A FUNCTIONAL GENOMICS STUDY OF EXTRACELLULAR PROTEASE PRODUCTION BY *ASPERGILLUS NIGER*

Machtelt Braaksma

A functional genomics study of extracellular protease production by *Aspergillus niger*

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op woensdag 15 december 2010

klokke 15:00 uur

door

Machtelt Braaksma

geboren te Stadskanaal

in 1977

PROMOTION COMMITTEE

Promotoren:	Prof. dr. P.J. Punt Prof. dr. C.A.M.J.J. van den Hondel
Co-promoter:	Dr. ir. M.J. van der Werf (DSM)
Other members:	Prof. dr. P.J.J. Hooykaas Prof. dr. J.H. de Winde (Technische Universiteit Delft) Prof. dr. A.K. Smilde (Universiteit van Amsterdam) Dr. M. Saloheimo (VTT Technical Research Centre of Finland)

The work described in this thesis was financially supported by the Kluyver Centre for Genomics of Industrial Fermentation and carried out at TNO Quality of Life.

Printed by: GVO drukkers & vormgevers B.V. | Ponsen & Looijen

CONTENTS

OUTLINE

The filamentous fungus *Aspergillus niger* has a long track record as a highly efficient producer of a wide variety of enzymes. Already soon after the development of fungal transformation systems this species was acknowledged for its potential as a production host for heterologous proteins. However, the production of homologous and especially heterologous proteins is often limited by the high levels of proteases produced by this fungus as well. **Chapter 1** reviews the role that protease activity plays in strain and process development of *A. niger* and other aspergilli. It discusses several approaches and techniques that have been applied to generate strains with reduced protease activity. Furthermore, it provides an outlook on how new research approaches, such as the -omics techniques, may play a role in understanding the proteolytic system of aspergilli.

The objective of the project described in this thesis is to study the complex induction of extracellular proteases in *A. niger* using information gathered with functional genomics technologies. A special emphasis is given to the requirements for performing a successful systems biology study and addressing the challenges met in analyzing the large, information-rich data sets generated with functional genomics technologies.

Chapter 2 of this thesis describes a systematic study of the influence of several environmental factors on the production of extracellular proteases of *A. niger* in controlled batch cultivations. Using a change-one-factor-at-a-time approach, the effect of pH and various medium components on protease production was investigated. Subsequently, a full two-level factorial design was applied with four environmental factors selected from the screening experiments that affected the protease production the most. Six protease-related quantitative phenotypes were calculated from these samples to study the individual and interaction effects of the tested environmental factors on each of these phenotypes. Samples generated in this full factorial experimental design were used for analysis with different functional genomics technologies (Chapter 3 to 5).

Chapter 3 presents an improved list of potential signal peptide directed proteins encoded by the *A. niger* genome. For the compilation of this list, the signal peptide predictions from *A. niger* were compared to those of the best homologs of three neighbouring *Aspergillus* species. In addition, a shotgun proteomics approach was used to determine the *A. niger* secretome under different experimental conditions.

Based on this analysis the complexity of the repertoire of secreted proteases was confirmed.

The effect of different quantitative phenotypes related to product formation on the information content of a metabolomics data set is investigated in **Chapter 4**. For this purpose, besides the production of secreted proteases the production of another industrially relevant product by *A. niger* was evaluated, i.e. the enzyme glucoamylase. For both products, different quantitative phenotypes associated with activity and productivity were defined and for each phenotype the relation with metabolome data was investigated. Results showed that, depending on the product studied, different quantitative phenotypes had the highest information content in relation to the metabolomics data set.

Chapter 5 describes the clustering of co-expressed genes using two DNA microarray data sets; one of these data sets was derived from the experiments described in Chapter 2. A set of conserved genes was used to construct gene co-expression networks for both the individual and combined data sets. By comparative analysis the existence of modules was revealed, some of which are present in all three networks. Subsequently, all protein-coding *A. niger* genes, including hypothetical and poorly conserved genes, were integrated into the co-expression analysis. We have used this two-step approach to relate the genes encoding hypothetical proteins to the identified functional modules.

In top-down systems biology the information gathered with functional genomics technologies is analyzed with multivariate data analysis tools, and can be used as a method to achieve unbiased selection and ranking of targets for both strain improvement and bioprocess optimization. **Chapter 6** discusses the key factors for a successful top-down systems biology approach.

CHAPTER 1

ASPERGILLUS AS A CELL FACTORY FOR PROTEIN PRODUCTION: CONTROLLING PROTEASE ACTIVITY IN FUNGAL PRODUCTION

Machtelt Braaksma and Peter J. Punt

This chapter has been published with minor modifications in:
The *Aspergilli*: Genomics, Medical Aspects, Biotechnology, and Research Methods (2008),
pp. 441-455. Edited by G. H. Goldman & S. A. Osmani. Boca Raton, FL: CRC Press

INTRODUCTION

Since ancient times micro-organisms have been used in a variety of traditional food processes (e.g., the production of alcoholic beverages, cheese, and bread). Fungi are applied in cheese-making and in traditionally oriental food such as soy-sauce, tempeh, and sake. However, the presence and role of these micro-organisms was for most processes only identified in recent times. Fungi, like *Aspergillus oryzae* in the production of sake, were discovered to play a key role in food production by the excretion of enzymes. In 1894, the first microbial enzyme that was commercial produced appeared on the market, called "takadiastase"; it was in fact fungal amylase produced by *A. oryzae* (Gwynne & Devchand, 1992). Nowadays, a large number of fungal enzymes are commercially available and their application extends well beyond their traditional use in food processes. Glucoamylase, α-amylases, cellulase, lipase, and protease are only a few examples of enzymes produced by filamentous fungi that are commercially available. *Aspergillus* species, and particularly *A. niger* and *A. oryzae*, play a dominant role in the production of many of these enzymes (for a list of commercial enzymes see the Association of Manufacturers and Formulators of Enzyme Products (AMFEP)[1]).

For the last two decades, filamentous fungi have also been explored as hosts for the production of heterologous proteins. Because of their established use as production host of homologous proteins aspergilli are the obvious expression system for heterologous proteins. The Danish company Novozymes A/S was in 1988 the first on the market with a non-native fungal lipase (Lipolase) produced from a genetically modified *A. oryzae* strain (Nevalainen & Te'-o, 2003). Since then, several species of *Aspergillus* have been used to express a wide variety of foreign genes (see also the list of commercial enzymes of the AMFEP). However, the production of heterologous as well as homologous proteins is often limited by the high levels of proteases also produced by the fungal host organism. This review will focus on the role of protease activity in strain and process development. Both classical mutagenesis and gene disruption techniques have been applied to generate strains with reduced protease activity. And indeed, production levels improved significantly when using protease deficient strains (e.g., tissue plasminogen activator (t-PA) production with a protease deficient *A. niger* strain [Wiebe *et al.*, 2001]). Controlling the culture conditions can result in a further improvement of the heterologous protein production (e.g., green fluorescent protein (GFP) production with a protease deficient *A. niger* strain at controlled pH [O'Donnell *et al.*, 2001]). However, the production levels for

[1] http://www.amfep.org/list.html; August 24, 2010

heterologous proteins are in most cases one to two orders of magnitude lower than for homologous proteins.

With the availability of the complete genome sequence of several *Aspergillus* strains (e.g., *A. flavus*[2]; *A. fumigatus*[3] [Nierman *et al.*, 2005]; *A. niger*[4]; *A. oryzae*[5] [Machida *et al.*, 2005]; *A. nidulans*[6] [Galagan *et al.*, 2005] and *A. terreus*[6]), homology searches for genes involved in the proteolytic systems of these organisms resulted in a much higher number of genes encoding protease activity than previously known. For example, for *A. niger* approximately 200 genes involved in proteolytic degradation were found in the genome (Pel *et al*., 2007). In comparison, before the genome sequence of *A. niger* was known, an extensive analysis of the proteolytic system of *A. niger* led to the identification of only eight protease genes (van den Hombergh, 1996). Given this very large gene potential, actual protease production and its regulation is expected to be very complicated.

The understanding of the regulation of the proteolytic system of *Aspergillus* strains is still only in its infancy. The involvement of several wide-domain regulatory systems (carbon catabolite repression, nitrogen metabolite repression, pH regulation [van den Hombergh, 1996]) and probably sulfur metabolite repression (VanKuyk *et al.*, 2000) in the overall regulation of protease expression in *Aspergillus* is suggested. This review gives state of the art in the protease research field and provides an outlook on new research approaches.

STRAIN DEVELOPMENT

Classical methods to screen for protease mutants

Mutagenesis by means of X-ray or UV irradiation and chemicals mutagenesis were discovered in the first half of the past century. Hara *et al.* (1992) describe the successful attempts of Iguchi (1955-1956) to isolate a mutant strain producing higher levels of protease compared to the parent strain. After X-ray irradiation a large number of isolates were screened in a laborious and time-consuming effort for a hyperproducing mutant. The screening procedure was greatly improved by the

[2] http://www.aspergillusflavus.org/genomics/; August 25, 2010

[3] http://www.sanger.ac.uk/projects/A_fumigatus/; August 25, 2010

[4] http://genome.jgi-psf.org/Aspni5/Aspni5.home.html; August 24, 2010

[5] http://www.bio.nite.go.jp/dogan/project/view/AO; August 25, 2010

[6] http://www.broadinstitute.org/science/projects/fungal-genome-initiative; August 25, 2010

method developed by Sekine in 1969 which enabled the screening of a large number of isolates (Hara *et al.*, 1992). Around colonies grown on casein-containing medium a halo (clear zone) was formed of which the diameter has a significant correlation with the protease production (see Fig. 1).

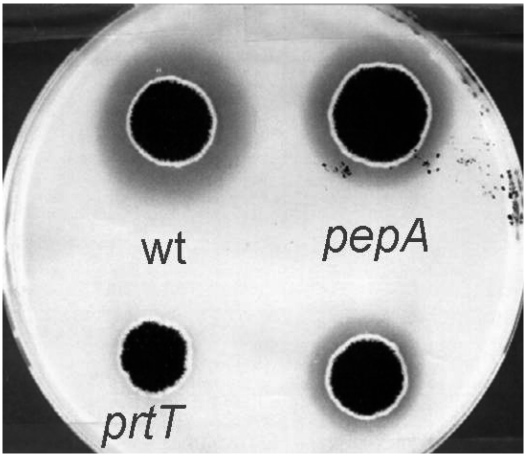


Fig. 1. Protease-deficient mutants (*pepA* and *prtT*) of *Aspergillus niger* show reduced degradation of casein compared to the wild type strain (WT).

