consensus sequence is generated. Third, the consensus sequence is searched against the *A. oryzae* genome by a PSI-BLAST [20]. Sequences where the hit has suitable statistical significance values are selected and extracted for protein function assignment by searching against the NR protein database [28] using BLAST [18] to verify its probable function.

The result clearly shows that there is a high probability for that the gene called "AO090003000859" encode D-xylose reductase. Based on searching of this gene in the *A. oryzae* genome database [21], the gene name AO090003000859 is only reported for general prediction and poorly characterized functions. Moreover, the exploration in other databases such as the Genbank, this gene name is only showed to have a region encoding aldo/keto reductase family proteins, but there is no evidence on the specific function of the gene. As a result from using GFAOP, the missing enzyme of D-xylose reductase is entered into the pathway. Our method results in an improved annotation of the genome using the context of the metabolic network. An iterative process was done for filling all the gaps in the whole metabolic network. Ultimately, 210 gaps in the metabolic network were closed using GFAOP. These gaps distributed with 86 gaps in lipid metabolism, 31 gaps in secondary metabolism, 34 gaps in amino acid metabolism, 23 gaps in nucleotide metabolism, 17 gaps in carbohydrate metabolism, 10 gaps in cofactor metabolism, and 9 gaps in energy metabolism.

Characteristics of the improved annotation and reconstructed metabolic network

The annotation process resulted in the improved annotated data shown in Table 2 where the data are compared with values in the *A. oryzae* genome database by Machida et al [21]. The results show that the number of improved annotated genes is 13,120 which are higher than the number of genes in the database [21]. Of these improved annotated data, the predicted genes and the putative functions are distributed into different groups. The first group contains new putative protein functions assigned to newly predicted genes, and it contains 398 new putative protein functions that are divided into 154 metabolic functions and 244 other functional groups. The second group contains hypothetical proteins assigned to newly predicted genes and it contains 648 hypothetical proteins. The third group is new putative protein functions assigned to proteins previously annotated as hypothetical proteins, and this group comprises 1,469 proteins of which 562 proteins have metabolic functions. The final group contains genes that is found to have the same putative protein function as previously reported in the database [21]. In total the hereby annotated genome of *A. oryzae* contains 5,391 protein functions of which 3,178 have metabolic functions. Even though the genome still contains 5,214 hypothetical proteins this is less than the 6,683 hypothetical proteins currently reported in the database [21], and our work therefore resulted in a substantial improvement of the genome annotation. An enhanced annotated data were mapped on the *A. oryzae* genome by using the Perl Scalable Vector Graphics (SVG) Module V2.33 [34]. Figure 3 shows an example of gene and EST mapping on the

Table 2: Statistical characteristics of improved annotation and metabolic reconstruction.

Characteristics of improved annotation	**Improved annotated data**	**Database**
Total protein-encoding genes	13,120	12,074
New putative protein functions to newly predicted genes	398	-
Metabolic functions	154	-
Other functional groups	244	-
Hypothetical proteins to newly predicted genes	648	-
New putative protein functions to previously hypothetical proteins	1,469	-
Metabolic functions	562	-
Other functional groups	907	-
Same putative protein functions	5,391	5,391
Metabolic functions	3,178	3,178
Other functional groups	2,213	2,213
Hypothetical proteins	5,214	6,683
Characteristics of network	***A. oryzae***	***A. nidulans***
Enzymes-encoding genes	1,314	666
Enzymes	729	466
Metabolites	1,073	733
Biochemical reactions	1,846 (1,053 Unique)	1,090 (676 Unique)
Cytosol	832	551
Mtochondria	172	103
Glyoxysome	-	5
Peroxisome	19	-
Extracellular	30	17
Transport reactions	281 (161 Unique)	118 (113 Unique)
Reactions with gene assignments	173 (53 Unique)	15 (12 Unique)
Reactions without gene assignments	108 (108 Unique)	103 (101 Unique)

An improved annotated data is compared with genome database of *A. oryzae* [21]. The reconstructed metabolic network of *A. oryzae* is compared with the reconstructed metabolic network of *A. nidulans* [14] contig of AP007151 which is a part of chromosome 1 of the *A. oryzae* genome. The complete genomic map is available as Additional file 2. The list of all ESTs and genes contained on the genomic map is presented as Additional file 3.

As previously mentioned the improved annotation resulted in a final reconstructed metabolic network that contains 729 enzymes, 1,846 (1,053 unique) biochemical reactions and 1,073 metabolites (Table 2). The large number of isoenzymes (indicated by the difference between total biochemical reactions and unique biochemical reactions) points to a very high degree of flexibility in the metabolic network of A. oryzae. The 1,053 unique biochemical reactions are distributed into 832 cytosolic, 172 mitochondrial, 19 perosixomal, and 30 extracellular reactions. There are 281 (161 unique) reactions that function as transport processes, and of these 173 (53 unique) are included on the basis of gene assignments whereas there are no annotated genes for 108 of the transport reactions. All the genes and functions involved in metabolism were inspected manually. A total of 1,314 genes without duplication represented as enzyme-encoding genes are included in the reconstructed network. This corresponded to about 10% of the 13,120 total predicted genes of A. oryzae. For model comparison, the metabolic network of A. nidulans [14] was chosen, and it shows that the metabolism of A. oryzae is much larger than that of A. nidulans as it contains a higher number of genes, enzymes, metabolites and reactions (see Table 2). A list of the reactions in the reconstructed metabolic network that comprised the genes, EC numbers and enzymes was hereby obtained (see Additional file 4). To illustrate a whole network, overall metabolic map of A. oryzae was

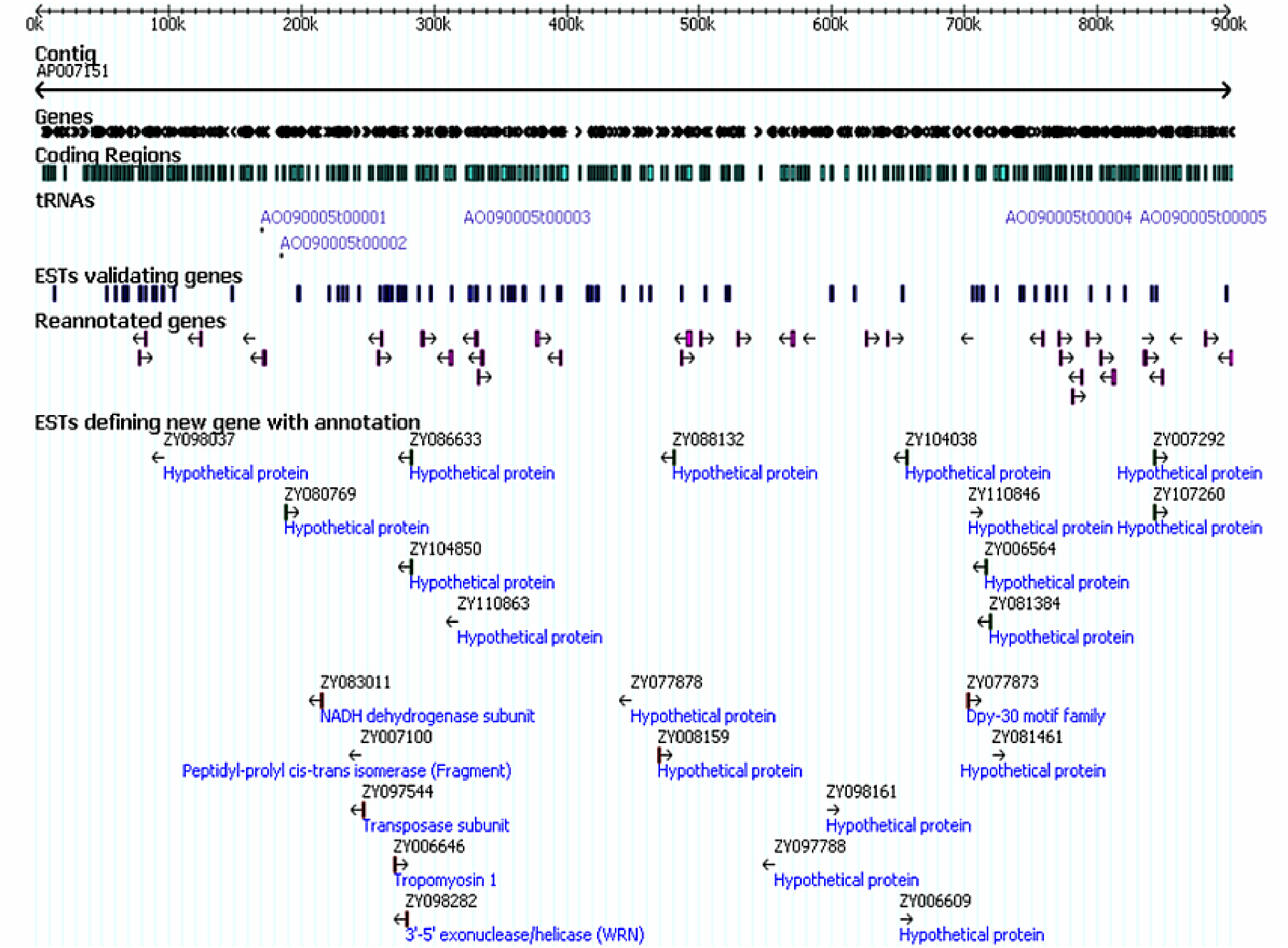


Figure 3
Gene and EST mapping on the *A. oryzae* genome. An example of how we map genes and ESTs on the AP007151 contig, a part of chromosome 1 of the *A. oryzae* genome. Along this contig, we mapped EST sequences defining new genes with annotation, EST sequences validating genes, and also re-annotated genes.

drawn as shown in Figure 4 (also see in Additional file 5 for full size) to represent a valuable link between genes, enzymes, metabolic reactions and metabolites. The complete metabolite list (with full name) is also given as Additional file 4.

Biomass growth simulation

Using the reconstructed metabolic network, a stoichiometric model was developed and subsequently used to simulate growth. A list of the reactions that comprise the stoichiometric model is presented as Additional file 4. To describe growth, biomass production is regarded as a drain of macromolecules and building blocks required to produce cellular components. The demands on each of these compounds are estimated based on the biomass composition. No drain of free metabolites or dilution of the metabolite pool due to biomass growth is considered [35]. The cellular composition considered for *A. oryzae* is based on the contents of the main biomass components of *A. oryzae* [36] as shown in Table 3 (see also Additional file 4 for the original data used to perform this analysis). In addition, concerning the biomass composition, the only parameters that have to be estimated are key energetic parameters: ATP requirement for non-growth associated purposes (mATP), ATP requirement for synthesis of biomass from macromolecules (K_{ATP}) and the operational P/O ratio. These parameters can not be determined independently, but if one of the parameters is known the others can be estimated from experimental data. The operational P/O ratio was assumed to be 1.5 [35], and mATP (mmol/gDW) was estimated to be 1.9 and K_{ATP} (mmol/gDW) was estimated by fitting model simulation with experimental data obtained at a specific growth rate of 0.1 h^{-1} [36] with glucose as the sole carbon source. The value of K_{ATP} was hereby estimated to be 49 mmoles ATP/g DW.

Assessment of model validation of A. oryzae

The model was evaluated by simulating *A. oryzae* cell growth on different carbon sources and comparison of the

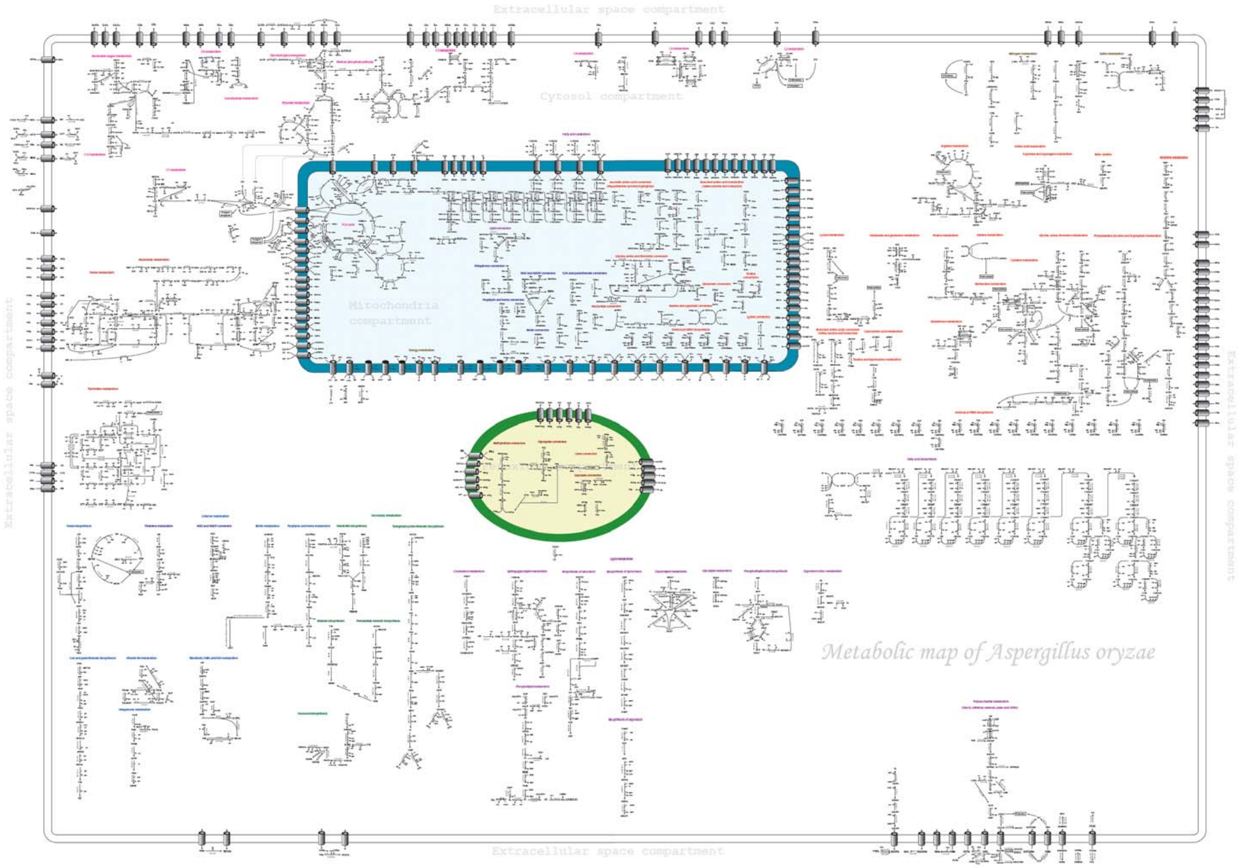

Extracellular space compartment
Cytosol compartment
Mitochondria compartment
Extracellular space compartment
Extracellular space compartment
Metabolic map of Aspergillus oryzae
Extracellular space compartment

Figure 4
Overall metabolic map of *A. oryzae*. A full size of metabolic map of *A. oryzae* is viewed in Additional file 5.

Table 3: Biomass composition in the metabolic model of *A. oryzae*

Biomass component	Average molecular weight[1] [g/mol]	Content[2] [g/100 g DW]		Stoichiometric coefficient[4] [mmol/g DW]
			Normalized[3]	
Proteins	134.58	40	47.1	3.50075
Carbohydrates	-	28	33	-
Glycogen	666.6	0.1	0.1	0.00212
Chitin	203.2	7	8.3	0.40759
Glucan	162.1	20.8	24.6	1.51453
RNA	341.9	5.3	6.2	0.18259
DNA	332.3	0.8	0.9	0.02836
Lipids	-	6.8	8	-
Triacylglycerol	954.96	2.12	2.49	0.02617
Free fatty acid	301.31	0.35	0.41	0.01365
Phosphatidylethanolamine	782.5	0.97	1.14	0.01468
Phosphatidylcholine	834.8	2.38	2.8	0.03356
Phosphatidylserine	827.3	0.4	0.47	0.00564
Phosphatidylamine	755.24	0.58	0.68	0.00903
D-Mannitol	182.2	3.3	3.9	0.21333
Glycerol	92.1	0.7	0.8	0.08952
Ash	-	15.1	-	-

[1]Average molecular weights (units: g/mol of monomers in polymer)
[2]For growth on glucose, using ammonia as the nitrogen source and for a specific growth rate of 0.1 h^{-1}
[3]Without considering ash
[4]In the equation representing biomass formation (units: mmol of monomers in polymer/g DW)

simulated data to the experimentally determined growth rate and biomass yield from literature data [37,38]. For each carbon source the substrate uptake rate was estimated from measurements of the substrate concentration in the medium, and this value is used as input to the model. From this input the flux distributions corresponding to optimal growth are calculated by maximizing the flux of the reaction leading to biomass. The validation results are shown in Figure 5 and Figure 6. From the results, Figure 5 indicates that the model can accurately predict the maximum specific growth rate (h^{-1}) during batch cultivations on different carbon sources (when the uptake rate of the carbon source is given as input). The accuracy is on average about 98% of the experimentally determined value. Figure 6 also shows that the model can accurately predict the biomass yield (gDW/mmol substrate) during chemostat cultivations on different carbon sources. The small deviation can be explained by kinetic

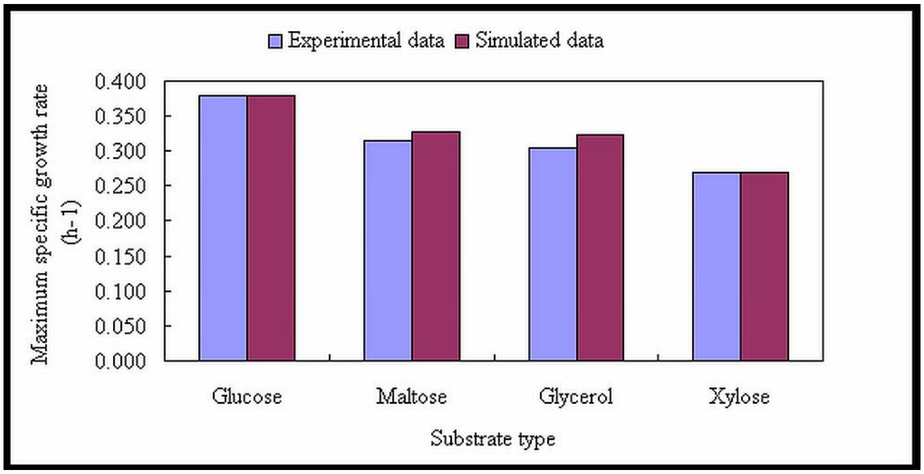


Figure 5
Model validation by experimental data. Comparison of the maximum specific growth rate (h^{-1}) between simulated data and experimental data. The experimental data were obtained from batch fermentation.

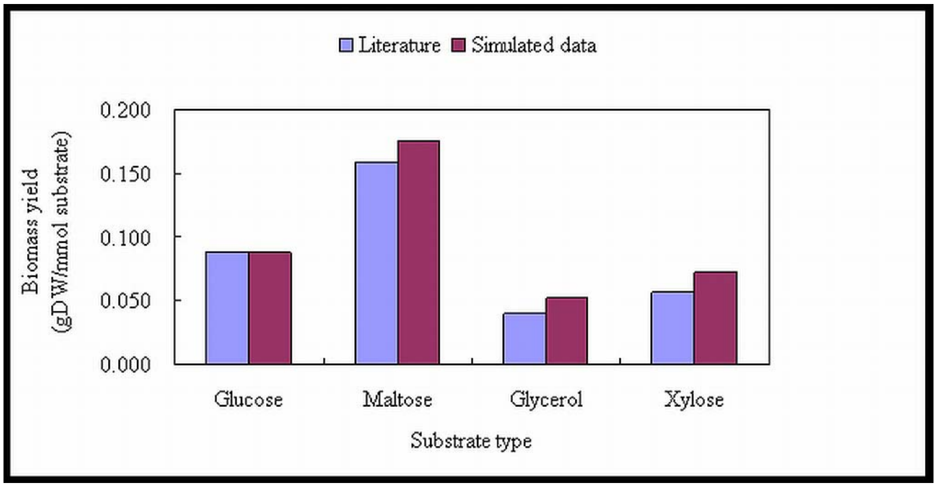


Figure 6
Model validation by literature. Comparison of biomass yield (gDW/mmol substrate) obtained by model simulation data and data from the literature [37, 38]. The biomass yield was obtained from chemostat fermentation.

or genetic regulation within the metabolism, which is not accounted for in the model [17].

Conclusion

A strategy for the improved annotation of the genome sequence of *A. oryzae* was developed. Using our assembled EST library, 1,046 EST sequences (about 12% of 9,038 EST sequences) were discovered as newly predicted genes and about 75% (6,773 of 9,038 EST sequences) were used to validate previously annotated genes. This indicates that the developed annotation strategy is a very useful approach for gene prediction. Applying a combination of various bioinformatics tools and databases, this annotation strategy was successfully applied for function assignment of genes. A high number of newly predicted genes were assigned with 398 new putative functions, and with new putative functions to 1,469 proteins previously annotated as hypothetical proteins. Therefore our analysis results in a substantially reduced number of hypothetical proteins. In particular, more enzyme-encoding genes could be assigned functions and this led to filling of 210 missing enzymes in the metabolic network. Applying the enhanced annotated genome, biochemical pathway databases, other related metabolic models, and the literature, a metabolic network was reconstructed. The network contains 729 enzymes, 1,314 enzyme-encoding genes (10% of 13,120 total predicted genes), 1,073 metabolites and 1,846 (1,053 unique) biochemical reactions. The 1,053 unique reactions are distributed into different compartments, with 831 reactions located in the cytosol, 173 reactions located in the mitochondria, 19 reactions located in the perosixome, and 30 reactions located in the extracellular space. Transport reactions between the different compartments and the extracellular space represents 281 (161 unique) reactions. This metabolic network was formulated to a stoichiometric model. The model was applied for Flux Balance Analysis (FBA) to obtain the flux distributions corresponding to maximized growth. A physiological study on different carbon sources of *A. oryzae* was performed to validate the genome-scale model, and the model is found to accurately predict the maximum specific growth rate and the biomass yield on different carbon sources. This indicates that the *A. oryzae* metabolic model is able to simulate the phenotypic behavior and the model will hereby serve as an important resource for gaining further insight into our understanding of the important cell factory *A. oryzae*.

Methods

An overview of the approach employed here for improved genome annotation of *A. oryzae* is depicted in Figure 7. For gene discovery and validation, we constructed EST library and performed sequencing as well as assembly as described in following section.

EST library construction

The EST sequences of *A. oryzae* strain A1560 were constructed from a normalized library and an un-normalized library. The normalized library was constructed by inserting cDNA of *A. oryzae* in pCMV-Sport6 plasmids between the *Mlu*I and the *Not*I sites (Vector – *Not*I – poly A (3' of insert) – 5' of insert – *Mlu*I – *Sal*I- vector). The plasmids were amplified in *Escherichia coli* EMDH10B-TONA (a recA strain). The un-normalized library was made by inserting cDNA of *A. oryzae* between the *EcoR1* and *NotI*

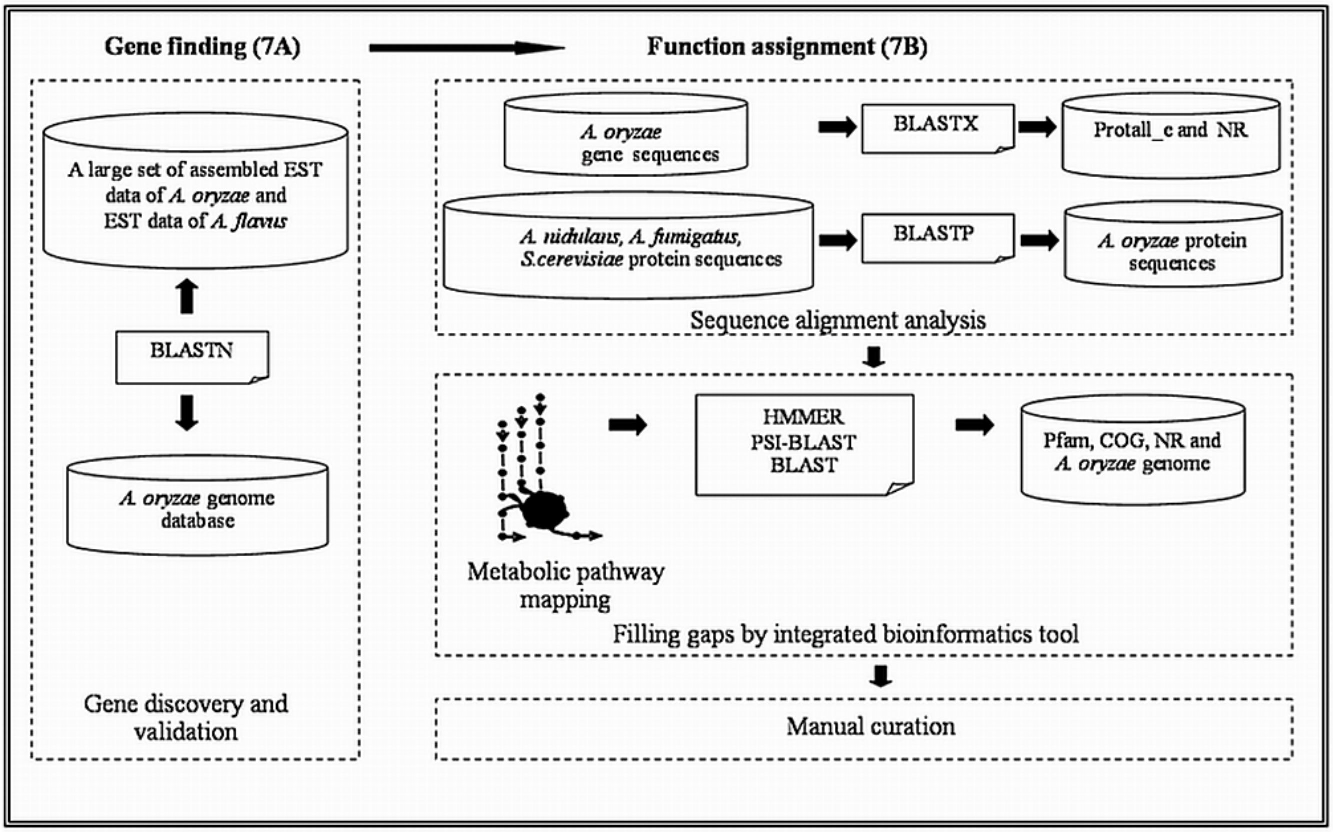

Gene finding (7A)
Function assignment (7B)
A large set of assembled EST data of *A. oryzae* and EST data of *A. flavus*
BLASTN
A. oryzae genome database
Gene discovery and validation
A. oryzae gene sequences
BLASTX
Protall_c and NR
A. nidulans, A. fumigatus, S.cerevisiae protein sequences
BLASTP
A. oryzae protein sequences
Sequence alignment analysis
Metabolic pathway mapping
HMMER
PSI-BLAST
BLAST
Pfam, COG, NR and *A. oryzae* genome
Filling gaps by integrated bioinformatics tool
Manual curation

Figure 7

Overview of annotation process of *A. oryzae* genome. Illustration of the annotation process, which is divided into two steps, namely gene finding (Figure 7A) and function assignment (Figure 7B).

sites in the vector pYES2. The plasmids were amplified in *E. coli* DH10B.

EST sequencing and assembly

The EST sequences were generated by sequencing on ABI 377 and ABI 3700 instruments from Applied Biosystems using BigDye terminators version 1 and 2. In total 23,072 EST sequences were produced. Quality clipping, vector removal, *E. coli* contamination removal and assembly were done with the phredPhrap package [39]. The sequences were assembled into 9,038 EST contigs.

Genome annotation process

The strategy of gene finding as shown in Figure 7A was carried out based on our assembled EST sequences of *A. oryzae* (see Additional file 1, also available online in Genbank database under accession number "EY424375–433412") together with public EST data of *A. flavus* [22]. Our assembled EST data of *A. oryzae* were compared to the genes previously identified [8] in the genome of *A. oryzae* strain RIB 40 by BLASTN [18]. The purpose of this comparison was to validate genes that were already annotated and to discover new genes that had not been annotated by Machida et al [8]. The 9,038 EST sequences were classified into four categories as outlined in Additional file 3 and described as follows. All sequences shorter than 300 bases were discarded from the analysis. If the length of an EST sequence was over 500 bps and the highest ranking hit had a score lower than 50 bits, then the EST sequence was categorized as a sequence that served as a newly predicted gene. If the length of the EST sequence was over 300 bps and the highest ranking hit had a score over 100 bits, then the EST sequence was categorized as validating an earlier identified gene [8]. If the highest ranking hit had a score lower than 100 bits, the EST sequence was classified as weakly validating a gene [8]. In the effort to predict new genes in the *A. oryzae* genome, *A. flavus* EST data from the TIGR database [22] was also used. The cut-off for gene discovery and validation was selected to be the identical as with our assembled EST data of *A. oryzae*. After performing gene finding, assignment of protein function was done. The main principle was performed based on sequence alignment analysis, metabolic pathway mapping, filling the gaps by integrated bioinformatics tool and lastly manual curation. The sequence alignment was done to assign putative function to newly predicted genes by BLASTX [18]. The newly predicted gene was searched against the NR protein database [28] and Protall_e protein database [Unpublished]. The assignment of putative protein function was transferred if the alignment length of the highest ranking hit was over 50 amino acids and the identity over 25%. The sequence alignment was done through pairwise comparison of protein sequences by BLASTP [18] between *A. oryzae* and other related fungi (i.e. *A. nidulans* strain FGSC-A4, *A. fumigatus* strain Af293, *S. cerevisiae* strain S288c) as shown in Figure 7B. The criteria for similarity searching were alignment length (bps) and identity (%), with the parameters depending on the type of fungus used for the comparison [40]. An estimated suitable cut-off for *S. cerevisiae* was an alignment length above 100 bps and an identity higher than 40%. For other related *Aspergillus* species, the cut-off was an alignment length above 200 bps and an identity higher than 40%. All cut-off values were determined by using sequences with known protein functions. After finishing the annotation process, the metabolic network of *A. oryzae* was reconstructed. At the beginning, an initial metabolic reaction list for *A. oryzae* was constructed by combination of *S. cerevisiae* [13], *A. nidulans* [14], and *A. niger* [15,16] metabolic models. In addition, data collection from metabolic pathway databases, such as KEGG [41] and BioCyc [42], of other organisms was integrated into this reaction list. The improved annotated genomic data (i.e. enzyme-encoding genes, enzyme functions, and EC numbers) were then mapped into the reaction list. In order to visualize all the metabolic reactions, overall metabolic map was drawn (see Figure 4 and Additional file 5 for full size). The improved annotated data were placed onto this map. At the end, gaps that existed in the metabolic network were then filled using an integrated bioinformatics tool that allowed for automatic searching for specific enzyme functions. Finally, manual curation of the model was done for finalizing the reconstruction process.

Metabolic network reconstruction

The metabolic network reconstruction aimed at representing the whole metabolism of *A. oryzae*, which consists of primary catabolism of carbohydrates, biosynthesis of amino acids, nucleotides, lipids, cofactors and production of Gibbs free energy required for biosynthesis, as well as of secondary metabolism. Combination of different types of information was essential to carry out a solid reconstruction. Information was collected from the improved annotated data of *A. oryzae*, biochemical pathways, publications on specific enzymes, online protein databases (e.g. Swiss-Prot database [43]) and also literature. In addition, there was physiological evidence for the presence of a reaction or pathway in *A. oryzae*, e.g. when there was information of presence of a specific enzyme activity or presence of a pathway involved in consumption of a given substrate or formation of a given metabolic product, then the underlying reaction was added to the model, even if there was no annotated gene supporting the presence of the reaction. In the processes of stoichiometry for cofactors as well as the information on reversibility or irreversibility for each reaction, these were verified and added as information into the reconstructed network. Different cellular compartments were considered and consequently biochemical reactions were distributed into four different compartments: the extracellular space, the cytosol, the mitochondria, and the peroxisome [44]. Identification of localization of each biochemical reaction was analyzed according to enzyme localization, which was performed by applying protein localization predictors. Herein, pTARGET [45] and CELLO [46] were selected to predict sub-cellular protein localization because they contain databases of known eukaryotic protein localizations. If there is no information on localization of a biochemical reaction or its corresponding enzyme, then by default this reaction was considered to occur in the cytosol. In addition, the reconstructed metabolic network included transport steps between the different intracellular compartments and between the cell and the environment.

Modeling and simulation based Flux Balance Analysis (FBA)

After the metabolic network was reconstructed, this was transformed into a mathematical framework to perform Flux Balance Analysis (FBA) [47]. This approach is based on conservation of mass under steady-state conditions. This conversion requires stoichiometry of metabolic pathways, metabolic demands and a few specific parameters. An optimal flux distribution can be obtained within the feasible region by using linear programming [48]. A reaction is selected as an objective function that is to be maximized or minimized. For physiologically meaningful results, the objective functions must be defined as the ability to produce the required components of cellular biomass for a specified uptake rate of a selected carbon source. By maximizing the flux towards biomass formation, a flux is obtained for each reaction in the metabolic network.

Model validation of A. oryzae *by physiological study on different carbon sources*

Model validation is an important step in the reconstruction process. In this study, the model was validated by simulating the rate of biomass formation on different carbon sources in batch experiments. Here the uptake rate of the carbon source was given as input to the simulations. Different carbon sources namely glucose (C6), maltose (C12), glycerol (C3) and xylose (C5), which were selected as they result in widely different physiological responses and parameters. The strain used for generating these data was *A. oryzae* wild type strain A1560, which was obtained from Novozymes A/S, Denmark. Three biological replicates were done for each carbon source. The fermentations were performed using an in-house fermenter with a working volume of 1.2 L, and operated at 34°C and pH was kept constant at 6 by adding 10% of H_3PO_4 or 10% NH_3 solution. The aeration flow rate was set at 1.2 L/min. The stirrer speed was controlled at 800 rpm for the first 4 hrs and later increased to 1100 rpm. The dissolved oxygen tension was initially calibrated at 100%. The concentrations of oxygen and carbon dioxide in the exhaust gas were measured by a gas analyzer (Magnos 4G for O_2, Uras 3G for CO_2, Hartmann & Braun, Germany). Biomass dry weight measurements were done as follows: A sample was filtered using nitrocellulose filters (pore size 0.45 μm, Munktell, Sweden), and the filter cake was therefore dried at 110°C overnight. Hereafter the filter was placed in a dessicator overnight, and subsequently, weighed. In addition, the extracellular concentration of sugars, organic acids, and polyols were measured by using high-performance liquid-chromatography (HPLC) on an Aminex HPX-87H, 300 mm*7.8 mm column. The column was kept at 45°C and eluted at 0.6 ml/min with 5 mM H_2SO_4.

Authors' contributions

WV carried out the improved annotation, performed the genome scale modeling and wrote the manuscript. PO performed the EST sequencing project, participated in the EST and gene mapping and supervised the bioinformatics work. KH supervised the fermentation process. SK provided the EST and gene mapping and supervised the bioinformatics work. JN supervised the whole work and assisted in manuscript preparation. All authors read and approved the final manuscript.

Additional material

Additional file 1

The assembled EST sequences of A. oryzae*. The FASTA file provides that 9038 assembled EST sequences. The title name of each sequence indicates dbEST ID, User ID and Genbank accession number.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-245-S1.fas]

Additional file 2

EST and gene mapping on A. oryzae *genome. The ZIP file provides that 48 separated files with Perl Scalable Vector Graphics.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-245-S2.zip]

Additional file 3

Improved annotated data in A. oryzae *genome. This file is created to extract gene and EST list existed on genomic map (Additional file 2).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-245-S3.xls]

Additional file 4

A reconstructed metabolic network of A. oryzae*. A list of all metabolic reactions is shown with EC numbers, enzyme names, gene names and annotation methods. Besides, a list of metabolite abbreviations and biomass compositions are shown also.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-245-S4.pdf]

Additional file 5

Metabolic map of A. oryzae *in Scalable Vector Graphics. EC numbers are shown for all reactions in the map. A list of abbreviations for the metabolite-names is available in Additional file 4.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-245-S5.pdf]

Acknowledgements

This research work was support by a stipend to Wanwipa Vongsangnak from Novozymes Bioprocess Academy (NBA) and Technical University of Denmark (DTU). The authors are grateful to José Manuel Otero, Roberto Olivares Hernandez, Chia-Wen Chang and Rawisara Ruenwai for discussions and critical revisions of the manuscript. Also, we thank Lone Vuholm and Jeanette Thomassen for valuable technical assistance.

References

1. Goldman GH, Osmani SA: **The Aspergilli.** In *Genomics, Medical Aspects, Biotechnology, and Research Methods (Mycology)* Taylor and Francis group: CRC Press; 2008.
2. Kitano H: **Computational systems biology.** *Nature* 2002, **420:**206-210.
3. Fisk DG, Ball CA, Dolinski K, Engel SR, Hong EL, Issel-Tarver L, Schwartz K, Sethuraman A, Botstein D, Cherry JM: **Saccharomyces cerevisiae S288C genome annotation: a working hypothesis.** *Yeast* 2006, **23:**857-865.
4. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, Selker EU, Archer DB, Penalva MA, Oakley BR, Momany M, Tanaka T, Kumagai T, Asai K, Machida M, Nierman WC, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B, Dyer P, Sachs MS, Osmani SA, Birren BW: **Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae.** *Nature* 2005, **438:**1105-1115.
5. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer PS, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, Gwilliam R, Haas B, Haas H, Harris D, Horiuchi H, Huang JQ, Humphray S, Jimenez J, Keller N, Khouri H, Kitamoto K, Kobayashi T, Konzack S, Kulkarni R, Kumagai T, Lafon A, Latge JP, Li WX, Lord A, Lu C, Majoros WH, May GS, Miller BL, Mohamoud Y, Molina M, Monod M, Mouyna I, Mulligan S, Murphy L, O'Neil S, Paulsen I, Penalva MA, Pertea M, Price C, Pritchard BL, Quail MA, Rabbinowitsch E, Rawlins N, Rajandream MA, Reichard U, Renauld H, Robson GD, de

Cordoba SR, Rodriguez-Pena JM, Ronning CM, Rutter S, Salzberg SL, Sanchez M, Sanchez-Ferrero JC, Saunders D, Seeger K, Squares R, Squares S, Takeuchi M, Tekaia F, Turner G, de Aldana CRV, Weidman J, White O, Woodward J, Yu JH, Fraser C, Galagan JE, Asai K, Machida M, Hall N, Barrell B, Denning DW: **Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus.** *Nature* 2006, **439:**502-502.