These classical methods to generate and screen for mutants with altered levels of excreted protease are still successfully applied. Nowadays, mutagenesis of spores is most often conducted with ultraviolet light irradiation, which is the less-aggressive than irradiation with X-rays. This approach has been applied to isolate several protease-deficient mutants in different aspergilli, such as *A. niger* (Mattern *et al.*, 1992; van den Hombergh *et al.*, 1995) and *A. nidulans* (Katz *et al.*, 1996). Also mutagenesis with mutagens such as nitrosoguanidine has been described (Kolattukudy *et al.*, 1993; Moralejo *et al.*, 2000). After mutagenesis the spores are plated on milk or gelatin-casein medium. Mutants with low proteolytic activity are screened for reduced degradation of casein which results in a reduced or no halo on those plates. In this way, Mattern *et al.* (1992) isolated *A. niger* mutants with residual extracellular proteolytic activities that vary from 2% to 80% of the protease activity of the parental strain. Katz *et al.* (1996) describe *A. nidulans* mutants with tenfold reduced levels of extracellular protease compared to the parental strain.

Molecular genetic methods to construct protease mutants

Protease genes

Clearly, the random mutagenesis approach results in potent production hosts, but the genetic basis of these mutants remains unknown and may have unwanted pleiotropic effects on fungal fermentation performance (e.g., gene expression, growth rate). Therefore, with the development of molecular genetic tools also a more targeted approach to obtain protease-deficient mutants became available.

The general strategy for this approach is the so-called reverse genetics. By separating proteins produced in culture medium by SDS-PAGE or chromatography and subsequently testing for protease activity (as determined, e.g., by protease activity on skim milk agarose) of the different bands or fractions several proteases can be identified. By determining the (partial) amino acids sequence the protein oligonucleotide probes corresponding to these sequences can be designed. These oligonucleotides or PCR fragments generated by using similar oligonucleotides are subsequently used to screen genomic libraries to clone the corresponding protease genes. With the resulting clones a disruption vector for the protease gene can be constructed for actual gene disruption. The more recent availability of genome databases makes it also possible to use obtained amino acid sequences directly to clone the corresponding genes by genome mining using sequence comparison algorithms such as BLASTX. However, even with knowledge of the genome sequence, an activity screen (most preferably based on proteolytic activity against the protein one wants to produce) is still necessary to identify which of all the protease genes present in the fungal genome is actually new and most active and thus the desired target for gene disruption. Berka *et al.* (1990) were the first to describe the construction of gene replacement vectors for *Aspergillus*, which were used to specifically delete the chromosomal DNA of the protease gene encoding the major extracellular acid protease aspergillopepsin A (PEPA) in *A. awamori*. Disruption of this *pepA* gene reduced extracellular proteolytic activity compared to the wild type. Similar results were achieved by disruption of the aspergillopepsin A gene in *A. niger* (Mattern *et al.*, 1992). Probes containing part of the coding region of this *pepA* gene were also used to screen the genomic library of an *A. nidulans* strain (VanKuyk *et al.*, 2000). And although *A. nidulans* appears to lack detectable acid protease activity, a clone which hybridized with the *pepA* gene was obtained. This aspartic protease gene, which was designated *prtB*, was only expressed at a very low level. Furthermore, homologues of the *pepA* gene have been cloned from other *Aspergillus* species, such as *A. fumigatus* (Lee & Kolattukudy, 1995), *A. oryzae* (Berka *et al.*, 1993) and *A. satoi* (Shintani & Ichishima, 1994).

In non-acid producing aspergilli, such as *A. nidulans*, neutral or alkaline proteases are responsible for the major part of the extracellular protease activity. Disruption of the gene coding for the dominant extracellular serine protease in *A. nidulans* strain resulted, when cultured under various medium limitations, in reduced levels of proteolytic activity under all culture conditions (VanKuyk *et al.*, 2000). Controlled batch fermentations of an *A. sojae* strain with a disruption of an alkaline protease gene resulted in about 40% reduction of proteolytic activity in comparison to the wild type (Heerikhuisen *et al.*, 2005). Shake flasks cultures with *A. oryzae* expressing the heterologous protein endoglucanase showed enhanced stability of this protein when an alkaline protease gene of the host strain was disrupted (Lehmbeck, 2001).

Not in all cases disruption of a protease gene results in decreased protease activity. Disruption of the serine protease gene (*sep*) in *A. flavus* led to a compensatory increase in the expression and production of a metalloproteinase gene (*mep20*) (Ramesh & Kolattukudy, 1996). Both wild type and mutant degraded elastin at the same rate. The authors concluded that the expression of the genes encoding both proteases is controlled by a common regulatory system and that the fungus has a mechanism to sense the status of the extracellular proteolytic activities.

An alternative method for reduction of expression of a particular gene is the use of antisense RNA. This approach was applied in an *A. awamori* strain used to express the heterologous protein thaumatin (Moralejo *et al.*, 2002). Even though an insertion in the *pepA* gene had resulted in an inactive PEPA protein, thaumatin was still degraded. Another protease, aspergillopepsin B (PEPB; previously believed to be a pepstatin-insensitive aspartyl protease, but more recently established to be a member of the newly discovered family of glutamic proteases [Fujinaga *et al.*, 2004]), was identified as the most likely protease responsible for this degradation. Expression of *pepB* antisense RNA improved thaumatin production with 30%. Nevertheless, thaumatin was still degraded, indicating the antisense mRNA had only a partial silencing effect on *pepB* gene expression. Disruption of the *pepB* gene resulted in a significant further increase of the thaumatin production. However, an advantage of gene silencing with respect to gene disruption is that it can be used to suppress the expression of complete gene families. Zheng *et al.* (1998) describe that the expression of antisense RNA of the structural gene of carboxipeptidase in *A. oryzae* did not only decrease the activity of that carboypeptidase, but also of two other carboypeptidases.

Yet another approach to obtain strains with low protease levels is disruption of proteases that proteolytically activate other protease precursor proteins which require processing for their activation. Disruption of such a protease gene will have a direct effect on the protease activity of one or more other proteases, as was described for *A. niger*. Disruption of the gene of an intracellular acid protease (PEPE) in *A. niger* did not only reduce the intracellular pepstatin-inhibitable aspartyl protease activity, but also intracellular serine protease and serine carboxypeptidase activities were significantly reduced in the Δ*pepE* strain (van den Hombergh *et al.*, 1997a). The transcription of these non-disrupted genes was not affected by the disruption of the single *pepE* gene. According to the authors this may indicate the presence of a cascade activation mechanism for several vacuolar proteases, triggered by the PEPE protein. A similar mechanism has been described for *Saccharomyces cerevisiae* (van den Hazel *et al.*, 1996).

In Table 1 described disruptions of protease genes in *Aspergillus* strains and the resulting residual proteolytic activities are summarized. In this table the construction of multiple disruptants can lead to further decrease of proteolytic activities. This was shown for a Δ*pepA*Δ*pepB*Δ*pepE* triple disruptant in *A. niger* (van den Hombergh *et al.*, 1997a) and disruption in *A. fumigatus* of both a gene encoding an extracellular serine alkaline protease and a gene encoding an extracellular metalloprotease (Jaton-Ogay *et al.*, 1994).

Protease regulators

Finally, a very efficient approach to generate strains with low protease levels is through disruption of genes that influence the expression of multiple protease genes. Two groups of regulatory genes have been described so far. In the first place, genes that encode specific regulators of protease genes; second, genes that encode wide-domain regulators. Interestingly, in the first group, to date, only one single gene has been identified both in fungi and yeast species. This gene is the *prtT* gene, as cloned from an UV-induced *A. niger* mutant (Punt *et al.*, 2008). This mutant was suggested to be a regulatory mutant as at least two proteases, including PEPA, were missing from the culture medium, while genetic data indicated the presence of a single semi-dominant mutation, not linked to the *pepA* gene (Mattern *et al.*, 1992). Recent analysis has indeed shown that the *prtT* gene is actually a regulatory gene encoding a member of the Zn-binuclear cluster family (Punt *et al.*, 2008). Interestingly, this gene is unique for *Aspergillus* species but actually absent in *A. nidulans*. With the disruption of the *prtT* gene in *A. niger* total protease activity was reduced to 20% of the wild type (Connelly & Brody, 2004).

Table 1. Effects on secreted protease activity of protease gene disruption strains in aspergilli

Species	**Name disrupted gene**	**Residual extracellular protease activity ***	**Reference**
	Extracellular serine protease (fam. S8)		
A. flavus	*sep*	100%	Ramesh *et al.*, 1996
A. fumigatus	*alp*	0-30%	Tang *et al.*, 1992; Monod *et al.*, 1993; Jaton-Ogay *et al.*, 1994
A. nidulans	*prtA*	10-50%	VanKuyk *et al.*, 2000
A. oryzae	*alp*	< WT	Lehmbeck, 2001
A. sojae	*alpA*	60%	Heerikhuisen *et al.*, 2005
	Vacuolar serine protease (fam. S8)		
A. oryzae	*pepC*	N/A	Christensen & Lehmbeck, 2000
	Extracellular aspartyl protease (fam. A1)		
A. awamori	*pepA*	<< WT	Berka *et al.*, 1990
A. fumigatus	*pep*	<< WT	Reichard *et al.*, 1997
A. niger	*pepA*	15-20%	Mattern *et al.*, 1992; van den Hombergh *et al.*, 1997a
	Vacuolar aspartyl protease (fam. A1)		
A. niger	*pepE*	~100%	van den Hombergh *et al.*, 1997a
A. oryzae	*pepE*	N/A	Christensen *et al.*, 2000
	Extracellular glutamic protease (fam. G1)		
A. awamori	*pepB*	< parent †	Moralejo *et al.*, 2002
A. niger	*pepB*	95%	van den Hombergh *et al.*, 1997a
	Extracellular metallo protease (fam. M35)		
A. nidulans	*pepI*	N/A	van den Hombergh & Visser, 1997b
	pepJ	N/A	van den Hombergh *et al.*, 1997b
A. oryzae	*npII*	< WT	Lehmbeck, 1999
	Extracellular metallo protease (fam. M36)		
A. fumigatus	*mep*	70%	Jaton-Ogay *et al.*, 1994
A. niger	*pepH*	< WT	van den Hombergh *et al.*, 1997b
A. oryzae	*npI*	N/A	Lehmbeck, 1999
	Multiple disruptants		
A. fumigatus	*alp, mep*	<< WT	Jaton-Ogay *et al.*, 1994
A. niger	*pepA, pepB*	10%	van den Hombergh *et al.*, 1997a
	pepA, pepE	~ Δ*pepA*	van den Hombergh *et al.*, 1997a
	pepB, pepE	~ Δ*pepB*	van den Hombergh *et al.*, 1997a
	pepA, pepB, pepE	<10%	van den Hombergh *et al.*, 1997a

* As determined with protease assays and expressed as percentage compared to the parent strain; N/A is data not available

† Parent strain is not the WT strain, but a classical *pepA*-deficient mutant

Besides regulatory genes specific for protease expression, wide-domain regulatory genes affect the expression of a broad spectrum of enzymes, including proteases, as a response to ambient pH (*pacC* gene), nitrogen source (*areA* gene) or carbon source (*creA* gene).