6. Baker SE: **Aspergillus niger genomics: Past, present and into the future.** *Medical Mycology* 2006, **44:**S17-S21.
7. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JAE, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EGJ, Debets AJM, Dekker P, van Dijck PWM, van Dijk A, Dijkhuizen L, Driessen AJM, d'Enfert C, Geysens S, Goosen C, Groot GSP, de Groot PWJ, Guillemette T, Henrissat B, Herweijer M, van den Hombergh J, van den Hondel C, van der Heijden R, van der Kaaij RM, Klis FM, Kools HJ, Kubicek CP, van Kuyk PA, Lauber J, Lu X, van der Maarel M, Meulenberg R, Menke H, Mortimer MA, Nielsen J, Oliver SG, Olsthoorn M, Pal K, van Peij N, Ram AFJ, Rinas U, Roubos JA, Sagt CMJ, Schmoll M, Sun JB, Ussery D, Varga J, Vervecken W, de Vondervoort P, Wedler H, Wosten HAB, Zeng AP, van Ooyen AJJ, Visser J, Stam H: **Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88.** *Nature Biotechnology* 2007, **25:**221-231.
8. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto KI, Arima T, Akita O, Kashiwagi Y, Abe K, Gomi K, Horiuchi H, Kitamoto K, Kobayashi T, Takeuchi M, Denning DW, Galagan JE, Nierman WC, Yu JJ, Archer DB, Bennett JW, Bhatnagar D, Cleveland TE, Fedorova ND, Gotoh O, Horikawa H, Hosoyama A, Ichinomiya M, Igarashi R, Iwashita K, Juvvadi PR, Kato M, Kato Y, Kin T, Kokubun A, Maeda H, Maeyama N, Maruyama J, Nagasaki H, Nakajima T, Oda K, Okada K, Paulsen I, Sakamoto K, Sawano T, Takahashi M, Takase K, Terabayashi Y, Wortman JR, Yamada O, Yamagata Y, Anazawa H, Hata Y, Koide Y, Komori T, Koyama Y, Minetoki T, Suharnan S, Tanaka A, Isono K, Kuhara S, Ogasawara N, Kikuchi H: **Genome sequencing and analysis of Aspergillus oryzae.** *Nature* 2005, **438:**1157-1161.
9. Gotoh O: **Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps.** *Bioinformatics* 2000, **16(3):**190-202.
10. Majoros WH, Pertea M, Antonescu C, Salzberg SL: **GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders.** *Nucleic Acids Research* 2003, **31(13):**3601-3604.
11. Asai K, Itou K, Ueno Y, Yada T: **Recognition of human genes by stochastic parsing.** *Pac Symp Biocomput* 1998, **3:**228-239.
12. Liu ET: **Integrative biology and systems biology.** *Molecular Systems Biology* 2005.
13. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13:**244-253.
14. David H, Özçelik , Hofmann G, Nielsen J: **Analysis of Aspergillus nidulans metabolism at the genome-scale.** *BMC Genomics* 2008, **9:**163.
15. Andersen MR, Nielsen ML, Nielsen J: **Metabolic model integration of the bibliome, genome,metabolome and reactome of Aspergillus niger.** *Molecular Systems Biology* 2008, **4:**178.
16. David H, Akesson M, Nielsen J: **Reconstruction of the central carbon metabolism of Aspergillus niger.** *European Journal of Biochemistry* 2003, **270:**4243-4253.
17. Borodina I, Nielsen J: **From genomes to in silico cells via metabolic networks.** *Current Opinion in Biotechnology* 2005, **16:**350-355.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Journal of Molecular Biology* 1990, **215:**403-410.
19. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
20. Altschul S, Madden T, Schaffer A, Zhang JH, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Faseb Journal* 1998, **12:**A1326-A1326.
21. **Aspergillus oryzae genome database** [http://www.bio.nite.go.jp/dogan/MicroTop?GENOME_ID=ao]
22. **Aspergillus flavus Gene Index** [http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=a_flavus]
23. **Aspergillus nidulans genome database** [http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans]
24. **Aspergillus fumigatus genome database** [http://www.sanger.ac.uk/Projects/A_fumigatus/]
25. **Saccharomyces genome database** [http://www.yeastgenome.org/index.shtml]
26. **Pfam database** [http://www.sanger.ac.uk/Software/Pfam]
27. **COG database** [http://www.ncbi.nih.gov/COG]
28. **Non-redundant protein database** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/]
29. Payne GA, Nierman WC, Wortman JR, Pritchard BL, Brown D, Dean RA, Bhatnagar D, Cleveland TE, Machida M, Yu J: **Whole genome comparison of Aspergillus flavus and A. oryzae.** *Medical Mycology* 2006, **44:**S9-S11.
30. Pain A, Böhme U, Berriman M: **Hot and sexy moulds!** *Nature reviews* 2006, **4:**244-245.
31. **Gene Ontology Database** [http://www.geneontology.org/GO.annotation.shtml]
32. McConkey GA, Pinney JW, Westhead DR, Plueckhahn K, Fitzpatrick TB, Macheroux P, Kappes B: **Annotating the Plasmodium genome and the enigma of the shikimate pathway.** *Trends in Parasitology* 2004, **20:**60-65.
33. Osterman A, Overbeek R: **Missing genes in metabolic pathways: a comparative genomics approach.** *Current Opinion in Chemical Biology* 2003, **7:**238-251.
34. **Perl Scalable Vector Graphics** [http://search.cpan.org/~ronan/]
35. Nielsen J: **Physiological engineering aspects of Penicillium chrysogenum.** : World Scientific Pub Co Inc; 1996.
36. Pedersen H, Carlsen M, Nielsen J: **Identification of enzymes and quantification of metabolic fluxes in the wild type and in a recombinant Aspergillus oryzae strain.** *Appl Environ Microbiol* 1999, **65(1):**11-19.
37. Prathumpai W, Gabelgaard JB, Wanchanthuek P, van de Vondervoort PJI, de Groot MJL, McIntyre M, Nielsen J: **Metabolic control analysis of xylose catabolism in Aspergillus.** *Biotechnology Progress* 2003, **19(4):**1136-1141.
38. Carlsen M, Nielsen J: **Influence of carbon source on alpha-amylase production by Aspergillus oryzae.** *Appl Microbiol Biotechnol* 2001, **57(3):**346-349.
39. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Research* 1998, **8:**186-194.
40. Rost B: **Twilight zone of protein sequence alignments.** *Protein Engineering* 1999, **12:**85-94.
41. **KEGG pathway database** [http://www.kegg.com]
42. **BioCyc database** [http://biocyc.org/server.html]
43. **Swiss-Prot database** [http://www.expasy.ch/sprot/]
44. Carson BD: **Microbodies in fungi. A review.** *Journal of industrial microbiology* 1990, **6:**1.
45. Guda C, Subramaniam S: **TARGET: a new method for predicting protein subcellular localization in eukaryotes.** *Bioinformatics* 2005, **21:**3963-3969.
46. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins-Structure Function and Bioinformatics* 2006, **64:**643-651.
47. Edwards JS, Covert M, Palsson B: **Metabolic modelling of microbes: the flux-balance approach.** *Environmental Microbiology* 2002, **4:**133-140.
48. Bonarius HPJ, Schmid G, Tramper J: **Flux analysis of underdetermined metabolic networks: The quest for the missing constraints.** *Trends in Biotechnology* 1997, **15:**308-314.

Paper 2

A trispecies *Aspergillus* microarray: Comparative transcriptomics
of three *Aspergillus* species

Mikael R. Andersen, Wanwipa Vongsangnak, Gianni Panagiotou,
Margarita P. Salazar, Linda Lehmann, and Jens Nielsen

Proceedings of the National Academy of Sciences 2008, 105: 4387–4392

A trispecies *Aspergillus* microarray: Comparative transcriptomics of three *Aspergillus* species

Mikael R. Andersen, Wanwipa Vongsangnak, Gianni Panagiotou, Margarita P. Salazar, Linda Lehmann, and Jens Nielsen*

Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark

Edited by Joan Wennstrom Bennett, Rutgers, The State University of New Jersey, New Brunswick, NJ, and approved January 10, 2008 (received for review October 19, 2007)

The full-genome sequencing of the filamentous fungi *Aspergillus nidulans*, *Aspergillus niger*, and *Aspergillus oryzae* has opened possibilities for studying the cellular physiology of these fungi on a systemic level. As a tool to explore this, we are making available an Affymetrix GeneChip developed for transcriptome analysis of any of the three above-mentioned aspergilli. Transcriptome analysis of triplicate batch cultivations of all three aspergilli on glucose and xylose media was used to validate the performance of the microarray. Gene comparisons of all three species and cross-analysis with the expression data identified 23 genes to be a conserved response across *Aspergillus* sp., including the xylose transcriptional activator XlnR. A promoter analysis of the up-regulated genes in all three species indicates the conserved XlnR-binding site to be 5′-GGNTAAA-3′. The composition of the conserved gene-set suggests that xylose acts as a molecule, indicating the presence of complex carbohydrates such as hemicellulose, and triggers an array of degrading enzymes. With this case example, we present a validated tool for transcriptome analysis of three *Aspergillus* species and a methodology for conducting cross-species evolutionary studies within a genus using comparative transcriptomics.

Aspergillus nidulans | *Aspergillus niger* | *Aspergillus oryzae* | XlnR

The *Aspergillus* genus of filamentous fungi has a long history as a work horse in the service of humankind. *Aspergillus oryzae* (koji mold) was first used for the preparation of food stuffs in China almost 2,000 years ago and was used for one of the first commercial preparations of enzymes in the late 19th century (1, 2). Since then, *Aspergillus niger* has also proven to be a high-yield producer of organic acids and enzymes, and today, both of these fungi are used as hosts for production of heterologous proteins (3). Since the 1950s, *Aspergillus nidulans* has been used as a model fungus (4) and has advanced the understanding of eukaryotic cellular physiology and genetics. Advancing our knowledge of these fungi as individual species and as a group holds interest for both fundamental biological sciences and applied biotechnology.

With the publication of the genome sequences of these three aspergilli (5–8), genome-wide systems biology studies in the aspergilli have been made a possibility. As a parallel to the Yeast 2.0 GeneChip that allows for transcriptome analysis of both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, we have designed a chip to facilitate system-wide studies of *A. nidulans*, *A. niger*, and *A. oryzae*.

Results

Array Design. An Affymetrix GeneChip was designed for genes from *A. nidulans*, *A. niger*, and *A. oryzae*. The chip has probes for 99.5% of the genes from the three aspergilli. Because of a selection of unique nonoverlapping probes of similar melting point, not all genes have the maximum of 11 probes. Details on the probe distribution can be found in supporting information (SI) Table 2.

To evaluate the effect of having probes for three species of the same genus on the same chip, we randomly picked one set of transcriptome data from each of the three species (see below for details on the experiments generating these data). For each experiment, the distribution of gene expression values for all three species was examined (Fig. 1). Expression values for the genes specific for the species from which the hybridized cRNA was isolated are higher than expression values for genes from the other two aspergilli. This is the case for samples from all three species. Even though *A. niger* and *A. oryzae* are more closely related to each other than *A. nidulans*, this lineage is not reflected in the shape and levels of the distributions. This reflects that probes not targeting the genes from the species of the hybridized sample are acting in an unspecific manner, much in the same way as probes for unexpressed genes from the same species. Exceptions to this pattern are genes with high expression values in the nonhybridized species. These are composed of constitutively highly expressed and conserved genes, specifically ribosome and histone components. Although this issue influences the expression levels measured, we do not believe this affects evaluation of differential expression between two sets of experiments, because the effect of the probes on the conserved genes will be the same between experiments.

Protein Comparison. To examine systems regulating transcription conserved in all three aspergilli, genes having homologues in all three species were identified by using a blastp-based comparison (see *Materials and Methods*). Using this approach, based on bidirectional best hits with an *e*-value cutoff of 1e-30, 5,561 ORFs were found to be conserved in all three species (tridirectional best hits, SI Table 3). The number of genes with bidirectional, unidirectional, and no hits are shown in SI Fig. 4. The three sets of 5,561 conserved genes (1:1:1 homologues) were used for the further analysis of the transcript data.

Fermentation Results. As a model example of experiments that can be conducted by using the presented microarray, cultures of *A. nidulans*, *A. niger*, and *A. oryzae* were prepared in well controlled bioreactors. All cultivations were batch cultures grown on defined salt medium with glucose or xylose as carbon sources. Each species had its own specific cultivation medium. For each of the three species, triplicate cultivations were performed on each

Author contributions: M.R.A., W.V., G.P., and J.N. designed research; M.R.A., W.V., G.P., M.P.S., and L.L. performed research; M.R.A. and W.V. contributed new reagents/analytic tools; M.R.A. analyzed data; and M.R.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The normalized and raw data values reported in this paper have been deposited with Gene Express Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE9298).

*To whom correspondence should be addressed. E-mail: jn@bio.dtu.dk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709964105/DC1.

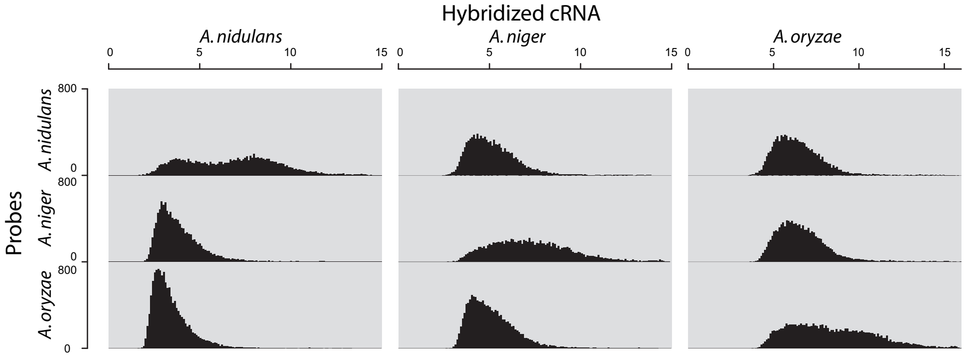


Fig. 1. Histograms of gene expression values. The distribution of log²-transformed gene expression values of the genes of the three *Aspergillus* sp. in three experiments. Each column shows an experiment where cRNA from the noted organism was hybridized to the chip. Each row shows the distribution for genes from a specific organism. The analysis was done with the statistical software R (31).

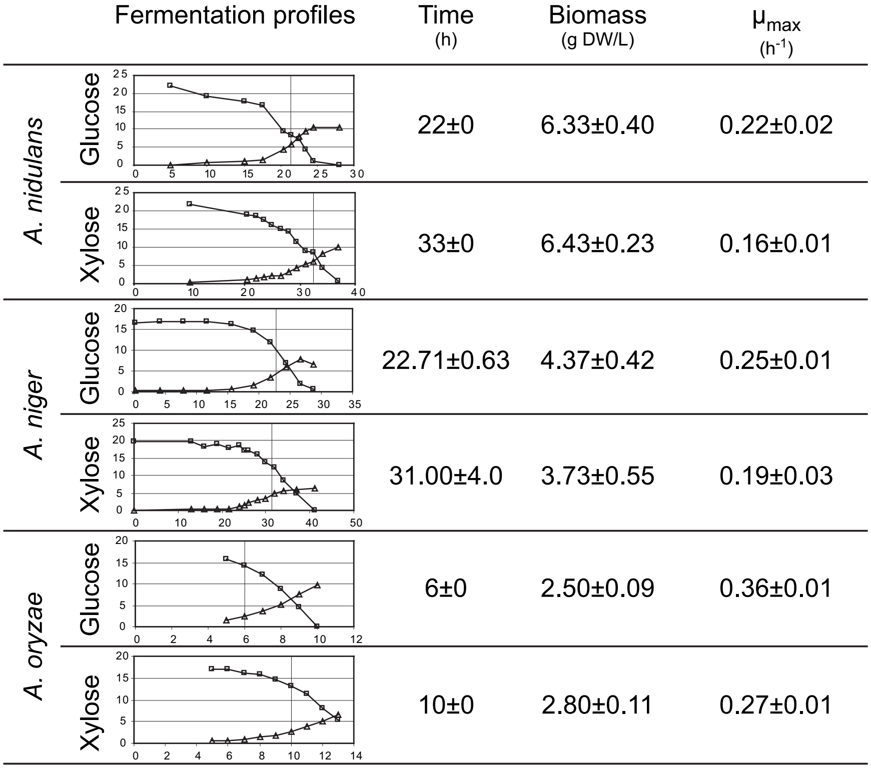
carbon source. Fig. 2 presents a summary of the six sets of triplicates. Dispersed filamentous growth was observed throughout the fermentation for all cultivations.

Transcriptome Analysis. For all three sets of glucose/xylose fermentations, statistical transcriptome analysis was performed. The significantly regulated genes in all three species were compared with the list of the 5,561 conserved genes and with each other (Fig. 3). This resulted in the identification of 23 conserved genes (Table 1) that are differentially regulated in all three species and 365 genes that are differentially expressed in only two of the aspergilli (Fig. 1). The 23 genes that are significant in all three can be seen as a conserved response across the *Aspergillus* genus.

A further inspection of the expression values of the 23 common genes revealed that the homologues are regulated in the same direction, with 22 of the genes being up-regulated on the xylose medium and only one gene being down-regulated.

The function of the genes in the less-annotated *A. nidulans* and *A. oryzae* was inferred from the well annotated *A. niger* genome sequence, based on the conserved sequences and responses (Table 1). The majority of the 23 common genes are enzymes and sugar transporters. Specifically, the entire D-xylose degradatory pathway (see SI Fig. 5 for an overview) was induced in all three species. A low-affinity glucose transporter (*mstC*) (9) was down-regulated, implying that this transporter has a higher affinity for glucose than xylose in all three species.

Interestingly, the xylanolytic transcriptional activator XlnR,



		Fermentation profiles	Time (h)	Biomass (g DW/L)	μ_{max} (h^{-1})
A. nidulans	Glucose		22±0	6.33±0.40	0.22±0.02
	Xylose		33±0	6.43±0.23	0.16±0.01
A. niger	Glucose		22.71±0.63	4.37±0.42	0.25±0.01
	Xylose		31.00±4.0	3.73±0.55	0.19±0.03
A. oryzae	Glucose		6±0	2.50±0.09	0.36±0.01
	Xylose		10±0	2.80±0.11	0.27±0.01

Fig. 2. Summary of fermentation parameters. For each of the *Aspergillus* sp., a profile of a representative replicate is shown. (□): Sugar concentration (g/liter). (△): Biomass concentration (g dry weight/liter). The vertical line shows the average time of sampling for transcriptome analysis. Time, sampling time for transcriptome analysis; Biomass, biomass concentration at the sampling time; μ_{max}, maximum specific growth rate. For all three values are shown average and standard deviation for the three replicates.

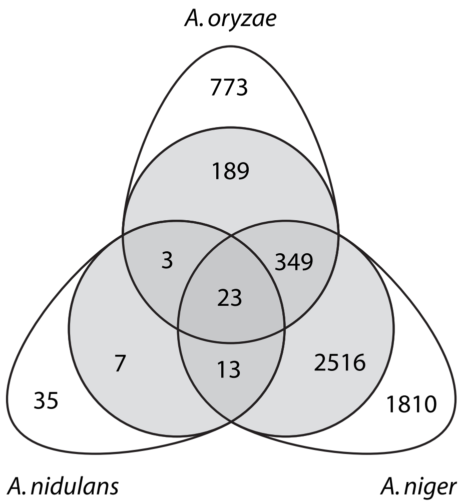


Fig. 3. Venn diagram of differentially expressed genes. The gray circles contain the genes that are significantly differentially expressed and conserved in all three *Aspergillus* species. The numbers on a white background are not conserved in all three species, but still differentially expressed in a single species.

previously described only in *A. niger* (10) and in *A. oryzae* as AoXlnR (11), has a homologue in *A. nidulans* (AN7610) that is significantly induced on xylose as well. This suggests that XlnR regulation is present in *A. nidulans* and functions in a manner similar to that reported for *A. niger* and *A. oryzae*.

***cis*-Acting Elements.** In recognition that one or more conserved transcriptional regulators might be active in all three species to produce the conserved response of the 23 genes, statistical promoter analysis was performed for all three species sets of 22 genes up-regulated on xylose. A 5′-GGNTAAA-3′ motif (motif A) was found to be significant (P = 3.6e-28) and present 102 times in the promoters of 46 of the 3 × 22 genes (Table 1). In 12 of the 22 conserved genes, the motif was present in the promoter region of all three species. Included in these 12 sets of genes is the xylose catabolic pathway. For some of the genes (L-arabitol dehydrogenase and D-xylose reductase), the motif is found at the same distance from the start codon for all three homologues (within 5–20 bp) and with a different third base in each of the species. A preference for a specific third base in any of the species across promoters could not be observed. This indicates evolutionary pressure for maintaining the motif but not the third base. Details on the location and sequence of the motifs are in SI Table 4.

The 5′-GGCTAAA-3′ motif has been reported to be the binding site for XlnR from *A. niger* and *A. oryzae* (10, 11). However, based on the statistical analysis, we are proposing that the 5′-GGNTAAA-3′ motif is indeed the XlnR motif that is conserved in *A. nidulans*, *A. niger*, and *A. oryzae*.

A separate analysis of the promoters that did not contain the XlnR-motif was carried out, but neither statistical analysis nor manual inspection for syntenic regions revealed any conserved domains. A similar analysis was performed for the down-regulated gene, but no significant motif was found.

Known XlnR-Induced Genes. As a further validation of the method and the quality of the array, we examined the expression of genes

Table 1. Twenty-three differentially expressed genes conserved in *A. nidulans*, *A. niger*, and *A. oryzae*

A. nidulans	*A. oryzae*	*A. niger*	*A. niger* annotation	Regulation	Motif A	Ng	O
AN0250	AO090001000069	JGI55668	Sugar transporter	Up	NdNgO		
AN0280	AO090005000767	JGI55419	Glucosyl hydrolase	Up			
AN0423	AO090003000859	JGI51997	D-xylose reductase (*xyrA*)	Up	NdNgO	12	
AN0942	AO090005001078	JGI46405	L-arabitol dehydrogenase	Up	NdNgO		
AN10124	AO090003000497	JGI213437	β-glycosidase	Up	Ng		
AN10169	AO090038000426	JGI177736	Short-chain dehydrogenase	Up	NdNgO		
AN1677	AO090023000688	JGI54541	Short-chain dehydrogenase	Up			
AN2359	AO090005000986	JGI205670	β-xylosidase (*xlnD*/*xylA*)	Up	NdNgO	10, 13	11, 14
AN3184	AO090012000809	JGI55604	Aldose 1-epimerase	Up	Ng		
AN3368	AO090010000208	JGI212893	Glycoside hydrolase	Up	NdNg		
AN3432	AO090020000042	JGI56084	Aldose 1-epimerase	Up	NdNgO		
AN4148	AO090009000275	JGI205766	Sugar transporter	Up	NgO		
AN4590	AO090011000483	JGI180923	Sugar transporter	Up	NgO		
AN5860	AO090026000494	JGI197162	Monosugar-transporter (*mstC*)	Down			
AN7193	AO090023000264	JGI55928	Aldo/keto reductase	Up	NdNgO		
AN7610	AO090012000267	JGI48811	XlnR	Up	NgO		
AN8138	AO090010000684	JGI212736	α-galactosidase	Up	Ng		
AN8400	AO090020000324	JGI199510	Sugar transporter	Up	NgO		
AN8790	AO090020000603	JGI209771	D-xylulokinase (*xkiA*)	Up	NdNgO	15	
AN9064	AO090038000631	JGI203198	Xylitol dehydrogenase (*xdhA*)	Up	NdNgO		
AN9173	AO090010000063	JGI194438	Sugar transporter	Up	NdNgO		
AN9286	AO090026000127	JGI56619	α-glucuronidase (*aguA*)	Up	NdNgO	13, 16	
AN9287	AO090701000345	JGI54859	Lipolytic enzyme	Up	NdNgO		

Genes marked with ″Up″ are up-regulated on xylose medium. The presence of motif A (5′-GGNTAAA-3′) in the promoter region of the *A. nidulans*, *A. niger*, and *A. oryzae* genes is marked with Nd, Ng, and O, respectively. Columns ″Ng″ and ″O″ give references to studies of XlnR induction in *A. niger* and *A. oryzae*. No references were found for *A. nidulans*. VanKuyk *et al.* (15) show that *xkiA* is induced on D-xylose in *A. niger* but does not depend on XlnR. The remaining references describe XlnR induction of the genes on D-xylose.

known to be induced by XlnR/AoXlnR on D-xylose. Genes found among the 22 sets of homologues have been marked with a reference in Table 1. Details for all genes are in SI Table 5. A multitude of genes are known from *A. niger*: xylanase B (*xlnB*/*xynB*) (10, 13), arabinoxylan arabinofuranohydrolase (*axhA*) (13), acetyl xylan esterase (*axeA*) (13), ferulic acid esterase A (*faeA*) (13, 17), endoglucanase A (*eglA*) (13), α- and β-galactosidase [*lacA* (18), and (*aglB*) (18)], cellobiohydrolase A (*cbhA*) (19), D-xylose reductase (*xyrA*) (12), β-xylosidase (*xlnD*) (10, 13), and α-glucuronidase (*aguA*) (13, 16). Additionally, D-xylulokinase (*xkiA*) (15) has been reported to be induced on xylose but not regulated by XlnR. All were found to be significantly induced in this study. Endoglucanase B and C (*eglB*/*C*) and cellobiohydrolase B (*cbhB*) are known to be induced by XlnR in *A. niger* but not when grown on D-xylose (13, 19, 20). These were not found to be significantly induced.

Fewer genes have been reported to be induced by XlnR on D-xylose in *A. oryzae*: β-xylosidase (*xylA*) (11, 14), endoxylanases F1 (*xynF1*) (14, 11, 21), G_1 (*xynG1*) (21), and G2 (*xynG2*) (21). All were significantly induced in *A. oryzae*. Endoglucanases A and B (*celA*/*B*) and cellobiohydrolases C and D (*celC*/*D*) are known to be induced by AoXlnR but not when *A. oryzae* is grown on D-xylose (21). These genes were not significantly induced.

No genes have been reported to be induced by XlnR in *A. nidulans*, but the xylanases X22 (*xlnA*) (22, 23), X24 (*xlnB*) (22, 23), and xylanase X34 (*xlnC*) (24) are known. Xylanases X22 and X24 are specific for acidic and alkaline medium, respectively, and only xylanase X24 was found to be induced. It had a *P* value of 0.0527 and was thus not included among the significant 81 genes. Xylanase X34 has been tested only for induction on xylan (24) but was found to be significantly induced on D-xylose.

Interestingly, xylanases are found to be significantly induced in all three species, but none of them are tridirectional best hits and are therefore not found in the core response of 22 genes. It thus seems that each species has a unique set of xylanases.

Discussion

In an application of the presented high-density microarray, we identify a carbon source-based response conserved in three aspergilli. The design of the study involving three different species, grown on three different defined minimal media, at three different values of pH, increases the likelihood of the found genes to be the true conserved "core" response to growth on xylose and not responses relying on an extra factor in addition to xylose (with the possible exception of abundant oxygen). We also believe this approach validates our argument that the xylanolytic transcriptional activator XlnR is a conserved system, even though it has not previously been studied in *A. nidulans*. Backed by the finding that the 5′-GGNTAAA-3′ motif is present and in some cases conserved as syntenic regions in all three species, we propose that the motif is indeed a XlnR motif and conserved in *A. nidulans*, *A. niger*, and *A. oryzae*. As a point of interest, a study of the homologous genes and their promoter regions in *A. fumigatus*, a known degrader of dead organic matter, showed a XlnR homologue to be present (Afu2g15620) and the presence of the 5′-GGNTAAA-3′ motif in several of the other homologues, including the entire xylose catabolic pathway (SI Table 6).

Upon further examination of the function of the up-regulated genes in *A. nidulans*, *A. niger*, and *A. oryzae*, it is interesting that the induction of L-arabitol dehydrogenase is found as a part of the core response. Because both the laboratory of Ronald de Vries (personal communication) and our laboratory have found minuscule amounts of arabinose in the commercial preparations of xylose, it might be an artifact. However, a further examination of the data shows that L-arabinose reductase (ORFs AN1679, JGI46249, and AO090009000031), the first step in L-arabinose degradation (see SI Fig. 5) is not significantly induced in any of the three species. Additionally, the XlnR motif is found in the promoter of L-arabitol dehydrogenase in all three homologues of the gene. This implies that this induction is not an artifact and is indeed triggered by xylose. One hypothesis might be that L-arabitol dehydrogenase has an affinity for xylitol as well. Another hypothesis is based on the observation that in nature, xylose is seldom encountered alone, but is a constituent of hemicellulose, along with arabinose, galactose, glucuronic acid, mannose, and other sugars (25, 26). It is thus likely that fungi have evolved coupled responses. This also poses an explanation for the conserved induction of the multitude of sugar transporters, glucuronidase, epimerases, α-galactosidase, and an array of glucoside hydrolases. It thus seems that the conserved xylose response is tailored to degrade complex carbohydrates such as hemicellulose, and xylose triggers this.

Another point of interest is that the promoter analysis suggests *xlnR* is not autoinducible in *A. nidulans*. Additionally, promoter analysis shows three sugar transporters in the core response to have only the XlnR motif in *A. niger* and *A. oryzae*. This might imply that another less- or non-conserved system is coregulating the xylose/xylanolytic response in *A. nidulans* along with XlnR. Another possibility is that XlnR can bind to variations of the motif and induces more of the genes than the statistical analysis indicates. van Peij *et al.* (13) demonstrate that for *A. niger*, the last base of the motif can vary. de Vries *et al.* (16) describe induction via a 5′-GGCTAR-3′ motif in *A. niger*. Marui *et al.* (11) describe the binding site to be 5′-GGCTA/GA-3′ for *A. oryzae*. It is thus likely that each species has versions of the motif that are not statistically significant because of the increased variation, and these facilitate induction by XlnR.

The low number of differentially expressed genes in *A. nidulans* might indicate a lower sensitivity. However, manual inspection of the expression values shows this is because of a higher level of variation between the replicates of the *A. nidulans* cultivations. Other in-house experiments with *A. nidulans* and the presented arrays have identified >2,500 significantly regulated genes and >1,000 in a single comparison (data not shown). Additionally, the validation of the array using statistical analysis is confirmed by the expression patterns of hemicellulytic genes known to be induced on D-xylose and/or by XlnR. These are in perfect accordance with the transcriptional analysis of this study.

With this study, we give access to a validated platform for analyzing transcription response in three different *Aspergillus* sp. This array allows transcriptome analysis of *A. oryzae*, which was previously unavailable, and is a publicly available Affymetrix-based platform for transcriptome studies in any of the three species. We hope this publication will spur an increase of transcriptome analysis in the individual fungi, thus adding to our knowledge base of this interesting genus of fungi. However, although we acknowledge the multitude of aspects that can be elucidated by traditional single-organism transcriptome analysis, we believe the biggest potential of the herein-presented microarray lies in studies of the multispecies type described in this study. We have demonstrated that data-analysis strategies, such as the blast-based strategy presented here, can add strength to conclusions and help identify systems and responses that are conserved across a genus. This possibility of studying the evolutionary depth of transcriptional regulation adds a new dimension to comparative transcriptomics.

Methods

Strains. The strains used were *A. nidulans* A4, *A. niger BO-1,* and *A. oryzae* A1560, obtained from Novozymes.

The *A. nidulans* stock culture was maintained on Sigma potato-dextrose-agar (PDA) at 4°C. *A. niger* was maintained as frozen spore suspensions at −80°C in 20% glycerol. *A. oryzae* stock culture was maintained on Cove-N-gly agar at 4°C.

Growth Media. *A. nidulans* batch cultivation medium: 15 g/liter $(NH_4)_2SO_4$, 3 g/liter KH_2PO_4, 2 g/liter MgSO4·7H_2O, 2 g/liter NaCl, 0.2 g/liter $CaCl_2$, and 1 ml/liter trace element solution. Trace element solution: 14.3 g/liter $ZnSO_4 \cdot 7H_2O$, 13.8 g/liter $FeSO_4 \cdot 7H_2O$, and 2.5 g/liter $CuSO_4 \cdot 5H_2O$. Carbon sources used were xylose or glucose monohydrate (20 g/liter). *A. niger* complex medium: 2 g/liter yeast extract, 3 g/liter tryptone, 10 g/liter glucose monohydrate, 20 g/liter agar, 0.52 g/liter KCl, 0.52 g/liter $MgSO_4 \cdot 7H_2O$, 1.52 g/liter KH_2PO_4, and 1 ml/liter of trace elements solution. Trace element solution: 0.4 g/liter $CuSO_4 \cdot 5H_2O$, 0.04 g/liter $Na_2B_4O_7 \cdot 10H_2O$, 0.8 g/liter $FeSO_4 \cdot 7H_2O$, 0.8 g/liter $MnSO_4 \cdot H_2O$, 0.8 g/liter $Na_2MoO_4 \cdot 2H_2O$, 8 g/liter $ZnSO_4 \cdot 7H_2O$. *A. niger* batch cultivation medium: mineral base, 1.0 g/liter $MgSO_4 \cdot 7H_2O$, 1 g/liter NaCl, 0.1 g/liter $CaCl_2 \cdot 2H_2O$, 0.05 ml/liter antifoam 204 (Sigma), and 1 ml/liter trace element solution. Trace element solution composition: 7.2 g/liter $ZnSO_4 \cdot 7H_2O$, 0.3 g/liter $NiCl_2 \cdot 6H_2O$, 6.9 g/liter $FeSO_4 \cdot 7H_2O$, 3.5 g/liter $MnCl_2 \cdot 4H_2O$, and 1.3 g/liter $CuSO_4 \cdot 5H_2O$. Carbon sources used were xylose or glucose monohydrate (20 g/liter). Nitrogen, sulfur and phosphate sources were 2.5 g/liter $(NH_4)_2SO_4$, 0.75 g/liter KH_2PO_4 (glucose medium), or 7.3 g/liter $(NH_4)_2SO_4$, 1.5 g/liter KH_2PO_4 (xylose medium). Concentrations were higher on the xylose medium to avoid nitrogen starvation. *A. oryzae* spore propagation medium (Cove-N-gly): 218 g/liter sorbitol, 10 g/liter glycerol 99.5%, 2.02 g/liter KNO_4, 25 g/liter agar and 50 ml/liter salt solution. Cove-N-gly salt solution: 26 g/liter LiCl, 26 g/liter $MgSO_4 \cdot 7H_2O$, 76 g/liter KH_2PO_4, 50 ml/liter trace element solution. Cove-N-gly trace element solution: 40 mg/liter $Na_2B_4O_7 \cdot 10H_2O$, 400 mg/liter $CuSO_4 \cdot 5H_2O$, 800 mg/liter $FePO_4 \cdot 2H_2O$, 800 mg/liter $MnSO_4 \cdot H_2O$, 800 mg/liter $NaMoO_4 \cdot 2H_2O$, 8 g/liter $ZnSO_4 \cdot 7H_2O$. *A. oryzae* medium for precultures (G2-GLY): 18 g/liter yeast extract, 24 g/liter glycerol 87%, 1 ml/liter pluronic PE-6100. *A. oryzae* batch cultivation medium: 2.4 g/liter $MgSO_4 \cdot 7H_2O$, 3.6 g/liter K_2SO_4, 1.2 g/liter citric acid monohydrate, 2.4 g/liter KH_2PO_4, 3 g/liter $(NH_4)_2HPO_4$, 1.2 g/liter pluronic acid (PE-6100) and 0.6 ml/liter trace element solution. Trace element solution: 14.3 g/liter $ZnSO_4 \cdot 7H_2O$, 8.5 g/liter $MnSO_4 \cdot H_2O$, 13.8 g/liter $FeSO_4 \cdot 7H_2O$, 2.5 g/liter $CuSO_4 \cdot 5H_2O$, 3 g/liter citric acid monohydrate (as a chelating agent), and 0.5 g/liter $NiCl_2 \cdot 6H_2O$. Carbon sources used were xylose or glucose monohydrate (15 g/liter).

Preparation of Inoculum. *A. nidulans* A4 fermenters were inoculated with spores to a final concentration of 6 × 10^9 spores per liter. The spores were cultivated on PDA at 37°C for 4–5 days and harvested by adding 20 ml of distilled water.

A. niger BO1 fermentations were initiated by spore inoculation to a final concentration of 2 × 10^9 spores/liter (glucose cultivations) or 5.7 × 10^9 spores/liter (xylose cultivations). The spores were propagated on complex media plates and incubated for 7–8 days at 30°C before harvest with 10 ml of Tween 80 0.01%.

A. oryzae A1560 fermenters were inoculated with ≈60 g of broth of *A. oryzae* A1560 cultured at 30°C for 24 h on G2-GLY liquid medium in shake flasks at 250 rpm (7 × *g*). The precultures were inoculated with 5 ml of spore solution harvested from mycelium grown on Cove-N-gly agar at 34°C for 3–4 days. Spores were harvested with Tween 80 0.1%.

Batch Cultivations. *A. nidulans* batch cultivations were performed in 1.5-liter bioreactors with a working volume of 1.2 liters. The bioreactors were equipped with two Rushton four-blade disk-turbine impellers rotating at 350 rpm. The pH was kept constant at 5.5 by addition of 2 M NaOH or HCl, and the temperature was maintained at 30°C. Air was used for sparging the bioreactor at a constant flow rate of 1 volume of gas per volume of liquid per minute (vvm).

A. niger batch cultivations on glucose medium were performed in 2-liter Braun fermentors with a working volume of 1.6 liters, equipped with three Rushton four-blade disk turbines. The bioreactor was sparged with air, and the concentrations of oxygen and carbon dioxide in the exhaust gas were measured in a gas analyzer. The temperature was maintained at 30°C. The pH was controlled by automatic addition of 2 M NaOH. Agitation and aeration were controlled throughout the cultivations. For inoculation of the bioreactor, the pH was set to 2.5, stirring rate to 100 rpm, and aeration to 0.1 vvm. After germination, the stirring rate was increased to 300 rpm and the air flow to 0.5 vvm. Eleven to 12 h after inoculation, the stirring rate was increased to 600 rpm and the air flow to 1 vvm. When the CO_2 in the exhaust gas reached a value of 0.1%, the stirring rate was set to 1,000 rpm, and the pH was gradually increased to 4.5.

A. niger batch cultivations on xylose medium were carried out in 5-liter reactors with a working volume of 4.5 liters. The bioreactors were equipped with two Rushton four-blade disk turbines, pH and temperature control, and no baffles. Inlet air was controlled with a mass flowmeter. The temperature was maintained at 30°C, and the pH was controlled by automatic addition of 2 M NaOH. The pH was initially set to 3.0 to prevent spore aggregation; only when spores started to germinate was the pH gradually increased to 4.5. Similarly, the stirring speed was initially set to 200 rpm and the aeration rate to 0.05 vvm. After germination, these parameters were progressively increased to 600 rpm and 0.89 vvm and kept steady throughout the rest of the fermentation.

A. oryzae batch cultivations were done in 2-liter fermenters with a working volume of 1.2 liters. The stirrer speed was kept at 800 rpm during the first 4 h and then increased to 1,100 rpm. The pH was controlled at 6 by addition of 4 M NaOH and 4 M HCl, and the temperature was maintained at 34°C. The aeration flow rate was set at 1.2 vvm. Dissolved oxygen tension was initially calibrated at 100%.

The concentrations of oxygen and carbon dioxide in the exhaust gas were monitored with a gas analyzer (1311 Fast Response Triple Gas, Innova combined with multiplexer controller for Gas Analysis MUX100, Braun Biotech).

Sampling. Cell dry weight was determined by using nitrocellulose filters (pore size 0.45 μm, Gelman Sciences). The filters were predried in a microwave oven at 150 W for 15 min or at 100°C for 24 h, cooled in a desiccator, and subsequently weighed. A known volume of cell culture was filtered, and the residue was washed with distilled water or 0.9% NaCl and dried on the filter for 15 min in a microwave oven at 150 W or at 100°C for 24 h and cooled in a desiccator. The filter was weighed again, and the cell mass concentration was calculated. These values were used to calculate maximum specific growth rates. For gene expression analysis, mycelium was harvested at the mid-late exponential phase by filtration through sterile Mira-Cloth (Calbiochem). At this point, *A. niger* mycelium was washed with a PBS buffer (8 g/liter NaCl, 0.20 g/liter KCl, 1.44 g/liter Na_2HPO_4, and 0.24 g/liter KH_2PO_4 in distilled water). The mycelium was quickly dried by squeezing and subsequently frozen in liquid nitrogen. Samples were stored at −80°C until RNA extraction.

Quantification of Sugars and Extracellular Metabolites. The concentrations of sugar in the filtrates were determined by using HPLC on an Aminex HPX-87H ion-exclusion column (BioRad). The column was eluted at 60°C with 5 mM H2SO4 at a flow rate of 0.6 ml/min. Metabolites were detected with a refractive index detector and a UV detector.

Extraction of Total RNA. *A. nidulans* and *A. niger*: 40–50 mg of frozen mycelium was placed in a 2 ml of microcentrifuge tube, precooled in liquid nitrogen containing three steel balls (two balls with a diameter of 2 mm and one ball with a diameter of 5 mm). The tubes were then shaken in a Retsch Mixer Mill, at 5°C for 10 min, until the mycelia were ground to powder. Total RNA was isolated from the powder using the Qiagen RNeasy Mini Kit, according to the protocol for isolation of total RNA from plant and fungi.

A. oryzae. Total RNA was purified by using the Promega RNAgents Total RNA Isolation system according to the protocol. For purification, ≈1 g of frozen mycelium was ground to a fine powder under liquid nitrogen using a ceramic mortar and pestle.

For all samples, the quality of the total RNA extracted was determined by using a BioAnalyzer 2100 (Agilent Technologies) and the quantity determined by using a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences). Total RNA was stored at −80°C until further processing.

Preparation of Biotin-Labeled cRNA and Microarray Processing. Fifteen micrograms of fragmented biotin-labeled cRNA was prepared from 5 μg of total RNA and hybridized to the 3AspergDTU GeneChip (available from Affymetrix on request, order no. 520520F) according to the Affymetrix GeneChip Expression Analysis Technical Manual (30).

cRNA was quantified in a spectrophotometer (as above). cRNA quality was assessed by using a BioAnalyzer. A GeneChip Fluidics Station FS-400 (fluidics protocol FS450_001) and a GeneChip Scanner 3000 were used for hybridization and scanning.

The scanned probe array images (.DAT files) were converted into .CEL files by using the GeneChip Operating Software (Affymetrix).

Analysis of Transcriptome Data. Affymetrix CEL-data files were preprocessed by using the statistical language and environment R (31) version 2.5. The probe intensities were normalized for background by using the robust multiarray average method (32) by using only perfect match (PM) probes. Normalization was performed subsequently by using the quantiles algorithm (33). Gene expression values were calculated from the PM probes with the medianpolish summary method (32). All statistical preprocessing methods were used by invoking them through the affy package (34).

Statistical analysis was applied to determine genes subject to differential transcriptional regulation. The limma package (35) was used to perform moderated Student's *t* tests between the two carbon sources for each of the three species. Empirical Bayesian statistics were used to moderate the standard errors within each gene and Benjamini–Hochberg's method (36) to adjust for multitesting. A cutoff value of adjusted $P < 0.05$ was set to assess statistical significance.