The *pacC* gene is expressed at alkaline pH and encodes a protein, which is able to activate the expression of other alkali-expressed genes and to prevent the expression of acid-expressed genes (Peñalva & Arst, Jr., 2002). In *A. nidulans* the expression of the major alkaline protease *prtA* gene is activated by PacC. However, disruption of the *pacC* coding region results in very poor growth, making this approach not very interesting to generate hosts for protein production (Tilburn *et al.*, 1995).

The gene *areA* is expressed in the absence of preferred nitrogen sources such as ammonium and encodes a protein that activates transcription of genes encoding enzymes (like proteases) involved in the utilizing of other resources (Ward *et al.*, 2005). Disruption of the *areA* gene in *A. oryzae* resulted in increased production of the heterologous protein chymosin due to reduced protease activity (Christensen & Hynes, 2000). Unfortunately, disruption of the *areA* gene in *A. niger* as well as *A. oryzae* also affected growth, even in culture medium with (low levels of) ammonium; this reduced growth was not noticed in *A. nidulans* (Christensen *et al.*, 1998; Lenouvel *et al.*, 2001).

The gene *creA* is expressed in the presence of preferred carbon sources such as glucose. The CreA protein represses the synthesis of enzymes (like proteases) involved in the catabolism of alternative carbon sources (Ruijter & Visser, 1997). However, attempts to disrupt the complete *creA* gene from *A. nidulans* resulted in lethal phenotypes (Dowzer & Kelly, 1991) or mutants with extremely severe effects on morphology (namely reduced growth rate and reduced conidiation) (Shroff *et al.*, 1997).

Altogether, the approach of using gene disruption of wide-domain regulatory genes seems unsuitable to generate proteases-deficient fungal host strains for protein production due to pleiotropic growth defects of this type of mutants. Specific mutation of these regulatory genes, alleviating the severe phenotypic effects of the complete knockout mutants could be used (Fraissinet-Tachet *et al.*, 1996). However, this approach relies on selection of specific spontaneous mutants making this approach not generally applicable.

The wide-domain regulatory mechanisms will be discussed in more detail later on in this chapter.

A novel and efficient method for isolation of protease-deficient fungi

Although both the classical screening approach and the gene-based approach have resulted in improved host strains, it is clear that both approaches have their limitations. The classical approach is very labor-intensive, whereas the disruption approach is limited by the availability of gene information. Therefore, we have developed a (direct) mutant selection approach, similar to those available for a number of other traits in filamentous fungi (*pyrG* [van Hartingsveldt *et al.*, 1987], *niaD* [Unkles *et al.*, 1989], *sC* [Buxton *et al.*, 1989]). This proprietary approach is based on a suicide substrate (SUI) to which protease mutants of fungi and yeasts are more resistant (SUI^{R}) than the parent strains (Punt *et al.*, unpublished results). The method can be used to select spontaneous mutants or mutants generated by mutagenesis by ultraviolet light irradiation. After a first round of selection the resulting mutants can be screened in a conventional milk halo screening. As shown in Table 2 the number of colonies resulting in a decreased halo formation is about 10% of the initial SUI^{R} strains even without UV-mutagenesis. In previous studies using milk halo screening after UV-mutagenesis only 0.1% of the surviving spores resulted in a reduced milk halo. With UV-mutagenesis prior to selection with the suicide substrate the efficiency of isolating protease-deficient mutants can be even further increased to over 50% (Punt *et al.*, unpublished result).

In Table 3 the analysis of a number of available and newly selected protease mutant strains is shown. Interestingly, also a mutant with a deficient intracellular protease gene (*pepE*), which results in no significant decrease of extracellular protease activity (van den Hombergh *et al.*, 1997a), can be selected with this method. From Table 3 it is also clear that, as is the case with virtually every method, not every type of protease mutant can be selected in this way. For example, a mutant lacking the major protease gene (*pepA*) in *A. niger*, which results in a residual extracellular protease activity of less than 20% (Mattern *et al.*, 1992; van den Hombergh *et al.*, 1997a), had no higher resistance against the suicide substrate than the wild type strain. Remarkably, with this approach also mutants with enhanced protease activity were selected (Punt *et al.*, unpublished results).

Table 2. Efficiency of isolation of protease-deficient mutants by spontaneous resistance to suicide substrate (SUI) compared to UV mutagenesis

Spontaneous resistance (SUI^R) of two *Aspergillus* species to suicide substrate *				
Strain	**No. of initial spores**	**No. of colonies SUI^R**	**Rescreen SUI^R**	**Reduced milk halo**
Aspergillus sp. section Nigri strain A	4×10^8	590	160/590	45/160
Aspergillus sp. section Nigri strain B	4×10^8	200	85/200	20/85
UV mutagenesis of *A. niger* † and *A. nidulans* ‡				
Strain	**No. of initial spores**	**Survival rate after UV mutagenesis**	**No. of spores screened for reduced milk halo**	**Reduced milk halo**
A. niger	5×10^4-1×10^5	10-20%	1×10^4	$7/1 \times 10^4$
A. nidulans	2.5×10^5-2.5×10^6	1-10%	2.5×10^4	$29/2.5 \times 10^4$

* Punt *et al*., unpublished results

† Mattern *et al*., 1992

‡ Katz *et al*., 1996

Table 3. Protease mutants show higher resistance to the suicide substrate than WT strains *

Species	**SUI (mg/l)**						**Residual protease activity**	
	0	**100**	**200**	**300**	**400**	**500**	**intracellular**	**extracellular**
A. niger WT	+	+	-	-	-	-	100%	100%
A. niger pepA	+	+	-	-	-	-	100% †	15-20% †‡
A. niger pepE	+	+	+	+/-	-	-	30% †	~100% †
A. niger prtT	+	+	+	+	+/-	-	N/A	<5% ‡
A. niger prtT/phmA §	+	+	+	+	+	+/-	N/A	<5% *

* Punt *et al*., unpublished results

† van den Hombergh *et al*., 1997a

‡ Mattern *et al*., 1992

§ The *A. niger prtT/phmA* mutant is a derivative of *A. niger prtT* that does not acidify its medium

FERMENTATION CONDITIONS

Strain improvement has proven to be a very useful tool for reducing the proteolytic degradation of especially heterologous proteins produced in the *Aspergillus* host strain. However, the large number of (extracellular) proteases able to degrade these heterologous proteins and the varying susceptibility of the produced heterologous proteins for the different proteases (Archer *et al.*, 1992; van den Hombergh *et al.*, 1995) makes one single (permanent) solution of the problem impossible. Therefore, an additional way to improve heterologous protein production can be the development of fermentation conditions repressing protease production. Although numerous empirical approaches have been followed to address the protease issue, only very few systematic studies have been performed. From these studies three environmental parameters have emerged which have been studied in somewhat more detail, that is, ambient pH, carbon catabolite control and nitrogen metabolite control.

pH regulation

Ambient pH was shown to be an environmental parameter greatly influencing the expression of proteases. Controlled fermentations with *A. niger* at pH 4 or pH 5 resulted in a significant decrease of protease activity at higher pH. When cultured at pH 6, protease activity was even further decreased (Braaksma *et al.*, 2009). Culture pH was also suggested to be a key player during the production of recombinant GFP by *A. niger* and *A. sojae* (Gordon *et al.*, 2000; Heerikhuisen *et al.*, 2005). GFP excreted by the recombinant *A. niger* strain was rapidly degraded, whereas in *A. sojae* significant amounts of extracellular GFP could be detected. Acidification of the culture medium of *A. niger* was suggested to be the cause for proteolytic degradation of GFP, as under identical conditions *A. sojae* did not significantly acidify. Maintaining the pH at 6 during the production of GFP with *A. niger* resulted in a tenfold increase of GFP levels compared to a culture controlled at pH 3 (O'Donnell *et al.*, 2001). This increase was due to reduced degradation of GFP by proteases. Also, production of the human cytokine interleukin-6 (Il-6) in a protease deficient strain and a derivative of that strain, which did not acidify, resulted in improved yield and stability of Il-6 in the non-acidifying host strain (Punt *et al.*, 2002).

The genes encoding the two major extracellular proteases of *A. niger*, *pepA* and *pepB*, were not expressed under alkaline conditions (Jarai & Buxton, 1994). On the other hand, the transcript levels of the major alkaline protease gene *prtA* produced by *A. nidulans* was elevated under alkaline conditions (Tilburn *et al.*, 1995). This was, however, not confirmed by similar experiments conducted by Katz *et al.* (1996), where nitrogen starvation appeared to override the repression of *prtA* by low culture pH (VanKuyk *et al.*, 2000). From these results we conclude that ambient pH is a regulator of protease expression. In *A. nidulans* pH regulation is mediated mainly by seven genes, *pacC*, *palA*, *palB*, *palC*, *palF*, *palH*, and *palI*, where *pacC* plays the key role in the regulation of gene expression by ambient pH (Tilburn *et al.*, 1995). The products of the pal genes transduce a signal able to trigger the PacC into an active form. This active PacC is able to activate the expression of alkali-expressed genes (including *prtA*) and to inhibit the expression of acid-expressed genes (Peñalva *et al.*, 2002). Homologues of the *pacC* gene and the *pal* genes have been identified in other aspergilli, such as *A. niger* (MacCabe *et al.*, 1996), *A. fumigatus* (Bignell *et al.*, 2005) and *A. oryzae*, as well as all major groups of ascomycetes (Peñalva *et al.*, 2002). The involvement of pH control in extracellular protease production was further confirmed by analysis of protease expression in PacC mutants of *A. nidulans* and *A. niger* (Tilburn *et al.*, 1995; Fraissinet-Tachet *et al.*, 1996). However, the expression of three vacuolar proteases in *A. niger* is not regulated by PacC, which may also be the case with intracellular proteases of other aspergilli (Fraissinet-Tachet *et al.*, 1996).

Carbon catabolite control

Growth on glucose or other favored carbon sources prevents the synthesis of enzymes involved in the utilization of other substrates, such as polysaccharides (Ward *et al.*, 2005). This seems to apply for fungal extracellular proteases as well. Unfortunately, literature about the effect of carbon source on protease production by aspergilli is scattered and in addition often rather outdated. However, a few examples of the repressing effect of glucose and other carbon sources on the levels of excreted proteases have been described. When mycelia from *A. nidulans* were transferred to a medium without carbon source, extracellular proteases were abundantly produced. When mycelia were transferred to medium with glucose, lactose, galactose, or glycerol, protease production was severely repressed (Katz *et al.*, 2000). Similarly, transferring experiments with *A. oryzae* showed a strong decrease of protease production when mycelia were transferred to medium with casein and glucose compared to medium with casein only (Fukushima *et al.*, 1989).

The expression of the two extracellular proteases *pepA* and *pepB* of *A. niger* was studied in the presence of various carbon sources (Jarai & Buxton, 1994). When cells were transferred to medium supplemented with glucose, expression of both protease genes was repressed. In the presence of the less favorable carbon source glycerol the *pepA* gene was derepressed and in medium without carbon source *pepA* and *pepB* were both strongly derepressed. Thus, protease expression is clearly affected by glucose (or carbon catabolite) repression. Repression may be caused by various other carbon sources, but glucose is suspected to be the most repressive. The repressor protein CreA plays a major role in carbon repression. CreA inhibits transcription of many target genes by binding to specific sequences in the promoter of these genes (Ruijter & Visser, 1997). The gene encoding this protein has been identified in several *Aspergillus* species, such as *A. nidulans* (Dowzer & Kelly, 1989), *A. oryzae* (Kim *et al.*, 2001) and *A. niger* (Drysdale *et al.*, 1993). With Northern blot analysis, protease expression in *creA* mutants of *A. niger* gave clear evidence for the involvement of carbon catabolite control (Fraissinet-Tachet *et al.*, 1996). Similarly, this was suggested by the fact that two of the isolated *A. nidulans* mutants, *xprF* and *xprG*, which carry a mutation in a hexokinase-like protein and an acid phosphatase, respectively, are thought to be involved in carbon catabolite repression and maybe also in nitrogen, sulfur, and phosphate regulation (Katz *et al.*, 2000; Katz *et al.*, 2006).

Nitrogen metabolite control

Similar as for the repression by glucose, the presence of preferred nitrogen sources such as ammonium suppress the production of enzymes, such as extracellular proteases, for utilizing other nitrogen sources (Ward *et al.*, 2005). For example, high concentrations of the preferred nitrogen source ammonium resulted in increased concentrations of bioactive tissue t-PA produced by *A. niger*, which was suspected to be due to less degradation of this heterologous protein (Wiebe *et al.*, 2001; Wiebe, 2003). Extracellular protease levels of *A. nidulans* were significant lower in a growth medium with ammonium compared to a nitrogen-free medium (VanKuyk *et al.*, 2000). The influence of nitrogen source on the expression of the *pepA* and *pepB* gene in *A. niger* was investigated by transferring cells to medium with and without ammonium. Cells grown with ammonia showed very low levels of both protease transcripts, whereas the levels of mRNA were much higher when cells were grown without ammonia (Jarai & Buxton, 1994).

The gene *areA* has been implicated in mediating the nitrogen metabolite control regulatory mechanism and it has been extensively studied in *A. nidulans* (Kudla *et al.*, 1990). The *areA* gene encodes a protein that activates transcription of many target genes by binding to specific sequences in the promoter of these genes. Homologues of this gene have also been identified in other *Aspergillus* species, such as *A. oryzae* (Christensen *et al.*, 1998) and *A. niger* (MacCabe *et al.*, 1998).

A study with an *A. niger* wild type strain and several different *areA* mutants (obtained by UV-mutagenesis and selection on chlorate plates) demonstrated that three intracellular protease genes were not controlled by AreA, because both wild type and *areA* mutants showed unaltered expression of these three genes (Fraissinet-Tachet *et al.*, 1996). The same study showed that three extracellular proteases were apparently regulated by AreA. However, the expression of the corresponding extracellular protease genes was not modulated in the same way in the different *areA* mutants, but depended on the combination of the protease gene and the particular *areA* mutation.

Sulfur and phosphorus metabolite repression

Already several decades ago the first studies on the effect of phosphorus and sulfur sources on protease expression in aspergilli were reported, but hardly any articles have been published on this subject since (Tomonaga *et al.* 1964; Cohen, 1972; Cohen, 1973; Cohen, 1981). Today, still little is known about sulfur and phosphorus metabolite repression in aspergilli and putative involvement in protease regulation. However, more recently a strong effect of sulfur limitation on the increase of protease activity for *A. nidulans* has been described (VanKuyk *et al.*, 2000). In addition, expression analyses of *prtA*, encoding the major extracellular protease in *A. nidulans*, showed a high transcript level when mycelia was transferred to sulfur-free medium (Katz *et al.*, 1996; Katz *et al.*, 1994).

Although the regulatory factors involved in sulfur metabolite repression are known (Natorff *et al.*, 1993; Natorff *et al.*, 2003), no information is available regarding protease gene expression. The regulatory factors involved in phosphorus metabolite repression are yet unknown. Identification of the role of these factors may help for a better understanding of the overall protease regulation.

Induction of protease by protein

The fact that in the presence of protein the production of proteases is stimulated has been applied for years in the production of extracellular proteases by the use of complex nitrogen and/or carbon sources (Singh & Vyas, 1977; Fukushima *et al.*, 1989; Srinivasan & Dhar, 1990; Singh *et al.*, 1994).

However, the opposite effect has also been described. Extracellular GFP could not be detected when the *A. niger* host strain was cultured on defined medium (Gordon *et al.*, 2000). When modified soya milk medium was used, fluorescence could be detected in the culture medium. The authors indicate that this was probably not due to a repressive effect of the soya milk protein, but due to the natural protease inhibitors that are present in the soya milk medium and the fact that the ambient pH can be maintained for longer at a value which limits protease induction than with defined medium. Another explanation is that the abundant availability of substrate for the proteases delayed the degradation of GFP.

The *A. niger pepA* and *pepB* protease genes were induced when mycelia was transferred to medium with elastin (Ruijter & Visser, 1997). Medium containing glucose next to elastin repressed expression of both proteases. Comparable experiments by Jarai and Buxton (1994) showed a somewhat different picture, as *A. niger* expressed *pepA* and *pepB* in the presence of glucose if BSA was also present. When additional ammonia or urea was supplemented both protease genes were repressed. These results suggest that induction by the presence of extracellular protein plays only a secondary role in the regulation of extracellular proteases. As for the sulfur and phosphorus regulation mechanisms little is known about the mechanism of specific induction of protease gene expression by external addition of proteins. It is also possible that protein itself is not an inducer, but that the added protein or its peptide degradation products, being a complex carbon and nitrogen source all in one, play a role in the wide-domain regulation mechanisms of nitrogen metabolite and carbon catabolite control.

Bioprocess engineering

Affecting protease production by the means of bioprocess engineering has also proved to be a successful means of controlling extracellular protease activity. However, again very little has been published on the subject. Immobilization of the cells of *A. niger* to materials like a metal-coated pad or Celite beads reduced secretion of extracellular protease and increased the secretion of glucoamylase (Liu *et al.*, 1998; Papagianni *et al.*, 2002). Manipulating the morphology of *A. niger* by means of inoculum levels (concentration of spores) or inoculum type (vegetative or spores) was also shown to affect protease levels (Xu *et al.*, 2000; Papagianni & Moo-Young, 2002). Growth of the mycelium in the form of (large) pellets resulted in lower specific protease activities and increased protein production compared with a filamentous morphology. Morphology clearly affects protease secretion as well as protein production, but the exact mechanism needs further investigation (Grimm *et al.*, 2005).

The effect of the bioprocess parameters agitation intensity, dissolved oxygen tension as well as initial glucose and yeast extract concentration on protease and heterologous protein production has been studied in *A. niger* (Wang *et al.*, 2003). However, altogether these studies should be considered as exploratory, as no systematic analysis was performed.

SYSTEMS BIOLOGY APPROACH

Strain development and optimization of fermentation conditions have improved the production of (heterologous) proteins by aspergilli to a considerable extend. However, the problem of proteases has in most cases been approached by trial-and-error, without taking the interaction between strain development and improvement of fermentation conditions in account (e.g., the best mutant may not be the best producer on the medium previously optimized for a precursor strain). Furthermore, the mechanism of induction and repression of protease production is far from completely understood. A more integrated approach is, therefore, desirable to come to a better understanding of the issue and from this to a solution that is also more generally applicable.

Recently developed techniques like (comparative) genomics, transcriptomics, proteomics, and metabolomics will very likely play a crucial role in understanding the proteolytic system of aspergilli. In addition to these –omics approaches we would also like to consider the role of the various physiological parameters involved in the fermentation process. These "physiomics" parameters such as pH, oxygenation, viscosity, agitation and so on add a further layer of data to be included in a full systems biology approach to study the proteolytic system of aspergilli.

Several articles reporting application of genomics techniques for research of *Aspergillus* strains have been published (e.g., Galagan *et al.*, 2005; Andersen *et al.*, 2008; Coutinho *et al.*, 2009). With the complete genome sequences of several *Aspergillus* strains open to the public (e.g., Machida *et al.*, 2005; Nierman *et al.*, 2005; Galagan *et al.*, 2005, Pel *et al.*, 2007) and more to be expected in the near future (for a recent overview, see Andersen & Nielsen, 2009), possibilities for studying these fungi on a systematic level are open for further research.

Transcriptomics is the most established of the genomics techniques. Several reviews discussing the results from these studies have already appeared (Breakspear & Momany, 2007; Andersen & Nielsen, 2009), illustrating the possibilities of this type of studies to elucidate complex biological processes in fungi.

The method for the identification of all proteins in complex mixtures is proteomic analysis. Initial approaches involved studying the proteins to be separated by one-dimensional (1D) SDS-PAGE. With the development of 2D gel electrophoresis, often coupled to mass spectrometry in order to identify the proteins, proteomic analysis has become a very powerful method for identification of proteins in complex mixtures. A few reviews on proteomics in filamentous fungi have been recently published (Carberry and Doyle, 2007; Kim *et al.*, 2007, Kim *et al.*, 2008).