Array Design. Initial probe design was done by using the OligoWiz 2.0 software (27, 28) from the CDS sequences of predicted ORFs from the genome sequences of *A. nidulans* FGSC A4 (5), *A. niger* ATCC 1015 (6), and *A. oryzae* RIB40 (8). For each gene, a maximum of 11 nonoverlapping perfect match probes were calculated by using the OligoWiz standard scoring of cross-hybridization, melting temperature, folding, position preference, and low complexity. A 3′ position preference for the probes was included in the computations. The probes were designed separately for each genome.

Pruning of the probe sequences to comply with Affymetrix recommendations was done by removing duplicate probe sequences and shortening probes that were not possible to synthesize in full length.

Also included on the chip were a number of Affymetrix standard controls, custom controls, an *A. oryzae* EST collection (courtesy of Novozymes), and probes for ORFs from the *Streptomyces. coelicolor* A3 (2, 29) genome.

Comparison of Protein Sequences. The amino acid sequences of the predicted ORFs from each of the three genomes were compared with those of the two others by using blastp (37) with an e-value cutoff of 1e-30. For each protein query sequence that gave one or more positive hits, the best hit was selected based on score (a unidirectional best hit). Bidirectional best hits were found by comparing the lists of best hits for two species against each other and selecting genes where the best hits paired up, thus giving a conservative set of 1:1 homologues for all pairwise comparisons. Tridirectional best hits were found by comparing the lists of bidirectional hits for all comparisons, and selecting the genes that had a 1:1:1 relationship in all comparisons between all three species (see SI Fig. 4).

Detection of Conserved Motifs. Conserved motifs were identified by using R 2.5 (31) with the cosmo package (38). Default settings were used with the following exceptions: a background Markov model was computed by using the intergenic regions from scaffold 1 of the *A. niger* ATCC genome sequence. Intergenic regions containing unknown bases (Ns) were pruned from the training set leaving 1.7 Mb in 1,214 sequences. The two-component-mixture model was used to search for conserved motifs. The maximum number of sites were increased to include all 102 sites. For all query sequences, 1,000 bp up-stream of the start codon of the gene was used, or, in the case of some *A. niger* genes, 1,000 bp upstream of the predicted transcription start. Only 120 bp was available of the AN4590 promoter.

P values were calculated as $P(X \leq n)$, with X being a Poisson-distributed stochastic variable with $\lambda = 0.418$ and n being the number of motifs found per kilobase. *lambda* was calculated as the number of the conserved motif found per kilobase of the intergenic training set.

ACKNOWLEDGMENTS. We thank Rasmus Wernersson for help with the addition of the data to the OligoWiz database, Michael L. Nielsen for revision of the manuscript, Scott E. Baker for good scientific discussions, and Lene Christiansen and Tina Johansen for indispensable technical assistance. This work was partially funded by Chr. Hansen, Novozymes, and the Lundbeck Foundation (funding of array design); Danish Research Agency for Technology and Production (M.R.A. and L.L.); Novozymes Bioprocess Academy and the Technical University of Denmark (W.V.); National Council of Research Conacyt-Mexico and the Mexican Ministry of Public Education SEP (M.P.S.); and the Danish Research Council for Technology and Production Sciences (G.P.). We thank the reviewers for their comments, which significantly improved the manuscript.

1. Baker SE, Bennett JW (2008) in *The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods*, eds Osmani SA, Goldman GH (CRC Press, Boca Raton, FL), pp 3–13.
2. Hjort CM (2003) in *Genetically Engineered Food*, ed Heller KJ (Wiley, VCH, Wennheim, Germany), pp 86–99.
3. Nevalainen KH, Te'o VSJ, Bergquist P (2005) Heterologous protein expression in filamentous fungi. *Trends Biotechnol* 23:468–474.
4. Pontecorvo G (1953) The genetics of *Aspergillus nidulans. Adv Genet* 5:141–238.
5. Galagan JE, *et al.* (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae. Nature* 438:1105–1115.
6. Baker S (2006) Aspergillus niger genomics: past, present and into the future. *Med Mycol* 44:S17–S21.
7. Pel HJ, *et al.* (2007) Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat Biotechnol* 25:221–231.
8. Machida M, *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae. Nature* 438:1157–1161.
9. Jørgensen TR, *et al.* (2007) Glucose uptake and growth of glucose-limited chemostat cultures of *Aspergillus niger* and a disruptant lacking MstA, a high-affinity glucose transporter. *Microbiology* 153:1963–1973.
10. van Peij N, Visser J, de Graaff L (1998) Isolation and analysis of *xlnR*, encoding a transcriptional activator co-ordinating xylanolytic expression in *Aspergillus niger. Mol Microbiol* 27:131–142.
11. Marui J, *et al.* (2002) A transcriptional activator, AoXlnR, controls the expression of genes encoding xylanolytic enzymes in *Aspergillus oryzae. Fungal Genet Biol* 35:157–169.
12. Hasper AA, Visser J, de Graaf LH (2000) The *Aspergillus niger* transcriptional activator XlnR, which is involved in the degradation of the polysaccharides xylan and cellulose, also regulates D-xylose reductase gene expression. *Mol Microbiol* 36:193–200.
13. van Peij NN, Gielkens MM, de Vries RP, Visser J, de Graaff LH (1998) The transcriptional activator XlnR regulates both xylanolytic and endoglucanase gene expression in *Aspergillus niger. Appl Environ Microbiol* 64:3615–3619.
14. Tsukagoshi N, Kobayashi T, Kato M (2001) Regulation of the amylolytic and (hemi-)cellulolytic genes in aspergilli. *J Gen Appl Microbiol* 47:1–19.
15. vanKuyk PA, de Groot MJL, Ruijter GJG, de Vries RP, Visser J (2001) The *Aspergillus niger* D-xululose kinase gene is co-expressed with genes encoding arabinan degrading enzymes, and is essential for growth on D-xylose and L-arabinose. *Eur J Biochem* 268:5414–5423.
16. de Vries RP, van de Vondervoort P, Hendriks L, van de Belt M, Visser J (2002) Regulation of the α-glucuronidase-encoding gene (aguA) from *Aspergillus niger. Mol Genet Genom* 268:96–102.
17. de Vries RP, Visser J (1996) Regulation of the feruloyl esterase (*faeA*) gene from *Aspergillus niger. Appl Environ Microbiol* 65:5500–5503.
18. de Vries RP, *et al.* (1996) Differential expression of three α-galactosidase genes and a single β-galactosidase gene from *Aspergillus niger. Appl Environ Microbiol* 65:2453–2460.
19. Gielkens MM, Dekkers E, Visser J, de Graaff LH (1996) Two cellobiohydrolase-encoding genes from *Aspergillus niger* require D-xylose and the xylanolytic transcriptional activator XlnR for their expression. *Appl Environ Microbiol* 65:4340–4345.
20. Hasper AA, Dekkers E, van Mil M, van de Vondervoort PJI, de Graaf LH (2002) EglC, a new endoglucanase from *Aspergillus niger* with major activity towards xyloglucan. *Appl Environ Microbiol* 68:1556–1560.
21. Marui J, Kitamoto N, Kato M, Kobayashi T, Tsukagoshi N (2002) Transcriptional activator, AoXlnR, mediates cellulose-inductive expression of the xylanolytic and cellulolytic genes in *Aspergillus oryzae. FEBS Lett* 528:279–282.
22. Pérez-Gonzalez JA, de Graaff LH, Visser J, Ramón D (1996) Molecular cloning and expression in *Saccharomyces cerevisiae* of two *Aspergillus nidulans* xylanase genes. *Appl Environ Microbiol* 62:2179–2182.
23. MacCabe AP, Orejas M, Pérez-González JA, Ramón D (1998) Opposite patterns of expression of two *Aspergillus nidulans* xylanase genes with respect to ambient pH. *J Bacteriol* 180:1331–1333.
24. MacCabe AP, Fernández-Espinar MT, de Graaff LH, Visser J, Ramón D (1996) Identification, isolation and sequence of the *Aspergillus nidulans xlnC* gene encoding the 34-kDa xylanase *Gene* 175:29–33.
25. Pettersen RC (1984) in *The Chemistry of Solid Wood*, ed Rowell RM (Am Chem Soc, Washington, DC), pp 57–126.
26. Carpita NC, Gibeaut DM (1993) Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J* 3:1–30.
27. Nielsen H, Wernersson R, Knudsen S (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. *Nucleic Acids Res* 31:3491–3496.
28. Wernersson R, Nielsen H (2005) OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res* 33:W611–615.
29. Bentley S, *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417:141–147.
30. Affymetrix (2007) *GeneChip Expression Analysis Technical Manual, P/N 702232.* (Affymetrix, Santa Clara, CA), Rev. 2.
31. R Development Core Team (2007) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna).
32. Irizarry R, *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
33. Bolstad B, Irizarry R, Åstrand M, Speed T (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.
34. Gautier L, Cope L, Bolstad B, Irizarry R (2004) Affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–315.
35. Smyth G (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article 3.
36. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing *J R Stat Soc B* 57:289–300.
37. McGinnis S, Madden T (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32:W20–W25.
38. Bembom O, Keles S, van der Laan M (2007) Supervised detection of conserved motifs in DNA sequences with cosmo *Stat Appl Genet Mol Biol* 6:Article 8.

Paper 3

Genome-wide analysis of maltose utilization and regulation in aspergilli

Wanwipa Vongsangnak*, Margarita Salazar*,
Kim Hansen, and Jens Nielsen
*Equal contribution

Microbiology, 2009

Microbiology Papers in Press. Published August 20, 2009 as doi:10.1099/mic.0.031104-0

Genome-wide analysis of maltose utilization and regulation in aspergilli

Wanwipa Vongsangnak[1,¤], Margarita Salazar[1,¤], Kim Hansen[2], and Jens Nielsen[1,*]

[1]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

[2]Novozymes A/S, DK-2880 Bagsvaerd, Denmark

*Correspondence to: Jens Nielsen, E-mail: nielsenj@chalmers.se, Telephone: +46 (31) 772 38 04, Fax: +46 (31) 772 38 01

[¤]Equal contribution

Received on May 22, 2009; revised on August 4, 2009; accepted on August 17, 2009

ABSTRACT

Maltose utilization and regulation in aspergilli is of great importance for cellular physiology and industrial fermentation processes. In *Aspergillus oryzae*, maltose utilization requires a functional *MAL* locus, each composed of three genes: *MALR* encoding a regulatory protein, *MALT* encoding maltose permease and *MALS* encoding maltase. Through a comparative genome and transcriptome analysis we show that the *MAL* regulon system is active in *A. oryzae* while it is not present in *Aspergillus niger*. In order to utilize maltose, *A. niger* requires a different regulatory system that involves the AmyR regulator for glucoamylase (*glaA*) induction. Analysis of reporter metabolites and subnetworks illustrate the major route of maltose transport and metabolism in *A. oryzae*. This demonstrates that overall metabolic responses of *A. oryzae* occur in terms of genes, enzymes, and metabolites when altering the carbon source. Although the amount of knowledge on maltose transport and metabolism is far from being complete in *Aspergillus* spp., our study not only helps to understand the sugar preference in industrial fermentation processes, but also indicates how maltose affects gene expression and overall metabolism.

INTRODUCTION

Aspergillus represents a large genus of filamentous fungi, and several species have a long history of application as cell factories in fermentation and food industries. In China and Japan, *A. oryzae* is used for production of alcoholic beverage, soy bean paste, soy sauce and rice vinegar as well as for production of various hydrolytic

enzymes, such as α-amylase (EC: 3.2.1.1) (Baker & Bennett, 2008). *A. niger* is largely exploited for the production of extracellular enzymes, such as glucoamylase (EC: 3.2.1.3) (Baker & Bennett, 2008; Magnuson & Lasure, 2004). Maltose utilization and regulation in aspergilli is of great value for academic and industrial research. Maltose is one of the most effective inducers for enzyme production in aspergilli such as α-amylase production by *A. oryzae* (Carlsen *et al.*, 1996; Carlsen & Nielsen, 2001), but also for glucoamylase production measured by enzymatic activity in *A. niger* and *Aspergillus nidulans* (Kato *et al.*, 2002; Larry *et al.*, 1972). However, compared to the yeast *Saccharomyces cerevisiae* little is known about maltose utilization, transport and regulation at the molecular level in aspergilli. Regulation of maltose transport and metabolism by yeast is well known (Klein *et al.*, 1996; Novak *et al.*, 2004). Maltose utilization in *S. cerevisiae* is under control of three general regulatory mechanisms: induction, glucose repression and glucose inactivation (Novak *et al.*, 2004). The presence of maltose in the environment is necessary for induction of synthesis of maltase and maltose transporter. The metabolism and regulation of maltose requires the presence of *MAL* loci, of which there are several identified in different strains of *S. cerevisiae*, but the *MAL6* locus is the most well studied (Klein *et al.*, 1996). The gene structure of the *MAL6* locus is composed of a cluster of three genes: *MAL61* (*MALT*) encoding maltose permease, *MAL62* (*MALS*) encoding maltase (EC: 3.2.1.20) and *MAL63* (*MALR*), encoding a transcriptional activator specifically activating expression of the *MALT* and *MALS* genes (Needleman *et al.*, 1984). Expression of both *MALT* and *MALS* is carbon catabolite repressed by glucose through the transcription factor Mig1 and coordinately induced by maltose (Klein *et al.*, 1996).

In recent years, available genome sequences of filamentous fungi such as *Aspergillus* species has allowed for studying metabolism in greater details, e.g. maltose metabolism and transport, as well as filling existing gaps at the molecular level. Since the release of ten different *Aspergillus* species genome sequences (Baker, 2006; Fedorova *et al.*, 2008; Galagan *et al.*, 2005; Machida *et al.*, 2005; Nierman *et al.*, 2006; Payne *et al.*, 2006; Pel *et al.*, 2007; Yu. J. *et al.*, 2005), several transcriptome analysis have been conducted with some of these sequences (Breakspear & Momany, 2007), and recently we developed an Affymetrix GeneChip that allows for transcriptome analysis of any of the three aspergilli: *A. oryzae*, *A. niger* and *A. nidulans* (Andersen *et al.*, 2008). In this study, we aim to identify the *MAL* gene cluster in different sequenced *Aspergillus* genomes using the gene structure of the *MAL*6 locus of *S. cerevisiae* as a model. We further validate the presence or absence of the *MAL* gene cluster in *A. oryzae* and *A. niger* by using our custom designed Affymetrix GeneChip for transcriptome analysis (Andersen *et al.*, 2008). To do this, we performed batch cultivations of two *Aspergillus* species (i.e. *A. oryzae* and *A. niger*) on glucose and maltose as carbon source, and we then performed comparative transcriptome analysis to examine expression of putative *MAL* gene clusters in *A. oryzae* and in *A. niger*. Furthermore, our study allowed for mapping the transcriptional response onto the metabolic network of *A. oryzae* when adjusting metabolism to a change of carbon source, from glucose to maltose.

METHODS

Strains

The strains used were *A. oryzae* wild type strain A1560 (an ancestor of strains used for commercial α-amylase production) and *A. niger* wild type strain BO1 (an ancestor of strains used for commercial glucoamylase production), both obtained from Novozymes (Carlsen & Nielsen, 2001; Pedersen *et al.*, 2000). *A. oryzae* stock culture was maintained on Cove-N-Gly agar at 4°C and *A. niger* stock culture was maintained as frozen spore suspensions at -80°C in 20% (v/v) glycerol.

Medium and inoculums

The details of spore propagation medium, pre-culture medium, and batch cultivation medium for *A. oryzae* or *A. niger* are described in Supplementary file 1. For initial fermentation process, *A. oryzae* A1560 fermenters were inoculated with ~60 g of broth of *A. oryzae* A1560 cultured at 30°C for 24 h on pre-culture liquid medium in shake flasks at 250 rpm. The pre-culture was inoculated with 5 ml of spore solution harvested from mycelium grown on spore propagation medium agar at 34°C for 3–4 days. Spores were harvested with Tween 80 0.1% (v/v). The *A. niger* fermenters were inoculated with spores ($5.7x10^9$ spores l^{-1}) previously propagated on spore propagation medium for 8 days at 30 °C. Spores were harvested by adding Tween 80 0.01% (v/v).

Batch cultivations

A. oryzae batch cultivations were done in 2 L bioreactors with a working volume of 1.2 L. The stirrer speed was kept at 800 rpm during the first 4 h and then increased to 1100 rpm. The pH was controlled at 6 by addition of 10% (v/v) of H_3PO_4 or 10% (v/v) NH_3 solution, and the temperature was maintained at 34°C. The aeration flow rate was set at 1.2 vvm (volume of gas per volume of liquid per minute). Dissolved oxygen tension was initially calibrated at 100%. The concentrations of oxygen (O_2) and carbon dioxide (CO_2) in the exhaust gas were monitored with a gas analyzer (Magnos 4G for O2, Uras 3G for CO2, Hartmann & Braun, Germany).

A. niger batch cultivations were carried out in 4.5 L bioreactors with a working volume of 4 L. The bioreactors were equipped with two Rushton four-blade disc turbines, pH and temperature control and no baffles. Inlet air was controlled with a mass flow meter. The concentrations of oxygen (O_2) and carbon dioxide (CO_2) in the exhaust gas were monitored with a gas analyzer (1311 Fast response Triple gas, Innova combined with multiplexer controller for Gas Analysis MUX100, B. Braun Biotech International). The temperature was maintained at 30 °C and the pH was controlled by automatic addition of 2 N NaOH. The pH was initially set to 3.0 to prevent spores aggregation. Once the spores started to germinate the pH was increased to 4.5 and kept constant through the cultivation. The stirring speed was initially set to 200 rpm and the aeration rate was set to 0.05 vvm. After germination, the stirring speed was increased to 600 rpm and aeration rate was raised to 0.89 vvm and kept steady throughout all the rest of the fermentation.

Sampling

Cell dry weight was determined by filtration. A known volume of cell culture was filtered, and then dried on the filter at 100°C for 24 h and cooled down in a desiccator. The filter with dried cell mass was weighed afterwards. The culture supernatant was obtained after centrifugation of original samples and subsequently frozen at -20°C for sugars and extracellular metabolites measurements. For gene expression analysis, mycelium was harvested at the early-mid exponential phase and then cultures were filtered through sterile filtration Miracloth (Calbiochem, San Diego, CA, USA). At this point, the mycelium was washed with distilled water or 0.9% (w/v) NaCl solution. The mycelium was quickly dried by squeezing and subsequently frozen in liquid nitrogen. Samples were stored at -80°C until RNA extraction.

Sugars and extracellular metabolites measurements

The concentration of sugars and extracellular metabolites were measured by HPLC analysis on an Aminex HPX-87H ion-exclusion column (BioRad, Hercules, CA) with previous filtration by 25 mm GD/X syringe filter, 0.45 µm pore size (Whatman, Inc, USA). The column was eluted at 45°C for *A. oryzae* or 60°C for *A. niger* with 5 mM H_2SO_4 at a flow rate of 0.6 ml min^{-1} Extracellular metabolites were detected with a refractive index detector and an UV detector.

Total RNA extraction

A. oryzae total RNA was extracted by using the Promega RNAgents Total RNA Isolation system, according to the protocol for purification of total RNA from fungi. For RNA extraction, ~1 g of frozen mycelium was ground to a fine powder under liquid nitrogen using a ceramic mortar and pestle. Total RNA of *A. niger* was isolated using the Qiagen RNeasy Mini Kit (QIAGEN Nordic, Ballerup, Denmark), according to the protocol for isolation of total RNA from plant and fungi. For RNA isolation, ~ 100 mg of frozen mycelium were placed in a 2 ml tube, pre-cooled in liquid nitrogen, containing three RNase-treated steel balls. The tubes were then shaken in a Mixer Mill, at 5°C for 10 minutes, until the mycelium was ground to powder, and thus ready for extraction of total RNA. For all samples, the quality of the total RNA extracted was determined by using a BioAnalyzer (2100 BioAnalyzer, Agilent Technologies Inc., Santa Clara, CA, USA) and the quantity determined by using a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Total purified RNA was stored at -80°C until further microarray processing.

Microarray manufacturing and design

Affymetrix arrays were used for the analysis of the transcriptome of *A. oryzae* and *A. niger* (Affymetrix company, Santa Clara, CA, USA). The arrays were packaged in an Affymetrix® GeneChip cartridge (49 format), and were processed with GeneChip reagents in the GeneChip® Instrument System. The design and selection of probes for interrogating the ORFs within the genome of *A. oryzae* and *A. niger* was performed by Andersen and coworkers (Andersen *et al.*, 2008). The array contains only perfect match (PM) probes which correspond to 25-base oligonucleotides perfect complementary to the transcript. Of the 13,120 putative genes identified in the genome of *A. oryzae* (Machida *et al.*, 2005; Vongsangnak *et al.*, 2008), 12,039 probe sets were used for microarray analysis. Of the 11,200 putative genes identified in the genome of *A. niger*, 11,122 probe sets were used for microarray analysis. Each of the probe sets were composed of 11 probes (whenever possible) of 25 oligomers (Andersen *et al.*, 2008).

Biotin-labeled cRNA and microarray processing

Biotin-labeled cRNA was prepared from ~ 5 µg of total RNA, according to the protocol described in the Affymetrix GeneChip® Expression Analysis Technical Manual (Affymetrix & GeneChip, 2007). The cRNA was cleaned before fragmentation by using the Qiagen RNeasy Mini Kit (protocol for RNA Cleanup), in order to guarantee the quality of cRNA samples for further processing. Biotin-labeled cRNA was quantified in a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Then, 20 µg of cRNA were fragmented following manufacturer protocol and ~ 15 µg of fragmented cRNA was hybridized to the *Aspergillus* Affymetrix chip (Andersen *et al.*, 2008) according to the Affymetrix GeneChip® Expression Analysis protocol (Affymetrix & GeneChip, 2007). Arrays were washed and stained using a GeneChip® Fluidics Station FS-400, and scanned on an Agilent GeneArray® Scanner.

Microarray data acquisition and analysis

Affymetrix CEL-data files were preprocessed using bioconductor (Gentleman *et al.*, 2004) and R package version 2.5.1 (R Development Core Team). Normalization was performed by using the qspline algorithm (Workman *et al.*, 2002). Normalized gene expression data set is presented in Supplementary file 2. The probe intensities were corrected for background by using the robust multiarray average method (Irizarry *et al.*, 2003) by using all the probes. Gene expression values were calculated from the probes associated with each gene with the medianpolish summary method (Irizarry *et al.*, 2003). All statistical preprocessing methods were invoked through the affy package (Gautier *et al.*, 2004) and R scripts (Dudoit *et al.*, 2003). Statistical analysis was applied to determine significantly different gene expressions. The limma package (Smyth *et al.*, 2005) was used to perform moderated Student's *t* tests for pairwise carbon source comparisons. Empirical Bayesian statistics were used to moderate the standard errors within each gene and Benjamini-Hochberg's method to adjust for multitesting (Benjamini & Hochberg, 1995). A cut-off value of adjusted $P<0.05$ was set to assess statistical significance.

Identification of *MAL* gene cluster based on protein sequence analysis

A cluster of three genes from the *S. cerevisiae MAL*6 locus was used as a scaffold model to identify the *MAL* gene cluster in 10 *Aspergillus* genome sequences. The three genes used were *MAL61*, *MAL62*, and *MAL63* from *S. cerevisiae* which amino acid sequences were extracted from GenBank database (http://www.ncbi.nlm.nih.gov/) with accession numbers P15685.1, P07265.1, and P10508.1. The complete set of three amino acid sequences of *S. cerevisiae* was further used as query and searched against the amino acid sequences of 10 different sequenced *Aspergillus* genomes by using BLASTP (Altschul *et al.*, 1990). The sequenced species included were: *A. oryzae* RIB40 (Machida *et al.*, 2005)*, A. niger* CBS 513.88 (Pel *et al.*, 2007), *A. niger* ATCC 1015 (version 3) (http://genome.jgi-psf.org/Aspni5/Aspni5.home.html), *A. nidulans* FGSC A4 (version 4) (Galagan *et al.*, 2005; Wortman *et al.*, 2009), *Aspergillus fumigatus* Af293 (Nierman *et al.*, 2006), *Aspergillus fumigatus* A1163 (Fedorova *et al.*, 2008), *Aspergillus flavus* NRRL 3357 (Payne *et al.*, 2006), *Aspergillus terreus* NIH2624 (www.broad.mit.edu/annotation/fungi/aspergillus_terreus), *Aspergillus clavatus* NRRL 1 (Fedorova *et al.*, 2008) and *Aspergillus fischeri* NRRL 181 (Fedorova *et al.*, 2008). An estimated expectation value cut-off of less than 1E-100, more than 40% identity, and more than 500 bps of alignment length was set to assess statistical significance for identification of any orthologues. For identification of *MAL* gene cluster based on synteny analysis, the expectation value cut-off of less than 1E-05, more than 20% identity, and more than 200 bps of alignment length was set to assess statistical significance.

Reporter metabolites and subnetwork analysis

The reporter metabolites and highly correlated metabolic subnetwork algorithm was applied as described by Patil and Nielsen (Patil & Nielsen, 2005). The analysis was run for *A. oryzae*, for the glucose versus maltose carbon source pairwise comparison. For this purpose, information on the topology of the reconstructed metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008) was used in combination with the adjusted p-values obtained from the Student's t-test analysis.

RESULTS

Comparative analysis between *MAL* gene cluster in *S. cerevisiae* and *Aspergillus* species

First we searched for the presence of the *MAL* gene cluster in 10 different sequenced *Aspergillus* genomes. For this purpose, the gene structure of the *MAL*6 locus of *S. cerevisiae* was used as a model and BLASTP was applied (See METHODS). The results showed that six different *Aspergillus* strains (i.e. *A. oryzae*, two strains of *A. fumigatus*, *A. flavus*, *A. clavatus*, and *A. fischeri*) contain at least one *MAL* gene cluster as illustrated in Fig. 1. Interestingly, *A. oryzae* and *A. flavus* contain at least two *MAL* gene clusters. Phylogenetic analysis suggests that events of gene duplication and horizontal gene transfer may have occurred (See Supplementary file 3). In contrast, we could not find any *MAL* cluster in *A. nidulans*, *A. terreus* and two strains of *A. niger* under the statistical constraints imposed. These results could suggest that these four *Aspergillus* strains most likely do not have the *MAL* regulon for maltose utilization. Notably, in all the sequenced *Aspergillus* genomes, we could identify multiple orthologue genes encoding maltase or α-glucosidase enzymes and maltose transporters as shown in Fig. 1. Statistical values of orthologous genes are presented in Supplementary file 3.

To prove the presence or absence of the *MAL* gene cluster at the transcriptional level, we further evaluated our results obtained from comparative genomics through transcriptomics analysis. In the following, we show an example of using our previously designed *Aspergillus* GeneChip (Andersen *et al.*, 2008) to validate the presence of *MAL* gene cluster in the *A. oryzae* genome and the absence of *MAL* gene cluster in the *A. niger* genome.

Growth physiology

To evaluate the physiology and validate the presence or absence of the *MAL* gene cluster as well as to analyze the regulatory response when adjusting metabolism to a change of carbon source, i.e. from glucose to maltose, we grew the two *Aspergillus* species in well-controlled bioreactors to perform reproducible fermentations.

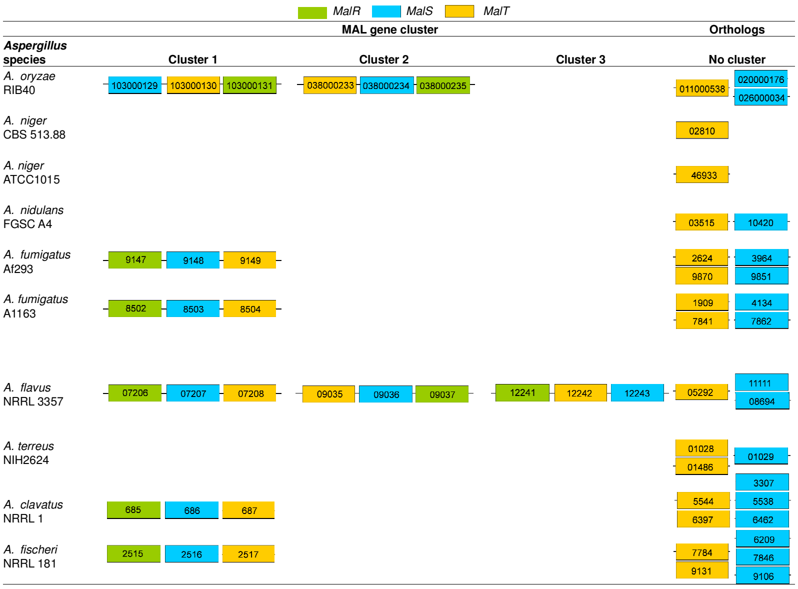


Fig. 1. - Diagram shows comparative sequence analysis of *MAL* gene cluster between *S. cerevisiae* and 10 different *Aspergillus* species. Values in each rectangle represent a shortened ORF name. For individual full name, the shorten ORF is prefixed by "AO090" for *A. oryzae* RIB40, "An02g" for *A. niger* CBS 513.88, "JGI" for *A. niger* ATCC 1015, "ANID_" for *A. nidulans* FGSC A4, "ORF" for *A. fumigatus* Af293 and A1163, *A. clavatus* NRRL 1 as well as *A. fischeri* NRRL 181, "AFL2T_" for *A. flavus* NRRL 3357, "ATET_" for *A. terreus* NIH2624.

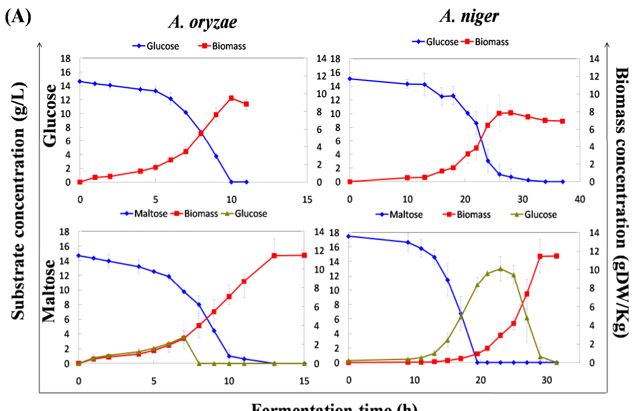
All batch cultivations of *A. oryzae* and *A. niger* were carried out on a defined, minimal medium with glucose or maltose as the sole carbon source. Each species had its own specific cultivation medium optimized for growth of those species (Andersen *et al.*, 2008; Pedersen *et al.*, 2000) (See METHODS section for details). For each of the two species, three biological replicate cultivations were performed on each carbon source. The results are shown in Fig. 2. Panel A shows the biomass growth and substrate concentration profiles of *A. oryzae* and *A. niger*. For both species, these two carbon sources were completely consumed. In comparison to *A. niger*, *A. oryzae* grew faster than *A. niger* in the two carbon sources used and they were completely consumed at different rates. Glucose was exhausted in 10 h and maltose in 12 h, at rates of 3.09±0.02 g $l^{-1}.h^{-1}$ and 2.46±0.02 g $l^{-1}.h^{-1}$, respectively. The maximum specific growth rate of *A. oryzae* on glucose was 0.38±0.01 h^{-1}, which is due to the high efficiency in the uptake and metabolism of this sugar. Slightly slower growth was achieved on maltose, where the maximum specific growth rate was 0.32±0.05 h^{-1}. As shown in Fig. 2, almost no accumulation of glucose was seen in the media during growth on maltose. With *A. niger,* glucose was exhausted in 32 h, and maltose was consumed after 19 h. As indicated in Fig. 2, *A. niger* growth on maltose was faster than on glucose, where a maximum specific growth rate of 0.22±0.01 h^{-1} was achieved. In the case of maltose consumption, there was an accumulation of glucose due to a very high extracellular glucosidase activity expressed by *A. niger*, which allowed the fungus to grow very fast on this carbon source at a maximum specific growth rate of 0.31±0.02 h^{-1}. Besides growth rates and biomass yields of *A. oryzae* and *A. niger*, transcriptional analysis (TA) sampling times and biomass yields at the specific TA sampling time were recorded for the two microorganisms on the two carbon sources (See Fig. 2, panel B).

Comparative transcriptome analysis

To further test our assumption obtained from comparative genomics for *MAL* regulon existing in *A. oryzae* or not present in *A. niger*, the genome-wide gene expression data obtained from glucose or maltose cultivations were pairwise compared for each species. To detect transcriptional changes in response to a change in the carbon source, Student's t-test statistics were used to identify significantly different gene expression levels with a p-value cut-off of 0.05. This cut-off p-value was adjusted by the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) for correction of multiple testing. Table 1 shows a list of genes that were significantly differentially expressed in *A. oryzae* between glucose and maltose (16 gene expression changes). In contrast, for *A. niger*, no genes were statistically differentially expressed when using glucose or maltose as carbon source.

Analysis of the *MAL* regulon

As shown in Table 1, 16 genes showed higher expression level on maltose compared to glucose in *A. oryzae*. It is suggested that these genes are induced during metabolism of maltose. Among these genes, 10 protein-encoding genes were functionally annotated and involved in polysaccharide and disaccharide metabolism, such as glucoamylase, maltose permease, maltase, sugar transporters and maltose-O-acetyltransferase (EC: 2.3.1.79 which acetylates oligosaccharides). In the yeast *S. cerevisiae*, regulation of maltose utilization occurs by the *MAL* regulon (Chow *et al.*, 1989) via the transcription activator (*MALR*), which induces maltose permease (*MALT*) and maltase (*MALS*) gene expression.



(B)

Strain	Carbon source	μ_{max} (h^{-1})	Ysx (g DW/g C source)	Sampling time for TA (h)	Biomass for TA (g DW/Kg)
A. oryzae	Glucose	0.38±0.01	0.51±0.01	6±0	2.50±0.09
	Maltose	0.32±0.05	0.49±0.05*	7±0	2.27±0.09
A. niger	Glucose	0.22±0.01	0.57±0.05	21±1	3.74±0.06
	Maltose	0.31±0.02	0.62±0.02*	24±0	3.55±0.51

Fig. 2. - Biomass and substrate concentration profiles with the different carbon sources (glucose and maltose)
(A) Batch cultivation with *A. oryzae* and *A. niger* (B) Physiological data for all cultivations, maximum specific growth rate, biomass yield, sampling time for transcriptional analysis (TA), and biomass concentration at the time of sampling for transcriptional analysis (TA) are given. Average values and standard deviations are calculated from three biological replicates. *Biomass yield was calculated based on glucose (g DW/g glucose)

Moreover, we also found up-regulated genes encoding maltose permease, AO090103000130 and AO090038000233, which are the functionally related orthologous genes of *MALT* in *S. cerevisiae* (See Fig. 1). The two *A. oryzae* genes orthologous to the *S. cerevisiae* *MALR* transcription factor, AO090103000131 and AO090038000235, were also up-regulated, but not statistically significant. This suggests that the mechanism behind the *MAL* regulon in *A. oryzae* is similar to *MALR* function in *S. cerevisiae*, where it is activated by maltose and repressed by glucose. Based on Regulatory Sequence Analysis Tools (RSAT) (Thomas-Chollier *et al.*, 2008) for identification of transcription factor binding site in upstream region of *MALR* in *A. oryzae*, we could putatively identify CreA binding site 5'-GCGGGG-3' (Cubero & Scazzocchio, 1994; Drysdale *et al.*, 1993) in upstream sequences where this site was found at base position 64 with respect to the start codon of AO090038000235. This result suggests that the presence of glucose represses *MALR* expression and the formation of active conformation of *MALR* protein as it occurs in *S. cerevisiae* (Klein *et al.*, 1996). According to the results of protein sequence analysis and synteny gene analysis (Sinha & Meller, 2007) of the *MAL* gene cluster, we could conclude that *A. oryzae* has two *MAL* regulons and each regulon contains two genes (i.e. AO090103000131 and AO090038000235) that are likely to be *MALR* transcriptional activators (See Fig. 1). From this significant evidence combined with the physiological response of *A. oryzae* growth on maltose (See Fig. 2), where *A. oryzae* continuously consumed maltose having almost no accumulated glucose over time, we propose that *A. oryzae* has systematic regulation of maltose utilization by these two *MAL* regulons, where *MALR* transcription factor induces maltose permeases (*MALT*) to transport extracellular maltose into the cell and *MALR* also induces maltase (*MALS*) that hydrolyzes intracellular maltose into glucose which is then channeled through glycolysis. Fig. 3 illustrates the proposed mechanisms for regulation of maltose utilization in *A. oryzae* (Panel A).

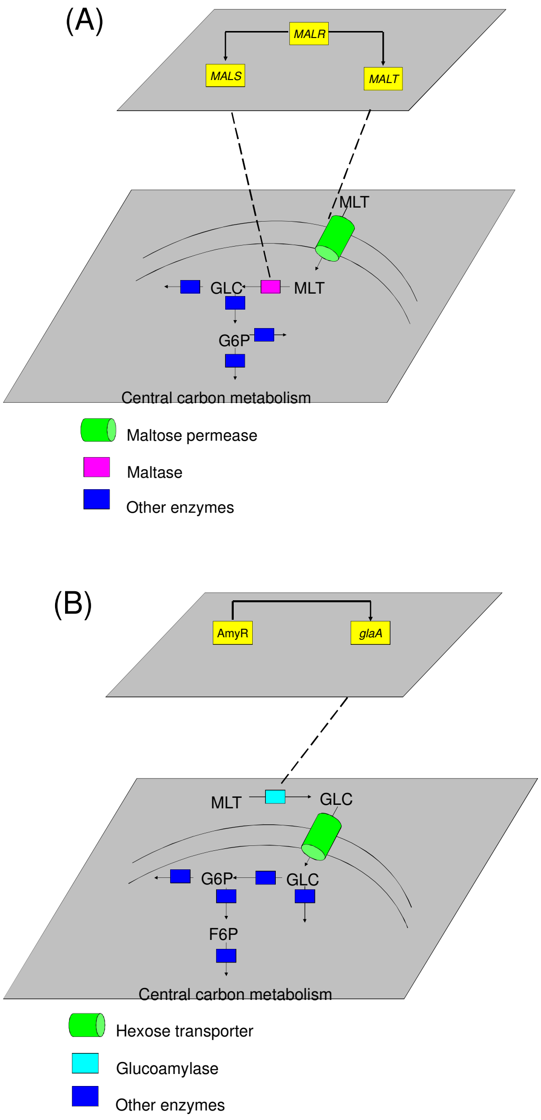


Fig. 3. - Diagram shows comparative maltose utilization and regulation in *A. oryzae* (A) and in *A. niger* (B)

Table 1 Pairwise carbon source comparison for gene expression analysis in *A. oryzae* under the cut-off p-value<0.05

Gene name	Protein function	Fold change	Up/Down	Adjust P-value
AO090103000130	Maltose permease	30.9	Up	2.99E-06
AO090103000129	Maltase	304.4	Up	2.03E-05
AO090701000246	Hypothetical protein	9.4	Up	1.87E-04
AO090038000233	Maltose permease	61.4	Up	2.82E-04
AO090038000234	Maltase	213.8	Up	2.82E-04
AO090012000888	Maltose-O-acetyltransferase	12.6	Up	1.24E-03
AO090003000576	Hypothetical protein	6.7	Up	5.83E-03
AO090003000321	Glucoamylase	9.3	Up	7.32E-03
AO090003000256	Hypothetical protein	3.1	Up	7.32E-03
AO090038000471	Maltase	11.7	Up	2.92E-02
AO090012000893	Hypothetical protein	2.0	Up	3.46E-02
AO090011000612	NAD binding Rossmann fold oxidoreductase	1.9	Up	3.46E-02
AO090023000245	Glucose transporter	9.1	Up	3.71E-02
AO090026000775	Monosaccharide transporter	4.2	Up	3.72E-02
AO090010000746	Glucoamylase	2.5	Up	3.90E-02
AO090012000896	Hypothetical protein	2.0	Up	4.39E-02

In contrast, we could not identify any *MAL* gene cluster in *A. niger* that is closely homologous to the one existing in *S. cerevisiae* (See Fig. 1). Furthermore, transcription data analysis of the pairwise comparison between maltose and glucose in *A. niger*, did not show any significant gene expression changes that can point out the presence of *MAL* cluster either. We therefore propose that maltose utilization in *A. niger* most likely do not involve a *MAL* regulon, but occurs through another regulatory system via AmyR regulator and the recent publication of genome-wide expression analysis in *A. niger* by Yuan and coworkers (Yuan *et al.*, 2008) supports these results. Fig. 3 illustrates the proposed mechanisms for regulation of maltose utilization in *A. niger* (Panel B).

Key metabolite identification and metabolic subnetworks analysis

In order to analyze the overall metabolic responses to changes in the carbon source, i.e. using glucose or maltose, we applied the reporter metabolites and subnetworks algorithm to identify key metabolites and to search for highly correlated metabolic subnetworks for the pairwise comparison (Patil & Nielsen, 2005). This analysis relied on the reconstructed genome-scale metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008), and therefore we demonstrated how these metabolic networks can be used to map regulatory responses in this *Aspergillus* spp. The top 15 high-scoring key metabolites for *A. oryzae* are listed in Table 2. To identify high-scoring metabolic subnetworks, we performed subnetwork analysis using the whole reaction set from the reconstructed metabolic network of *A. oryzae*. Fig. 4 shows the list of key genes-encoding enzymes and transporters comprising the subnetwork of *A. oryzae* investigated upon a change of carbon source, from glucose to maltose.