One of the more recent functional genomics tools is metabolomics, the analysis of all intracellular and extracellular metabolites. Already in the mid-1990s a method to extract intermediary metabolites from *A. niger* has been described by Ruijter and Visser (1996), and glycolytic intermediates were analyzed using an automated spectrophotometer. Since then, analytical platforms for metabolite detection have gone through major developments (van der Werf *et al.*, 2005; Koek *et al.*, 2006; Coulier *et al.*, 2006; Oldiges *et al.*, 2007). However, while based on these methods the potential for large-scale quantitative studies in aspergilli is present, relatively little has been published on metabolomics involving *Aspergillus* species (e.g., Frisvad *et al.*, 2008; KousKoumvekaki *et al.*, 2008).

As is clear from the indicated studies, all functional genomics tools are still under development, with identification of expressed genes or proteins as the major challenge for transcriptomics and proteomics, respectively. However, for all genomics tools extracting relevant biological information from the overwhelming amount of data resulting from these tools is perhaps the biggest challenge. Focusing on the biggest changes in gene expression or protein or metabolite concentration does not automatically lead to the identification of the most important parameter in a biological process (van der Werf, 2004). The choice for a data pretreatment method and a data analysis method greatly affects the outcome (van den Berg *et al.*, 2006). The final goal will be to combine the results distilled from the high-throughput functional genomics methods with information from small-scale studies focusing on particular cellular functions and systems in order to construct a biological network of all protein and genetic interactions. A comprehensive collection of experimentally observed interactions has been put together for the best-studied eukaryote, the budding yeast *S. cerevisiae*, but it is suggested that there are probably many more interactions to be discovered (Reguly *et al.*, 2006). For *Aspergillus*, the study of complex biological networks, among which are also the proteolytic systems, is still in its infancy and will provide the scientific community with a huge challenge on the road to a more complete understanding of this type of organism.

CHAPTER 2

THE EFFECT OF ENVIRONMENTAL CONDITIONS ON EXTRACELLULAR PROTEASE ACTIVITY IN CONTROLLED FERMENTATIONS OF *ASPERGILLUS NIGER*

Machtelt Braaksma, Age K. Smilde, Mariët J. van der Werf and Peter J. Punt

This chapter has been published with minor modifications in:
Microbiology (2009), 155: 3430-3439

Supplementary data are available with the online version of this paper at http://mic.sgmjournals.org/

ABSTRACT

Proteolytic degradation by host proteases is one of the key issues in the application of filamentous fungi for non-fungal protein production. In this study the influence of several environmental factors on the production of extracellular proteases of *Aspergillus niger* was investigated systematically in controlled batch cultivations. Of all factors investigated in a series of initial screening experiments, culture pH and nitrogen concentration in particular strongly affected extracellular protease activities. For instance, at culture pH 4, protease activity was higher than at culture pH 5, and protease activity increased with increasing concentrations of ammonium as nitrogen source. Interestingly, an interdependence was observed for several of the factors studied. These possible interaction effects were investigated further using a full factorial experimental design. Amongst others, the results showed a clear interaction effect between nitrogen source and nitrogen concentration. Based on the observed interactions, the selection of environmental factors to reduce protease activity is not straightforward, as unexpected antagonistic or synergistic effects occur. Furthermore, not only were the effects of the process parameters on maximum protease activity investigated, but for five other protease-related phenotypes were studied as well, such as maximum specific protease activity and maximum protease productivity. There were significant differences in the effect of the environmental parameters on the various protease-related phenotypes. For instance, pH significantly affected final levels of protease activity, but not protease productivity. The results obtained in this study are important for the optimization of *A. niger* for protein production.

INTRODUCTION

Aspergillus species such as *A. niger* and *A. oryzae* are known for their exceptional ability to secrete large amounts of homologous enzymes. For decades they have been commonly exploited as commercial production organisms for a variety of enzymes. With the development of transformation systems for these industrially important members of the genus (Buxton *et al.*, 1985; Kelly & Hynes, 1985; van Hartingsveldt *et al.*, 1987; Iimura *et al.*, 1987; Unkles *et al.*, 1989), the expression of large quantities of heterologous proteins seemed within reach as well. And indeed, nowadays *Aspergillus* species dominate the list of host organisms for the commercial production of enzymes from fungal origin (according to the Association of Manufacturers and Formulators of Enzyme Products[1]). Also, proteins from non-fungal origin, such as chymosin, lysozyme, lactoferrin, interleukin-6 and antibody fragments, have been successfully expressed in several *Aspergillus* species (Yoder & Lembeck, 2004). However, thus far most of these products have only been produced at a laboratory scale, as the production levels, often not more than several tens of milligrams per litre, are too low to be commercially interesting.

The reason for the relatively poor production levels of non-fungal proteins in aspergilli is not completely understood. A combination of inefficient (post)translational steps or proteolytic degradation by extracellular proteases probably affects secreted heterologous protein levels (Yoder & Lembeck, 2004). To date, this latter problem has mainly been approached by disruption or silencing of protease-encoding genes (Berka *et al.*, 1990; van den Hombergh *et al.*, 1997a; Zheng *et al.*, 1998; Moralejo *et al.*, 2002; Braaksma & Punt, 2008) or protease regulator genes (Punt *et al.*, 2008). With this approach, significant reduction of proteolytic activity is achieved with subsequent improvement of heterologous protein production levels (Moralejo *et al.*, 2000; Wang *et al.*, 2008). The recent sequencing of the genomes of several *Aspergillus* species, including *A. niger* (the complete genome sequences of two *A. niger* strains are available, see Pel *et al.* (2007) and the DOE Joint Genome Institute[2]), has created the possibility of identification and disruption of new protease genes (Wang *et al.*, 2008). However, approximately 200 genes involved in proteolytic degradation have been identified in *A. niger* (Pel *et al.*, 2007). Due to this high number of putative proteases, the construction of production hosts essentially free of extracellular protease activities seems unrealistic. It is likely that in such strongly altered strains, other cellular processes will be affected as well, making these multiple protease-deficient mutants unsuitable for robust production conditions.

[1] http://www.amfep.org/list.html; August 24, 2010

[2] http://genome.jgi-psf.org/Aspni5/Aspni5.home.html; August 24, 2010

The role of extracellular proteases in fungi is to degrade proteins into small peptides or amino acids to supply the cell with nutrients when the preferred carbon or nitrogen sources are not available to the cell. Several wide-domain regulatory systems involved in the adaptation of the overall metabolism of nutrients in the cell are implicated in the regulation of extracellular protease expression. Complementary to strain improvement, manipulation of environmental conditions can help to reduce protease secretion and thus improve heterologous protein production, but this has not been investigated systematically (Braaksma & Punt, 2008). Examples of bioprocess parameters which have been investigated for their influence on extracellular protease activity include fungal morphology manipulation (Liu *et al.*, 1998; Xu *et al.*, 2000; Papagianni *et al.*, 2002; Papagianni & Moo-Young, 2002), pH control (O'Donnell *et al.*, 2001), oxygen enrichment and cultivation temperature (Li *et al.*, 2008). Studies on the effect of medium components have mainly focused on the derepression of protease genes when transferring mycelia to medium lacking either carbon or nitrogen source (Cohen, 1981; Jarai & Buxton, 1994). In this study, the effect of pH and various medium components on extracellular protease activity levels in controlled batch cultures with *A. niger* N402 was investigated systematically.

METHODS

Strain and culture media

Aspergillus niger N402 used in this study is a *cspA1* (conferring short conidiophores) derivative of ATCC 9029 (Bos *et al.*, 1988). Stock cultures of this strain were maintained at -80 °C as conidial suspensions in 20% (v/v) glycerol.

Minimal medium (MM) (Bennett & Lasure, 1991) contained 7 mM KCl, 11 mM KH_2PO_4, 2 mM $MgSO_4$, 76 nM $ZnSO_4$, 178 nM H_3BO_3, 25 nM $MnCl_2$, 18 nM $FeSO_4$, 7.1 nM $CoCl_2$, 6.4 nM $CuSO_4$, 6.2 nM Na_2MoO_4 and 134 nM EDTA. This medium was supplemented with the appropriate carbon source or nitrogen source as indicated in Table 1. To prevent foaming, 1% (v/v) antifoam (Struktol J 673) was added to the medium and, when necessary, additional antifoam was added during the cultivation.

Pre-cultivations

For inoculation of the batch cultivations, baffled 500 ml Erlenmeyer flasks were inoculated with 10^6 spores ml^{-1}. The flasks were incubated at 30 °C in a rotary shaker at 125 r.p.m. until approximately half the amount of carbon source was consumed, which took 4-7 days. Each flask contained 100 ml MM (pH 6.5) supplemented with carbon source and nitrogen source, identical to the medium in the batch cultivations.

Batch cultivations

For the screening of the environmental parameters involved in extracellular protease production, as listed in Table 1, cultivation was carried out in 3.3 l BioFlo 3000 bioreactors (New Brunswick Scientific) with a working volume of 2 l. The cultivations of the full two-level factorial design (Table 2) were carried out in 6.6 l BioFlo 3000 bioreactors with a working volume of 5 l. The bioreactors were equipped with two six-blade Rushton turbines and one pitched blade impeller between both Rushton turbines rotating at 400 r.p.m. at the start of the cultivation. When the dissolved oxygen tension dropped below 20%, the agitation was automatically increased to a maximum of 1000 r.p.m., maintaining the dissolved oxygen tension at 20%. Air was used for sparging the bioreactor at a constant flow of 0.25 VVM [vol. gas (vol. liquid)$^{-1}$ min^{-1}]. The pH was controlled at the set value (Table 1 and 2) by automatic addition of 8 M KOH and 1.5 M H_3PO_4, and the temperature was maintained at 30 °C. The controlled batch cultures were inoculated with 4% (v/v) pre-culture.

Cell dry weight determination

For the quantification of cell dry weight (DWT), a known volume of cell culture was filtered though a dried, pre-weighted filter paper, followed by washing with distilled water twice and then drying at 110 °C for 24 h.

Analysis of carbon source concentration

Enzymatic kits were used to analyze glucose (ABX Pentra), sucrose and fructose (Sigma). Lactose concentration was analyzed by incubating 4x-diluted culture samples with an equal volume of 10% (v/v) *β*-galactosidase (Roche) in 0.1 M citrate buffer, pH 6.6, at 37 °C for 10 min, to convert lactose to free glucose and galactose. Glucose concentration was determined as measure for lactose concentration; with correction for free glucose present before incubation with *β*-galactosidase. All these assays were automated on a COBAS MIRA Plus autoanalyzer (Roche Diagnostic Systems). Xylose concentration was measured with the dinitrosalicylic method for quantification of reducing sugars (Sumner & Somers, 1949). The culture sample was 10-50x-diluted and 1 ml of this sample was incubated with 1.5 ml DNS-reagent (1% 3,5-dinitrosalicylic acid, 1.6% NaOH, 30% potassium sodium tartrate) at 100 °C for 5 min, cooled to room temperature and the absorbance was measured at 540 nm.

Analysis of ammonium concentration

Ammonium was assayed by the phenol-hypochlorite colorimetric assay according to Weatherburn (1967). This assay was automated on a COBAS MIRA Plus autoanalyzer.

Preparation of dimethyl BSA for protease assay

N,N-Dimethyl BSA was prepared by a modification of the procedure described by Lin *et al.* (1969). BSA fraction V (20 g) was dissolved in 2 litre 0.1 M borate buffer, pH 9.0, and then cooled to 0 °C. The solution was rapidly stirred, and 4 g of sodium borohydride was added. Formaldehyde (40 ml) was then added in 1.3 ml increments over a period of 30 min. A few minutes after the last addition of formaldehyde, the solution was acidified to pH 6.0 by the addition of 50% acetic acid and dialyzed against deionized water. The desalted protein was lyophilized and stored at -20 °C as a fluffy white powder.

Protease assay

Extracellular proteolytic activities were measured according to a modified procedure as described by Holm (1980) using *N,N*-dimethylated BSA as substrate. The procedure was fully automated using a COBAS MIRA Plus autoanalyzer. Proteolytic activity of cleared culture supernatants was determined by incubating 2 or 8 μl sample with 75 μl 0.5% (w/v) *N,N*-dimethylated BSA in 0.25 M sodium acetate buffer, pH 4.0, for 17.5 min at 37 °C. As a blank, samples were incubated with sodium acetate buffer without *N,N*-dimethylated BSA. The reaction was stopped by the addition of 185 μl 0.1 M borate buffer, pH 9.3, with 0.5 g/l Na_2SO_3. Simultaneously, 5 μl 1x-diluted 2,4,6,-trinitrobenzene sulfonic acid (TNBSA, Pierce) was added. TNBSA reacts with the free amino acid groups, resulting in a yellow colour, which was measured at 405 nm after 3 min. Glycine was used as the standard. One unit of protease activity was defined as that amount of enzyme which in 1 min under the given standard conditions produces a hydrolysate of which the absorption at 405 nm is equal to 1 μmol glycine l^{-1}. Proteolytic activities were determined at pH 6 (0.25 M MES buffer, pH 6.0) and pH 8 (0.25 M MOPS buffer, pH 8.0) as well, but protease activities were very low (results not shown).

Statistical analysis

Before statistical analysis, the curves for DWT and protease activity were corrected for noise and possible outliers by using a smoothing algorithm based on penalized least-squares (Eilers, 2003). The degree of smoothing depended on the value of the penalty (λ) and the derivative that was used. Several combinations of restrictions and derivatives (first- and second-order) were considered in order to find the most appropriate smoothing. For the analysis of the full two-level factorial design experiments, six different phenotypes were defined to express protease activity. With analysis of variance (ANOVA) the effect of an environmental factor or a combination of factors on protease activity levels was evaluated for each of the six protease-related phenotypes. Each ANOVA model contained all main effects (i.e., the effect of either pH, carbon source, nitrogen source or nitrogen concentration on the protease-related phenotype) and the interaction effects of two and three environmental factors. Interaction between all four environmental factors was not included in the models, for this effect was not significant for any of the six individual protease-related phenotypes. An effect was considered significant when the *P*-value was below 0.05. As a measure for the relative contribution of each effect to variation in protease activity, η^2 was calculated as the sum of squares of each effect relative to the total sum of squares. Both smoothing and ANOVA were performed using Matlab Version 7.5.0.342 R2007b (The Mathworks).

RESULTS

Screening of environmental parameters involved in extracellular protease production by *A. niger*

The effect of various environmental factors (Table 1) on extracellular protease production by *A. niger* N402 was investigated in controlled batch cultivations with a change-one-factor-at-a-time approach. Tested variables included carbon source, nitrogen source, nitrogen concentration and pH.

To investigate the effect of carbon sources on production of extracellular protease, six different carbon sources were tested at a culture of pH 4 in a minimal medium containing 70.6 mM sodium nitrate as the nitrogen source. Fig. 1 depicts the concentration profiles of carbon source and biomass as well as protease activity as assayed at pH 4 for a controlled batch cultivation with glucose as carbon source. Protease activity was assayed at pH 6 and pH 8 as well. However, at these pH values hardly any or no proteolytic activity was detected (results not shown), as was also reported by van Noort *et al.* (1991). *A. niger* N402 grew exponentially until the carbon source was completely consumed at approximately 96 h; after this, biomass concentration started to decline. Before glucose depletion, extracellular protease activity had already started to rise and increased rapidly until approximately 18 h after the carbon source in the medium was completely utilized. Near the end of the culture period, the rate of increase in extracellular protease activity decreased. This cultivation with glucose as carbon source was carried out in quadruplicate and used as a reference culture in this study. Physiological parameters for growth and protease activity of this cultivation condition and all other environmental conditions of this screening design are summarized in Table 1.

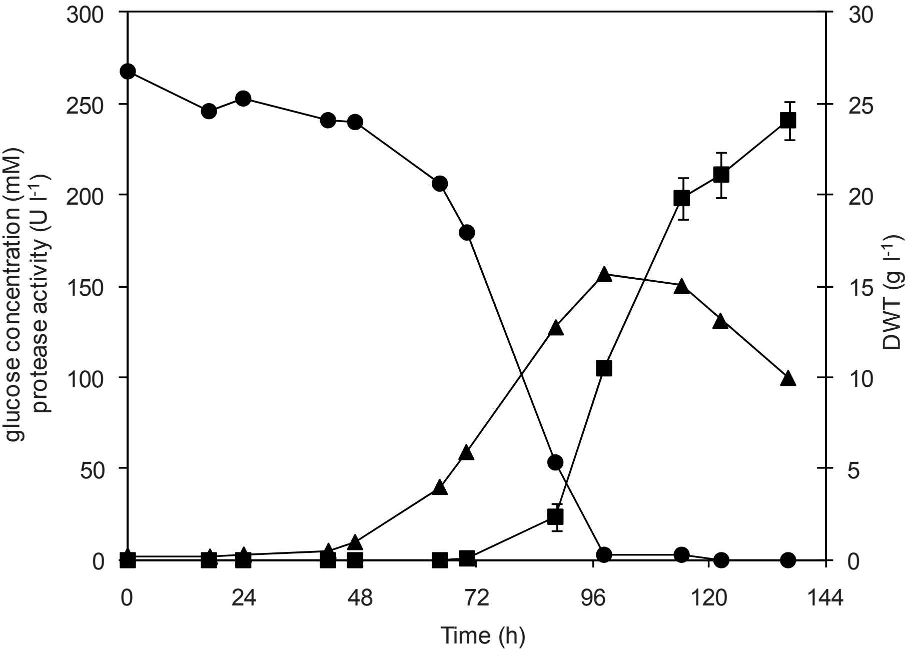


Fig. 1. Profile for extracellular protease production by *A. niger* during controlled batch cultivation with glucose, sodium nitrate and at pH 4. ⍰, Glucose; ■, protease activity assayed at pH 4; ▲, biomass concentration. Protease activity was measured in duplicate and the results are expressed as means±SD.

With xylose as carbon source, profiles for growth and protease secretion (data not shown) showed similar trends as for glucose, although the maximum protease activity and specific protease activity were significantly lower with xylose (Table 1). When sucrose is used as a carbon source, it is first converted into glucose and fructose, after which glucose is consumed prior to fructose. Growth did not stall during the switch from glucose to fructose and was comparable to the cultivations with glucose or xylose. Both maximum protease activity and specific protease activity were considerably lower compared to growth on glucose (Table 1).

Table 1. Fermentation results obtained under the different environmental conditions used in the screening design

Experiment name	Carbon source	Carbon source level (mM) *	pH	Nitrogen source	Nitrogen source level (mM)	Max. biomass DWT_{max} (g l^{-1})	Max. protease activity at pH 4 (U l^{-1})	Specific protease activity (U g^{-1}) †
Glucose/$NaNO_3$ ‡	Glucose	277.5	4	$NaNO_3$	70.6	12.2 (±11.7%)	239 (±15%)	26.3 (±20.7%)
Sucrose	Sucrose	138.8	4	$NaNO_3$	70.6	14.3	171	18.5
Xylose	Xylose	333.0	4	$NaNO_3$	70.6	13.0	177	15.8
Citric acid	Citric acid	277.5	4	$NaNO_3$	70.6	2.2 §	17 §	7.9 §
Lactose	Lactose	138.8	4	$NaNO_3$	70.6	9.6 §	37 §	3.8 §
Proline	Proline	333.0	4	$NaNO_3$	70.6	13.4 §	56 §	4.2 §
Glucose/Proline	Glucose	277.5	4	Proline	70.6	14.4	150	26.6
Glucose/NH_4Cl ‡	Glucose	277.5	4	NH_4Cl	70.6	12.9 (±17.6%)	199 (±16%)	19.4 (±8.1%)
0.5NH_4Cl	Glucose	277.5	4	NH_4Cl	35.3	8.6	77	10.3
2NH_4Cl	Glucose	277.5	4	NH_4Cl	141.2	17.9	258	19.5
4NH_4Cl	Glucose	277.5	4	NH_4Cl	282.4	11.4	369	32.3
8NH_4Cl	Glucose	277.5	4	NH_4Cl	564.8	10.9	469	43.9
4$NaNO_3$	Glucose	277.5	4	$NaNO_3$	282.4	11.4	366	32.2
8$NaNO_3$	Glucose	277.5	4	$NaNO_3$	564.8	13.5	389	28.9
pH 6/$NaNO_3$	Glucose	277.5	6	$NaNO_3$	70.6	1.1	13	12.8
pH 5/$NaNO_3$	Glucose	277.5	5	$NaNO_3$	70.6	4.2	49	25.0
pH 5/NH_4Cl	Glucose	277.5	5	NH_4Cl	70.6	8.4	93	11.1

* The concentration of carbon was equal under all conditions.

† Specific protease activity was calculated as the maximum protease activity (in U l^{-1}) divided by the dry weight concentration (in g l^{-1}) at the time point that the maximum activity was reached.

‡ These experiments were performed in quadruplicate. Results are presented as means (±RSD).

§ With these carbon sources growth was very slow and therefore these cultivations were stopped while biomass concentrations and protease activity were still increasing.