Table 2 List of significant key metabolites from the pairwise carbon source comparison in *A. oryzae* (p-value cut-off < 0.01)

Maltose Versus Glucose	
Key metabolite	P-value
Maltose	0.00E+00
Maltose (Extracellular)	0.00E+00
Maltotriose	3.04E-14
Glucose	1.18E-11
Glucose (Extracellular)	1.44E-09
Glycogen (Extracellular)	2.50E-06
Starch (Extracellular)	2.50E-06
D-fructose (Extracellular)	3.73E-05
2-keto-myo-inositol	4.22E-05
D-galactose	7.41E-05
D-aspartate	1.16E-04
D-arabinono-1,4-lactone	1.60E-04
H2O (Extracellular)	4.88E-04
D-Mannose (Extracellular)	5.85E-04
Alpha, alpha-trehalose (Extracellular)	9.11E-04

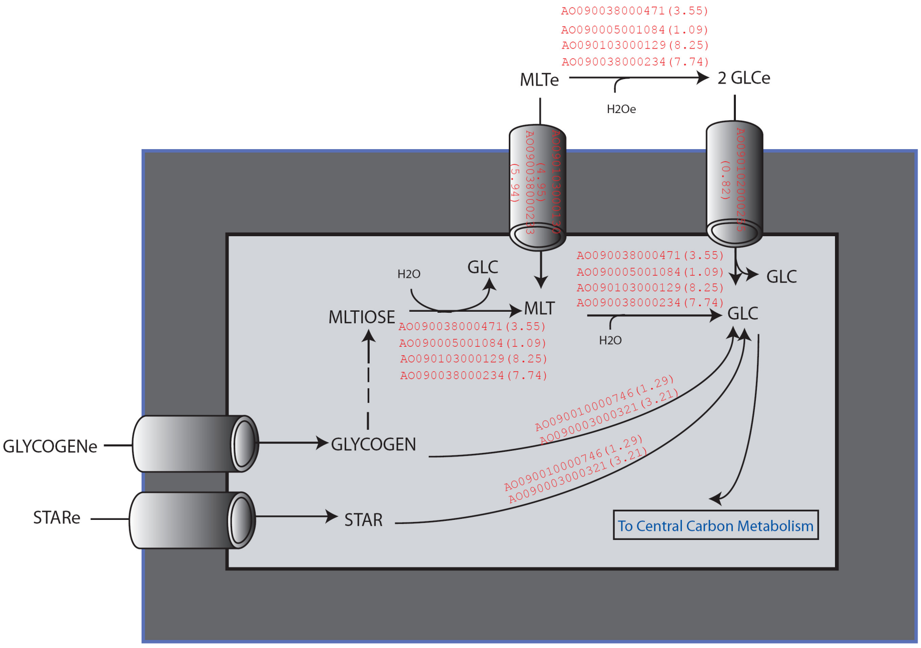


Fig. 4. - Small metabolic subnetwork from pairwise carbon source comparison in *A. oryzae*

Key metabolites, key enzymes and transporters in the small subnetwork identified for a change of carbon source, from glucose to maltose. Gene names marked red are up-regulated upon the change and numerical numbers in parenthesis marked red are $\log_2$ fold-change value upon the change of carbon source. The abbreviations of metabolites are described as follows: GLC, glucose; MLT, maltose; MLTIOSE, maltotriose; STAR, starch; GLYCOGEN, glycogen; H2O, water. Extracellular metabolites are designated by subscript 'e'.

DISCUSSION

Maltose is of economical relevance due to the fact that it is one of the products of starch hydrolysis and it is used as an inducer for protein production in *Aspergillus* species. In this work, we conducted comparative genomics and transcriptomics in *A. oryzae* or *A. niger* in order to investigate maltose uptake, metabolism and regulation. In *S. cerevisiae*, maltose utilization and regulation occurs through the use of the *MAL* regulon. It is composed of three genes, one maltase, one maltose transmembrane transporter, as well as a third gene that encodes a positive regulator, *MALR* (Novak et al., 2004). Interestingly, some bacteria such as *Lactococcus lactis* (Andersson & Rådström, 2002) and *Streptomyces coelicolor* (van Wezel *et al.*, 1997) also make use of the *MAL* regulatory system involving the *MALR* transcriptional activator of their corresponding operons for maltose utilization. Through comparative genomics analysis we were able to significantly identify two *MAL* clusters in *A. oryzae*, but no *MAL* cluster in *A. niger*. Besides, we also found one *MAL* cluster in *A. clavatus*, *A. fumigatus* and *A. fischeri* as well as three *MAL* clusters in *A. flavus*.

We further confirmed the presence of *MAL* clusters in *A. oryzae* at the transcriptional level. The results from genome-wide expression analysis identified a subset of 16 genes to be significantly up-regulated during growth on maltose. Among them, we found *MALS* encoding maltase and *MALT* encoding maltose permease. In addition, according to gene expression data, we also found higher expression level of *MALR* encoding a transcriptional activator in maltose utilization, but the expression changes were not statistically significant. Based on our integrated data analysis, we could conclude that *A. oryzae* can utilize maltose via the *MAL* regulon system.

We further investigated the transcriptional responses to change in the carbon source in batch cultivations, when using glucose or maltose through the use of reporter metabolites and subnetwork analysis (Patil & Nielsen, 2005) together with the reconstructed metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008). The transcriptional responses were in general consistent with the changes expected at the phenotypic level, which indicate that the regulation at the transcriptional level plays a significant role in the overall regulation during growth on these two carbon sources.

In comparison to *A. oryzae*, comparative genome analysis showed that *A. niger* does not have any *MAL* cluster. We further evaluated expression level of the *MAL* gene cluster by transcriptome analysis. The results showed absence of a *MAL* gene cluster in *A. niger*. Our results are consistent with previous transcriptome studies in *A. niger* using the wild type strain N402 (ATCC 9029) (Yuan *et al.*, 2008) that did not identify up-regulation of components of the

MAL regulon when comparing gene expression data on xylose versus maltose cultivations.

Moreover, we did not find any significant genes in *A. niger* when comparing maltose to glucose, although *A. niger glaA* is proposed to be strongly induced on maltose, as described earlier by (Fowler *et al.*, 1990). In order to utilize maltose, we propose that *A. niger* makes use of another regulatory system via AmyR regulator. Yuan *et al.* (Yuan *et al.*, 2008) suggested that AmyR is an important transcription factor that is found in *A. niger* and that *amyR* itself is induced by the presence of maltose. In addition, their studies indicated that *amyR* gene transcription regulation takes place in *A. niger* and showed that a disruption of the AmyR transcription factor resulted in low levels of extracellular enzymes i.e. acid α-amylase (AamA), α-glucosidases (AgdA and a putative AgdB) and glucoamylase (GlaA) converting maltose to glucose and consequently activating a stress response due to low glucose levels. The low availability of glucose transferred a signal to down-regulate glucose transporters (Yuan *et al.*, 2008). This is in accordance with our transcriptome results and the physiological response obtained in *A. niger* maltose cultivations (See Fig. 2), where maltose cleavage occurred faster than glucose uptake and metabolism leading to a high extracellular accumulation of glucose over time. We therefore support the conclusions from previous studies where it is stated that *A. niger* has regulation of maltose utilization by the AmyR transcriptional activator. It activates genes encoding known extracellular starch degrading enzymes, such as *aamA, glaA, agdA* and a putative α-glucosidase *agdB* encoding gene. The *glaA* gene product, an extracellular glucoamylase, can convert extracellular maltose to extracellular glucose and then glucose can be taken up by glucose transporters.
Interestingly, maltose has been found to be better than glucose for glucoamylase production, but it has been reported as well that it is a strain dependent phenomenon, where for some strains there is no difference between the use of any of the two carbon sources, glucose or maltose (Schrickx *et al.*, 1993). Furthermore, studies of *glaA* regulation in the *A. niger* strain ATCC 10864 showed starch, maltose and glucose as positive inducers of glucoamylase production (Fowler *et al.*, 1990). In contrast, previous studies with the *A. niger* BO1 strain used in this study have reported no difference between maltose and glucose as carbon source with respect to glucoamylase productivity as well as identical mRNA levels for growth on maltose or glucose in chemostat cultivations (Pedersen *et al.*, 2000). Based on findings in the literature (Yuan *et al.*, 2008) and our findings, we suggest that *A. niger* utilize maltose by means of extracellular hydrolysis by glucoside hydrolases such as glucoamylase followed by glucose uptake and metabolism.

From comparative genome and transcriptome analysis, we showed that maltose utilization and regulation of *A. oryzae* is very similar to that found in the yeast *S. cerevisiae* with sugar degradation pathways where a number of enzymes and proteins are involved. Our analysis can help to understand how maltose is utilized and regulated in *Aspergillus* species and to convey improvements in industrial practice for protein production in the future.

ACKNOWLEDGEMENTS

The authors would like to thank Lone Vuholm and Anne Kejser Jensen at Novozymes as well as Tina Johansen, Pia Friis and Lene Christiansen at Technical University of Denmark for technical assistance. Finally, we would like to thank Novozymes Bioprocess Academy and the Chalmers University of Technology Foundation for financial support to WV and National Council of Research Conacyt-Mexico and Chalmers University of Technology Foundation for financial support to MS.

DATA DEPOSITION

Normalized gene expression data were deposited at the GEO database (http://www.ncbi.nlm.nih.gov/geo/), with accession numbers GPL5975 (platform), GSM348998-GSM349003 (samples of *A. oryzae*), GSM349007-GSM349012 (samples of *A. niger*), and GSE13868 (series).

SUPPLEMENTARY FILES

The following additional data files are available with the online version of this paper.

Supplementary file 1
File format: DOC
Description: This file provides the details of medium compositions for *A. oryzae* and *A. niger*.

Supplementary file 2
File format: XLS
Description: This file provides the normalized intensities of maltose and glucose conditions considering all the categories of the three biological replicated experiments in *A. oryzae* (Table S1) and *A. niger* (Table S2).

Supplementary file 3
File format: PDF
Description: This file provides Supplementary figure (Fig. S1) and table (Table S1) for statistical details of comparative sequence analysis of *MAL* gene cluster between *S. cerevisiae* and 10 different *Aspergillus* species. Statistical values are presented: E-value, % identity and alignment length. Besides, this file also provides Supplementary figures (Fig. S2-Fig. S5). It shows phylogenetic tree for *MAL* gene cluster.

REFERENCES

Affymetrix & GeneChip (2007). Affymetrix Genechip Expression Analysis Technical Manual. *P/N 702232, Affymetrix, Santa Clara, CA, Revision 2*.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403-410.

Andersen, M. R., Vongsangnak, W., Panagiotou, G., Margarita, P. S., Lehmann, L. & Nielsen, J. (2008). A tri-species *Aspergillus* microarray - advancing comparative transcriptomics. *Proc Nat Acad Sci USA* **105**, 4387-4392.

Andersson, U. & Rådström, P. (2002). Physiological function of the maltose operon regulator, MalR, in *Lactococcus lactis*. *BMC Microbiol* **2: 28**.

Baker, S. E. (2006). *Aspergillus niger* genomics: Past, present and into the future. *Med Mycol* **44**, S17-S21.

Baker, S. E. & Bennett, J. (2008). An Overview of the Genus *Aspergillus. The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods, eds Osmani SA, Goldman GH*, CRC Press, Boca Raton, FL, pp. 3-13.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B-Methodol* **57**, 289-300.

Breakspear, A. & Momany, M. (2007). The first fifty microarray studies in filamentous fungi. *Microbiology-Sgm* **153**, 7-15.

Carlsen, M., Nielsen, J. & Villadsen, J. (1996). Growth and alpha-amylase production by *Aspergillus oryzae* during continuous cultivations. *Journal of Biotechnology* **45**, 81-93.

Carlsen, M. & Nielsen, J. (2001). Influence of carbon source on alpha-amylase production by *Aspergillus oryzae. Appl Microbiol Biotechnol* **57**, 346-349.

Chow, T. H. C., Sollitti, P. & Marmur, J. (1989). Structure of the Multigene Family of Mal Loci in *Saccharomyces. Mol Gen Genet* **217**, 60-69.

Cubero, B. & Scazzocchio, C. (1994). Two different, adjacent and divergent zinc finger binding sites are necessary for CreA mediated carbon catabolite repression in the proline gene cluster of *Aspergillus nidulans. Embo Journal* **13**, 407-415.

Drysdale, M. R., Kolze, S. E. & Kelly, J. M. (1993). The *Aspergillus niger* carbon catabolite repressor gene, creA. *Gene* **130**, 241-245.

Dudoit, S., Gendeman, R. C. & Quackenbush, J. (2003). Open source software for the analysis of microarray data. *Biotechniques*, 45-51.

Fedorova, N., Khaldi, N., Joardar, V. & other authors (2008). Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus. PLoS Genet* **4**.

Fowler, T., Berka, R. M. & Ward, M. (1990). Regulation of the glaA gene of *Aspergillus niger. Curr Genet* **18**, 537-545.

Galagan, J. E., Calvo, S. E., Cuomo, C. & other authors (2005). Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105-1115.

Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315.

Gentleman, R. C., Carey, V. J., Bates, D. M. & other authors (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Kato, N., Murakoshi, Y., Kato, M., Kobayashi, T. & Tsukagoshi, N. (2002). Isomaltose formed by alpha-glucosidases triggers amylase induction in *Aspergillus nidulans. Current Genetics* **42**, 43-50.

Klein, C. J. L., Olsson, L., Ronnow, B., Mikkelsen, J. D. & Nielsen, J. (1996). Alleviation of glucose repression of maltose metabolism by MIG1 disruption in *Saccharomyces cerevisiae. Applied and Environmental Microbiology* **62**, 4441-4449.

Larry, L., Barton, I., Carl, E., Georgi & David, R. L. (1972). Effect of maltose on glucoamylase formation by *Aspergillus niger. J Bacteriol* **111**, 771-777.

Machida, M., Asai, K., Sano, M. & other authors (2005). Genome sequencing and analysis of *Aspergillus oryzae. Nature* **438**, 1157-1161.

Magnuson, J. K. & Lasure, L. L. (2004). Organic Acid Production by Filamentous Fungi. *Advances in Fungal Biotechnology for Industry, Agriculture, and Medicine, eds Jan and Lene Lange,*, Kluwer Academic/Plenum Publishers.

Needleman, R. B., Kaback, D. B., Dubin, R. A., Perkins, E. L., Rosenberg, N. G., Sutherland, K. A., Forrest, D. B. & Michels, C. A. (1984). MAL6 of *Saccharomyces*: A complex genetic locus containing three genes required for maltose fermentation. *Proc Natl Acad Sci U S A* **81**, 2811-2815.

Nierman, W. C., Pain, A., Anderson, M. J. & other authors (2006). Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus. Nature* **439**, 502-502.

Novak, S., Zechner-Krpan, V. & Maric, V. (2004). Regulation of maltose transport and metabolism in Saccharomyces cerevisiae. *Food Technology and Biotechnology* **42**, 213-218.

Patil, K. R. & Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci U S A* **102**, 2685-2689.

Payne, G. A., Nierman, W. C., Wortman, J. R. & other authors (2006). Whole genome comparison of *Aspergillus flavus* and *A. oryzae. Med Mycol* **44**, S9-S11.

Pedersen, H., Beyer, M. & Nielsen, J. (2000). Glucoamylase production in batch, chemostat and fed-batch cultivations by an industrial strain of *Aspergillus niger. Appl Microbiol Biotechnol* **53**, 272-277.

Pel, H. J., de Winde, J. H., Archer, D. B. & other authors (2007). Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat Biotechnol* **25**, 221-231.

Schrickx, J. M., Krave, A. S., Verdoes, J. C., Vandenhondel, C., Stouthamer, A. H. & Vanverseveld, H. W. (1993). Growth and product formation in chemostat and recycling cultures by *Aspergillus niger* N402 and a glucoamylase overproducing transformant, provided with multiple copies of the glaA gene. *Journal of General Microbiology* **139**, 2801-2810.

Sinha, A. U. & Meller, J. (2007). Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* **8**.

Smyth, G. K., Michaud, J. & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075.

Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). RSAT: regulatory sequence analysis tools. *Nucleic Acids Res* **36**, W119-W127.

van Wezel, G. P., White, J., Young, P., Postma, P. W. & Bibb, M. J. (1997). Substrate induction and glucose repression of maltose utilization by *Streptomyces coelicolor* A3(2) is controlled by malR, a member of the lacl-galR family of regulatory genes. *Molecular Microbiology* **23**, 537-549.

Vongsangnak, W., Olsen, P., Hansen, H., Krogsgaard, S., Nielsen, J. & (2008). Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC Genomics* **9**.

Workman, C., Jensen, L. J., Jarmer, H. & other authors (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**.

Wortman, J. R., Gilsenan, J. M., Joardar, V. & other authors (2009). The 2008 update of the *Aspergillus nidulans* genome annotation: A community effort. *Fungal Genetics and Biology* **46**, S2-S13.

Yu. J., Cleveland. T., Nierman. W. & Bennett. J. (2005). *Aspergillus flavus* genomics: gateway to human and animal health, food safety, and crop resistance to diseases. *Rev Iberoam Micol* **22**, 194-202.

Yuan, X. L., van der Kaaij, R. M., van den Hondel, C., Punt, P. J., van der Maarel, M., Dijkhuizen, L. & Ram, A. F. J. (2008). *Aspergillus niger* genome-wide analysis reveals a large number of novel alpha-glucan acting enzymes with unexpected expression profiles. *Mol Genet Genomics* **279**, 545-561.

Paper 4

Uncovering transcriptional regulation of glycerol metabolism in Aspergilli through genome-wide gene expression data analysis

Margarita Salazar[*], Wanwipa Vongsangnak[*],
Gianni Panagiotou, Mikael R. Andersen, and Jens Nielsen
[*]Equal contribution

Molecular Genetics and Genomics, 2009

Mol Genet Genomics
DOI 10.1007/s00438-009-0486-y

ORIGINAL PAPER

Uncovering transcriptional regulation of glycerol metabolism in Aspergilli through genome-wide gene expression data analysis

Margarita Salazar · Wanwipa Vongsangnak · Gianni Panagiotou · Mikael R. Andersen · Jens Nielsen

Received: 19 June 2009 / Accepted: 13 September 2009

Abstract Glycerol is catabolized by a wide range of microorganisms including *Aspergillus* species. To identify the transcriptional regulation of glycerol metabolism in *Aspergillus*, we analyzed data from triplicate batch fermentations of three different Aspergilli (*Aspergillus nidulans*, *Aspergillus oryzae* and *Aspergillus niger*) with glucose and glycerol as carbon sources. Protein comparisons and cross-analysis with gene expression data of all three species resulted in the identification of 88 genes having a conserved response across the three Aspergilli. A promoter analysis of the up-regulated genes led to the identification of a conserved binding site for a putative regulator to be 5′-TGCGGGGA-3′, a binding site that is similar to the binding site for Adr1 in yeast and humans. We show that this Adr1 consensus binding sequence was over-represented on promoter regions of several genes in *A. nidulans*, *A. oryzae* and *A. niger*. Our transcriptome analysis indicated that genes involved in ethanol, glycerol, fatty acid, amino acids and formate utilization are putatively regulated by Adr1 in Aspergilli as in *Saccharomyces cerevisiae* and this transcription factor therefore is likely to be cross-species conserved among *Saccharomyces* and distant *Ascomycetes*. Transcriptome data were further used to evaluate the high osmolarity glycerol pathway. All the components of this pathway present in yeast have orthologues in the three Aspergilli studied and its gene expression response suggested that this pathway functions as in *S. cerevisiae*. Our study clearly demonstrates that cross-species evolutionary comparisons among filamentous fungi, using comparative genomics and transcriptomics, are a powerful tool for uncovering regulatory systems.

Keywords *Aspergillus* species · Glycerol metabolism · Transcriptional regulation

Introduction

Glycerol is becoming of considerable importance in industrial fermentation processes as it is a major by-product from biodiesel production; and hereby represents a cheap carbon source for bio-based production of chemicals. Glycerol is a non-fermentable carbon source which can be utilized by many yeast species, including *Saccharomyces cerevisiae* and filamentous fungi such as *Aspergillus nidulans*, *Aspergillus oryzae*, and *Aspergillus niger*. Therefore, it would be valuable to identify regulatory nodes that control glycerol consumption in industrial relevant Aspergilli in order to convert this by-product into chemicals or proteins for further commercialization.

Communicated by S. Hohmann.

M. Salazar and W. Vongsangnak contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-009-0486-y) contains supplementary material, which is available to authorized users.

M. Salazar · W. Vongsangnak · J. Nielsen (✉)
Department of Chemical and Biological Engineering,
Chalmers University of Technology,
412 96 Göteborg, Sweden
e-mail: nielsenj@chalmers.se

M. Salazar
e-mail: margarita.salazar@chalmers.se

G. Panagiotou · M. R. Andersen
Department of Systems Biology,
Center for Microbial Biotechnology,
Technical University of Denmark,
2800 Kongens Lyngby, Denmark

Published online: 26 September 2009

In the yeast *S. cerevisiae* and Aspergilli, glycerol degradation occurs via a two-step glycerol phosphorylative pathway. In the first step, glycerol is converted to glycerol-3-phosphate by glycerol kinase (EC 2.7.1.30), product of the gene *GUT1*, YHL032C in *S. cerevisiae*, AN5589.3, JGI45434 and AO090001000509 in *A. nidulans*, *A. niger* and *A. oryzae*, respectively. Glycerol-3-phosphate then crosses the outer mitochondrial membrane, where it is oxidized to glycerone phosphate by the inner mitochondrial membrane enzyme, FAD^+-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5), which is encoded by the gene *GUT2*, YIL155C in *S. cerevisiae*, JGI55910 in *A. niger*, AN1396.3 in *A. nidulans* and AO090005001646 in *A. oryzae* (David et al. 2006; Ronnow and Kielland-Brandt 1993). Finally, glycerone phosphate enters the cytosol, where it is used either in the glycolytic or in the gluconeogenic pathways.

During evolution Aspergilli have, like many other organisms, developed efficient regulatory systems for controlling transcription, translation, protein stability and activity. These regulatory systems allow the cell to shift between the use of different metabolic pathways and hereby utilize preferred carbon and energy sources to obtain maximal growth. An example of this phenomenon is glucose repression, where the presence of rapidly fermentable sugars, such as glucose, represses a large number of genes required for utilization of less preferred carbon sources (Felenbok and Kelly 1996; Ruijter and Visser 1997; Strauss et al. 1995). In *S. cerevisiae*, many genes encoding proteins involved in metabolic adaptation showed increased expression during a shift from fermentative growth (glucose based) to respiratory growth (ethanol or glycerol based). Genes including *ALD2* (YMR170C), encoding a cytoplasmic aldehyde dehydrogenase and *ACS1* (YAL054C), encoding an acetyl-CoA synthetase isoform, are up-regulated on non-fermentative carbon sources. These two enzymes direct the products of alcohol dehydrogenase into the tricarboxylic acid and glyoxylate cycles. Increased expression of *PCK1* (YKR097W), encoding phosphoenolpyruvate carboxykinase, and *FBP1* (YLR377C), encoding fructose biphosphatase is also observed (DeRisi et al. 1997). These enzymes reverse the direction of metabolites within the glycolytic pathway to favor the production of glucose-6-phosphate (DeRisi et al. 1997). Increased transcript levels from genes involved in the general stress response as well as up-regulation of genes involved in the electron transport–oxidative phosphorylation systems and the mitochondrial translation system have also been observed during a shift from glucose to glycerol consumption (Roberts and Hudson 2006).

Here, we aimed at the identification of global regulatory patterns of gene expression during a metabolic shift from repressed to derepressed conditions. First, we conducted well-controlled batch cultivations with either glucose or glycerol as a carbon source with the three *Aspergillus* species, *A. nidulans*, *A. oryzae* and *A. niger*. Subsequently, we analyzed the transcriptome and used these data to identify a conserved regulatory response among the three *Aspergillus* species, which was found to be consistent with the response reported previously for *S. cerevisiae* (Young et al. 2003). In yeast, these gene expression changes appear to be general features of the adaptation to respiratory growth. However, they also show the aspects of an adaptation process which is carbon source specific and seems to be cross-species conserved among *S. cerevisiae* and other *Ascomycetes* species.

Materials and methods

Strains

The strains used in this study were *A. niger* BO1 and *A. oryzae* A1560, both obtained from Novozymes (Carlsen and Nielsen 2001; Pedersen et al. 2000), and the data for *A. nidulans* were obtained using the strain FGSC A4. BO1 is a glucoamylase over-producer and A1560 is an α-amylase producer. *A. niger* was maintained as frozen spore suspensions at −80°C in 20% glycerol. *A. oryzae* stock culture was maintained on Cove-N-Gly agar at 4°C and *A. nidulans* stock culture was maintained on Sigma potato-dextrose-agar (PDA) at 4°C.

Growth medium

The composition of the batch cultivation medium for *A. niger* was the following: 20 $g\,L^{-1}$ glucose monohydrate or glycerol, respectively, 7.3 $g\,L^{-1}$ $(NH_4)_2SO_4$, 1.5 $g\,L^{-1}$ KH_2PO_4, 1.0 $g\,L^{-1}$ $MgSO_4{\cdot}7H_2O$, 1 $g\,L^{-1}$ NaCl, 0.1 $g\,L^{-1}$ $CaCl_2{\cdot}2H_2O$, 0.05 $mL\,L^{-1}$ antifoam 204 (Sigma) and 1 $mL\,L^{-1}$ of trace elements solution. Trace elements solution composition ($g\,L^{-1}$): 7.2 g $ZnSO_4{\cdot}7H_2O$, 1.3 g $CuSO_4{\cdot}5H_2O$, 0.3 g $NiCl_2{\cdot}6H_2O$, 3.5 g $MnCl_2{\cdot}4H_2O$ and 6.9 g $FeSO_4{\cdot}7H_2O$. Complex media composition: 10 $g\,L^{-1}$ glucose monohydrate, 2 $g\,L^{-1}$ yeast extract, 3 $g\,L^{-1}$ tryptone, 0.52 $g\,L^{-1}$ KCl, 0.52 $g\,L^{-1}$ $MgSO_4{\cdot}7H_2O$, 1.52 $g\,L^{-1}$ KH_2PO_4, 20 $g\,L^{-1}$ agar and 1 $mL\,L^{-1}$ of trace elements solution. The trace elements solution used in this case contained ($g\,L^{-1}$): 0.4 g $CuSO_4{\cdot}5H_2O$, 0.04 g $Na_2B_4O_7{\cdot}10H_2O$, 0.8 g $FeSO_4{\cdot}7H_2O$, 0.8 g $MnSO_4{\cdot}H_2O$, 0.8 g $Na_2MoO_4{\cdot}2H_2O$ and 8 g $ZnSO_4{\cdot}7H_2O$. In the case of *A. oryzae*, several kinds of media were required. *A. oryzae* spore propagation medium (Cove-N-Gly): 218 $g\,L^{-1}$ sorbitol, 10 $g\,L^{-1}$ glycerol 99.5%, 2.02 $g\,L^{-1}$ KNO_3, 25 $g\,L^{-1}$ agar and 50 $mL\,L^{-1}$ salt solution. Cove-N-Gly salt solution: 26 $g\,L^{-1}$ KCl, 26 $g\,L^{-1}$ $MgSO_4{\cdot}7H_2O$, 76 $g\,L^{-1}$ KH_2PO_4, 50 $mL\,L^{-1}$ trace elements solution. Cove-N-Gly trace elements solution: 40 mg L^{-1} $Na_2B_4O_7·10H_2O$, 400 mg L^{-1} $CuSO_4·5H_2O$, 1,200 mg L^{-1} $FeSO_4·7H_2O$, 700 mg L^{-1} $MnSO_4·H_2O$, 800 mg L^{-1} $Na_2MoO_4·2H_2O$, 10 g L^{-1} $ZnSO_4·7H_2O$. *A. oryzae* medium for pre-cultures (G2-GLY): 18 g L^{-1} yeast extract, 24 g L^{-1} glycerol 87%, 1 mL L^{-1} pluronic PE-6100. *A. oryzae* batch cultivation medium: 2.4 g L^{-1} $MgSO_4·7H_2O$, 3.6 g L^{-1} K_2SO_4, 1.2 g L^{-1} citric acid monohydrate, 2.4 g L^{-1} KH_2PO_4, 3 g L^{-1} $(NH_4)_2HPO_4$, 1.2 g L^{-1} pluronic acid (PE-6100) and 0.6 mL L^{-1} trace elements solution. Trace elements solution composition (g L^{-1}): 14.3 g L^{-1} $ZnSO_4·7H_2O$, 8.5 g L^{-1} $MnSO_4·H_2O$, 13.8 g L^{-1} $FeSO_4·7H_2O$, 2.5 g L^{-1} $CuSO_4·5H_2O$, 3 g L^{-1} citric acid monohydrate (as a chelating agent), and 0.5 g L^{-1} $NiCl_2·6H_2O$. Carbon sources used were glucose monohydrate or glycerol (15 g L^{-1}). *A. nidulans* batch cultivation medium is reported in Panagiotou et al. (2008).

Preparation of inoculum

Aspergillus niger BO1 and *A. nidulans* A4 fermentations were inoculated with spores propagated on complex media plates, respectively, incubated for 8 days at 30°C or 4–5 days at 37°C in the case of *A. nidulans*. The same stock of spores was used to inoculate all plates. *A. niger* spores were harvested by adding 10 mL of Tween 80 0.01%. *A. nidulans* spores were suspended in 20 mL of distilled water. *A. niger* and *A. nidulans* cultivations were inoculated with a spore suspension to obtain a final concentration of 6×10^9 spores L^{-1} and the same solution of spores was used to inoculate the three replicates. *A. oryzae* A1560 fermenters were inoculated with approximately 60 g of broth of *A. oryzae* A1560 previously cultured at 30°C for 24 h on G2-GLY liquid medium in shake flasks at 250 rpm. The pre-cultures were inoculated with 5 mL of spore suspension harvested from sporulated mycelium grown on Cove-N-Gly agar at 34°C for 3–4 days. Spores were harvested with Tween 80 0.1%.

Batch fermentations

To determine the physiological characteristics of the strains, batch cultivations with the three Aspergilli were carried out. *A. niger* fermentations were performed in 5 L bioreactors with a working volume of 4.5 L. Reactors were equipped with two Rushton four-blade disk turbines, pH and temperature control. The temperature was maintained at 30°C and the pH was controlled by automatic addition of 2 N NaOH. The pH was initially set to 3.0 to prevent spore aggregation and only when spores started to germinate the pH was increased to 4.5 and kept constant through the cultivation. Likewise, the stirring speed was initially set to 200 rpm and the aeration rate to 0.05 vvm (volume of gas per volume of liquid per minute). After germination, these parameters were increased to 600 rpm and 0.9 vvm and kept steady throughout the rest of the fermentation. *A. oryzae* batch fermentations were conducted in 2 L reactors with a working volume of 1.2 L. The stirrer speed was kept at 800 rpm during the first 4 h and then increased to 1,100 rpm. The pH was kept at 6 by addition of 10% of H_3PO_4 or 10% NH_3 solution, and the temperature was maintained at 34°C. The aeration rate was set at 1.2 vvm. *A. nidulans* batch fermentations were performed in 1.5 L bioreactors with a working volume of 1.2 L. The bioreactors were equipped with two disk-turbine impellers rotating at 350 rpm. The pH was kept constant at 5.5 by addition of 2 N NaOH or HCl and the temperature was maintained at 30°C. Air was used for sparging the bioreactor at a constant flow rate of 1 vm. The concentrations of oxygen and carbon dioxide in the exhaust gas were monitored with a gas analyzer (1311 Fast response Triple gas, Innova combined with multiplexer controller for Gas Analysis MUX100, B. Braun Biotech International) for *A. nidulans* and *A. niger*. For fermentations with *A. oryzae*, the exhaust gas was measured with a gas analyzer (Magnos 4G for O_2, Uras3G for CO_2, Hartmann & Braun, Germany). In all cases, dissolved oxygen tension was initially calibrated at 100%.

Sampling

For quantification of cell mass and extracellular metabolites, a known volume of cell culture was withdrawn from the reactor, filtered and washed. The culture filtrates were frozen at −20°C for subsequently sugar and extracellular metabolite analysis. Cell dry weight was determined using nitrocellulose filters (pore size 0.45 μm, Gelman Sciences). The filters were pre-dried in a microwave oven at 150 W for 15 min, cooled in a desiccator and subsequently weighed. A known volume of cell culture was filtered and the residue was washed with distilled water and dried on the filter for 15 min in a microwave oven at 150 W or in an oven at 100°C for 24 h. The filter was weighed again and the cell mass concentration was calculated. For gene expression analysis, mycelium was harvested in the mid exponential phase from each of the three biological replicates. The cultures were filtered through sterile Miracloth (Calbiochem, San Diego, CA, USA) and washed with a suitable amount of 0.9% NaCl solution for *A. nidulans* and *A. niger* or distilled water for *A. oryzae*. The mycelium was quickly dried by squeezing and subsequently frozen in liquid nitrogen. Samples were stored at −80°C until RNA extraction.

Sugars and extracellular metabolites quantification

The concentration of sugars and extracellular metabolites in the filtrates were determined using high pressure liquid chromatography (HPLC) on an Aminex HPX-87H ion-exclusion column (BioRad, Hercules, CA, USA). The column was eluted at 60°C for *A. niger* and *A. nidulans* and at 45°C for *A. oryzae* with 5 mM H_2SO_4 at a flow rate of 0.6 mL min^{-1}. Metabolites were detected with a refractive index detector and an UV detector.

Total RNA extraction

Aspergillus niger and *A. nidulans* total RNA was isolated using the Qiagen RNeasy Mini Kit (QIAGEN Nordic, Ballerup, Denmark), according to the protocol for isolation of total RNA from plant and fungi. For the purpose, approximately 100 mg of frozen mycelium were placed in a 2 mL tube, pre-cooled in liquid nitrogen, containing three RNase-treated steel balls (two balls with a diameter of 2 mm and one ball with a diameter of 5 mm). The tubes were subsequently shaken in a Mixer Mill, at 5°C for 10 min, until the mycelium was ground to powder and thus ready for extraction of total RNA. *A. oryzae* total RNA was purified using the Promega RNAgents Total RNA Isolation system following manufacturer's recommendations. For purification, approximately 1 g of frozen mycelium was ground to powder using a ceramic mortar and pestle. Mycelium was kept in liquid nitrogen throughout the grinding processing. All samples were inspected for good quality of total RNA extracted with a BioAnalyzer (2100 BioAnalyzer, Agilent Technologies Inc., Santa Clara, CA, USA). RNA quantification was performed in a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden) and total RNA was stored at −80°C until further processing.

Microarray manufacturing and design

Affymetrix arrays were used for the analysis of the transcriptome data of *A. nidulans*, *A. oryzae and A. niger* (Affymetrix, Santa Clara, CA, USA). The arrays were packaged in an Affymetrix® GeneChip cartridge (49 format), and were processed with GeneChip reagents in the GeneChip® Instrument System. The design and selection of probes for interrogating gene expression levels based on the genomes of *A. nidulans* FGSC A4 (BROAD Institute database. *Aspergillus nidulans* genome database), *A. oryzae* RIB40 (DOGAN database, *Aspergillus oryzae* genome database) and *A. niger* ATCC 1015 (JGI database, *Aspergillus niger* genome database) was performed in our previous work (Andersen et al. 2008). The arrays contain a maximum of 11 non-overlapping perfect match (PM) probes of 25 oligomers length per gene. 11,122 probe sets were represented in the microarray for *A. niger*, 12,039 probe sets plus an EST collection (courtesy of Novozymes) for *A. oryzae* and 10,656 probe sets for *A. nidulans*.

Preparation of biotin-labeled cRNA and microarray processing

Biotin-labeled cRNA was prepared from approximately 5 µg of total RNA, according to the protocol described in the Affymetrix GeneChip® Expression Analysis Technical Manual (Affymetrix). All samples were prepared in the same manner. The cRNA was cleaned before fragmentation using the Qiagen RNeasy Mini Kit (protocol for RNA Cleanup), in order to guarantee good-quality cRNA samples for subsequent processing. Biotin-labeled cRNA was quantified in a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden) and 20 µg was fragmented following the manufacturer recommendations. Approximately 15 µg of fragmented cRNA was hybridized to the 3AspergDTU Affymetrix GeneChip (Andersen et al. 2008) following the Affymetrix GeneChip® Expression Analysis protocol. Arrays were washed and stained using a GeneChip® Fluidics Station FS-400, and scanned on an Agilent GeneArray® Scanner 3000. The scanned probe array images (.DAT files) were converted into .CEL files using the Affymetrix GeneChip Operating Software.

Transcriptome analysis

Affymetrix CEL-data files were preprocessed using the statistical language R version 2.7.1 (R Development Core Team 2007) and Bioconductor version 2.2 (Gentleman et al. 2004). The probe intensities were normalized for background by using the robust multi-array average method with perfect match (PM) probes only (Irizarry et al. 2003). Subsequent normalization was performed using the qspline algorithm (Workman et al. 2002). Gene expression indexes were calculated from the PM probes with the median polish summary method (Irizarry et al. 2003). All statistical preprocessing methods were implemented in affy package (Gentleman et al. 2004) using R scripts (Dudoit et al. 2003). Statistical analysis was applied to identify differential gene expression levels based on three replicates for each condition. Moderated Student's *t*-tests between the two carbon sources for each *Aspergillus* spp. was conducted by using limma package (Smyth 2004). Empirical Bayesian statistics were used to moderate the standard errors within each gene and Benjamini–Hochberg's method to adjust for multiple testing (Benjamini and Hochberg 1995). Unless otherwise stated, a cut-off of adjusted *P* value <0.05 was used to assess for statistical significance.

Protein sequence comparisons

A cross-comparison between the amino acid sequences of the predicted ORFs from each of the three *Aspergillus* genomes, based on DOGAN (DOGAN database, *Aspergillus oryzae* genome database), JGI (JGI database, *Aspergillus niger* genome database) and BROAD Institute databases (BROAD Institute database, *Aspergillus nidulans* genome database) using the BLASTP algorithm, was applied (Altschul et al. 1990). The *A. oryzae* genome sequence (NBRC 100959), *A. nidulans* FGSC A4 version 3.1 and *A. niger* ATCC 1015 version 1.0 were used. An estimated expectation value cut-off of 1E-30 was set to assess for statistical significance. The best hit, based on the score, was selected for the case in which the protein query produced more than one hit. Bi-directional best hits were found by comparing the lists of best hits for two species against each other (i.e. *Niger_Oryzae*, *Oryzae_Niger*) and selecting those genes where the best hit in the other organism was the same best hit, thus giving a conservative set of 1:1 homologues for all three pair-wise comparisons. Tri-directional best hits were found by comparing the three lists of bi-directional hits (*Niger_Oryzae*, *Nidulans_Oryzae*, *Niger_Nidulans*) and selecting the genes that had a 1:1:1-relationship in all comparisons between all three species. The full subset of tri-directional homologues is given in Supplementary Table 1.

Detection of conserved regulatory elements

Several bioinformatics tools were applied for the detection of conserved regulatory elements. As a first step, pattern recognition was conducted in Regulatory Sequence Analysis Tools (RSAT) (van Helden et al. 1998) using the option of oligo-analysis. The method is based on the detection of over-represented oligonucleotides. The statistical significance of a site was assessed based on pre-computed tables of oligonucleotide frequencies observed in all non-coding sequences from *A. oryzae* and *A. nidulans* genomes, respectively, as these two organisms are supported by the application. In the case of *A. niger*, our own frequency table was calculated based on the intergenic regions from scaffold 1 of the *A. niger* ATCC 1015 genome sequence for 6, 7 and 8 base pairs (bps) oligonucleotides. Intergenic regions containing unknown bases (N's) were removed from the training set leaving 1,214 sequences. The motif recognition was computed by running the analysis with a 1,000 bps upstream region counted from the start codon of each gene or predicted transcription start site in the case of *A. niger*. A subset of 243 promoters, 3 times 81 promoters for each of the species, was analyzed. Statistical analysis was conducted to find consensus motifs in the subsets of 81 up-regulated conserved genes as well as in the 5 down-regulated conserved genes in the 3 Aspergilli species. The analysis was done considering a different length of consensus patterns, ranging from 6 to 8 bps for each *Aspergillus*. After having a number of probable consensus conserved motifs; these were further inspected using R 2.7.1 and Cosmo package (Bembom et al. 2007). Default settings were used and the program was run for different pattern lengths. A background Markov model was computed using the intergenic regions from scaffold 1 of the *A. niger* ATCC 1015 genome sequence as previously reported (Andersen et al. 2008). The two component mixture (TCM) model was used to search for a conserved motif where the maximum number of sites was increased to include all 174 binding sites. Finally, a more refined search for potential transcription factor binding sites in the subset sequences was done with the pattern search program Patch using TRANSFAC 6.0 public sites (http://www.gene-regulation.de/).