With citric acid, lactose and proline the lag phase was long (4-5 days) and subsequent growth was slow. Therefore, these cultivations were stopped before the stationary phase was reached, suggesting that the carbon source was not yet completely consumed. This was confirmed for lactose, where 40% of the carbon source was not consumed at the time point when the cultivation was stopped. Protease activity appeared much earlier in cultures with less preferred carbon sources (data not shown), at a time point when excess carbon source was still present, whereas with glucose, protease activity increased after glucose depletion.

We further investigated the effects of different nitrogen sources, i.e., proline, ammonium chloride and sodium nitrate, on extracellular protease levels at pH 4 with glucose as carbon source. In addition to cultivation with nitrate, cultivation with ammonium was performed in quadruplicate and used as a reference culture as well. Although DWT_{max} was in the same range for cultures grown with any of the three nitrogen sources, maximum protease activity was especially lower with proline, whereas with ammonium the specific protease activity was lower (Table 1).

Based on our analysis, ammonium was limiting for at least a period of time during the cultivation in medium containing 70.6 mM ammonium chloride (result not shown). To analyze the effect of the ammonium concentration, we tested a variety of initial ammonium chloride concentrations. From this analysis it was clear that ammonium was limiting only at an initial ammonium chloride concentration of 141.2 mM or below (results not shown). Regarding the relation to protease activity, with an increasing concentration of ammonium chloride maximum protease activity increased as well (Table 1).

Based on the results obtained with ammonium chloride, the effect of an increase in the nitrogen concentration by a factor of 4-8 was also studied for sodium nitrate. Nitrate is not expected to be limiting in these cultures, although its concentration was not actually determined. In comparison to the reference culture with sodium nitrate, protease activities increased considerably at elevated nitrate concentrations (Table 1). However, the highest concentration tested did not result in further increase of protease activity, as was the case with ammonium.

The last process parameter tested in the screening design was culture pH. Using a minimal medium containing 277.5 mM glucose and 70.6 mM sodium nitrate, an increase in the culture pH from 4 to 5 resulted in a fivefold decrease of maximum protease activity to 49 U l^{-1}, and DWT_{max} decreased threefold to 4.2 g l^{-1} (see Table 1). The specific protease activity, however, was equal at both culture pH values. At pH 6, maximum protease activity and DWT_{max} were even more severely affected (13 U l^{-1} and 1.1 g l^{-1}, respectively) than at culture pH 5. Notwithstanding the concomitant decrease of biomass formation, the specific protease activity at pH 6 was approximately 50% lower compared to cultures controlled at pH 4. It was striking that biomass formation was so severely affected by an increase in pH, although the rate of carbon consumption was normal. This suggests increased production of carbon dioxide or other carbon metabolites, such as organic acids, at elevated pH.

The effect of the increase in culture pH varied for different nitrogen sources. When pH was increased from pH 4 to 5, maximum protease activity and DWT_{max} were significantly lower when nitrate was used instead of ammonium (see Table 1). At pH 4 the differences between the two nitrogen sources were marginal. This discrepancy in response to pH suggests that, for growth and protease activity, pH and nitrogen source may be interdependent.

To regulate pH at the indicated values, KOH was added to the various cultures. The final concentration of K^+ varied between 288 mM and 820 mM. However, no clear correlation was found between $[K^+]$, pH and protease activity at the time of maximum protease activity.

Analysis of the interaction effects between environmental factors on protease production

In the screening experiments, large variations in both maximum protease activity and specific protease activity were observed for the different culture conditions. There were also indications that the effects of some of the environmental factors were dependent on a combination of factors, e.g., nitrogen source and nitrogen concentration, or nitrogen source and pH. However, the screening approach applied, in which a single factor was changed while keeping all other factors constant, is unsuitable for identifying interactions among environmental factors. A full factorial design, on the other hand, is effective in assessing the contribution of a single environmental factor on the response studied as well as possible interaction effects between these factors (Lundstedt *et al.*, 1998; Kennedy & Krouse, 1999). In this type of design, each factor is considered at two or more levels and the experiments are carried out at each possible combination of these levels. A full two-level factorial design was applied with four environmental factors from the screening experiments. For each factor, two levels were selected (Table 2). The resulting 2^4 full factorial design was performed with eight replicates that were randomly selected.

Table 2. Conditions of the full factorial design used in this study

Experiment name	pH	Carbon source *	Nitrogen source	Nitrogen level †
4 G 4NO_3	4	Glucose	$NaNO_3$	Low
4 G 8NO_3	4	Glucose	$NaNO_3$	High
4 G 4NH_4 ‡	4	Glucose	NH_4Cl	Low
4 G 8NH_4 ‡	4	Glucose	NH_4Cl	High
4 X 4NO_3	4	Xylose	$NaNO_3$	Low
4 X 8NO_3 ‡	4	Xylose	$NaNO_3$	High
4 X 4NH_4 ‡	4	Xylose	NH_4Cl	Low
4 X 8NH_4	4	Xylose	NH_4Cl	High
5 G 4NO_3 ‡	5	Glucose	$NaNO_3$	Low
5 G 8NO_3	5	Glucose	$NaNO_3$	High
5 G 4NH_4	5	Glucose	NH_4Cl	Low
5 G 8NH_4	5	Glucose	NH_4Cl	High
5 X 4NO_3 ‡	5	Xylose	$NaNO_3$	Low
5 X 8NO_3 ‡	5	Xylose	$NaNO_3$	High
5 X 4NH_4 ‡	5	Xylose	NH_4Cl	Low
5 X 8NH_4	5	Xylose	NH_4Cl	High

* Glucose and xylose were used at 277.5 and 333.0 mM, respectively. The concentration of carbon was equal under both conditions.

† $NaNO_3$ and NH_4Cl were used at 282.4 mM (low) or 564.8 mM (high).

‡ These cultivations were performed in duplicate.

As in the screening experiments, the response to the various experimental factors was sometimes different for protease activity and specific protease activity. Therefore, six phenotypes to express protease activity (Table 3) were evaluated in the analysis of the experiments of the full 2^4 factorial design. In addition to maximum protease activity (see A in Fig. 2), the maximum rate of protease production, i.e., the maximum protease productivity (see B in Fig. 2), was also considered. However, for secreted products, the concentration (or activity) and rate of production also depend on the biomass concentration. Therefore, maximum specific protease activity and maximum specific protease productivity were included as well. These two phenotypes can be calculated using the DWT at the time point of maximum protease activity (see A1 in Fig. 2) or maximum protease productivity (see B1 in Fig. 2), respectively. However, these phenotypes were reached while biomass concentration was declining, thus making maximum specific protease activity and maximum specific protease productivity strongly dependent on the degree of lysis. Therefore, both phenotypes were also calculated in relation to the maximal biomass concentration reached, DWT_{max} (see A2 and B2 in Fig. 2).

Table 3. The six protease-related phenotypes evaluated in this study (see also Fig. 2)

Protease-related phenotype	**Description**
Maximum protease activity (max. act.)	Maximum extracellular protease activity measured during cultivation
Maximum specific protease activity – 1 (max. spec. act. – 1)	Maximum activity divided by the DWT at the time point that the maximum activity was reached
Maximum specific protease activity – 2 (max. spec. act. – 2)	Maximum activity divided by DWT_{max}
Maximum protease productivity (max. prod.)	Maximum increase of extracellular protease activity per unit of time
Maximum specific protease productivity – 1 (max. spec. prod. – 1)	Maximum productivity divided by the DWT at the time point that the maximum productivity was reached
Maximum specific protease productivity – 2 (max. spec. prod. – 2)	Maximum productivity divided by DWT_{max}

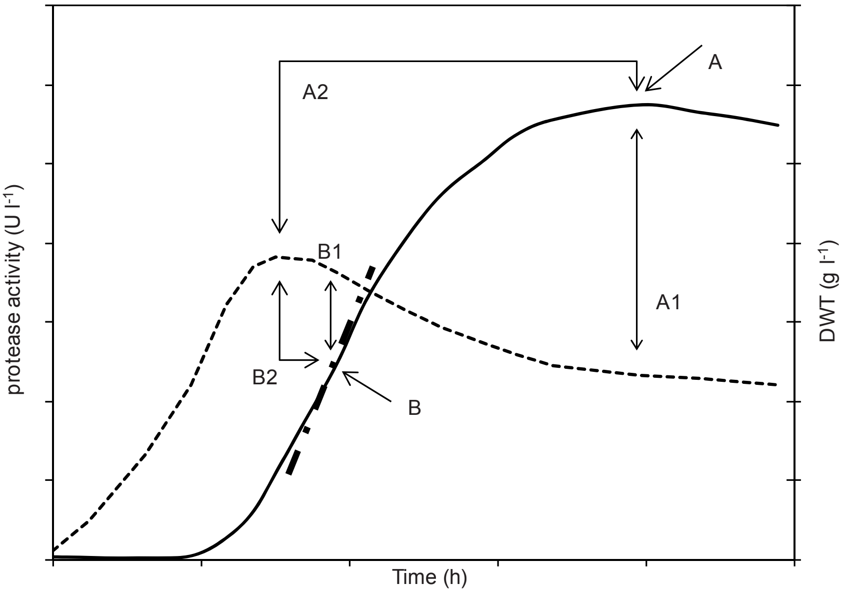


Fig. 2. Schematic representation of extracellular protease production by *A. niger* during controlled batch culture to explain the various protease-related phenotypes as described in Table 3. Solid line, protease activity; dashed line, biomass concentration. A, Maximum protease activity; A1, maximum specific protease activity – 1; A2, maximum specific protease activity – 2; B, maximum protease productivity; B1, maximum specific protease productivity – 1; B2, maximum specific protease productivity – 2.

For each of these six protease-related phenotypes, similar trends of protease activity or protease productivity were observed under the different environmental conditions (Table 4). In general, trends showed that protease activity and protease productivity were high in cultures at pH 4 and low in cultures at pH 5. An exception to this were the cultures performed at pH 5 in the presence of high ammonium concentrations, which resulted in relative high values. A closer look at the six protease-related phenotypes revealed subtle differences in reaction to the different process conditions, as illustrated for maximum specific protease activity and maximum specific protease productivity (Fig. 3). For both phenotypes, the experiments were displayed in the same order, showing a clear difference in response. More specifically, in the experiments with high ammonium concentrations, maximum specific protease activities were lower at pH 5 compared to pH 4, while for the same experiments the maximum specific protease productivities were clearly highest at pH 5.