GO term enrichment analysis

GO term enrichment analysis was conducted with the *A. niger* ATCC 1015 conserved up-regulated genes list (81 genes) and with the significantly differentially expressed genes list (P value < 0.05) using R 2.7.1 (R development Core Team 2007) with BioConductor (Gentleman et al. 2004) and the topGO-package v. 1.2.1 with the elim algorithm to remove local dependencies between GO terms (Alexa et al. 2006). GO term assignments were based on automatic annotation of the *A. niger* ATCC 1015 version 1.0 gene models, a cut-off of P value <0.05 was used to assess significance.

Results

Protein comparisons

To identify conserved regulatory systems between the three *Aspergillus* species as well as to exploit both similarities and differences at the protein level, genes having orthologues in the three species were identified using a BLASTP based comparison (Altschul et al. 1990). Initially, *A. nidulans*, *A. oryzae* and *A. niger* genome-wide protein sequences were compared among each other in order to obtain tri-directional homologues as described in "Materials and methods". By defining a threshold of E value of 1E-30, 5,190 orthologues were found to be conserved in all three species (List of tri-directional homologues, Supplementary Table 1). The set of conserved genes (1:1:1 orthologues) was used for further analysis of the transcriptome data.

Fermentation results

In order to analyze the conserved transcriptional response toward a change of carbon source from glucose to glycerol and to have a complete dataset of transcriptome data in all three *Aspergillus* species; we collected fermentation data from our previous work and conducted new fermentation experiments under the same conditions required for each strain. *A. oryzae* cultivations carried out on glucose are part of our published results in the report by Andersen et al. (2008), *A. niger* batch fermentations on glucose are from the report of Vongsangnak et al. (2009) and *A. nidulans* batch fermentations data on glucose and glycerol are from the report of Panagiotou et al. (2008). Fermentations on glycerol with *A. oryzae* and *A. niger* were conducted specifically for this study under the conditions defined in "Materials and methods". Each *Aspergillus* species had its own specific cultivation medium and all fermentations were run in three biological replicates. A summary of all fermentation results are shown in Fig. 1. Figure 1a shows the substrate and biomass concentration profiles for each *Aspergillus* spp. Figure 1b shows the statistics of the physiological characterization data. In all three species, both carbon sources were completely exhausted; nevertheless, glucose was consumed at a faster consumption rate than glycerol. For *A. oryzae*, the maximum specific growth rate on glucose was 0.38 ± 0.004 h^{-1} while it was 0.30 ± 0.004 h^{-1} on glycerol. In the case of *A. nidulans*

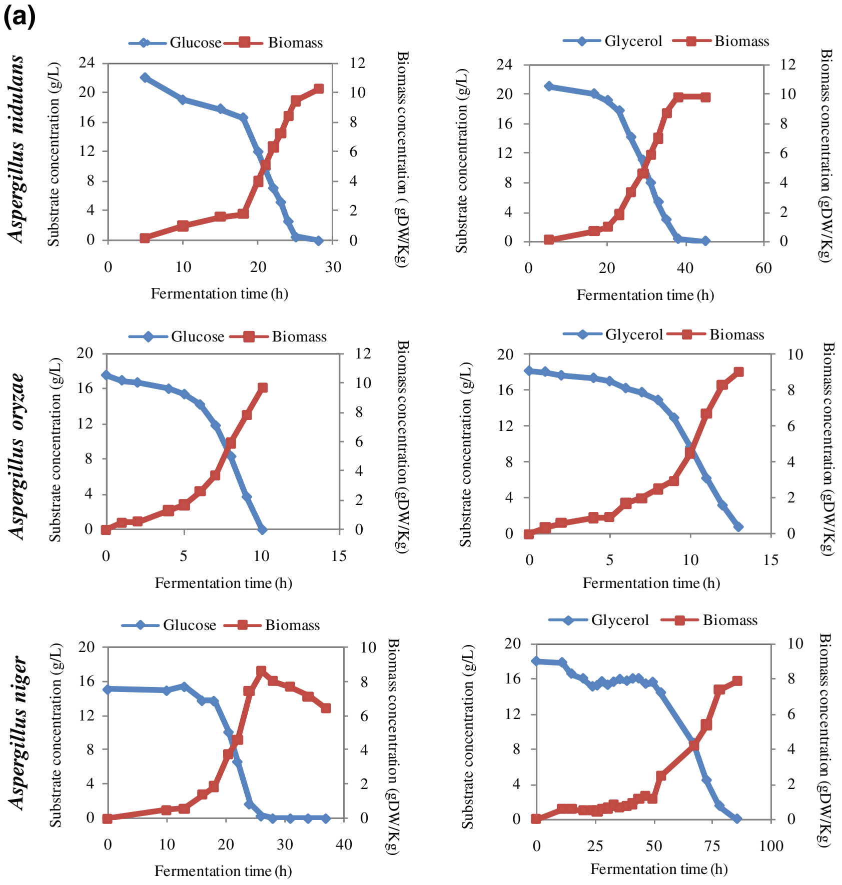


Fig. 1 Summary of batch fermentation parameters of *A. nidulans*, *A. oryzae* and *A. niger* grown on glucose or glycerol as sole carbon source. **a** Fermentation profiles of a representative replicate. *Filled diamonds* Substrate concentration (g/L). *Filled squares* Biomass concentration (g DW/kg). All fermentations were performed in three biological replicates. **b** Summary of batch cultivations statistics. For all cultivations, maximum specific growth rate (μ_{max}), biomass yield (Y_{sx}), time of sampling for transcriptome analysis (TA), and biomass concentration at the time of sampling for transcriptome analysis (TA) are given. *nd* Not determined

(b)

Strain	**Carbon source**	μ_{max} (h^{-1})	**Ysx** (g DW/g substrate)	**Time of sampling (h)**	**Biomass concentration (g DW/kg)**
A. nidulans	Glucose	0.23±0.020	0.47±nd	~ 22	6.33±0.40
	Glycerol	0.11±0.010	0.42±nd	nd	6.50±0.50
A. oryzae	Glucose	0.38±0.004	0.54±0.013	~6	2.50±0.09
	Glycerol	0.30±0.004	0.52±0.008	~8	2.44±0.05
A. niger	Glucose	0.22±0.015	0.57±0.053	~21	3.74±0.06
	Glycerol	0.05±0.007	0.40±0.022	~36	0.88±0.29

and *A. niger*, the growth rates differences on the two carbon sources were more prominent. For *A. nidulans*, the maximum specific growth rate on glucose was double that on glycerol; and for *A. niger*, growth on glycerol was four times slower when compared to glucose (Fig. 1b). The difference on growth rates in the three Aspergilli might be due to several reasons, but one of them is most likely that glycerol is a favorite carbon source for the *A. oryzae* strain A1560 used in this study (Vongsangnak et al. 2008) besides the cultivation conditions, which have been optimized for growth (Novozymes's fermentation conditions). In addition to the maximum specific growth rates and biomass yields, sampling times and biomass yields at the time of sampling for transcriptome analysis (TA) were recorded for the three *Aspergillus* species (Fig. 1b).

Transcriptome data analysis

Genome-wide gene expression data were analyzed for all three sets of glucose and glycerol batch fermentations for the three *Aspergillus* species. A *t*-test pair-wise comparison for each *Aspergillus* spp. on glycerol versus glucose (cut-off *P* value <0.05) revealed 904, 1,145 and 3,058 significantly differentially expressed genes for *A. nidulans*, *A. oryzae* and *A. niger*, respectively. From the subset of significantly differentially expressed genes in each *Aspergillus*, there were 69 metabolic genes in *A. nidulans* (Supplementary Fig. 1), 182 metabolic genes in *A. oryzae* (Supplementary Fig. 2) and 278 metabolic genes in *A. niger* (Supplementary Fig. 3). The in-house reconstructed *A. oryzae* metabolic map was used as a visualization tool for plotting the gene expression changes for each *Aspergillus* spp. (Vongsangnak et al. 2008). The genes mapped to the metabolic pathways illustrated in Supplementary Fig. 4 (details in Supplementary Figs. 1 to 3) were significantly up-regulated on glycerol media in *A. niger* and *A. oryzae*. In a previous transcriptome study in *A. nidulans* (David et al. 2006), but using another DNA array platform, we found the same number of metabolic genes to be differentially expressed when comparing glucose to glycerol growth with the same *P* value cut-off. Surprisingly, the transcriptional regulator CreA was only found in the subset of conserved and differentially expressed genes between *A. niger* and *A. nidulans*. CreA was not found in the subset of differentially expressed genes in *A. oryzae* and thereby, not captured in the conserved regulatory response.

Subsequently, these three subsets of significant genes in all three species were compared to the list of 5,190 conserved genes in the three Aspergilli (Supplementary Table 1) as well as with each other. This resulted in the identification of 88 conserved genes that were differentially expressed in all three species (Fig. 2). Among them, 81 genes were up-regulated during growth on glycerol, 5 genes were down-regulated and 2 genes did not show a clear trend (Supplementary Table 2). In order to identify significant tendencies in the transcriptome profiles of the three *Aspergillus* species, a GO term enrichment analysis was conducted on the subset of the 81 up-regulated and conserved genes. Among the over-represented GO terms from biological processes we found pyruvate metabolism, amino acid metabolism, specifically tyrosine, valine and the aromatic amino acids metabolism, as well as GO terms for genes involved in gluconeogenesis, the hexose biosynthetic process and the alcohol and monosaccharide biosynthetic process (details in Supplementary Table 4). Individual examination of the up-regulated genes showed the presence of genes involved in fatty acid metabolism, amino acid metabolism, ribosome biogenesis and peroxisomal biogenesis (Table 1).

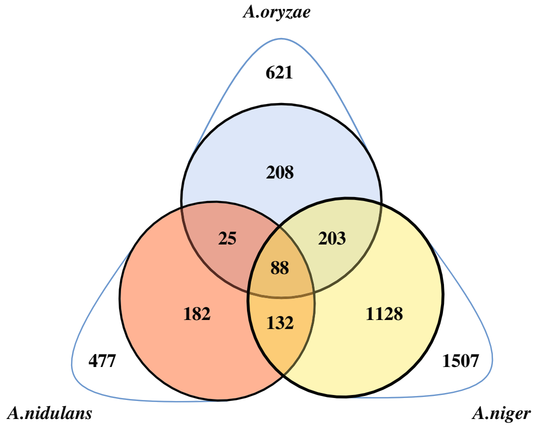


Fig. 2 Venn diagram of significantly differentially expressed genes from glycerol versus glucose by pair-wise comparison for each *Aspergillus* species. The *colored overlapping middle area* contains the genes that are significantly differentially expressed and conserved in all three *Aspergillus* species. The *numbers on a white background* represent the non-conserved genes in all three Aspergilli, but still differentially expressed in a single species. Adjusted *P* value cut-off <0.05 (color figure online)

The obtained response of 88 differentially expressed genes in the 3 species suggests a conserved regulatory response across the *Aspergillus* genus, from which, there were several genes with unknown function (Supplementary Table 2), whereas the function of several of the other proteins is only inferred. From the subset of the 48 annotated genes (BROAD Institute database, *Aspergillus nidulans* genome database, DOGAN database, *Aspergillus oryzae* genome database and JGI database, *Aspergillus niger* genome database), there were at least two predicted transporters of the major facilitator superfamily (MFS); namely, AN10075.3 and AN6703.3 and its corresponding orthologous genes in *A. oryzae* and in *A. niger* that were up-regulated on glycerol (Table 1). The rest, which corresponds to 40 genes, are predicted as hypothetical proteins which results were obtained from the genome annotation and whose function has not been confirmed by any experimental evidence supported by a publication. Interestingly, the

Table 1 Conserved gene expression response of *A. nidulans*, *A. oryzae* and *A. niger* on glycerol medium. 88 differentially expressed genes were conserved in all three Aspergilli

ORF			Functional annotation
A. nidulans	*A. oryzae*	*A. niger*	
AN6723	AO090005000447	JGI209864	2,3-Dihydroxybenzoate carboxylyase (*dhbD*)/metal-dependent hydrolase of the TIM-barrel fold
AN1897	AO090003000210	JGI38851	Homogentisate 1,2-dioxygenase
AN6399	AO090005000599	JGI52058	Bleomycin hydrolases
AN1899	AO090003000208	JGI199151	4-Hydroxyphenylpyruvate dioxygenase
AN0051	AO090120000371	JGI184789	Isopenicillin *N* synthase and related dioxygenases
AN3741	AO090005000125	JGI44729	**Alcohol dehydrogenase (*alcB*), class V, EC number 1.1.1.1**[a]
AN1896	AO090003000211	JGI199148	Fumarylacetoacetase, EC number 3.7.1.2
AN5860	AO090026000494	JGI197162	MFS monosaccharide transporter (major facilitator superfamily) (*mstC*)$^{\mathrm{DOWN}}$
AN8130	AO090102000470	JGI138792	NAD/NADP transhydrogenase beta subunit
AN7451	AO090001000717	JGI37620	NAD-dependent glutamate dehydrogenase, EC number 1.4.1.2
AN10981	AO090701000398	JGI213851	Bifunctional GTP cyclohydrolase II/3, 4-dihydroxy-2-butanone-4-phosphate synthase
AN2133	AO090102000246	JGI53518	UPRTase, uracil phosphoribosyltransferase (*furA*), EC number 2.4.2.9
AN1726	AO090001000555	JGI43391	3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)
AN4691	AO090020000497	JGI180570	Dehydrogenases with different specificities
AN5690	AO090005000103	JGI57198	Copper amine oxidase (AO-I and AO-II), EC number 1.4.3.6
AN8496	AO090009000063	JGI213815	Ribulose-5-phosphate 4-epimerase and related epimerases and aldolases
AN10075	AO090026000207	JGI124797	Permease of the major facilitator superfamily
AN0554	AO090023000467	JGI55742	**Aldehyde dehydrogenase (*aldA*), EC number 1.2.1.3**[a]
AN5833	AO090011000917	JGI185892	Propionate/acetate CoA ligase/acyl-CoA synthetase
AN4901	AO090003000638	JGI197415	Glutaminase A, EC numer 3.5.1.2
AN0485	AO090023000254	JGI205368	Phosphatidylinositol transfer protein PDR16 and related proteins
AN4687	AO090020000492	JGI181451	3-Methylcrotonyl-CoA carboxylase
AN1733	AO090001000549	JGI208879	Delta-1-pyrroline-5-carboxylate dehydrogenase, EC numer 1.5.1.12
AN9138	AO090038000537	JGI54468	Amidase/acetamidase, EC number 3.5.1.4
AN4170	AO090003000144	JGI52919	Carbon catabolite repression protein CreD$^{\mathrm{DOWN}}$
AN4659	AO090011000447	JGI121695	Acyl-CoA synthetase/AMP-binding domain protein
AN10030	AO090020000517	JGI200187	Subtilisin-related protease/Vacuolar protease B (PepC)/Serine protease (Alp2)
AN8559	AO090023000349	JGI192202	Branched chain alpha-keto acid dehydrogenase E1, beta subunit
AN7641	AO090701000307	JGI204355	Copper amine oxidase (maoN, AO-I, AO-II), EC number 1.4.3.6
AN4779	AO090020000332	JGI209032	NIPSNAP1 protein
AN6703	AO090005000420	JGI180069	Permeases of the major facilitator superfamily
AN7529	AO090001000612	JGI178113	Metal-dependent amidase/aminoacylase/carboxypeptidase
AN4688	AO090020000493	JGI209685	Isovaleryl-CoA dehydrogenase, EC number 1.3.99.10
AN5669	AO090009000195	JGI55680	Succinyl-CoA:alpha-ketoacid-CoA transferase
AN6985	AO090206000019	JGI53716	Ribulokinase, EC number 2.7.1.47
AN3184	AO090012000809	JGI55604	Aldose 1-epimerase, EC number 5.1.3.3
AN9075	AO090038000620	JGI212771	NADPH: quinone reductase
AN8163	AO090102000483	JGI54341	Short-chain dehydrogenase/reductase SDR
AN8242	AO090102000588	JGI188214	Lipase
AN4201	AO090009000486	JGI187366	Acyl-CoA synthetase/AMP-binding domain protein
AN0129	AO090120000287	JGI211917	Protein tyrosine phosphatase Pps1, EC number 3.1.3.48
AN4245	AO090001000449	JGI120161	Ceramidase
AN10520	AO090023000421	JGI44810	Alpha/beta hydrolase

Table 1 continued

ORF			Functional annotation
A. nidulans	*A. oryzae*	*A. niger*	
AN4102	AO090009000356	JGI56782	Beta-glucosidase A and related glycosidases (*bglA*), EC number 3.2.1.21
AN1918	AO090003000174	JGI208685	Phosphoenolpyruvate carboxykinase (ATP), EC number 4.1.1.49
AN9384	AO090124000014	JGI51356	Cytochrome P450 alkane hydroxylase
AN3639	AO090003001008	JGI189170	Dihydrolipoamide transacylase (alpha keto acid dehydrogenase E2 subunit)
AN0942	AO090005001078	JGI46405	L-arabitol dehydrogenase (*ladA*), EC number 1.1.1.12

[a] The enzymes highlighted in bold have been proved to be regulated by Adr1 in the yeast *S. cerevisiae*

Genes unmarked mean up-regulated on glycerol medium. Genes marked with "DOWN" mean down-regulated on glycerol medium. Genes with functional annotation available are listed. Complete list of conserved genes and log2 ratios are listed in Supplementary Table 2

presence of a glycerol transport system was demonstrated earlier in *A. nidulans* where glycerol uptake defective mutants were selected at high concentrations of glycerol (Visser et al. 1988). The *glcC* mutant discussed in this publication was defective in glycerol uptake by mutation on a possible glycerol transporter which was mapped to linkage group VI (Visser et al. 1988). In our case, the two *A. nidulans* MFS genes, AN10075.3 and AN6703.3, are mapped to linkage group VIII, Contig 6:256957−258913+, and linkage group I, Contig 112:48095−49693+, respectively (BROAD Institute database, *Aspergillus nidulans* genome database). Therefore, it seems unlikely that we are addressing the same glycerol transporters.

A closer look at the transcriptome results showed differences on the preference of glycerol utilization pathways in *A. nidulans*, *A. niger* and *A. oryzae*. In naturally glycerol utilizing fungi, glycerol (GL) can be phosphorylated either into glycerol-3-phosphate (GL3P) and further oxidized by the FAD^+-dependent glycerol-3-phosphate dehydrogenase into glycerone phosphate (T3P2), which then enters glycolysis. In the other pathway, glycerol can be converted through NAD^+/$NADP^+$ glycerol dehydrogenases into glycerone (GLYN) and further phosphorylated by glycerone kinase into T3P2. A simplified scheme of the metabolic pathways leading to or from glycerol is illustrated in Fig. 3.

Detection of conserved motifs

One or more conserved transcriptional regulators were suspected to be up-regulating the subset of 81 genes or down-regulating the subset of 5 down-regulated genes within the group of 88 genes having a conserved transcriptional response. Statistical promoter analysis was conducted for all 3 data sets of 81 up-regulated genes on glycerol medium. By inspecting the upstream sequences of each up-regulated orthologues dataset, giving a subset of 243 promoters (3 × 81 promoters), we found the most over-represented pattern to be "TGCGGGGA" (reverse complement, TCCCCGCA). The corresponding logo plot is shown in Fig. 4. The same analysis was conducted with the subset of down-regulated genes, but no consensus *cis*-acting regulatory element was found. Based on a literature search, it was proposed that TGCGGGGA is the consensus binding sequence of the transcriptional activator Adr1, which has been found to regulate several pathways in *S. cerevisiae* (Young et al. 2003) and in humans (Das and Baez 2008). The consensus binding sequence of Adr1 in humans is GCGGGGA, which regulates the transcription of *psen1* (gene encoding presenilin 1) (Das and Baez 2008), a transmembrane protein that functions as a part of the gamma-secretase protease complex. In *S. cerevisiae*, Adr1 is known

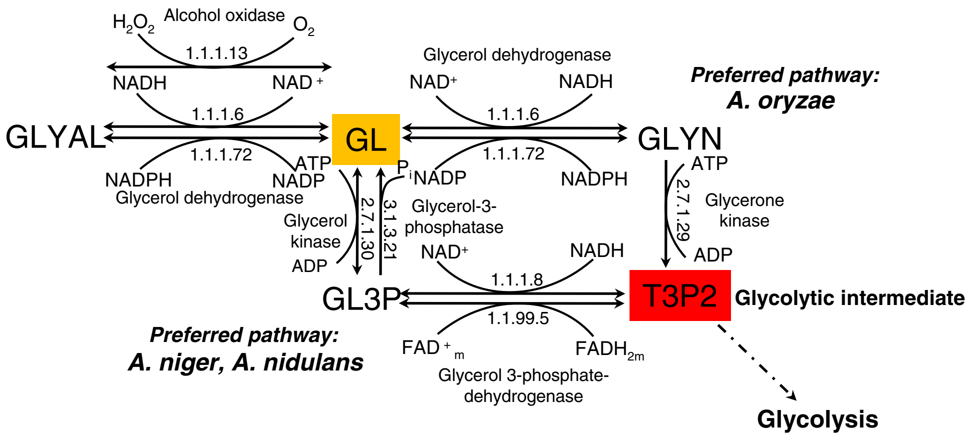


Fig. 3 Glycerol utilization pathways in *Aspergillus* species leading to the production of the glycolytic intermediate glycerone phosphate. The abbreviation of metabolites is described as follows: *GL* glycerol, *GLYAL* D-glyceraldehyde, *GLYN* glycerone, *GL3P* sn-glycerol 3-phosphate, *T3P2* glycerone phosphate

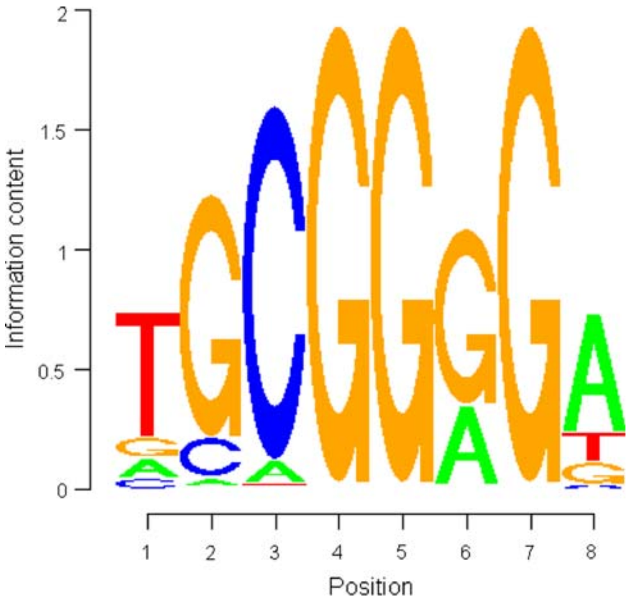


Fig. 4 Logo plot of the over-represented motif from the 81 promoter regions of *A. nidulans*, *A. oryzae* and *A. niger* genes significantly up-regulated on glycerol medium. The nucleotides representing the sequence are stacked on top of each other for each position in the aligned sequences. The height of each nucleotide is made proportional to its frequency, and the nucleotides are sorted so that the most common is on top. The height of the entire stack is then adjusted to signify the information content of the sequences at that position (Schneider and Stephens 1990). The "*x* axis" indicates the position of the corresponding nucleotide (*A*, *T*, *C* or *G*). "*y* axis" represents the information content of the corresponding nucleotide at each position in a bits scale, where 2 is the maximum value

to regulate several pathways including glycerol metabolism and fatty acid metabolism (Young et al. 2003). The consensus binding sequence is TTGG(A/G)GA, where according to Cheng et al. (1994) there are only four essential base pairs: GG(A/G)G. From the subset of 81 up-regulated genes, 24 genes contained the motif TGCGGGGA in all three *Aspergillus* species and 30 of the total of 72 genes had it more than once, while in general, it was located at an average position of 650 bps upstream from the start codon. The location of the Adr1 promoter binding sites are summarized in Supplementary Table 5. From this subset of 24 up-regulated genes, 5 of them had orthologues in *S. cerevisiae*; namely, *ADH2* (YMR303C) or *alcB* in *Aspergillus*; *ALD5* (YER073W) or *aldA* in *Aspergillus*, *ACS2* (YLR153C), *CCC1* (YLR220W) and *PUT2* (YHR037W). This conserved up-regulatory response in all three *Aspergillus* spp. suggests that these genes might be activated by a common cross-species conserved transcription factor, which could be Adr1 as it is found to be in *S. cerevisiae* regulating, i.e. *ADH2* and *ALD5* (Young et al. 2003) (Table 1).

Amino acid metabolism is also regulated by Adr1 in *S. cerevisiae* (Young et al. 2003). A list of enzymes involved in amino acid metabolism and thought to be regulated by Adr1 in the three Aspergilli is shown in Table 2. Bag7, Car2 and Put4, orthologues were significantly up-regulated on glycerol in *A. niger* and to a lower extent in *A. oryzae* (Supplementary Table 3). Of particular interest are Put2 orthologues which were up-regulated on glycerol in the three *Aspergillus* species and therefore, captured as a conserved transcriptional response. Put2 is a mitochondrial delta-1-pyrroline-5-carboxylate dehydrogenase involved in utilization of proline as sole nitrogen source (SGD database, *Saccharomyces* Genome Database).

Stress response relevant components

Further inspection of transcriptome data in all three *Aspergillus* species pinpointed signaling pathways involved in stress response and several histidine kinases regulating a number of processes as differentially expressed. Osmotic stress components are summarized in Supplementary Table 6 and the proposed mechanism in *S. cerevisiae* and Aspergilli is shown in Fig. 5a, b, respectively. Interestingly, there are 15 histidine kinases (HKs) reported to exist in *A. nidulans* (Suzuki et al. 2008). However, we found orthologues for only 9 of these kinases under the statistical constraints imposed in the three Aspergilli, implicating that *A. nidulans* might have more histidine kinases regulating more processes besides osmoadaptation, i.e. sexual development (Appleyard et al. 2000; Blumenstein et al. 2005) than *A. oryzae* and *A. niger*, which are presumed to be asexual fungi. We maintained the histidine kinases classification described by Kobayashi et al. (2007), where based on amino acid sequence similarity, *Aspergillus* histidine kinases were classified into nine families. According to these authors, for i.e. in *A. oryzae* there are three predicted histidine kinases belonging to family 3, whereas for *A. nidulans* there is solely one (Kobayashi et al. 2007). Therefore, this has its implications when considering *A. nidulans* histidine kinase list as a query for finding the best hit on the other two *Aspergillus* species. The *A. niger* histidine kinase orthologues JGI183029 (*tcsA*), JGI39736 (*nikA*), JGI41708 (*fphA*), JGI174132 (*hk-8-4*)* and JGI54610 (*hk-9*)* (details in Supplementary Table 7) were all significantly up-regulated on glycerol compared to glucose fermentations and these results are in accordance to the up-regulation seen in the osmotic stress components listed in Supplementary Table 6.

Discussion

Glycerol utilization

Glycerol metabolism in *Aspergillus* species is not studied as well as in the yeast *S. cerevisiae*. In particular, regulation

Table 2 Systems regulated by Adr1 and its orthologues in *A. nidulans*, *A. oryzae* and *A. niger*

Gene	*S. cerevisiae* ORF	*A. nidulans* ORF	*A. oryzae* ORF	*A. niger* ORF	Annotation
General regulation of systems by Adr1					
ADY2[a]	YCR010C	AN5226.3	AO090005001538	JGI53176	Acetate transporter[b,c]
ADH2[a]	YMR303C	AN3741.3	AO090005000125	JGI44729	Alcohol dehydrogenase[b,c]
ACS1[a]	YAL054C	AN5626.3	No hit[d]	JGI214348	Acetyl-CoA synthetase isoform[b,c]
GUT1[a]	YHL032C	AN5589.3	AO090001000509	JGI45434	Glycerol kinase[b,c]
ALD4[a]	YOR374W	AN0554.3	AO090023000467	JGI55742	Mitochondrial aldehyde dehydrogenase[b,c]
CIT3[a]	YPR001W	AN8275.3[e]	AO090102000627[e]	JGI202801[e]	Citrate synthase[b,c]
GIP2[a]	YER054C	AN2425.3	AO090026000188	JGI206783	Putative regulatory subunit of Glc7[b,c]
ICL2[a]	YPR006C	AN8755.3	AO090120000179	JGI42171	2-Methylisocitrate lyase[b,c]
ETR1[a]	YBR026C	AN9401.3	AO090124000077	JGI206405	2-Enoyl thioester reductase[b,c]
FDH2[a]	YPL275W	AN6525.3	AO090023000508	JGI124156	NAD$^+$-dependent formate dehydrogenase[b,c]
BAG7	YOR134W	AN7650.3	AO090701000375	JGI119642	Rho GTPase activator (related to Sac7)[b]
Regulation of amino acid metabolism					
CAR2	YLR438W	AN1810.3	AO090023000546	JGI54525	L-ornithine transaminase (OTAse)
PUT4	YOR348C	AN3359.3	AO090010000119	JGI43857	Proline permease
BAT1	YHR208W	AN4323.3	AO090023000123	JGI190990	Branched amino acid aminotransferase
Regulation of fatty acids, β-oxidation and peroxisome biogenesis					
SPS19	YNL202W	AN7770.3	AO090701000656	JGI48719	2,4-Dienoyl-CoA reductase (NADPH)[b,c]
POX1[a]	YGL205W	AN6752.3	AO090005000479	JGI181397	Fatty acyl-CoA oxidase[b,c]
POT1[a]	YIL160C	AN1050.3	AO090012000715	JGI38275	3-Ketoacyl-CoA thiolase[b,c,f]
CTA1[a]	YDR256C	AN5918.3	AO090011000540	JGI206591	Catalase[b,c,f]
PXA1	YPL147W	AN10078.3	AO090003000864	JGI196686	Peroxisome ABC transporter[b,c]
PXA2	YKL188C	AN1014.3	AO090012000602	JGI177847	Peroxisome ABC transporter[b,c]
FAA2	YER015W	AN8280.3	AO090102000633	JGI188673	Long-chain fatty acyl-CoA synthetase[b,c]
PEX1	YKL197C	AN5991.3	AO090011000621	JGI48950	Pts1 and Pts2 protein import[g]
PEX6	YNL329C	AN2925.3	AO090005001500	JGI41300	Pts1 and Pts2 protein import[g]
PEX13	YLR191W	AN1511.3	AO090005000629	JGI57403	Pts1 and Pts2 protein import[g]
PEX3	YDR329C	AN2281.3	AO090009000627	JGI192954	Peroxisome biogenesis[g]
PEX5	YDR244W	AN10215.3	AO090005000623	JGI193981	Pst1 protein import receptor[g]
PEX7	YDR142C	AN0880.3	AO090005001175	JGI35474	Pst2 protein import receptor[g]
PEX11	YOL147C	AN1921.3	AO090003000168	JGI55954	Peroxisome proliferation[g]
ANT1	YPR128C	AN0257.3	AO090102000637	JGI36158	Mitochondrial ATP carrier[g]

[a] Promoters were shown to bind Adr1 in *S. cerevisiae* by chromatin immunoprecipitation assays (Young et al. 2002, 2003)

[b] According to SGD database

[c] According to Young et al. 2003

[d] Hits below BLASTP significance threshold and not tri-directional in the three Aspergilli, *E*-value of 1E-30

[e] Orthologues are below BLASTP threshold, *E*-value of 1E-30. Nevertheless, they are conserved in all three *Aspergillus* spp

[f] According to Young et al. 2002

[g] Adapted from Hynes et al. 2008

of the biosynthesis and breakdown of glycerol are less studied in Aspergilli when compared to the metabolism of other carbon sources such as ethanol (David et al. 2006). Only a few previous studies have reported glycerol utilization pathways, for instance, Hondmann et al. (1991) in *A. nidulans* and later David et al. (2006). Through biochemical and genetical analyses, Hondmann et al. (1991) demonstrated that glycerol is catabolized by a glycerol kinase and a mitochondrial FAD$^+$-dependent glycerol-3-phosphate dehydrogenase and that the levels of both enzymes are controlled by carbon catabolite repression and by specific induction in *A. nidulans*. Interestingly, the data obtained from our transcriptome analysis confirmed that the catabolic pathway via glycerol-3-phosphate is the major route for glycerol catabolism in *A. nidulans* and in *A. niger*. In another study, a glycerol kinase mutant of *A. niger* showed weak growth on

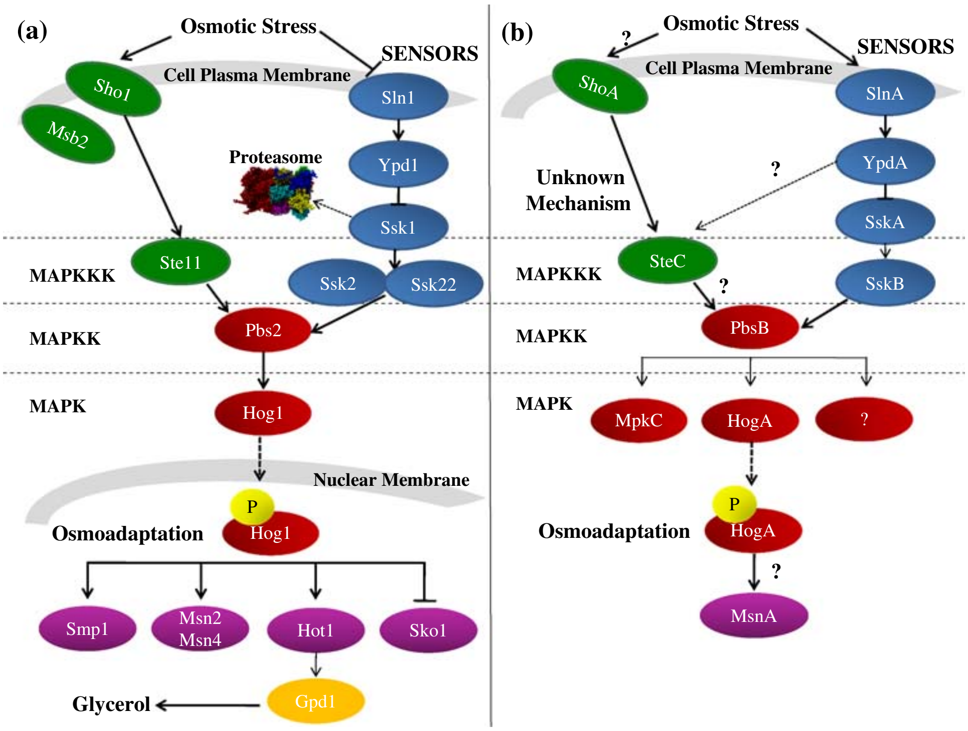


Fig. 5 Osmotic stress response models in *S. cerevisiae* and *Aspergillus* species. **a** HOG pathway mechanism in *S. cerevisiae*. The HOG pathway consists of two osmosensor branches, Sln1 and Sho1. In yeast, the HOG pathway is activated by the remaining branch if either Sln1 or Sho1 is intercepted. **b** HOG pathway mechanism proposed for *Aspergillus* species; i.e. *A. nidulans*, *A. oryzae* and *A. niger*. The osmotic stress response results in activation of several components and ultimately in MsnA activation

glycerol and it was demonstrated that phosphorylation is an important step in glycerol catabolism (Witteveen et al. 1990). Enzymatic analysis of both the mutant and the parental strain showed that at least three different glycerol dehydrogenases were formed under different physiological conditions and sustained growth on glycerol (Witteveen et al. 1990). In our study, the gene encoding glycerol kinase as well as the gene encoding the FAD$^+$-dependent glycerol-3-phosphate dehydrogenase were significantly up-regulated on glycerol compared to glucose media in *A. niger* (JGI45434 and JGI55910, respectively) and to a less extent in *A. oryzae* (AO090001000509, and AO090005001646, respectively). The *A. niger* glycerol kinase gene had a log$_2$-fold change of 3.3 and the FAD$^+$-dependent glycerol-3-phosphate dehydrogenase gene had a log$_2$-fold change of 2. The transcriptional response was not as evident in *A. nidulans* orthologues (AN5589.3 and AN1396.3). In this case, the gene expression changes were not statistically significant. Nevertheless, previous transcriptome analysis reports have shown their up-regulation on glycerol media compared to glucose using another DNA microarrays platform (David et al. 2006). Here, they reported a significant up-regulation of the *A. nidulans* gene encoding the glycerol kinase (AN5589.3), as well as the gene encoding the FADH-dependent glycerol-3-phosphate dehydrogenase (AN1396.3) of 2.9-fold and 2.5-fold, respectively (David et al. 2006).

An alternative pathway is also involved in the catabolism of glycerol. In fact, a gene that was identified to encode a putative NADPH-dependent glycerol dehydrogenase (AN7193.3, AO090023000264 and JGI55928 in *A. nidulans*, *A. oryzae* and *A. niger*, respectively), was up-regulated on glycerol in the three *Aspergillus* species. The major up-regulation occurred in *A. oryzae* and to a lower degree in *A. niger* and *A. nidulans*, where the gene expression changes were not captured as statistically significant. This up-regulation was previously reported by David et al. (2006) in *A. nidulans*. Furthermore, the NAD$^+$-dependent glycerol dehydrogenase encoding gene (*gldB*) (AN5563.3, JGI196413 and AO090009000563) was down-regulated in all three species, being AO090009000563 from *A. oryzae* the most affected.

In *A. oryzae*, the most statistically significant up-regulated gene was the one encoding the enzyme glycerone kinase, AO090120000396, which showed a log$_2$ fold change up-regulation on glycerol of 5 (Supplementary Table 3). The *A. niger* orthologue was also up-regulated on glycerol while the *A. nidulans* orthologue was not assessed as significantly differentially expressed, and consequently it was not captured as an up-regulated conserved response as shown in Table 1. It is likely that both pathways leading to the glycolytic intermediate T3P2 are involved in glycerol utilization in *Aspergillus* species. Nevertheless, the most active pathway in *A. oryzae* is probably the one using glycerol dehydrogenase and glycerone kinase to produce glycerone phosphate. In contrast, in *A. niger* and *A. nidulans*, the pathway using glycerol kinase and the FAD$^+$-dependent glycerol-3-phosphate dehydrogenase is most likely to be the dominant (see Fig. 3). Another study in *A. niger* supporting our transcriptome results has shown that glycerol accumulated in a glycerol kinase mutant (Witteveen and Visser 1995), which was able to synthesize glycerol, but not able to catabolize it, suggesting that the activity of this pathway is important for glycerol catabolism.

Regulation of gene expression by the Adr1 transcription factor

Adr1-dependent genes encode enzymes involved in channeling metabolites into acetyl-CoA and NADH production in *S. cerevisiae*. From transcriptome data analysis, genes listed in Table 1 and mapped to metabolic pathways illustrated in Supplementary Fig. 4 (details in Supplementary Figs. 1 to 3) are suspected to be expressed in an Adr1-dependent manner in *Aspergillus* species based on the fact that many of these genes were significantly up-regulated on glycerol media. According to *S. cerevisiae* studies, Adr1-dependent genes are mainly present in pathways leading from ethanol, glycerol, lactate and the oxidation of fatty acids to the formation of acetyl-CoA, generating NADH in the process (Young et al. 2003).

Adr1 regulates a diversity of genes which fall into a variety of functional classes in *S. cerevisiae*, such as non-fermentative carbon metabolism, amino acid transport and metabolism, peroxisomal biogenesis, meiosis and sporulation, transcriptional regulation and signal transduction (Young et al. 2003). Many of the regulated genes had potential Adr1 binding sites in their promoters. Chromatin immunoprecipitation experiments showed that Adr1 bound to the promoters of genes such as *ADH2*, *ACS1*, *CTA1*, *POT1* and *GUT1* (Young et al. 2002) and *ADY2*, *ALD4*, *POX1*, *CIT3*, *GIP2*, *ICL2*, *ETR1* and *FDH2* (Young et al. 2003), demonstrating that Adr1 regulates directly those genes. Interestingly, in our study, the *Aspergillus* orthologues of these genes were up-regulated on glycerol media. A number of these orthologues have the Adr1 binding motif (TGCGGGGA) present several times in their promoter regions (see Supplementary Table 5). In *A. nidulans*, mutants affecting glycerol uptake have been constructed (Visser et al. 1988); however, a transcription factor responsible for its metabolic regulation has not been identified yet. We therefore propose that the Adr1 transcription factor is conserved through evolution, present in *Aspergillus* species and furthermore, responsible in concert with other transcription factors for regulation of several pathways such as glycerol metabolism as found in *S. cerevisiae*.

Interestingly, uptake of fatty acyl-CoA by peroxisomes is also Adr1 dependent in *S. cerevisiae* (Young et al. 2003). Fatty acids are first activated to their corresponding acyl-CoA thioesters in an ATP-dependent manner (Maggio-Hall and Keller 2004). *S. cerevisiae* houses only peroxisomal β-oxidation, while in contrast the filamentous fungus *A. nidulans* houses both peroxisomal and mitochondrial β-oxidation, similar to mammals and plants (Hynes et al. 2008). In our microarray study, several components of fatty acid β-oxidation and peroxisome biogenesis have been significantly up-regulated on glycerol, particularly in *A. oryzae* and *A. niger*. Sps19, Pox1, Pot1 and a Cta1 orthologue in *A. oryzae*, AO090011000540, encoding a catalase, which breaks down hydrogen peroxide in the peroxisomal matrix formed by acyl-CoA oxidase during fatty acid β-oxidation were up-regulated on glycerol (SGD database, *Saccharomyces* Genome Database). Besides in *S. cerevisiae*, Pxa1 and Pxa2 encode subunits of the ABC transporter responsible for uptake of long-chain fatty acyl-CoA derivatives into peroxisomes (Shani and Valle 1996). Similarly, we could identify a statistically significant up-regulation of Pxa2 orthologues on glycerol media in *A. niger* and *A. oryzae* and of Pxa1 orthologue in *A. oryzae*. Likewise in a similar transcriptome study in *S. cerevisiae*, Pxa1 and Pxa2 showed dependence on Adr1 (Young et al. 2003). In addition, medium chain fatty acids are converted into their acyl-CoA derivative by a peroxisomal acyl-CoA ligase encoded by Faa2, which is activated by Adr1 in yeast (Young et al. 2003) and where we found its corresponding orthologues in *A. niger* and *A. oryzae* significantly up-regulated on glycerol. This suggests that fatty acids β-oxidation might be also regulated by Adr1 in *Aspergillus* species in cooperation with other transcription factors such as FarA and FarB previously found in *A. nidulans* (Hynes et al. 2006).

Stress response

In yeast and fungi, glycerol has a role in regulating the osmotic pressure in the cells (Blomberg and Adler 1992). In the yeast *S. cerevisiae*, the high osmolarity glycerol (HOG) pathway is activated by increased external osmolarity and it consists of two upstream osmo-sensing input branches, Sln1 and Sho1, and downstream kinases, which constitute the mitogen-activated protein kinase, kinase, kinase (MAPKKK) signaling cascade pathway (Hohmann 2002). The main components are shown in Fig. 5a); namely Ssk2/Ssk22 (MAPKKK), Pbs2 (MAPKK) and Hog1 (MAPK) (Maeda et al. 1995). Activation of the HOG pathway results in the induction of genes required for osmotic adaptation in *S. cerevisiae*, for instance, glycerol biosynthesis genes such as glycerol-3-phosphate dehydrogenase (*GPD1*) (Albertyn et al. 1994) and glycerol-3-phosphatase (*GPP2*) (Norbeck et al. 1996). Likewise, glycerol has been shown to be involved in the osmotic response in *A. nidulans* (Beever and Laracy 1986). All the components of the HOG pathway in yeast have orthologues in *A. nidulans* (Furukawa et al. 2005; Han and Prade 2002) and in other filamentous fungi, such as *Neurospora crassa*, proteins homologous to the yeast HOG components have also been previously described (Fujimura et al. 2003).

Here, we conducted a homology search to find the orthologous genes in the other two *Aspergillus* species. The results are summarized in Supplementary Table 6 and the proposed consensus osmo-sensing signaling cascade mechanism for these three *Aspergillus* species is depicted in Fig. 5b). We found *SHO1* orthologues in *A. oryzae* and *A. niger* which were significantly up-regulated on glycerol. In addition, *SLN1* orthologues were identified as differentially regulated in the three *Aspergillus* species, but none of the expression changes was statistically significant. A previous study showed that the sensor protein SlnA (AN1800.3) was significantly up-regulated and possibly induced when glycerol was the sole carbon source when compared to glucose or ethanol (David et al. 2006). Interestingly, *YPD1* orthologues in *A. nidulans* and *A. niger*, AN2005.3 and JGI214261 were down-regulated. Additionally, *SSK1*, *SSK2* and *PBS2* orthologues were up-regulated in one or more As*pergillus* species, but the only statistically significant up-regulation was for *A. niger*. *HOG1* orthologue genes in the three *Aspergillus* were up-regulated with exception of the *A. niger HOG1* homologue. Interestingly, the other *HOG1* orthologue in *A. nidulans*, *mpkC* (Furukawa et al. 2005) and a putative *hogA*, were up-regulated as well as two other putative *hogA* orthologues in *A. niger*, JGI187878 and JGI207710.

As it can be seen from the gene expression changes, the most affected *Aspergillus* species in relation to glycerol was *A. niger*, where almost all the HOG pathway components were statistically significantly up-regulated. The response in the other two *Aspergillus*, *A. oryzae* and *A. nidulans*, was not as evident, nevertheless all the HOG pathway components are conserved. It is quite interesting that the genes are transcriptionally regulated as most signal transduction pathways are mainly activated through modifications at the post-translational level.

A study by Furukawa et al. (Furukawa et al. 2005) in *A. nidulans* proposed that activation of HogA depended on the two-component signaling pathway, but not on ShoA. Instead, they proposed that PbsB could activate another Hog1 MAPK orthologue, named MpkC, when *mpkC* was over-expressed in *A. nidulans*. They further proposed that SskA regulates the *A. nidulans* HOG pathway in response to a variety of external stimuli and suggested that unlike yeast Sho1, it might not be involved in osmoresponsive activation of HogA. Interception of the HOG pathway did not cause a drastic sensitivity to high osmolarity, implying that *A. nidulans* had one or more unknown osmoresponsive pathways or unidentified mechanisms for osmoadaptation (Furukawa et al. 2005). In contrast to our work, these authors did not detect *mpkC* transcript levels in *A. nidulans*, wild type or *hogA* deletion mutants under normal or stress conditions. In contrast, we were able to identify a statistically significant up-regulation of MpkC orthologue in *A. niger* on glycerol while earlier transcriptome studies showed *mpkC* up-regulation in *A. nidulans* as well (David et al. 2006).

The other components of the Sho1 osmo-sensing branch, Ste11 in *S. cervisiae* or its orthologue in *A. nidulans*, SteC (Wei et al. 2003) as well as the coding genes of Cdc24, Cdc42, Ste20 and Ste50 were also significantly up-regulated in *A. niger*. This suggests a partial activation of both osmo-sensing pathways in *A. nidulans*, *A. oryzae* and *A. niger*, through ShoA and SlnA. The osmoresponsive mechanism can be generalized to be a conserved mechanism in the three *Aspergillus* species due to the gene expression patterns obtained; where in general, the *hogA* orthologues in the three Aspergilli were up-regulated on glycerol. In *S. cerevisiae*, phosphorylation of Hog1 is fast and causes its immediate translocation to the nucleus (Posas et al. 1996), where it phosphorylates several transcription factors, for instance, Msn2, Msn4, Sko1 and Hot1. Interestingly, in our study MsnA was significantly up-regulated in *A. niger*.

Our results have shown that comparative genomics and transcriptomics of closely related species such as three *Aspergillus* species can help in elucidating unknown regulatory mechanisms as well as in improving functional annotation. This example applied to glycerol metabolism is of biotechnological importance as glycerol is a major by-product from biodiesel production; and thereby represents a cheap carbon source which could be exploited for bio-based production of chemicals. Our identification of a conserved regulatory element that controls glycerol metabolism is important for future engineering of Aspergilli as cell factories for sustainable chemical production.

Data deposition

Normalized gene expression values, obtained from batch fermentations on glucose and glycerol medium for *A. nidulans*, *A. oryzae* and *A. niger*, were deposited at the GEO database (GEO database, Gene Expression Omnibus database), with accession numbers GPL5975 (platform), GSM393134-GSM393151 (samples) and GSE15702 (series).

Acknowledgments The authors would like to thank Tina Johansen, Pia Friis and Lene Christiansen at Technical University of Denmark for assistance with the experimental work and Dr. Kim Hansen for supervising the fermentations with *A. oryzae*, Lone Vuholm and Anne Kejser Jensen at Novozymes for technical assistance. We would like to thank National Council of Research Conacyt-Mexico and Chalmers University of Technology for financial support to MS; Novozymes Bioprocess Academy and Chalmers University of Technology for financial support to WV; Danish Research Council for Technology and Production Sciences and Lundbeck Foundation for financial support to GP and Danish Research Agency for Technology and Production for financial support to MRA. We thank Dr. Gerald Hofmann for revision of the manuscript and good scientific discussion.

References

Affymetrix: GeneChip expression analysis technical manual, with specific protocols for using the GeneChip hybridization, wash, and stain kit (2007) P/N 702232, Affymetrix, Santa Clara, CA, Revision 2

Albertyn J, Hohmann S, Thevelein JM et al (1994) GPD1, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic-stress in *Saccharomyces cerevisiae*, and its expression is regulated by the high osmolarity glycerol response pathway. Mol Cell Biol 14:4135–4144

Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22:1600–1607

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andersen MR, Vongsangnak W, Panagiotou G et al (2008) A trispecies *Aspergillus* microarray: comparative transcriptomics of three *Aspergillus* species. Proc Natl Acad Sci USA 105:4387–4392

Appleyard M, McPheat WL, Stark MJR (2000) A novel 'two-component' protein containing histidine kinase and response regulator domains required for sporulation in *Aspergillus nidulans*. Curr Genet 37:364–372

Beever RE, Laracy EP (1986) Osmotic adjustment in the filamentous fungus *Aspergillus nidulans*. J Bacteriol 168:1358–1365

Bembom O, Keles S, van der Laan MJ (2007) Supervised detection of conserved motifs in DNA sequences with cosmo. Stat Appl Genet Mol Biol 6:Article 8

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc B Methodol 57:289–300

Blomberg A, Adler L (1992) Physiology of osmotolerance in fungi. Adv Microb Physiol 33:145–212

Blumenstein A, Vienken K, Tasler R et al (2005) The *Aspergillus nidulans* phytochrome FphA represses sexual development in red light. Curr Biol 15:1833–1838

Carlsen M, Nielsen J (2001) Influence of carbon source on alpha-amylase production by *Aspergillus oryzae*. Appl Microbiol Biotechnol 57:346–349

Cheng C, Kacherovsky N, Dombek KM et al (1994) Identification of potential target genes for Adr1p through characterization of essential nucleotides in UAS1. Mol Cell Biol 14:3842–3852

Das HK, Baez ML (2008) ADR1 interacts with a down-stream positive element to activate PS1 transcription. Front Biosci 13:3439–3447

David H, Hofmann G, Oliveira A et al (2006) Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans*. Genome Biol 7(11):R108

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686

Dudoit S, Gendeman RC, Quackenbush J (2003) Open source software for the analysis of microarray data. Biotechniques (Suppl):45–51

Felenbok B, Kelly JM (1996) Regulation of carbon metabolism in mycelia fungi. In: Marzluf G, Brambl R (eds) The Mycota: III: Biochemistry and molecular biology. Springer, Berlin, pp 369–380

Fujimura M, Ochiai N, Oshima M et al (2003) Putative homologs of SSK22 MAPKK kinase and PBS2 MAPK kinase of *Saccharomyces cerevisiae* encoded by os-4 and os-5 genes for osmotic sensitivity and fungicide resistance in *Neurospora crassa*. Biosci Biotechnol Biochem 67:186–191

Furukawa K, Hoshi Y, Maeda T et al (2005) *Aspergillus nidulans* HOG pathway is activated only by two-component signalling pathway in response to osmotic stress. Mol Microbiol 56:1246–1261

Gentleman RC, Carey VJ, Bates DM et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10):R80

Han KH, Prade RA (2002) Osmotic stress-coupled maintenance of polar growth in *Aspergillus nidulans*. Mol Microbiol 43:1065–1078

Hohmann S (2002) Osmotic stress signaling and osmoadaptation in yeasts. Microbiol Mol Biol Rev 66(2):300–372

Hondmann DHA, Busink R, Witteveen CFB et al (1991) Glycerol catabolism in *Aspergillus nidulans*. J Gen Microbiol 137:629–636

Hynes MJ, Murray SL, Duncan A et al (2006) Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus *Aspergillus nidulans*. Eukaryotic Cell 5:794–805

Hynes MJ, Murray SL, Khew GS et al (2008) Genetic analysis of the role of peroxisomes in the utilization of acetate and fatty acids in *Aspergillus nidulans*. Genetics 178:1355–1369

Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264

Kobayashi T, Abe K, Asai K et al (2007) Genomics of *Aspergillus oryzae*. Biosci Biotechnol Biochem 71:646–670

Maeda T, Takekawa M, Saito H (1995) Activation of yeast Pbs2 MAPKK by MAPKKKs of by binding of an SH3-containing osmosensor. Science 269:554–558

Maggio-Hall LA, Keller NP (2004) Mitochondrial beta-oxidation in *Aspergillus nidulans*. Mol Microbiol 54:1173–1185

Norbeck J, Pahlman AK, Akhtar N et al (1996) Purification and characterization of two isoenzymes of DL-glycerol-3-phosphatase from *Saccharomyces cerevisiae*—identification of the corresponding GPP1 and GPP2 genes and evidence for osmotic regulation of Gpp2p expression by the osmosensing mitogen-activated protein kinase signal transduction pathway. J Biol Chem 271:13875–13881

Panagiotou G, Andersen MR, Grotkjaer T et al (2008) Systems analysis unfolds the relationship between the phosphoketolase pathway and growth in *Aspergillus nidulans*. PLoS ONE 3:3847

Pedersen H, Beyer M, Nielsen J (2000) Glucoamylase production in batch, chemostat and fed-batch cultivations by an industrial strain of *Aspergillus niger*. Appl Microbiol Biotechnol 53:272–277

Posas F, WurglerMurphy SM, Maeda T et al (1996) Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 ''two-component'' osmosensor. Cell 86:865–875

R Development Core Team (2007) A Language and Environment for Statistical Computing

Roberts GG, Hudson AP (2006) Transcriptome profiling of *Saccharomyces cerevisiae* during a transition from fermentative to glycerol-based respiratory growth reveals extensive metabolic and structural remodeling. Mol Genet Genomics 276:170–186

Ronnow B, Kielland-Brandt MC (1993) *GUT2*, a gene for mitochondrial glycerol 3-phosphate dehydrogenase of *Saccharomyces cerevisiae*. Yeast 9:1121–1130

Ruijter GJG, Visser J (1997) Carbon repression in Aspergilli. FEMS Microbiol Lett 151:103–114

Schneider TD, Stephens RM (1990) Sequence logos—a new way to display consensus sequences. Nucleic Acids Res 18:6097–6100

Shani N, Valle D (1996) A *Saccharomyces cerevisiae* homolog of the human adrenoleukodystrophy transporter is a heterodimer of two half ATP-binding cassette transporters. Proc Natl Acad Sci USA 93:11901–11906

Smyth G (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article 3

Strauss J, Mach RL, Zeilinger S et al (1995) Cre1, the carbon catabolite repressor protein from *Trichoderma reesei*. FEBS Lett 376:103–107

Suzuki A, Kanamaru K, Azuma N et al (2008) GFP-Tagged expression analysis revealed that some histidine kinases of *Aspergillus nidulans* show temporally and spatially different expression during the life cycle. Biosci Biotechnol Biochem 72:428–434

van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827–842

Visser J, Vanrooijen R, Dijkema C et al (1988) Glycerol uptake mutants of the hyphal fungus *Aspergillus nidulans*. J Gen Microbiol 134:655–659

Vongsangnak W, Olsen P, Hansen K et al (2008) Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. BMC Genomics 9:14

Vongsangnak W, Salazar M, Hansen K et al (2009) Genome-wide analysis of maltose utilization and regulation in aspergilli. Microbiology (in press)

Wei HJ, Requena N, Fischer R (2003) The MAPKK kinase SteC regulates conidiophore morphology and is essential for heterokaryon formation and sexual development in the homothallic fungus *Aspergillus nidulans*. Mol Microbiol 47:1577–1588

Witteveen CFB, Visser J (1995) Polyols pools in *Aspergillus niger*. FEMS Microbiol Lett 134:57–62

Witteveen CFB, Vandevondervoort P, Dijkema C et al (1990) Characterization of a glycerol kinase mutant of *Aspergillus niger*. J Gen Microbiol 136:1299–1305

Workman C, Jensen LJ, Jarmer H et al (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 3(9)

Young ET, Kacherovsky N, Van Riper K (2002) Snf1 protein kinase regulates Adr1 binding to chromatin but not transcription activation. J Biol Chem 277:38095–38103

Young ET, Dombek KM, Tachibana C et al (2003) Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem 278:26146–26158

Paper 5

Analysis of genome-wide co-expression and co-evolution of *Aspergillus oryzae* and *Aspergillus niger*

Wanwipa Vongsangnak, Intawat Nookaew,
Margarita Salazar, and Jens Nielsen

Manuscript in submission

Analysis of genome-wide co-expression and co-evolution of *Aspergillus oryzae* and *Aspergillus niger*

Wanwipa Vongsangnak, Intawat Nookaew, Margarita Salazar and Jens Nielsen*

Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

*Correspondence: Jens Nielsen. Email: nielsenj@chalmers.se

ABSTRACT

Analysis of co-expressed genes in response to different perturbations at the genome-level can provide new insight into global regulatory structures. Here we performed a cross-species comparative investigation by exploring genomes and transcriptional co-expressions profiles in *Aspergillus oryzae* and *Aspergillus niger*. Based on analysis of conserved co-expressed genes, fatty acid catabolism via beta-oxidation, fatty acid transport, the glyoxylate bypass and peroxisomal biogenesis were identified as core co-evolved pathways between the two species. The occurrence of co-expression patterns, allowed for identification of DNA regulatory motifs and putative corresponding transcription factors, and we hereby show that comparative transcriptome analysis between two closely related fungi allows for identification of how genes involved in the utilization of fatty acids, peroxisomal biogenesis and the glyoxylate bypass are regulated. Interestingly, "CCTCGG" was identified as a core binding site for putative FarA and FarB transcription factors that govern the underlined biological processes. Phylogeny and domain architecture analysis of amino acid sequences of FarA and FarB in 8 species of aspergilli, clearly indicate that these proteins are evolutionarily conserved across *Aspergillus* species as well as they are conserved in other fungi.

INTRODUCTION

The recently completed genome sequences of several *Aspergillus* species provide a good opportunity to better understand the biology of fungi through comparative genomics. Comparison of DNA sequences in these fungal genomes enables us to identify genes and their functions as well as regulations based on evolutions (Fedorova et al. 2008; Galagan et al. 2005; Nierman et al. 2006). To translate information in DNA sequences to biological functions, functional genomics has emerged as a research field that involves systematic analysis of gene function via omics technologies (Brent, 2000; Hieter & Boguski, 1997). Transcriptomics has been widely used in many studies with high reliability and reproducibility (Shi et al. 2008). Transcriptomics can be used for identification of co-expressed genes which provide hints toward inferring gene function based on the concept of guilt-by-association, i.e. co-expressed genes are likely to serve similar purposes and to be regulated by similar mechanisms (Altman & Raychaudhuri, 2001; Schulze & Downward, 2001).

Several publications have shown that clustering genes according to their mRNA expression profiles often share common upstream sequence motif (s) (Brazma et al. 1998; Tavazoie et al. 1999) allowing for inferring the transcription factor (s) that potentially bind the motif (s) and probably regulate the genes in the cluster. This strategy has been widely applied in several organisms such as *Saccharomyces cerevisiae* (Young et al. 2003) (Wolfsberg et al. 1999), mammalian cells (Oldham et al. 2006), *Arabidopsis thaliana* (Vandepoele et al. 2009) and also filamentous fungi i.e. *Aspergillus nidulans* (David et al. 2006). *A. niger* and *A. oryzae* (Andersen et al. 2008b; Salazar et al. 2009; Vongsangnak et al. 2009). Although several transcriptome data sets from previous studies have been publicly available, there are few insightful studies on transcriptional regulation systems in aspergilli. Thus there is an opportunity to enrich the information content in already published data with the objective to perform analysis that leads to underline the core biological processes on both genotype and expression in different aspergilli and hereby identify the possible coexistence of DNA regulatory motifs and transcription factors.

In this study, we aim to carry out cross-species analysis of genome-wide transcriptional co-expression patterns under different growth conditions, specifically to identify concerted transcriptional changes of genes that clearly reflect cellular adaptations. Our analysis first focus on comparative analyses between *A. oryzae* and *A. niger* at the genome and the transcriptome levels. The results were combined to reconstruct core metabolic pathways in the two species which are co-evolved and have similar environmental responses. Our study demonstrates that most of the genes in these two species have similar environmental responses in terms of gene expression patterns, but there are a few of these responses that can be found to be due to conserved regulatory motifs between the two species.

METHODS

Pair-wise protein sequence comparisons

The complete set of amino acid sequences of the open reading frames from *A. oryzae* RIB 40 (Machida et al. 2005) (version 1) and *A. niger* ATCC 1015 (version 3) (http://genome.jgi-psf.org/Aspni5/Aspni5.home.html) were compared against each other by using BLASTP (Altschul et al. 1990) to identify their homologues. An estimated expectation values of cut-off of 1E-30, alignment length of 200 amino acids, and percentage identity of 40 (%), were set to evaluate statistical significance of conserved orthologues. Bidirectional BLASTP was applied to obtain a conservative set of 1:1 orthologues between the two *Aspergillus* species. The full subset of gene orthologues is listed in Supplementary file 1.

Microarray data acquisition and analysis

Affymetrix CEL-data files were preprocessed using the bioconductor package (Gentleman et al. 2004) on the R software version 2.9.0 (R Development Core Team). Of the 13,120 putative genes identified in the genome of *A. oryzae* (Machida et al. 2005; Vongsangnak et al. 2008), 12,039 probe sets were used for microarray analysis. Of the 11,200 putative genes identified in the genome of *A. niger*, 11,122 probe sets were used for microarray analysis. Normalization was performed by using the qspline algorithm (Workman et al. 2002). The probe intensities were corrected for background and gene expression values were calculated by using Probe Logarithmic Intensity Error (PLIER) estimation method (Seo & Hoffman, 2006). All statistical analyses were invoked through the affy package (Gautier et al. 2004) of R scripts (Dudoit et al. 2003). In order to conduct an initial characterization and quality assessment of all microarray data sets, we applied Principal Component Analysis (PCA) based on Singular Value Decomposition (SVD) to visualize and detect global variation of gene expression data. Statistical analysis with multiple testing corrections (Benjamin-Hogberg method) by one-way ANOVA test to the four different carbon sources dataset (i.e. glucose, maltose, glycerol, and xylose) was applied to determine significantly different gene expressions in *A. oryzae* or *A. niger*. A cut-off value of adjusted p-value < 0.075 was considered to ensure identification of statistically significant mRNA levels for both *Aspergillus* species.

Clustering analysis

Consensus clustering algorithm was used (Grotkjaer et al. 2006) in order to identify similar expression patterns among 4 different carbon sources as mentioned above. The algorithm was implemented in the MATLAB toolbox called ClusterLustre. In data preprocessing process, Pearson's correlation was used for similarity measurement of average expression values from three biological replicates in each carbon source. This allowed us to cluster gene expression of *A. oryzae* and *A. niger*, simultaneously. Then clustering was performed by a K-means algorithm.

Evaluation of function and metabolic pathway enrichment

The functional categories of gene ontology (GO) were retrieved from the Joint Genome Institute (JGI) database for *A. niger* and Uniprot protein database for *A. oryzae*. The reporter algorithm was firstly employed to uncover the important GO terms across conditions by integration of GO with significance values of the transcriptome data. The results of significant reporter GO (p-value < 0.05) were found in Supplementary file 2. We also evaluated whether there was over-representation of genes associated with a metabolic pathway within a cluster of genes by using the genome-scale metabolic models *iWV1314* (Vongsangnak et al. 2008) and *iMA871* (Andersen et al. 2008a) for *A.oryzae* and *A. niger* respectively. These enrichment evaluations were performed using standard hypergeometric tests. We assessed significant categories using a p-value <0.05.

Detection of DNA regulatory motif and transcription factor

DNA regulatory motifs were identified by using R 2.9.0 with the cosmo package (Bembom et al. 2007). A background Markov model was pre-computed by using the intergenic regions from the *Aspergillus* sequences by following the previous work (Andersen et al. 2008b). The two-component-mixture (TCM) model approach was employed to search for the most over-represented motif in each cluster of gene expression using a thousand base-pairs of up-stream sequences of the relevant genes of both species simultaneously. The up-stream sequences of *A. oryzae* and *A. niger* co-expressed genes were extracted from the Broad Institute database (http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html). To obtain a biological meaning from the motif identification results, the obtained over-represented motifs were queried against known or predicted *Aspergillus* consensus motifs or other fungal functional consensus sequences from public databases or the literature to identify potential associated transcription factors.

Conserved domain architecture and phylogeny analysis of FarA and FarB proteins in aspergilli

To identify conservation of the proteins FarA and FarB in aspergilli, we performed comparative sequence analysis of these proteins. Initially, the known amino acid sequences of FarA and FarB of *A. nidulans* original Glasgow strain (Hynes et al. 2006) were extracted from GenBank database (http://www.ncbi.nlm.nih.gov/) with accession numbers ABD51992.1 and ABD51993.1, respectively. The complete set of the two amino acid sequences of *A. nidulans* original Glasgow strain was used as query for searching against the amino acid sequences of 10 different sequenced *Aspergillus* genomes by using BLASTP (Altschul et al. 1990). The sequenced species included were: *A. oryzae* RIB40 (Machida et al. 2005), *A. niger* CBS 513.88 (Pel et al. 2007), *A. niger* ATCC 1015 (version 3.0) (http://genome.jgipsf.org/Aspni5/Aspni5.home.html), *A. nidulans* FGSC A4 (version 4) (Galagan et al. 2005; Wortman et al. 2009), *Aspergillus fumigatus* Af293 (Nierman et al. 2006), *A. fumigatus* A1163 (Fedorova et al. 2008), *Aspergillus flavus* NRRL 3357 (Payne et al. 2006), *Aspergillus terreus* NIH2624 (www.broad.mit.edu/annotation/fungi/aspergillus_terreus), *Aspergillus clavatus* NRRL 1 (Fedorova et al. 2008) and *Aspergillus fischeri* NRRL 181 (Fedorova et al. 2008). An estimated expectation value cut-off of less than 1E-100, more than 40% identity, and more than 200 amino acids of alignment length was set to assess statistical significance for identification of orthologous genes.

The evolution of Far proteins of the aspergilli was inferred using the UPGMA method (Sneath & Sokal, 1973). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (50000 replicates) (Felsenstein, 1985). The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl & Pauling, 1965) and were in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset. Phylogeny analyses were conducted in software MEGA4 (Tamura et al. 2007). In order to identify conserved protein domain structure for FarA and FarB, we applied conserved domain database (CDD) containing domain models imported from a number of reliable databases (Pfam, SMART, COG, PRK, TIGRFAM) (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) for domain analyses.

RESULTS

Comparative analysis of co-evolved genes in *A. oryzae* and *A. niger*

As a basis for our cross-species comparative investigation at the genome level of *A. oryzae* and *A. niger*, genes having orthologues in the two species were identified by using a BLASTP-based comparison (See METHODS). Using this approach 5,928 conserved genes (1:1 orthologues) were found (See Supplementary file 1) and this gene-set was used to identify transcriptional co-expression described in the following.

Identification of transcriptional co-expression

Quality evaluation of multiple microarray data sets

In our analysis, we were using biological triplicate transcriptome data for *A. niger* and *A. oryzae* grown on glucose, xylose, maltose and glycerol, all data published by our group earlier (Andersen et al. 2008b; Salazar et al. 2009; Vongsangnak et al. 2009). As these data were generated as part of different studies, we first checked the quality of the data, in particularly with respect to whether the three replicated experiments on the four different carbon sources for the two species grouped together. For this we performed Singular Value Decomposition (SVD)-based Principle Component Analysis (PCA) (See METHODS). The results are shown in Figure 1. It is seen that there is a high reproducibility of transcriptome data sets from the independent biological replicates and they are also well-separated with respect to the different conditions for each of the two species. The data set can therefore be confidently used for further analysis.

Statistical transcriptome data analysis

All twelve data sets from the four different carbon sources of *A. oryzae* or *A. niger*, were used for performing a one-way Analysis Of VAriance (ANOVA) using normalized transcriptome data from all the replicated experiments. This allowed capturing the genes with the largest variance response across the different carbon sources leading to identification of 2,115 and 4,957 significantly differentially expressed genes for *A. oryzae* and *A. niger*, respectively.

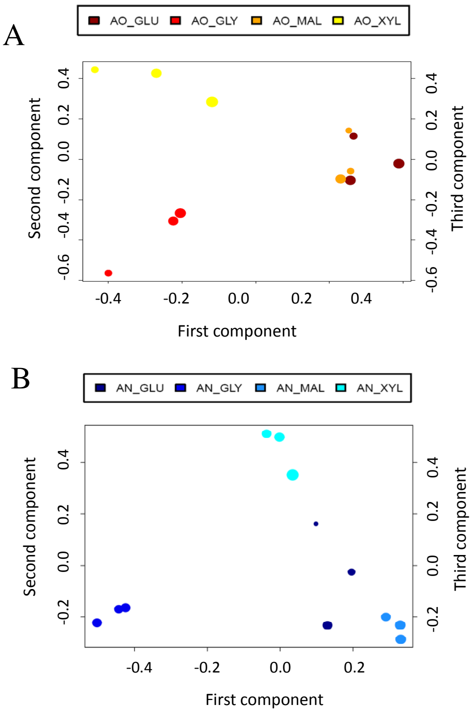


Figure 1 PCA analysis by pseudo-three dimensions plot of the first three eigen vectors from SVD for quality assessment of all three replicated transcriptome data on four different carbon sources; GLU: Glucose, GLY: Glycerol, MAL: Maltose, XYL: Xylose. (A) *A. oryzae* data (B) *A. niger* data

A list of these significant genes is presented in Supplementary file 3. The significantly regulated genes in both species were cross compared with the list of the 5,928 conserved genes for the two species. This resulted in the identification of 687 significantly regulated genes in each species that are orthologues between the two species (see Figure 2). All significant values were integrated into the GO network and the reporter algorithm was used to identify the important GOs across carbon sources for both *A. oryzae* and *A. niger* (See Supplementary file 2).

Gene clustering

The statistical significant genes of the two species, which were identified from the one-way ANOVA analysis as reported above, were clustered according to their similar gene expression patterns across the four different carbon sources. Using a consensus clustering method (Grotkjaer et al. 2006), six patterns of environmental response for both aspegilli were achieved as shown in Table 1, with the details of genes in each cluster reported in Supplementary file 4. Evaluation of metabolic pathway enrichment in all six clusters is presented in Table 1. Commonly, genes with highly correlated expression profiles are likely to have related functions and possible even common transcriptional regulation if it is based on evolutionary sequence conservation. We therefore combined the analysis of co-evolved genes of *A. oryzae* and *A. niger* as described in the previous section with the occurrence of co-expression pattern to underline core conserved cellular processes in both genotype and expression. As shown in Table 1, interestingly, there are only two clusters, named cluster 1 and cluster 2 that have a significant number of genes with occurrence of co-expressions and conserved orthologues. Following both criteria, 238*2 genes in cluster 1 and 23*2 genes in cluster 2, were identified.

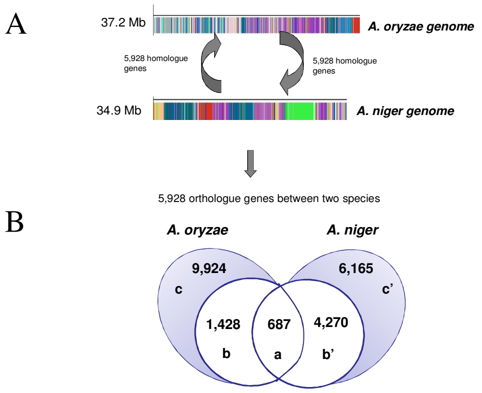
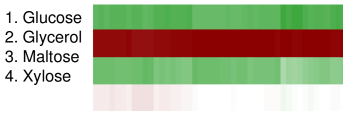
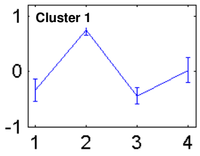
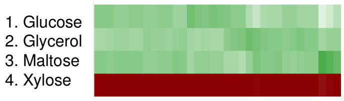
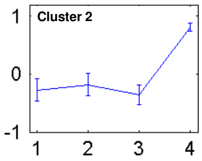
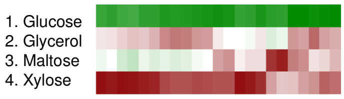
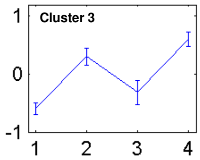
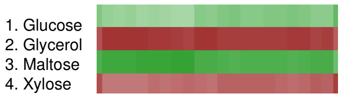
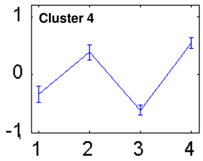
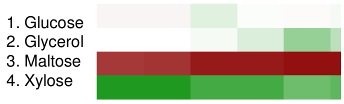
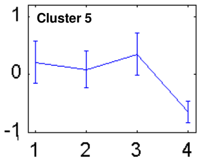
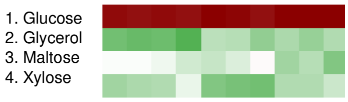
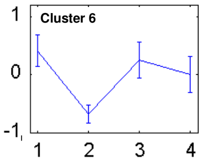


Figure 2 Illustration of the steps of the data integration approach applied. (A) Comparative genomics of *A. oryzae* and *A. niger* where the genome sequences of both species were compared. (B) Integration of genomics and transcriptomics analysis resulted in the identification of significantly regulated genes with conserved sequences. Venn diagram inner area, a, denotes the number of differentially expressed and conserved genes present in both species. b and b' corresponds to the number of genes differentially expressed in either *A. oryzae* or *A. niger*, respectively. c and c' corresponds to the number of genes present in each *Aspergillus* specie, some of them conserved, but not differentially expressed, a+b+c represents the total number of genes present in the genome of *A. oryzae*; a+b'+c' represents the total number of genes present in the genome of *A. niger*.

Table 1 Gene expression profiles and cluster patterns of *A. oryzae* and *A. niger*

Gene expression profiles	Cluster patterns*	Number of co-expressed genes in cluster	Over-represented metabolic pathway
		Total gene number: 777 (238)$^{\#}$ *A. oryzae* : 290 genes *A. niger* : 487 genes	Glycerol metabolism Glycolysis and gluconeogenesis Propanoate and butanoate metabolism Polysaccharide metabolism Valine/leucine/isoleucine metabolism Phenylalanine/tyrosine/tryptophan biosynthesis
		Total gene number: 172 (23)$^{\#}$ *A. oryzae :* 120 genes *A. niger* : 52 genes	Arabinose and xylose metabolism Pentose phosphate pathway Polysaccharide metabolism
		Total gene number: 152 (2)$^{\#}$ *A. oryzae :* 138 genes *A. niger* : 14 genes	Polysaccharide metabolism Pyruvate metabolism
		Total gene number: 160 (7)$^{\#}$ *A. oryzae* : 100 genes *A. niger* : 60 genes	Phenylalanine/tyrosine/tryptophan biosynthesis
		Total gene number: 30 (1)$^{\#}$ *A. oryzae :* 7 genes *A. niger* : 23 genes	ND
		Total gene number: 83 (14)$^{\#}$ *A. oryzae :* 32 genes *A. niger* : 51 genes	Polysaccharide metabolism Carbohydrates transport Other compounds transport

*Cluster patterns: The *x* axis represents the four different carbon sources investigated: 1 - Glucose; 2 - Glycerol, 3 - Maltose; 4 - Xylose; the *y* axis represents normalized gene expression intensities.
$^{\#}$Number of orthologous genes across two aspergilli
ND: Not Detectable

Identification of core cellular processes of co-expression and co-evolution for cross-species of the aspergilli

As mentioned above, there are many genes in cluster 1 and cluster 2 that are orthologues and co-expressed in the two fungi. Considering cluster 1, it is clear that genes were up-regulated in response to growth on glycerol, which is also found by the identification of genes associated with the metabolic pathway of glycerol metabolism. The cluster contains a lot of co-evolved and co-expressed genes that has been annotated with similar functions, i.e. genes encoding enzymes involved in fatty acid catabolism by the beta-oxidation pathway, fatty acid transport (e.g. mitochondria carnitine-acylcarnitine carrier protein and peroxisomal long-chain acyl-CoA transporter), the glyoxylate bypass, peroxisomal biogenesis and function. These were also found in reporter GO, such as peroxisome (GO:0005777), glyoxylate cycle (GO:0006097), lipid metabolic process (GO:0006629), and glycerol metabolic process (GO:0006071) (See Supplementary file 2). The results indicates that there is a co-regulation of metabolic pathways involved in glycerol and fatty acid catabolism, probably due to the co-existence of these compounds as triacylglycerides and phospholipids in nature. In fungi, the beta-oxidation pathway has been studied for localization by Shen et al. (Shen & Burger, 2009), based on a large scale *in silico* screening of localization prediction for all relevant enzymes in more than 50 fungal species, the results showed this pathway mainly take place in the mitochondria and the peroxisome.

To evaluate whether the two aspergilli contains co-evolved pathways, we mapped the 238 conserved co-expressed genes identified in cluster 1 onto the genome-scale metabolic networks of *A. oryzae* (Vongsangnak et al. 2008) and *A. niger* (Andersen et al. 2008a). Once gene-metabolic pathway mapping were performed, the results showed that conserved co-expressed genes are involved in fatty acid catabolism by beta-oxidation. Besides, we also found enzymes/protein functions involved in the glyoxylate bypass and perixisomal protein functions (peroxins) to be conserved. The common pathways and core protein functions for the two *Aspergillus* species are illustrated in Figure 3.

For the genes in cluster 2, the pattern indicates up-regulation of genes in response to growth on xylose, and not surprisingly it is found that many of these genes are associated with the arabinose and xylose metabolism pathway (Table 1) and reporter GO terms like xylan catabolic process (GO:0045493), transaldolase activity (GO:0004801), and D-xylulose reductase activity (GO:0046526). We also performed mapping of the 23 conserved genes onto the metabolic networks. As expected, we found co-evolved pathways which are mainly involved in pentose metabolism, especially the xylose degradation pathway, and nucleotide sugar metabolism (See Supplementary file 5). This result showed good agreement with our previous dedicated study of the conserved response in three *Aspergillus* species to growth on xylose (Andersen et al. 2008b).

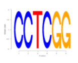
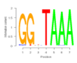
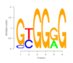
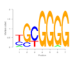
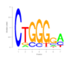
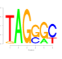
Analysis of DNA regulatory motif and transcription factor underlying co-expression

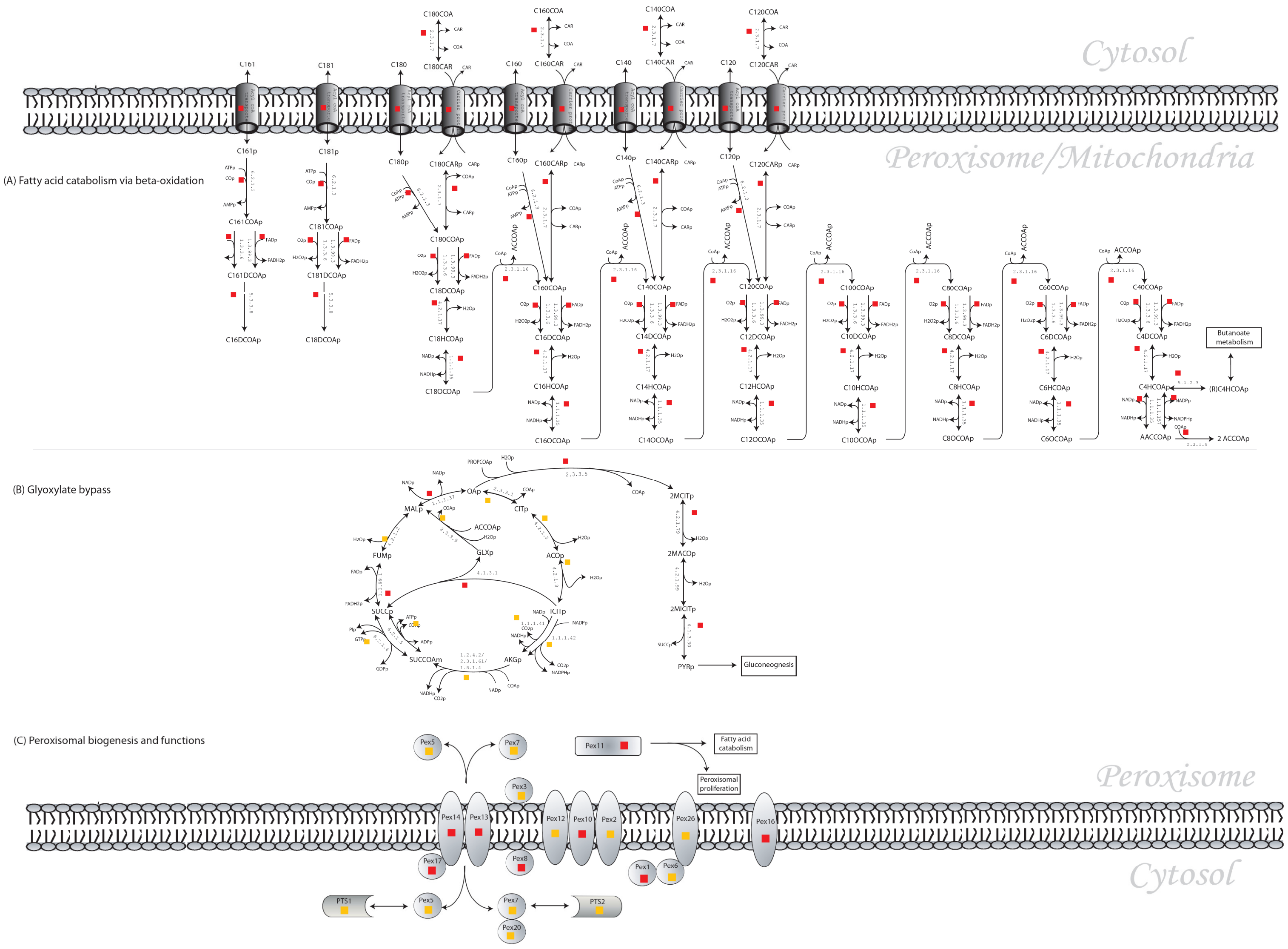
Genes with similar expression profiles will often have their promoter regions bound by common transcription factors at specific motifs and potentially regulated through common regulatory mechanisms. By promoter sequence analysis, we sought potential *cis*-regulatory motifs in the upstream DNA sequences and further searched for the corresponding transcription factors that underlie the transcriptional co-expression patterns. The 1,000 base pairs (bp) of the upstream regions from the start codon of relevant genes in each cluster were analyzed to find the most over-represented common motif (See METHODS for details). For example, for the genes belonging to cluster 1, a 1 kb upstream sequence from the start codon was scanned to identify an over-represented pattern. This was done for all the genes from the *A. oryzae* genome (290 genes) and *A. niger* genome (487 genes). Hereby we identified the motif, “CCTCGG” (reverse complement, CCGAGG) for this cluster. Several other common motifs of other gene in the other co-expression clusters were also found and the corresponding logo plots of all detected motifs are presented in Table 2.

In order to analyze a transcription factor that potentially bind to the identified DNA motifs, we compared the motifs with known or predicted *Aspergillus* consensus motifs from public databases or the literature. We found that four out of six of the over-represented motifs are highly conserved to fungal species which allowed for identification of putative transcription factors as summarized in Table 2. The identified motifs were consistent with known binding site of known transcription factor in aspergilli, such as FarA (Hynes et al. 2006), FarB (Hynes et al. 2006), XlnR (Andersen et al. 2008b), CreA (Chamalaun-Hussey, 1996) and Adr1 (Salazar et al. 2009).

As described above, the core cellular processes with co-evolution and co-regulation found in cluster 1, are likely regulated by the FarA and the FarB proteins that can potential bind to the over-represented motif pattern “CCTCGG”. There are few studies on these proteins that are existing in both the genome of *A. oryzae* and *A. niger*. However, a number of studies have been reported that the Far protein family governs transcriptional activators controlling the utilization of fatty acids in several fungi (Hynes et al. 2006). An elegant study presented by Hynes et al. (Hynes et al. 2006), showed that these transcription proteins FarA and FarB can bind to DNA sequences at 5′ region of a large number of genes involved in fatty acids catabolism related processes in *A. nidulans.* Following their results, they concluded that FarA and FarB mainly induce genes via binding to the 6-bp core sequence “CCTCGG” in the 5’ regions. They also found that genes involved in catabolism of fatty acids have a high enrichment of the core motif on their promoters. From their conclusion, we further performed our analysis of occurrence of “CCTCGG” sequence in the upstream regions of core genes that have transcriptional co-expression and co-evolution in cluster 1, and which have cellular processes related to those reported by Hynes and coworkers (Hynes et al. 2006). The results clearly showed that this pattern was also enriched in genes involved in fatty acid catabolism, glyoxylate bypass and peroxisome biogenesis (See Figure 4). Based on this, we postulate that the core cellular processes are all conserved at the genetic, transcriptional and regulatory level in *A. oryzae* and *A. niger*.

Table 2 List of identified putative DNA regulatory motifs and transcription factors

Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
DNA regulatory motif						
Potential transcription factor	FarA FarB	XlnR	CreA	Adr1	Unknown	Unknown

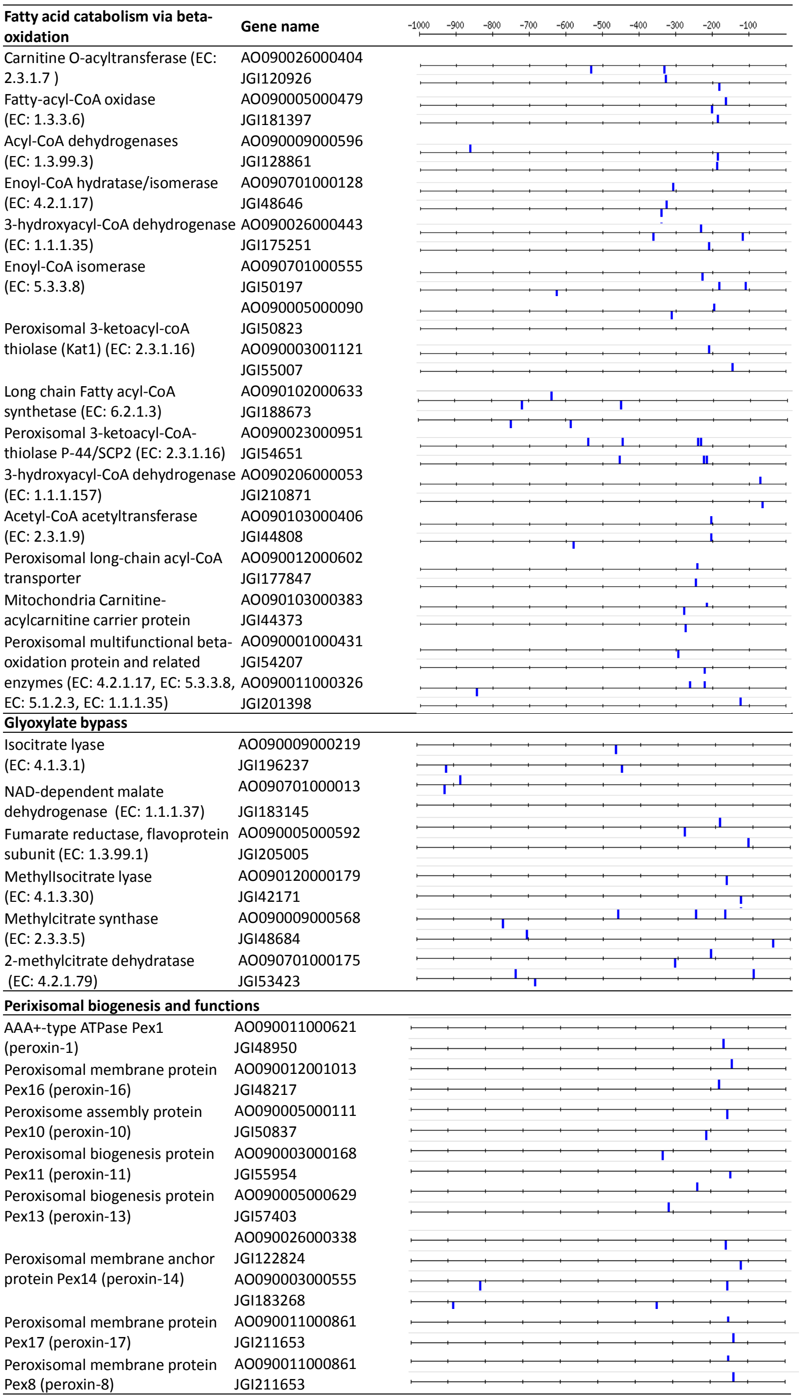


(A) Fatty acid catabolism via beta-oxidation

(B) Glyoxylate bypass

(C) Peroxisomal biogenesis and functions

Figure 3 Co-evolved pathways and functions identified between *A. oryzae* and *A. niger*
(A) Fatty acid catabolism via beta-oxidation, (B) Glyoxylate bypass, (C) Peroxisomal biogenesis and functions. Red boxes refers to conserved orthologous genes with co-expression profiles for cross-species of *A. oryzae* and *A. niger*. Yellow boxes indicates that genes with conserved orthologous for cross-species of *A. oryzae* and *A. niger* but not co-expression profiles.



Fatty acid catabolism via beta-oxidation	Gene name
Carnitine O-acyltransferase (EC: 2.3.1.7)	AO090026000404 JGI120926
Fatty-acyl-CoA oxidase (EC: 1.3.3.6)	AO090005000479 JGI181397
Acyl-CoA dehydrogenases (EC: 1.3.99.3)	AO090009000596 JGI128861
Enoyl-CoA hydratase/isomerase (EC: 4.2.1.17)	AO090701000128 JGI48646
3-hydroxyacyl-CoA dehydrogenase (EC: 1.1.1.35)	AO090026000443 JGI175251
Enoyl-CoA isomerase (EC: 5.3.3.8)	AO090701000555 JGI50197
Peroxisomal 3-ketoacyl-coA thiolase (Kat1) (EC: 2.3.1.16)	AO090005000090 JGI50823 AO090003001121 JGI55007
Long chain Fatty acyl-CoA synthetase (EC: 6.2.1.3)	AO090102000633 JGI188673
Peroxisomal 3-ketoacyl-CoA-thiolase P-44/SCP2 (EC: 2.3.1.16)	AO090023000951 JGI54651
3-hydroxyacyl-CoA dehydrogenase (EC: 1.1.1.157)	AO090206000053 JGI210871
Acetyl-CoA acetyltransferase (EC: 2.3.1.9)	AO090103000406 JGI44808
Peroxisomal long-chain acyl-CoA transporter	AO090012000602 JGI177847
Mitochondria Carnitine-acylcarnitine carrier protein	AO090103000383 JGI44373
Peroxisomal multifunctional beta-oxidation protein and related enzymes (EC: 4.2.1.17, EC: 5.3.3.8, EC: 5.1.2.3, EC: 1.1.1.35)	AO090001000431 JGI54207 AO090011000326 JGI201398
Glyoxylate bypass	
Isocitrate lyase (EC: 4.1.3.1)	AO090009000219 JGI196237
NAD-dependent malate dehydrogenase (EC: 1.1.1.37)	AO090701000013 JGI183145
Fumarate reductase, flavoprotein subunit (EC: 1.3.99.1)	AO090005000592 JGI205005
Methyllsocitrate lyase (EC: 4.1.3.30)	AO090120000179 JGI42171
Methylcitrate synthase (EC: 2.3.3.5)	AO090009000568 JGI48684
2-methylcitrate dehydratase (EC: 4.2.1.79)	AO090701000175 JGI53423
Perixisomal biogenesis and functions	
AAA+-type ATPase Pex1 (peroxin-1)	AO090011000621 JGI48950
Peroxisomal membrane protein Pex16 (peroxin-16)	AO090012001013 JGI48217
Peroxisome assembly protein Pex10 (peroxin-10)	AO090005000111 JGI50837
Peroxisomal biogenesis protein Pex11 (peroxin-11)	AO090003000168 JGI55954
Peroxisomal biogenesis protein Pex13 (peroxin-13)	AO090005000629 JGI57403
Peroxisomal membrane anchor protein Pex14 (peroxin-14)	AO090026000338 JGI122824 AO090003000555 JGI183268
Peroxisomal membrane protein Pex17 (peroxin-17)	AO090011000861 JGI211653
Peroxisomal membrane protein Pex8 (peroxin-8)	AO090011000861 JGI211653

Figure 4 Positions of FarA/FarB binding sites (CCTCGG) motif in the 5'regions of core genes involved in fatty acid catabolism, glyoxylate bypass and peroxisome biogenesis that have transcriptional co-expression in the two species.

For individual gene name, the ORF is prefixed by "AO" for *A. oryzae* RIB40, "JGI" for *A. niger* ATCC1015.

Conserved FarA and FarB transcription factors in aspergilli

Both FarA and FarB proteins are classified as $Zn(II)_2Cys_6$ transcription factors. FarA is required for induction by both short- and long-chain fatty acids, while FarB is likely required only for short-chain fatty acid induction in *A. nidulans* (Hynes et al. 2006). As shown in Figure 5, we found highly conserved protein sequences of FarA and FarB across 8 species, *A. nidulans* (two strains), *A. oryzae* and *A. niger* (two strains), *A. flavus*, *A. clavatus*, *A. terreus*, *A. fischeri* and *A. fumigatus* (two strains). As shown in the phylogenic tree, considering known conserved domains analysis, the FarA protein contains both of Zn_2-Cys_6 binuclear cluster domain (PF00172) and a fungal specific transcription factors domain (PF04082) that are conserved for all 8 species. For the FarB protein, we could not identify a conserved domain of the Zn_2-Cys_6 binuclear cluster (PF00172) in the *A. niger* strain ATCC 1015 and in *A. terreus*, while we found this domain in the other aspergilli. According to the conserved domain analysis, both proteins FarA and FarB have the similar architecture of the two conserved domains, and from these result, we can conclude that these transcription factors are evolutionary conserved among aspergilli. We further evaluated whether the FarA and FarB proteins are conserved among other fungi, and here it is found to be highly homologue to genes identified in *Penicillium* spp, *Fusarium* spp, *Neurospora* spp, *Sclerotinia* spp, *Ajellomyces* spp, *Paracoccidioides* spp, *Coccidioides* spp, *Talaromyces* spp and *Microsporum* spp (See more information in Supplementary file 6), whereas there is no highly conservation to yeast genes.

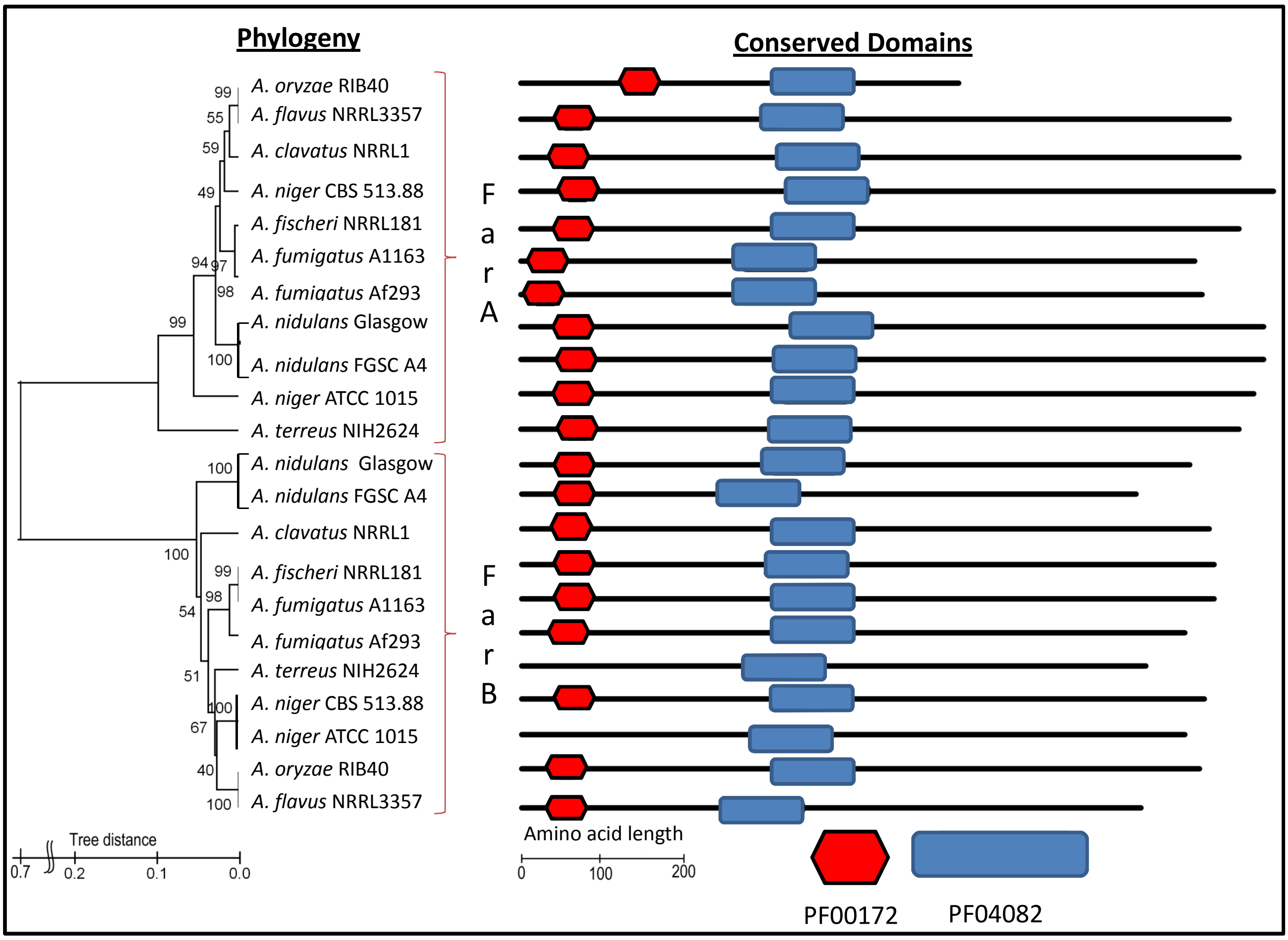


Figure 5 Phylogeny and conserved domains analysis of FarA and FarB proteins in 8 *Aspergillus* species

DISCUSSION

Fatty acid catabolism

Fatty acid catabolism occurs via the beta-oxidation pathway, in which short-chain and long-chain fatty acids are first activated to the corresponding acyl coenzyme A (acyl-CoA) and then oxidized by repeating enzymatic steps to release acetyl-CoA and acyl-CoA shortened by two carbons, which can go through further cycles of beta-oxidation. The acetyl-CoA released is converted to C_4 compounds via the glyoxylate bypass, comprising the enzymes isocitrate lyase (EC: 4.1.3.1) and malate synthase (EC: 2.3.3.9). With cross-species analysis of transcriptional co-expression patterns, we found that isocitrate lyase is a common enzyme between *A. oryzae* and *A. niger*. On the contrary, malate synthase did not have any co-expression, even though it is conserved at the genomics level, i.e. we found orthologus genes of *A. oryzae* (AO090009000557) and of *A. niger* (JGI48680). However, at the transcriptional level the *A. oryzae* gene expression show a highly significant change with glycerol whereas the *A. niger* did not show any significant change. Considering motif identification, we also searched for the 'CCTCGG' binding site in both of the orthologous genes. The results indicated a single motif exists in the 5' upstream sequences of both species as shown in Figure 4. This suggests that the gene-encoding malate synthase of *A. oryzae* and *A. niger* is possibly controlled via the same regulator (FarA/FarB), and the fact that malate synthase was not found to have a significant changed expression in *A. niger* that can be a result of a false negative in the high-throughput analysis. We further identified a wide range of the enzymes that are also subject to induction by glycerol besides the beta-oxidation pathway, i.e. acetyl-carnitine transferase (EC: 2.3.1.7), long-chain acyl-CoA transporter, acyl-CoA synthetase (EC: 6.2.1.3), fatty-acyl-CoA oxidase (EC: 1.3.3.6), enoyl-CoA hydratase (EC: 4.2.1.17), 3-hydroxyacyl-CoA dehydrogenase (EC: 1.1.1.35), and enoyl-CoA isomerase (EC: 5.3.3.8) and carnitine-acylcarnitine carrier protein.

Peroxisomal biogenesis and peroxin

A broad range of peroxisomal proteins (peroxins) in cluster 1 have been identified as being both co-evolved and co-expressed in the two aspergilli. Peroxin-11 (Pex11p) is normally required for peroxisome proliferation (Titorenko & Rachubinski, 2001), and this allows the cell to regulate a number of organelles in the cell, which may increase the efficiency of certain metabolic routes significantly (Kiel & van der Klei, 2009). This peroxin localizes around integral protein components of the peroxisomal membrane. Like found in *A. nidulans*, we found that the *pex11* gene of *A. oryz*ae and *A. niger* were potentially induced by glycerol. This induction depends on the transcription factors FarA and FarB (Hynes et al. 2006; Kiel & van der Klei, 2009), where a FarA/FarB binding site can be identified upstream of *pex11* (Figure 4). In addition, our analysis showed the presence in *A. oryzae* and *A. niger* of various peroxins, such as the receptor docking complex, consisting of the peroxins Pex13p, Pex14p and Pex17p. These proteins are involved in peroxisomal biogenesis and form a protein complex present at the peroxisome membrane. Furthermore, we identified peroxin-10 (pex10p) that contains the RING-finger motif that is a characteristic element of E3-ubiquitin ligase (Kiel & van der Klei, 2009). Another important peroxin found is peroxin-1 (Pex1p), which is AAA+-type ATPase associated with a variety of cellular activities. In addition we found Pex16p, which is an integral membrane protein that may play a role in the formation of the peroxisomal membrane (Kiel & van der Klei, 2009). Through this study, we found the presence of the *Pex* gene family in the two *Aspergillus*'s genomes and this suggests that the peroxins involved in peroxisome biogenesis and proliferation are conserved in these fungi and probably induced by glycerol via FarA and FarB proteins, indicating the importance of peroxisomes in cellular metabolism.

Regulation of transcription factors

Aspergilli are very versatile in their ability to use diverse carbon sources. Not surprisingly, we found several transcription factors controlling carbon metabolism from this study. Our finding indicates that the conserved CCTCGG motif from gene cluster 1 is present in the upstream region of genes involved in lipid metabolism and peroxisome biogenesis of *A. niger* and *A. oryzae*. These indicate that regulation of these genes is governed by the FarA and FarB transcriptional activators. One may speculate why there is regulation of these pathways by glycerol, but this is likely due to the fact that fatty acids are generally available as triacylglycerides and phospholipids that contain glycerol as the backbone. We also found that XlnR is a transcriptional activator that regulates xylanolytic and cellulolytic enzymes (Gielkens et al. 1999; Marui *et al.*, 2002a; Marui et al. 2002b; Tamayo et al. 2008; van Peij et al. 1998a; van Peij et al. 1998b). We found the conserved "GGNTAAA" motif over-represented upstream of genes in cluster 2, and in particular in the upstream region of genes involved in pentose metabolism, the pentose phosphate pathway and nucleotide sugar metabolism of *A. niger* and *A. oryzae* (See Supplementary file 5). We also found creA over-represented as a putative transcription factor for genes in cluster 3. CreA is the major mediator of carbon catabolite repression (Dowzer & Kelly, 1989; Dowzer & Kelly, 1991), its binding was enriched in the genes for utilization of glucose in *A. oryzae* as also found in metabolic pathway enrichment analysis in Table 1. Another transcription factor identified was Adr1. We observed the conserved TGCGGGG motif to be present in the upstream sequence of genes in cluster 4, in particular for genes involved in ethanol utilization, lactate metabolism, amino acid metabolism and fatty acid metabolism.

From this study, one can obtain a better understanding of the complex relationships between co-expression of genes. We found that FarA and FarB are conserved regulators of aspergilli that govern regulation of co-evolved and co-expressed genes related with core biological processes. Our work therefore improved functional annotation and the reconstruction of gene regulatory networks in aspergilli.

ACKNOWLEDGEMENTS

The authors would like to thank the Chalmers Foundation for financial support and the National Council of Research Conacyt-Mexico for financing a stipend to MS.

SUPPLEMENTARY FILES

Supplementary file 1

File format: PDF

Description: List of 5,928 conserved orthologous genes between *A. oryzae* and *A. niger* (See Table S1.1).

Supplementary file 2

File format: PDF

Description: List of reporter GO for *A. oryzae* (See Table S2.1) and *A. niger* (See Table S2.2) under cut-off (p-value < 0.05).

Supplementary file 3
File format: EXCEL
Description: List of gene expression values under cut-off from one-way ANOVA test from four different carbon sources with three biological replicates of *A. oryzae* (See Table S3.1) and *A. niger* (See Table S3.2).

Supplementary file 4
File format: PDF
Description: List of 687 genes distributed into six clusters from consensus clustering (See Table S4.1). The genes are sorted according to ascending cluster number.

Supplementary file 5
File format: PDF
Description: Conserved genes with co-expressions of *A. oryzae* and *A.niger* from gene cluster 2 are mapped in co-evolved pathway (Figure S5.1). Besides, this file also provides the detection of presence of GGNTAAA motif in the 5' upstream regions of the relevant genes in the co-evolved pathways (Table S5.1).

Supplementary file 6
File format: PDF
Description: This file provides statistical details of comparative sequence analysis of FarA (See Table S6.1) and FarB (See Table S6.2) proteins between *A. nidulans* original Glasgow strain (Hynes et al. 2006) and different fungi. Statistical values are presented: E-value and %identity.

REFERENCES

Altman, R. B. & Raychaudhuri, S. (2001). Whole-genome expression analysis: challenges beyond clustering. *Current Opinion in Structural Biology* **11**, 340-347.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403-410.

Andersen, M. R., Nielsen, M. L. & Nielsen, J. (2008a). Metabolic model integration of the bibliome, genome,metabolome and reactome of *Aspergillus niger. Mol Syst Biol* **4**, 178.

Andersen, M. R., Vongsangnak, W., Panagiotou, G., Margarita, P. S., Lehmann, L. & Nielsen, J. (2008b). A tri-species *Aspergillus* microarray - advancing comparative transcriptomics. *Proc Nat Acad Sci USA* **105**, 4387-4392.

Bembom, O., Keles, S. & van der Laan, M. (2007). Supervised detection of conserved motifs in DNA sequences with cosmo. *Stat Appl Genet Mol Biol* **6**, Article 8.

Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Research* **8**, 1202-1215.

Brent, R. (2000). Genomic biology. *Cell* **100**, 169-183.

Chamalaun-Hussey, N. (1996). Characterization of DNA-binding by the CreA protein of *Aspergillus nidulans*. *PhD thesis*, The university of Adelaide, Department of Genetics.

David, H., Hofmann, G., Oliveira, A. P., Jarmer, H. & Nielsen, J. (2006). Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans*. *Genome Biol* **7**.

Dowzer, C. E. & Kelly, J. M. (1989). Cloning of the creA gene from *Aspergillus nidulans*: a gene involved in carbon catabolite repression. *Current Genetics* **15**, 457-459.

Dowzer, C. E. & Kelly, J. M. (1991). Analysis of the creA gene, a regulator of carbon catabolite repression in *Aspergillus nidulans*. *Molecular and Cellular Biology* **11**, 5701-5709.

Dudoit, S., Gendeman, R. C. & Quackenbush, J. (2003). Open source software for the analysis of microarray data. *Biotechniques*, 45-51.

Fedorova, N., Khaldi, N., Joardar, V. & other authors (2008). Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet* **4**.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783-791.

Galagan, J. E., Calvo, S. E., Cuomo, C. & other authors (2005). Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105-1115.

Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315.

Gentleman, R. C., Carey, V. J., Bates, D. M. & other authors (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**.

Gielkens, M. M., Dekkers, E., Visser, J. & de Graaff, L. H. (1999). Two cellobiohydrolase-encoding genes from *Aspergillus niger* require D-xylose and the xylanolytic transcriptional activator XlnR for their expression. *Applied and Environmental Microbiology* **65**, 4340-4345.

Grotkjaer, T., Winther, O., Regenberg, B., Nielsen, J. & Hansen, L. K. (2006). Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. *Bioinformatics* **22**, 58-67.

Hieter, P. & Boguski, M. (1997). Functional genomics: It's all how you read it. *Science* **278**, 601-602.

Hynes, M. J., Murray, S. L., Duncan, A., Khew, G. S. & Davis, M. A. (2006). Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus *Aspergillus nidulans*. *Eukaryotic Cell* **5**, 794-805.

Kiel, J. & van der Klei, I. J. (2009). Proteins involved in microbody biogenesis and degradation in *Aspergillus nidulans*. *Fungal Genetics and Biology* **46**, S62-S71.

Machida, M., Asai, K., Sano, M. & other authors (2005). Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**, 1157-1161.

Marui, J., Kitamoto, N., Kato, M., Kobayashi, T. & Tsukagoshi, N. (2002a). Transcriptional activator, AoXlnR, mediates cellulose-inductive expression of the xylanolytic and cellulolytic genes in *Aspergillus oryzae*. *FEBS Letters* **528**, 279-282.

Marui, J., Tanaka, A., Mimura, S., de Graaff, L. H., Visser, J., Kitamoto, N., Kato, M., Kobayashi, T. & Tsukagoshi, N. (2002b). A transcriptional activator, AoXlnR, controls the expression of genes encoding xylanolytic enzymes in *Aspergillus oryzae*. *Fungal Genetics and Biology* **35**, 157-169.

Nierman, W. C., Pain, A., Anderson, M. J. & other authors (2006). Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **439**, 502-502.

Oldham, M. C., Horvath, S. & Geschwind, D. H. (2006). Conservation and evolution of gene colexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17973-17978.

Salazar, M., Vongsangnak, W., Panagiotou, G., Andersen, M. & Nielsen, J. (2009). Uncovering transcriptional regulation of glycerol metabolism in Aspergilli through genome-wide gene expression data analysis. *Molecular Genetics and Genomics* **In press**.

Schulze, A. & Downward, J. (2001). Navigating gene expression using microarrays - a technology review. *Nature Cell Biology* **3**, E190-E195.

Seo, J. & Hoffman, E. P. (2006). Probe set algorithms: is there a rational best bet? *BMC Bioinformatics* **7: 395**.

Shen, Y. Q. & Burger, G. (2009). Plasticity of a key metabolic pathway in fungi. *Funct Integr Genomics* **9**, 145-151.

Shi, L. M., Perkins, R. G., Fang, H. & Tong, W. D. (2008). Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol* **19**, 10-18.

Sneath, P. & Sokal, R. R. (1973). Numerical taxonomy: the principles and practice of numerical classification. *San Francisco: Freeman*, 573.

Tamayo, E. N., Villanueva, A., Hasper, A. A., de Graaff, L. H., Ramon, D. & M., O. (2008). CreA mediates repression of the regulatory gene xlnR which controls the production of xylanolytic enzymes in *Aspergillus nidulans*. *Fungal Genetics and Biology* **45**, 984-003.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281-285.

Titorenko, V. I. & Rachubinski, R. A. (2001). The life cycle of the peroxisome. *Nat Rev Mol Cell Biol 2:* **2**, 357–368.

van Peij, N. N., Visser, J. & de Graaf, J. H. (1998a). Isolation and analysis of xlnR, encoding a transcriptional activator coordinating xylanolytic expression in *Aspergillus niger*. *Molecular Microbiology* **27**, 131-142.

van Peij, N. N. M. E., Gielkens, M. M. C., de Vries, R. P., Visser, J. & de Graaff, L. H. (1998b). The transcriptional activator XlnR regulates both xylanolytic and endoglucanase gene expression in *Aspergillus niger*. *Applied and Environmental Microbiology* **64**, 3615-3619.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. & Van de Peer, Y. (2009). Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology* **150**, 535-546.

Vongsangnak, W., Olsen, P., Hansen, H., Krogsgaard, S., Nielsen, J. & (2008). Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC Genomics* **9**.

Vongsangnak, W., Salazar, M., Hansen, H. & Nielsen, J. (2009). Genome-wide analysis of maltose utilization and regulation in aspergilli. *Microbiology, special issue in Fungal Physiology* **In press**.

Wolfsberg, T. G., Gabrielian, A. E., Campbell, M. J., Cho, R. J., Spouge, J. L. & Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Research* **9**, 775-792.

Workman, C., Jensen, L. J., Jarmer, H. & other authors (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**.

Young, E. T., Dombek, K. M., Tachibana, C. & Ideker, T. (2003). Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. *J Biol Chem* **278**, 26146-26158.

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins, in Evolving Genes and Proteins. *edited by V Bryson and HJ Vogel Academic Press, New York*, 97-166.

Paper 6

Protein production by *Aspergillus oryzae*

Wanwipa Vongsangnak,
Kim Hansen and Jens Nielsen

Manuscript in preparation

Protein production by *Aspergillus oryzae*

Wanwipa Vongsangnak[1], Kim Hansen[2] and Jens Nielsen[1*]

[1]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

[2]Novozymes A/S, DK-2880 Bagsværd, Denmark

*Correspondence to: Jens Nielsen, Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden, Tel.: +46 (31) 772 38 04, Fax: +46 (31) 772 38 01; e-mail: nielsenj@chalmers.se

INTRODUCTION

Aspergillus oryzae is used for the production of various industrial enzymes and it has a very high natural protein secretion capacity, which enables high level production of many fungal enzymes that find applications in the technical, feed and food industries. In connection with sustainable development of chemicals and biofuel productions, raw materials such as starches, celluloses, and hemicelluloses are widely used in production of e.g. ethanol and lactic acid. In order to degrade starch to glucose, there is a need for efficient enzyme biocatalysts. *A. oryzae* seems to be a producing organism that secretes significant amounts of α-amylases that break polysaccharides into sugars which are further fermented by yeast and lactic acid bacteria. Therefore a demand for α-amylase production is growing at a fast pace. Despite the industrial importance of *A. oryzae* as mentioned, there is relatively little known information about its fundamental process of protein production. Such knowledge is quite important for optimization of an industrial enzyme fermentation processes. For example, information about which genes/pathways are key players for high level protein production.

In this study, the aim was to perform integrative data analysis (i.e. genomes, transcriptomes, metabolic networks, interactomes and fluxes) for diagnosis of industrial enzyme production in fermentation process. Our integrative approach involves comparative transcriptome analysis of a high-producing strain of α-amylase with a reference strain. We compared the *A. oryzae* transformant strain CF1.1 from Novozymes A/S that contains multiple additional gene copies of the α-amylase production strain with the wild type strain A1560. To identify fundamental metabolic process of protein production, we further combined the genome-scale metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008) with the transcriptome data. Moreover, analysis of the global regulatory structure underlying protein synthesis and secretion was also analyzed. We further reconstructed a global interaction network of *A. oryzae* based on identification of putative components through comparative genomics and interactomics (e.g. protein-protein interaction) between *A. oryzae* and *S. cerevisiae*. The reconstructed interaction network was used for identification of key proteins and co-regulated modules in the transcriptional response to high-level protein production. Additionally, flux calculation was performed for analysis of amino acid consumption for protein production. Here we summarize the integration of these multiple data dimensions for diagnosis of the protein production process. We present possible key players/targets (i.e. genes, enzymes, proteins, metabolites and pathways) in response to protein production that may lead us to further improve industrial enzyme fermentation processes.

METHODS

Strains

The two strains used were *A. oryzae* strain A1560 (wild type strain) and multiple additional gene copies of the α-amylase producing strain CF1.1 (transformant strain), both strains obtained from Novozymes (Carlsen & Nielsen, 2001; Pedersen *et al.*, 1999). *A. oryzae* stock culture was maintained on Cove-N-Gly agar at 4°C.

Medium and inoculum compositions

For both strains of *A. oryzae*, spore propagation medium (Cove-N-Gly): 218 g/L sorbitol, 10 g/L glycerol 99.5%, 2.02 g/L KNO_3, 25 g/L agar and 50 ml/L salt solution. Cove-N-Gly salt solution: 26 g/L KCl, 26 g/L $MgSO_4.7H_2O$, 76 g/L KH_2PO_4, 50 ml/L trace element solution. Cove-N-Gly trace elements solution: 40 mg/L $Na_2B_4O_7.10H_2O$, 400 mg/L $CuSO_4.5H_2O$, 1200 mg/L $FeSO_4.7H_2O$, 700 mg/L $MnSO_4.H_2O$, 800 mg/L $Na_2MoO_4.2H_2O$, 10 g/L $ZnSO_4.7H_2O$. Pre-cultures medium (G2-GLY): 18 g/L yeast extract, 24 g/L glycerol (87%), 1 ml/L pluronic PE-6100. Batch cultivation medium: 2.4 g/L $MgSO_4.7H_2O$, 3.6 g/L K_2SO_4, 1.2 g/L citric acid monohydrate, 2.4 g/L KH_2PO_4, 3 g/L $(NH_4)_2HPO_4$, 1.2 g/L pluronic acid (PE-6100) and 0.6 ml/L trace element solution. Trace elements solution: 14.3 g/L $ZnSO_4.7H_2O$, 8.5 g/L $MnSO_4.H_2O$, 13.8 g/L $FeSO_4.7H_2O$, 2.5 g/L $CuSO_4.5H_2O$, 3 g/L citric acid monohydrate as a chelating agent, and 0.5 g/L $NiCl_2.6H_2O$. Carbon sources used were glucose monohydrate or maltose monohydrate (15 g/L). Fermenters were inoculated with ~60 g of broth of *A. oryzae* cultured at 30°C for 24 h on G2-GLY liquid medium in shake flasks at 250 rpm. The pre-cultures were inoculated with 5 ml of spore solution harvested from mycelium grown on Cove-N-Gly agar at 34°C for 3–4 days. Spores were harvested with Tween 80 0.1%.

Batch cultivations

Batch cultivations were done in 2 L bioreactors with a working volume of 1.2 L. The stirrer speed was kept at 800 rpm during the first 4 h and then increased to 1100 rpm. The pH was controlled at 6 by addition of 10% (v/v) of H_3PO_4 or 10% (v/v) NH_3 solution, and the temperature was maintained at 34°C. The aeration flow rate was set at 1.2 vvm (volume of gas per volume of liquid per minute). Dissolved oxygen tension was initially calibrated at 100%. The concentrations of oxygen (O_2) and carbon dioxide (CO_2) in the exhaust gas were monitored with a gas analyzer (Magnos 4G for O2, Uras 3G for CO2, Hartmann & Braun, Germany).

Sampling

Cell dry weight was determined by filtration. A known volume of cell culture was filtered by 70 mm paper filter (Munktell, GlycksBo, Sweden), and then dried at 100°C for 24 h and cooled down in a desiccator. The filter with dried cell mass was weighed afterwards. The culture supernatant was obtained after centrifugation of original samples and subsequently frozen at -20°C for sugars and extracellular metabolites measurements. For gene expression analysis, mycelium was harvested at the early-mid exponential phase and then cultures were filtered. At this point, the mycelium was washed with distilled water. The mycelium was quickly dried by squeezing and subsequently frozen in liquid nitrogen. Samples were stored at -80°C until RNA extraction.

Sugars and extracellular metabolites measurements

The concentration of sugars and extracellular metabolites were measured by HPLC analysis on an Aminex HPX-87H ion-exclusion column (BioRad, Hercules, CA) with previous filtration by 25 mm GD/X syringe filter, 0.45 µm pore size (Whatman, Inc, USA). The column was eluted at 45°C with 5 mM H_2SO_4 at a flow rate of 0.6 ml min^{-1}. Extracellular metabolites were detected with a refractive index detector and an UV detector.

Assay of α-amylase activity

The α-amylase activity was measured by the Fungamyl analysis method according to the cobas fara/PNP method: SOP EAL-SM-0216.02 from Novozymes A/S protocol. The reagents used from Roche Diagnostics GmbH, Mannheim kit 11555685. The analysis of amylase activity was performed by kinetic assay for 10 minutes with OD-measurement every minute at 405 nm in microplate reader (i.e. Thermomax, Molecular Devices). Fungamyl activity was calculated by Softmax® Pro Software on the basis of Vmax rate (mOD/min) by comparison to a linear fungamyl standard curve. The unit of α-amylase activity was expressed in FAU/ml.

Total RNA extraction

Total RNA was extracted by using the Promega RNAgents Total RNA Isolation system, according to the protocol for purification of total RNA from fungi. For RNA extraction, ~1 g of frozen mycelium was ground to a fine powder under liquid nitrogen using a ceramic mortar and pestle. For all samples, the quality of the total RNA extracted was determined by using a BioAnalyzer (2100 BioAnalyzer, Agilent Technologies Inc., Santa Clara, CA, USA) and the quantity determined by using a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Total purified RNA was stored at -80°C until further microarray processing.

Microarray manufacturing and design

Affymetrix arrays were used for the analysis of the transcriptome of *A. oryzae* (Affymetrix company, Santa Clara, CA, USA). The arrays were packaged in an Affymetrix® GeneChip cartridge (49 format), and were processed with GeneChip reagents in the GeneChip® Instrument System. The array contains only perfect match (PM) probes which correspond to 25-base oligonucleotides perfect complementary to the transcript. Of the 13,120 putative genes identified in the genome of *A. oryzae* (Machida *et al.*, 2005; Vongsangnak *et al.*, 2008), 12,039 probe sets were used for microarray analysis. Each of the probe sets were composed of 11 probes (whenever possible) of 25 oligomers (Andersen *et al.*, 2008).

Biotin-labeled cRNA and microarray processing

Biotin-labeled cRNA was prepared from ~ 5 µg of total RNA, according to the protocol described in the Affymetrix GeneChip® Expression Analysis Technical Manual (Affymetrix & GeneChip, 2007). The cRNA was cleaned before fragmentation by using the Qiagen RNeasy Mini Kit (protocol for RNA Cleanup), in order to guarantee the quality of cRNA samples for further processing. Biotin-labeled cRNA was quantified in a spectrophotometer (Amersham Pharmacia Biotech, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). Then, 20 µg of cRNA were fragmented following manufacturer protocol and ~ 15 µg of fragmented cRNA was hybridized to the *Aspergillus* Affymetrix chip (Andersen *et al.*, 2008) according to the Affymetrix GeneChip® Expression Analysis protocol (Affymetrix & GeneChip, 2007). Arrays were washed and stained using a GeneChip® Fluidics Station FS-400, and scanned on an Agilent GeneArray® Scanner.

Microarray data acquisition and analysis

Affymetrix CEL-data files were preprocessed using bioconductor (Gentleman *et al.*, 2004) and R package version 2.9.0 (R Development Core Team). Normalization was performed by using the qspline algorithm (Workman *et al.*, 2002). Normalized gene expression data set is presented in Supplementary file 1. The probe intensities were corrected for background by using the robust multiarray average method (Irizarry *et al.*, 2003) by using all the probes. Gene expression values were calculated from the probes associated with each gene with the medianpolish summary method (Irizarry *et al.*, 2003). All statistical preprocessing methods were invoked through the affy package (Gautier *et al.*, 2004) and R scripts (Dudoit *et al.*, 2003). Statistical analysis was applied to determine significantly different gene expressions. The limma package (Smyth *et al.*, 2005) was used to perform moderated Student's *t* tests. Empirical Bayesian statistics were used to moderate the standard errors within each gene and Benjamini-Hochberg's method to adjust for multi testing (Benjamini & Hochberg, 1995). A cut-off value of adjusted $P<0.05$ was set to assess statistical significance. For pair-wise strains comparison on the individual carbon source, moderated Student's *t* test was done. To study strains background effect (i.e. transformant strain CF1.1 or wild type strain A1560) and to avoid influence of carbon source, the list of genes with significance under cut-off from individual carbon source was overlapped. To study carbon sources effect (i.e. maltose or glucose) on α-amylase production, pairwise carbon sources comparison was also performed on the individual strain.

Reporter metabolites and subnetwork analysis

The reporter metabolites and highly correlated metabolic subnetwork algorithm was applied (Patil & Nielsen, 2005). The pairwise comparison analysis was run for *A. oryzae*, for the transformant strain CF1.1 versus wild type strain A1560. For this purpose, information on the topology of the reconstructed metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008) was used in combination with the adjusted p-values obtained from the Student's t-test analysis.

Pairwise protein sequence comparisons

The complete set of amino acid sequences of the predicted genes from *A. oryzae* (Machida *et al.*, 2005) was used as a sequence database and *S. cerevisiae* was used as a sequence query by applying BLASTP for sequence comparisons (Altschul *et al.*, 1990). An estimated expectation value cut-off of 1E-10, alignment length of

100 amino acids (aa) and percentage of identity of 25 (%) was set to assess statistical significance. The result is a list of orthologous genes (see Supplementary file 2) and represents the scaffold for the reconstruction of protein-protein interaction network.

Reconstruction of protein-protein interaction network

The reconstruction of protein-protein interaction graph was performed based on the following assumption: if there is experimental evidence that proteins a_1 and b_1 in an organism X, interact and proteins a_2 and b_2 of another organism Y are the best homologous to a_1 and b_1, respectively, thus proteins a_2 and b_2 may interact (Jonsson *et al.*, 2006). In this study, *S. cerevisiae* was selected as reference organism (organism X) since it is closely related fungal species to *A. oryzae* (organism Y) and a large data set of protein-protein interactions is available. We retrieved data sets of *S. cerevisiae* (version 2.0.51) from BIOGRID database (http://www.thebiogrid.org/).

Identification of a key protein

Reporter features algorithm (Oliveira *et al.*, 2008) was applied to identify a key protein. Reporter features is a hypothesis-driven method for analysis of transcriptome data. It combines the topology of biological network with the different gene expression data and allows the identification of key biological feature (i.e. key protein) around which transcriptional changes are significantly concentrated. The applied reporter feature algorithm was based on reconstructed protein-protein interaction graph of *A. oryzae* obtained from previous step.

In order to use the algorithm, a pairwise t-test analysis between two strains was performed to obtain statistical values. Then, gene expression data sets were combined with the reconstructed graph to identify key proteins that are significantly affected in response to change of strain backgrounds. Gene expression, in the form of P-value and Z-score was mapped onto the "gene nodes" of the graph. The final score of each key protein can be calculated based on the score of its neighbors "gene nodes". The reporter proteins showed a Z-score and number of neighbor genes (n) under the cut-off (Oliveira et al., 2008) that were applied for biologically meaningful study. In our case, the selected cut-offs of Z-scores were more than 1, and number of neighbor genes (n) were more than 25.

Flux calculations

The specific consumption rate of each of the 20 amino acids was calculated for the two strains in terms of flux (as mmol $gDW^{-1}h^{-1}$). To calculate this flux, stoichiometry of each amino acid consumption for synthesis of total protein for biomass (Vongsangnak *et al.*, 2008) and stoichiometry of each amino acid consumption for synthesis of α-amylase (Pedersen *et al.*, 1999), specific growth rate and specific rate of product formation (α-amylase) were used for calculation.

Result

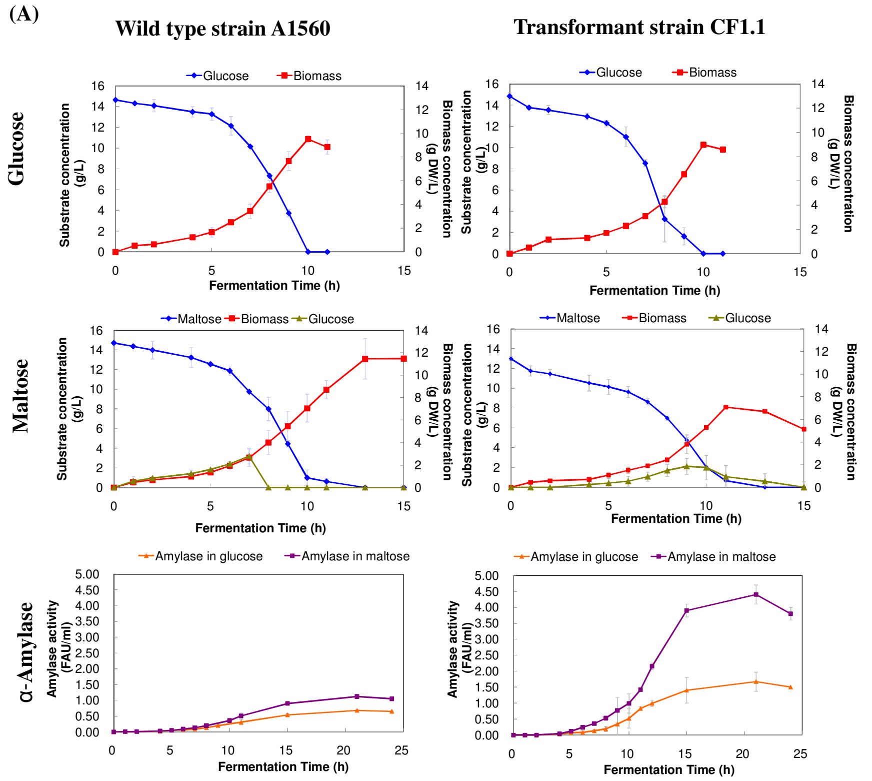
Growth physiology

The growth physiology of the two *A. oryzae* strains was examined in well-controlled bioreactors. Batch cultures were carried out using the same defined salt medium with glucose or maltose as the carbon source. Three biological replicates cultivations were performed for each strain. The results illustrated in Figure 1 (panel A) present profiles of the growth and α-amylase enzyme activities of the two strains (wild type strain A1560 and transformant strain CF1.1) for each carbon source (glucose or maltose). Panel B summarizes the physiology data for the batch cultures. Comparison of the fermentation profiles of the two strains showed that the wild type strain A1560 has a slightly higher specific growth rate than the transformant strain CF1.1, but the maximum activities of extracellular α-amylase enzyme produced were higher for the transformant strain CF1.1 than for the wild type strain A1560. For growth on glucose the increase was approximately 2.3-fold higher for transformant strain CF1.1, whereas for the maltose medium the enzyme production was approximately 4-fold higher for the transformant strain CF1.1 than for the wild type strain A1560 (See Figure 1). These results raised an important question: what are the key players that cause increased level of protein production of transformant strain CF1.1 (compared to wild type strain A1560)? To find out the key players/targets and their functional role, comparative transcriptome analysis was performed as described in the following section.

Comparative transcriptome analysis

To identify the key players for α-amylase production in *A. oryzae*, the genome-wide gene expression data obtained from wild type A1560 and transformant CF1.1 cultivations were pair-wise compared for each of the two carbon sources (either glucose or maltose). To detect transcriptional changes in response to the strain background, Student's t-test statistics was used to identify significantly different gene expression levels with a p-value cut-off of 0.05. To avoid influence of carbon source, the list of conserved genes that respond to strain background for each carbon source was overlapped. Figure 2 shows 2,560 overlapped genes that were significantly differentially expressed in the two *A. oryzae* strains. Among these 2,560 genes, 1,916 (~75%) were up-regulated genes in the transformant strain CF1.1. List of these genes is presented in Supplementary file 3.

Based on gene classification of *A. oryzae* from the DOGAN database, we found that 474 out of the 2,560 genes were involved in the functional category of protein synthesis and secretion. These results are highly reasonable as α-amylase is a protein and therefore not surprisingly the process of synthesis of protein is significantly changed in the transformant strain CF1.1. In addition, we also classified the protein synthesis into sub-functional categories (See Supplementary file 3). Hereby we found many genes involved in RNA processing and translation as well as post-translation modification and secretion processes. A number of genes with relative difference in gene expressions between the transformant CF1.1 and the wild type A1560 were also found (e.g genes encoding protein functions involved in energy metabolism, amino acid metabolism, DNA processing and transcription, and cellular development process). The results are shown in Figure 2. For list of genes with general functions or unknown function, see in Supplementary file 3.



(B)

A. oryzae strain	Carbon sources	Sampling time (h)	µmax (h-1)	Biomass (gDW/L)	Maximum α-Amylase activity (FAU/ml)
A1560	Glucose	6	0.38±0.01	2.50±0.09	0.68±0.21
	Maltose	7	0.33±0.02	2.24±0.09	1.12±0.11
CF1.1	Glucose	6	0.34±0.01	2.35±0.24	1.53±0.25
	Maltose	8	0.30±0.03	2.42±0.12	4.40±0.18

Figure 1 Fermentation profiles of *A. oryzae*

(A) Growth and enzyme activity profiles of two strains (i.e. wild type strain A1560 and transformant strain CF1.1) during growth on glucose and maltose as carbon sources.

(B) Overview of time for sampling for transcriptome analysis, biomass concentration at the sampling time, the maximum specific growth rate, and the maximum enzyme activity for the two strains grown on the two different carbon sources. For all values, average values and standard deviations for the three replicates are shown.

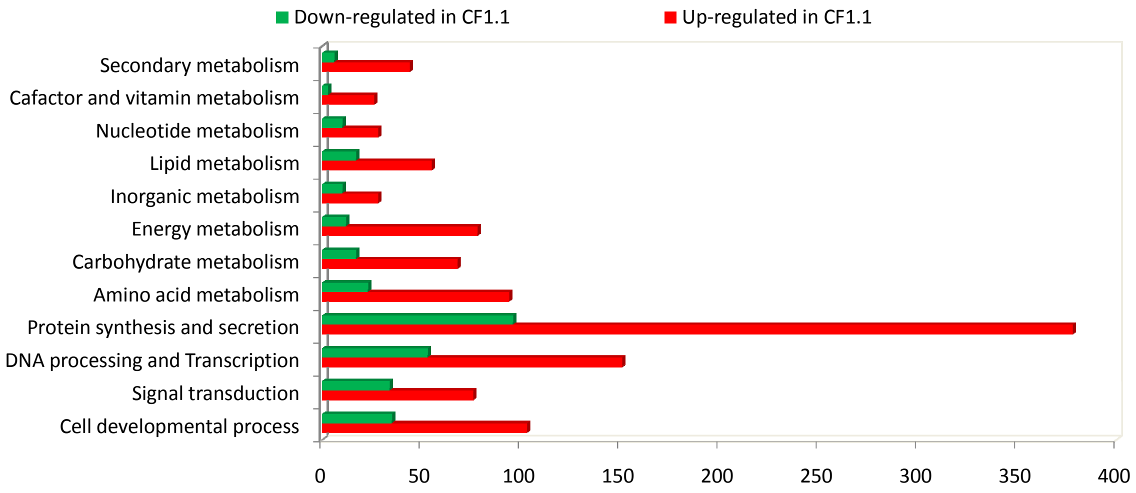


Figure 2 Bar graph represents significantly differentially expressed genes between transformant strain (CF1.1) and wild type strain (A1560), distributed into different functional categories

Key metabolites and metabolic subnetwork analysis in response to protein production

The reporter metabolites and subnetworks algorithm was applied to identify key metabolites involved in protein production and to search for highly correlated metabolic sub-networks for the pairwise comparison (Patil & Nielsen, 2005). This analysis relies on the reconstructed genome-scale metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008). Here, we demonstrated how a metabolic network can be used to msap global regulatory responses for protein production in *A. oryzae.* The top 25 high-scoring key metabolites for *A. oryzae* are listed in Table 1. The fact that tRNA (both cytosol and mitochondria tRNA) was identified as a key metabolites is biologically reasonable, as charged tRNAs are precursors for protein synthesis. Of the 25 metabolites, 20 key metabolites are involved in purine and pyrimidine nucleotide biosynthesis, namely 5-phospho-α-ribosyl-1-pyrophosphate (PRPP), mitochondrial and cytosol pyrophosphate (PPI), inosine monophosphate (IMP), xanthosine monophosphate (XMP), guanosine monophosphate (GMP), mitochondrial and cytosol adenosine monophosphate (AMP), adenosine triphosphate (ATP), deoxy-uridine monophosphate (dUMP), deoxy-guanosine monophosphate (dGMP), guanine, adenine, cytosine, ADP-ribose, 3',5'-cyclic deoxy-adenosine monophosphate (cdAMP), 3',5'-cyclic inosine monophosphate (cIMP), 3',5'-cyclic adenosine monophosphate (cAMP), 3',5'-cyclic guanosine monophosphate (cGMP), and 3',5'-cyclic cytosine monophosphate (cCMP). The results are in agreement with classical molecular biology, where formation of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) is very important in protein production. Besides, we found that ferricytochrome C and ferrocytochrome C are involved in energy metabolism and also nicotinate-D-ribonucleotide is involved in nicotinate and nicotinamide metabolism. In addition to key metabolites, we also identified key enzymes or transporters in response to increased protein production. We performed metabolic sub-network analysis using the whole reaction set from the reconstructed metabolic network of *A. oryzae* (Vongsangnak *et al.*, 2008). Figure 3 captures key genes encoding enzymes in nucleotide metabolism (purine and pyrimidine biosynthesis) and key genes encoding enzymes involved in amino acid metabolism that are significantly changed in the metabolic sub-networks identified from pair-wise strains comparison in *A. oryzae* (wild type strain A1560 vs. transformant strain CF1.1).

Table 1 Reporter metabolites analysis

High α-amylase producer strain (CF1.1) versus reference strain (A1560)	
Key metabolite	**P-value**
Pyrophosphate (PPI)	4.62E-04
Guanosine monophosphate (GMP)	8.81E-04
tRNA	1.02E-03
Pyrophosphate (PPI), mitochondria	2.68E-03
Inosine monophosphate (IMP)	3.98E-03
Adenosine monophosphate (AMP), mitochondria	4.04E-03
Adenosine monophosphate (AMP)	5.02E-03
tRNA, mitochondria	6.85E-03
xanthosine monophosphate (XMP)	9.43E-03
Ferricytochrome C, mitochondria	1.11E-02
Ferrocytochrome C, mitochondria	1.11E-02
Adenosine triphosphate (ATP)	1.22E-02
3',5'-cyclic adenosine monophosphate (cAMP)	1.23E-02
Adenine	1.30E-02
5-phospho-α-ribosyl-1-pyrophosphate (PRPP)	1.52E-02
deoxy-uridine monophosphate (dUMP)	1.63E-02
Nicotinate-D-ribonucleotide	1.67E-02
Cytosine	1.75E-02
deoxy-guanosine monophosphate (dGMP)	2.02E-02
Guanine	2.27E-02
ADP-ribose	2.89E-02
3',5'-cyclic deoxy-adenosine monophosphate (cdAMP)	2.99E-02
3',5'-cyclic inosine monophosphate (cIMP)	2.99E-02
3',5'-cyclic guanosine monophosphate (cGMP)	2.99E-02
3',5'-cyclic cytosine monophosphate (cCMP)	2.99E-02

Reporter metabolite analysis identifies metabolites around which the most significant transcriptional changes occur. The algorithm uses the pairwise t-test analysis referring to strain background effect as an input. The *P*-value gives a measure of significance and all results with $P < 0.03$ are reported.

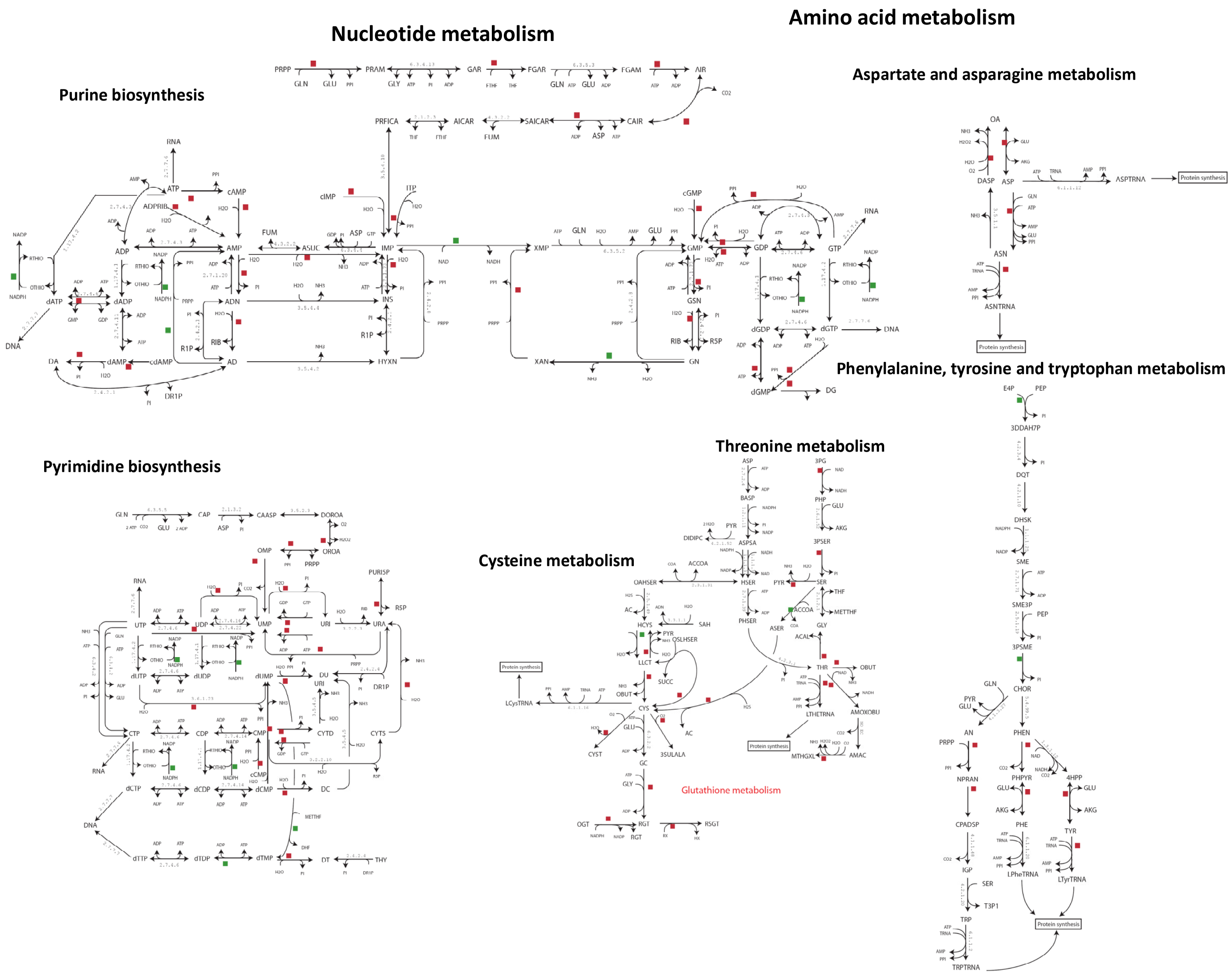


Figure 3 Parts of *A. oryzae* metabolisms in which genes are significantly changed in the metabolic sub-networks identified from pair-wise strain comparison. Red box represents up-regulated genes and green box represents down-regulated genes

Identification of key proteins in respond to protein production

In order to find key proteins regulating the global transcriptional response to the increased protein production level, the reporter features algorithm (Oliveira *et al.*, 2008) was applied. The reporter protein identification is based on reconstructed networks covering each protein-protein interaction combined with gene expression data. In order to apply the algorithm, we first needed to reconstruct a global protein-protein interaction network of *A. oryzae* (See Methods). Each protein pairs from yeast *S. cerevisiae* BIOGRID database was used as a query interaction and searched against *A. oryzae* genes obtained from pairwise protein sequences comparisons (See Methods). Considering only 1:1 orthologous genes of *A. oryzae* and *S. cerevisiae*, we identified 3,514 genes. We searched 140,849 protein pairs of *S. cerevisiae* used as a query interaction database against the 3,514 genes of *A. oryzae* to reconstruct a putative protein-protein interaction network (hypothesis presented by Jonsson et al and as described in Methods. The reconstructed interaction network of *A. oryzae* contains 2,704 individual proteins with 48,483 putative interactions of protein pairs (See Supplementary file 4). In order to identify key proteins, the gene expression data set from comparative transcriptome analysis were combined with the reconstructed network of *A. oryzae*.

Applying the reporter features algorithm with specific thresholds (See Methods), we could identify 33 proteins (see Figure 4) that are possible to be the key targets in gene expression regulation in response to increased protein production. These 33 proteins can be divided into five functional groups, namely 4 proteins involved in transcription, 8 proteins involved in RNA processing and translation, 6 proteins involved in the proteasome, 7 proteins involved in post-translation modification and protein secretion, and 8 proteins involved in cell cycle and structure. In the following, we discuss two interesting cases of how increased enzyme production causes a global response in *A. oryzae* and hereby may impact the overall physiology of the organism.

Analysis of general amino acid control

Not surprisingly, the key proteins are involved in processes of protein synthesis since the pairwise t-test analysis from two strains comparison was used as input for the network analysis. The applied reporter features algorithm can identify regulatory hot-spots in bio-molecular interaction networks that are significantly affected in response to different conditions. An important regulatory hot-spot is the *S. cerevisiae* protein kinase GCN2 or the *Aspergillus* homologue CpcC involved in RNA processing and translation process (See Figure 4). GCN2/CpcC is known as a sensor for amino acid abundance. It usually enhances the sensitivity of translation of the transcription activator GCN4 (as named in *S. cerevisiae*) or CpcA (as named in *Aspergillus* species), leading to transcriptional induction of multiple genes encoding amino acid biosynthetic enzymes upon amino acid starvation. In *S. cerevisiae*, this phenomenon is called “general control of amino acids”, whereas in *Aspergillus* species, this event is named “cross pathway control of amino acid biosynthesis”. Evidently, if the presence of the general control/the cross pathway control of amino acid biosynthesis occurs by the GCN2/CpcC protein kinase, then increased expression of multiple enzymes in different amino acid biosynthetic pathways is found. Since the reporter feature analysis identified GCN2/CpcC as one of the regulatory hot-spots, we hypothesize that the cross pathway control of amino acid biosynthesis in *A. oryzae* is likely to occur in connection with increased protein production.

To test our hypothesis, we used the amino acid biosynthetic enzymes that are known to be under general amino acid control in yeast, fungi and bacteria as a query list and searched against our comparative transcriptome data between the wild type strain A1560 and the transformant strain CF1.1 in order to see if we could find these enzymes. As expected, we found several enzymes subject that are targets for cross pathway control in *A. oryzae*, and this indicates that amino acid starvation is likely to occur in the transformant CF1.1 as most of the genes encoding amino acid enzymes were up-regulated, such as multiple enzymes in tyrosine, tryptophan, ariginine, histidine, lysine, isoleucine, valine, and general aromatic amino acids biosynthesis. A list of the enzymes subject to the cross pathway control is given in Table 2.

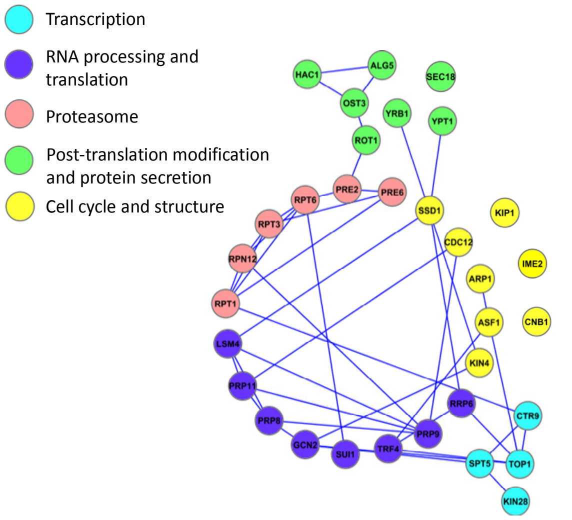


Figure 4 An interaction network of 33 key proteins identified as a global response to increased protein production. The connectivity among the proteins (the nodes) is based on the interactions stored at the BioGRID database of the yeast *S. cerevisiae*. The network was drawn by using Cytoscape (http://www.cytoscape.org).

Analysis of occurrence of Unfolded Protein Response (UPR)

HAC1 was also identified as one of the key proteins in the protein-protein interaction network. This protein is a key regulatory component of the UPR pathway that is activated in response to poor protein folding that leads to block in the protein secretion pathway, which is obviously an important step for protein production. In eukaryotic cells, the synthesized proteins are folded and assembled in the endoplasmic reticulum (ER). The ER provides an oxidising environment in which protein folding is assisted by a number of molecular chaperones and folding enzymes. Protein folding in the ER can be compromised by several endogenous and exogenous factors such as changing environmental conditions or genetic perturbations. This event leads to the accumulation of unfolded proteins within the ER and this lead to ER stress conditions. To maintain homeostasis of ER functions, the cell reacts to the accumulation of unfolded proteins in the ER by inducing a pathway known as the UPR. The UPR pathway has been studied in *A. oryzae* and four key components of this pathway are: (1) The HAC1 protein is a transcription activator that up-regulates the transcription of various target-genes of the UPR pathway; (2) The Bip protein is a chaperone of the HSP70 class that plays an important role in the UPR; (3) The Pdi is a luminal ER enzyme that catalysts the mechanism of disulfide bond formation; (4) The Ppi is an enzyme of catalyzing the cis-trans isomerisation of a peptide bond on the N-terminal side of proline residues in polypeptides.

Since the reporter feature algorithm revealed that HAC1 is an important protein in response to the protein secretion, we reasoned that the UPR pathway is activated in the α-amylase over-producing strain CF1.1. To test our assumption, three UPR-relevant genes known to be controlled by HAC1 in *A. oryzae* were selected, and our transcriptome results showed that the UPR pathway is very likely to be active, since the three target genes in the UPR pathway were up-regulated in the transformant strain CF1.1 (See Supplementary file 3), namely AO090003000257 gene-encoding Bip protein, AO090001000733 gene-encoding Pdi protein, AO090023000811 gene-encoding Ppi protein.

Analysis of amino acid consumptions

We performed comparative analysis of amino acid consumptions in terms of metabolic flux calculation for synthesis of protein content in biomass and α-amylase between the two strains (wild type A1560 and transformant CF1.1). Based on the amino acid composition in biomass protein and α-amylase, the specific consumption rate of each of the 20 amino acids (as mmol $gDW^{-1}h^{-1}$) was calculated for the two strains, the results are summarized in Table 3 (See more results in supplementary file 5). In terms of flux calculation, we found that in particularly four amino acids are drained substantially more in the CF1.1 strain due to the over-production of α-amylase, and the biosynthesis of these amino acids could be possible targets for further increasing the protein production by the transformant strain CF1.1. These four candidate amino acids are tyrosine, aspartate, cysteine and threonine.

DISCUSSION

Integrated data analysis as a scaffold for analysis of industrial enzyme fermentation process

We demonstrated that by performing integrated data analysis, i.e., genomics, transcriptomics, interactomes (protein–protein interactions), metabolic networks and flux analysis, it is possible to identify key regulatory pathways involved in protein production. From comparative transcriptome analysis of two strains (wild type A1560 and transformant CF1.1), we identified that several key processes involved in protein synthesis and secretion are affected at the transcriptional level in response to high-level protein production. We found key metabolites and key enzymes in nucleotide metabolism (purine and pyrimidine biosynthesis) used for synthesis of DNA and RNA (i.e. mRNA, tRNA and rRNA). In addition, we found several amino acid biosynthetic enzymes whose genes are significantly changed in the metabolic sub-networks with respect to protein production, such as tyrosine, aspartate, cysteine and threonine (See Figure 3). The results obtained from flux calculations support that these four amino acids (See Table 3) are playing key role for increased α-amylase production. To find out key regulatory steps, a global interaction network (protein-protein interaction) of *A. oryzae* was reconstructed. The reporter feature algorithm was applied to this network and hereby 33 proteins were identified that are possible key targets regulating gene expression in response to increased level of protein production. 2 proteins out of the 33 proteins, namely GCN2/CpcC and HAC1, suggest that the limiting step for production of α-amylase is control of general amino acids upon starvation and lack of folding capacity in the ER resulting in an UPR. In addition to these 2 proteins, the other key proteins also imply that other steps in protein production are limited, such as transcription, RNA processing and translation, post-translational modification, and proteasome degradation.

In addition to a comparative strain study, we also provided evidence for carbon source induction for α-amylase production. We studied influence of maltose on induction of α-amylase. Our transcriptome analysis of transformant strain CF1.1 showed that maltose is an inducible carbon source. The results from strain CF1.1 showed consistency with our previous publication of strain A1560 (Vongsangnak *et al.*, 2009) that gene function is specific to maltose degradation enzymes and also genes encoding secreted carbohydrases (See more results in Supplementary file 6). For example, there are genes encoding α-amylases, α-glucosidases, maltose permeases, and glucoamylases.

From this study, one can obtain a better understanding of the complex relationship of biological processes in response to high level protein production. Our work therefore revealed the key players/targets (i.e. genes, enzymes, proteins, metabolites and pathways) in response to α-amylase production that may lead us to further improvements of industrial enzyme fermentation processes. We believe that the integrated data analysis can be a scaffold for identifying possible limiting steps for protein production and hereby strategies for strain improvement or process optimization of *A. oryzae* in relation to industrial enzyme production.

Table 2 List of enzymes subject to general amino acid control/cross pathway control targets (P-value<0.05)

Gene name	Enzyme name	Common name	Up/down
Tryptophan biosynthesis			
AO090012000581	Anthranilate synthase (Multifunctional protein)	TRP2	Up
AO090012000581	indole-3-glycerol-phosphate synthase	TRP3	Up
AO090012000581	phosphoribosylanthranilate isomerase	TRP1	Up
AO090003001011	Anthranilate phosphoribosyltransferase	TRP4	Up
AO090005001315	Tryptophan synthase beta chain	TRP5	Up
Arginine biosynthesis			
AO090026000498	Acetylglutamate kinase	ARG2	Up
AO090026000498	Acetylglutamate synthase	ARG6	Up
AO090020000418	Argininosuccinate lyase	ARG4	Up
AO090701000214	Multifunctional pyrimidine synthesis protein CAD (includes carbamoyl-phophate synthetase, aspartate transcarbamylase, and glutamine amidotransferase)	CPA1	Down
Histidine biosynthesis			
AO090206000105	Histidinol-phosphatase	HIS2	Up
AO090012000450	Histidinol phosphate aminotransferase	HIS5	Up
Lysine biosynthesis			
AO090026000245	Transaminases (Aromatic aminotransferases)		Up
AO090001000516	Alpha-aminoadipate reductase and related enzymes	LYS2	Up
Isoleucine and valine biosynthesis			
AO090166000076	Acetolactate synthase, large subunit	ILV2	Up
Leucine biosynthesis			
AO090010000218	Isoleucyl-tRNA synthetase	ILS1	Up
Tyrosine biosynthesis			
AO090003000301	Tyrosine decarboxylase		Up
AO090001000383	Tyrosinase		Up

Table 3 Consumption of amino acids required for synthesis protein contents in biomass and α-amylase for two strains (wild type A1560 and transformant CF1.1)

	% fluxes (mmol gDW^{-1}h^{-1})				
	Wild type A1560		Transformant CF1.1		CF1.1/A1560
Amino acid	Protein contents for biomass	α-amylase	Protein contents for biomass	α-amylase	Ratio
Glycine	12.4362	0.1548	11.1860	0.2838	0.1032
Serine	8.7318	0.1350	7.8540	0.2475	0.1282
Cysteine	1.4553	0.0342	1.3090	0.0627	0.1948
Alanine	12.5685	0.1404	11.3050	0.2574	0.0926
Valine	8.4672	0.1170	7.6160	0.2145	0.1145
Leucine	9.1287	0.1278	8.2110	0.2343	0.1161
Phenylalanine	4.1013	0.0522	3.6890	0.0957	0.1055
Tryptophan	2.3814	0.0378	2.1420	0.0693	0.1316
Tyrosine	3.7044	0.1314	3.3320	0.2409	0.2940
Histidine	2.6460	0.0270	2.3800	0.0495	0.0846
Aspartate	6.0858	0.1548	5.4740	0.2838	0.2109
Methionine	1.8522	0.0342	1.6660	0.0627	0.1531
Isoleucine	5.9535	0.1062	5.3550	0.1947	0.1479
Lysine	7.5411	0.0756	6.7830	0.1386	0.0831
Threonine	6.3504	0.1476	5.7120	0.2706	0.1927
Asparagine	6.0858	0.0972	5.4740	0.1782	0.1324
Arginine	5.8212	0.0378	5.2360	0.0693	0.0538
Glutamate	10.5840	0.0450	9.5200	0.0825	0.0352
Glutamine	10.5840	0.0720	9.5200	0.1320	0.0564
Proline	6.2181	0.0756	5.5930	0.1386	0.1008

ACKNOWLEDGEMENTS

The authors would like to thank Lone Vuholm and Anne Kejser Jensen at Novozymes, Pia Friis and Lene Christiansen at Technical University of Denmark for technical assistance. We would like to thank Sergio Bordel for assisting in flux calculation. Also, we would like to thank Marija Cvijovic for helpful discussion. We would like to thank Chalmers University of Technology Foundation for financial support.

SUPPLEMENTARY FILES

Supplementary file 1

File format: XLS
Description: This file provides the normalized intensities of two strains of *A. oryzae* (transformant strain CF1.1 and wild type strain A1560) on glucose condition (See Table S1.1) or on maltose condition (See Table S1.2) considering all the categories of the three biological replicated experiments.

Supplementary file 2
File format: XLS
Description: This file provides Supplementary table (Table S1) for statistical details of comparative protein sequence analysis between *S. cerevisiae* and *A. oryzae*. Statistical values are presented: E-value, % identity and alignment length.

Supplementary file 3
File format: XLS
Description: List of 2,560 significantly regulated expressed genes between two strains of *A. oryzae* (CF1.1 and A1560) with annotation and statistical values indicated in each gene (adjust P-value and Log$_2$-fold change). The genes are sorted according to functional category.

Supplementary file 4
File format: SIF
Description: A protein-protein interaction data of *A. oryzae*

Supplementary file 5
File format: XLS
Description: Flux calculation of amino acid consumption between wild type strain A1560 and transformant strain CF1.1.

Supplementary file 6
File format: XLS
Description: List of significant genes obtained from comparative transcriptome analysis between maltose and glucose for transformant strain CF1.1

REFERENCES

BIOGRID database. [http://www.thebiogrid.org/].

Abe, K. & Gomi, K. (2007). Food products fermented by *Aspergillus oryzae*. In *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods*, pp. 428-438. Edited by S. A. Osmani & G. H. Goldman. CRC Press.

Affymetrix & GeneChip (2007). Affymetrix Genechip Expression Analysis Technical Manual. *P/N 702232, Affymetrix, Santa Clara, CA, Revision 2*.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403-410.

Andersen, M. R., Vongsangnak, W., Panagiotou, G., Margarita, P. S., Lehmann, L. & Nielsen, J. (2008). A tri-species *Aspergillus* microarray - advancing comparative transcriptomics. *Proc Nat Acad Sci USA* **105**, 4387-4392.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B-Methodol* **57**, 289-300.

Carlsen, M. & Nielsen, J. (2001). Influence of carbon source on alpha-amylase production by *Aspergillus oryzae*. *Appl Microbiol Biotechnol* **57**, 346-349.

Dudoit, S., Gendeman, R. C. & Quackenbush, J. (2003). Open source software for the analysis of microarray data. *Biotechniques*, 45-51.

Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315.

Gentleman, R. C., Carey, V. J., Bates, D. M. & other authors (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**.

Hinnebusch, A. G. (2005). Translational Regulation of GCN4 and the General Amino Acid Control of Yeast. *Annu Rev Microbiol* **59**, 407–450.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Jonsson, P. F., Cavanna, T., Zicha, D. & Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* **7**, 12.

Machida, M., Asai, K., Sano, M. & other authors (2005). Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**, 1157-1161.

Niederberger, P., Miozzari, G. & Hutter, R. (1981). Biological role of the general control of amino acid biosynthesis in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **1**, 584-593.

Oliveira, A. P., Patil, K. R. & Nielsen, J. (2008). Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *Bmc Systems Biology* **2**, 16.

Patil, K. R. & Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci USA* **102**, 2685-2689.

Pedersen, H., Carlsen, M. & Nielsen, J. (1999). Identification of enzymes and quantification of metabolic fluxes in the wild type and in a recombinant *Aspergillus oryzae* strain. *Applied and Environmental Microbiology* **65**, 11-19.

Smyth, G. K., Michaud, J. & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075.

Vongsangnak, W., Olsen, P., Hansen, H., Krogsgaard, S., Nielsen, J. & (2008). Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC Genomics* **9**.

Vongsangnak, W., Salazar, M., Hansen, H. & Nielsen, J. (2009). Genome-wide analysis of maltose utilization and regulation in aspergilli. *Microbiology,* **In press**.

Workman, C., Jensen, L. J., Jarmer, H. & other authors (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**.

Ye, L. & Pan, L. (2008). A Comparison of the Unfolded Protein Response in Solid-State with Submerged Cultures of *Aspergillus oryzae*. *Bioscience Biotechnology and Biochemistry* **72**, 2998-3001.

SYSBIOMICS of Aspergilli:
SYStems Biology, **BI**oinformatics and **OMICS** analysis **of Aspergilli** cell factories

Aspergili are one of the most important fungal species. Three members of the genus *Aspergillus* are widely used as industrial workhorses in production of enzymes and chemicals and as key models for basic scientific work, namely *Aspergillus oryzae*, *Aspergillus niger* and *Aspergillus nidulans*, are particularly interested in this study. Due to their wide applications, it is valuable to gain understanding of their metabolism, regulation and evolution with respect to genotypes and phenotypes, as this may lead to improve industrial fermentation processes for desired product formation (e.g. enzymes). We therefore applied three approaches for this investigation, namely SYStems biology, BIoinformatics and OMICS analysis (SYSBIOMICS).

Firstly, we developed BIoinformatics strategies to improve the genome annotation of *A. oryzae* and this improved annotation was used to reconstruct a high quality genome-scale metabolic network that could be used for mathematical modeling of the physiology and for OMICS data integration, which are the core of SYStems biology. Secondly, we designed a tri-*Aspergillus* DNA microarray chip to monitor the global regulation response at the transcriptional level. This DNA chip has been exploited to reveal conserved regulatory responses through evolution in the three aspergilli in response to changes in carbon source. This resulted in mapping of key regulatory points of metabolism in these fungi, and it showed that SYSBIOMICS analysis of transcriptional data can lead to reconstruction of how carbon metabolism is regulated. Lastly, we also applied the SYSBIOMICS concept to identify possible key players/targets associated with protein production in a high producing strain of *A. oryzae*. This analysis may enable diagnosis and improvement of industrial process of protein production.

In conclusion, through a number of studies it has been demonstrated in this thesis that SYSBIOMICS can find wide applications in industrial biotechnology and assist in improving industrial process required for sustainable production of enzymes and chemicals in the future.

Wanwipa Vongsangnak

Systems Biology
Department of Chemical and Biological Engineering
Chalmers University of Technology
Sweden