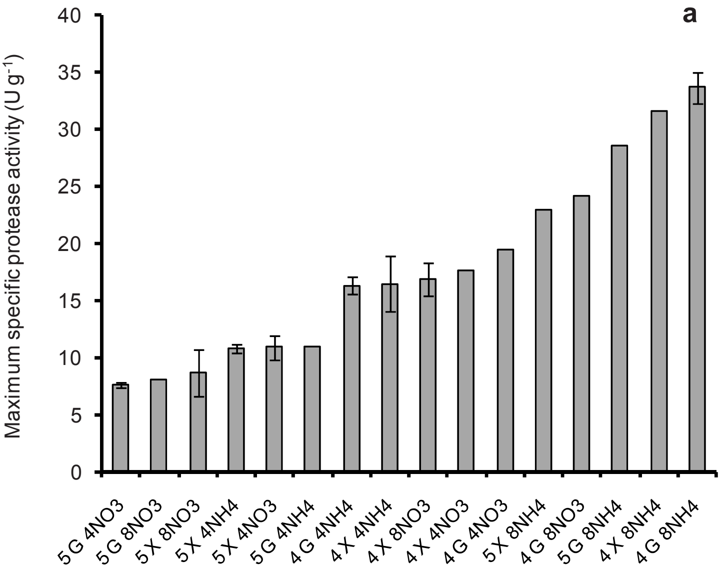
To establish the contribution of the various environmental factors to protease activity and possible interaction effects between these factors, analysis of variance (ANOVA) was performed for each of the six individual protease-related phenotypes (Table 5; for a complete overview of the results and the interaction plots, see the online supplementary data). In the main, nitrogen source and nitrogen concentration had a large relative contribution (η^2) to protease-related phenotypes (Table 5). The contribution of pH is substantial as well, but only to maximum protease activity and maximum specific protease activity.

However, caution is necessary in the interpretation of these main effects, for they cannot be interpreted without taking interaction effects into account. For instance, from Table 5 nitrogen concentration comes up as a significant main effect for all protease-related phenotypes. When only looking at this main effect, comparing the mean protease activities of experiments with either high or low nitrogen concentrations, one might conclude that an increase in nitrogen concentration affects protease activity in all cases. However, this was only true for ammonium, while nitrogen concentration had little effect on the protease-related phenotypes with nitrate as the nitrogen source. This difference in response to nitrogen concentration for the two nitrogen sources tested points towards an interaction effect of nitrogen source and nitrogen concentration. Similarly, for nitrogen source the main effect cannot be viewed without taking into account that the effect is dependent on pH. With the exception of maximum protease activity, protease-related phenotypes were more affected by a change in culture pH when nitrate was used as a nitrogen source instead of ammonium.

Table 4. Values of the six protease-related phenotypes (Table 3) under the experimental conditions of the full factorial design

Experiment **name**	**Max. biomass** **DWT_{max} (g l^{-1})**	**Max. act.** **(U l^{-1})**	**Max. spec.** **act. – 1 (U g^{-1})**	**Max. spec.** **act. – 2 (U g^{-1})**	**Max. prod.** **(U l^{-1} h^{-1})**	**Max. spec.** **prod. – 1 (U g^{-1} h^{-1})**	**Max. spec.** **prod. – 2 (U g^{-1} h^{-1})**
4 G 4NO_3	15.4	300	30.0	19.4	5.5	0.39	0.36
4 G 8NO_3	13.5	326	34.6	24.1	6.1	0.46	0.45
4 G 4NH_4 *	19.1 (±0.5%)	312 (±5%)	25.1 (±12.5%)	16.3 (±4.8%)	9.5 (±7.5%)	0.51 (±5.01%)	0.50 (±7.23%)
4 G 8NH_4 *	19.4 (±3.5%)	652 (±0%)	51.9 (±3.5%)	33.7 (±3.9%)	13.7 (±16.5%)	0.71 (±19.93%)	0.71 (±19.93%)
4 X 4NO_3	12.8	225	23.2	17.7	6.2	0.50	0.49
4 X 8NO_3 *	14.9 (±9.7%)	249 (±1%)	23.0 (±5.7%)	16.8 (±8.6%)	6.3 (±24.8%)	0.44 (±35.70%)	0.43 (±34.02%)
4 X 4NH_4 *	17.4 (±7.5%)	286 (±7%)	25.7 (±14.1%)	16.5 (±14.6%)	6.3 (±12.9%)	0.37 (±17.02%)	0.36 (±20.31%)
4 X 8NH_4	19.4	613	53.9	31.7	13.5	0.70	0.70
5 G 4NO_3 *	9.1 (±41.4%)	70 (±39%)	12.4 (±19.2%)	7.7 (±2.9%)	1.7 (±46.4%)	0.21 (±7.94%)	0.18 (±5.49%)
5 G 8NO_3	11.4	93	12.0	8.1	1.7	0.15	0.15
5 G 4NH_4	15.5	170	20.4	11.0	3.8	0.25	0.25
5 G 8NH_4	16.5	472	43.9	28.5	16.7	1.08	1.01
5 X 4NO_3 *	10.7 (±5.0%)	117 (±4%)	14.9 (±3.4%)	10.9 (±9.3%)	3.3 (±36.2%)	0.32 (±42.22%)	0.31 (±40.80%)
5 X 8NO_3 *	11.5 (±1.9%)	100 (±21%)	11.5 (±28.0%)	8.7 (±22.7%)	2.6 (±48.7%)	0.23 (±50.68%)	0.23 (±50.38%)
5 X 4NH_4 *	17.2 (±3.6%)	187 (±1%)	16.2 (±17.0%)	10.8 (±4.1%)	4.5 (±1.4%)	0.27 (±0.01%)	0.26 (±2.19%)
5 X 8NH_4	16.7	381	40.6	22.9	14.7	0.89	0.88

* These cultivations were performed in duplicate. Results are presented as means (±RSD).



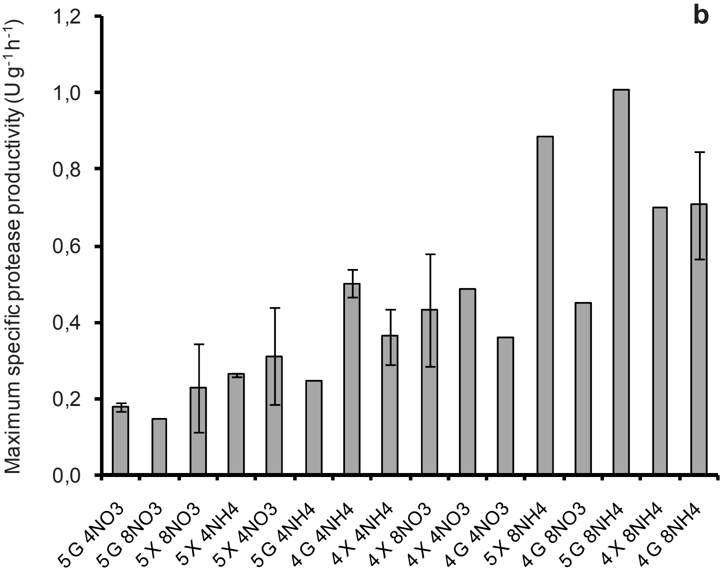


Fig. 3. (a) Maximum specific protease activity and (b) maximum specific protease productivity under the experimental conditions of the full factorial design. To illustrate the differences between the two protease-related phenotypes under identical environmental conditions, the ordering of the experiments is identical in both figures. For cultivations that were performed in duplicate the results are presented as means (±RSD).

However, the contribution of the two-way interaction effect between nitrogen source and nitrogen concentration was considerably higher than for pH in combination with nitrogen source (see Table 5).

Additionally, significant three-way interaction effects were detected (Table 5). For example, the combination of pH, nitrogen source and nitrogen concentration showed a significant interaction effect for maximum protease productivity and maximum specific protease productivity. For instance, in the case of ammonium as nitrogen source, the effect of nitrogen concentration on the maximum protease productivity was larger at pH 5 than at pH 4. On the other hand, the effect of nitrogen concentration was negligible at both pH values with nitrate as nitrogen source. However, in general the contribution of the interaction effects between three environmental factors to variation in protease activity was small.

DISCUSSION

Degradation of secreted proteins by native extracellular proteases is one of the key factors hindering the successful application of filamentous fungi in non-fungal protein production. Approaches to overcome this problem have mainly focussed on strain improvement (Berka *et al.*, 1990; Mattern *et al.*, 1992; van den Hombergh *et al.*, 1995; van den Hombergh *et al.*, 1997b; Zheng *et al.*, 1998; Moralejo *et al.*, 2000; Wiebe *et al.*, 2001; Moralejo *et al.*, 2002). In addition, the use of fungal strains with growth characteristics (e.g., optimal pH) more favorable to the stability of these non-fungal proteins has been evaluated (e.g., Heerikhuisen *et al.*, 2005). However, unlike *A. niger*, these strains were shown to produce high levels of protease at pH values higher than pH 4. Reduction of protease secretion by means of manipulation of the environmental conditions has obtained relatively little attention. In this study the influence of several environmental factors on the extracellular protease activity levels of *A. niger* N402 was systematically investigated in batch cultures. After an initial screening design to select important environmental parameters that influence protease activity, a full factorial design was applied to determine the contribution of each environmental factor to the induction of protease activity. An important additional advantage of a full factorial design is that it can be used to identify possible interaction effects between the environmental factors tested (Lundstedt *et al.*, 1998; Kennedy *et al.*, 1999). In this study, the existence of significant interaction effects between several of the environmental factors was established. To our knowledge, interaction effects between environmental factors in relation to protease secretion of *A. niger* have not been reported before.

Table 5. ANOVA analyses of the main and interaction effects of the full factorial design

An effect with a *P*-value <0.05 is considered significant (shown in bold).

Source of variation*	Max. act.		Max. spec. act. – 1		Max. spec. act. – 2		Max. prod.		Max. spec. prod. – 1		Max. spec. prod. – 2	
	P-value	η^2	*P*-value	η^2	*P*-value	η^2	*P*-value	η^2	*P*-value	η^2	*P*-value	η^2
Main effects												
pH	**<.0001**	**0.22**	**<.0001**	**0.18**	**<.0001**	**0.24**	**0.0034**	**0.05**	0.1242	0.02	0.0801	0.03
C	**0.0015**	**0.01**	**0.0296**	**0.01**	**0.024**	**0.01**	0.7208	0.00	0.8125	0.00	0.9159	0.00
N	**<.0001**	**0.30**	**<.0001**	**0.26**	**<.0001**	**0.18**	**<.0001**	**0.42**	**0.0003**	**0.27**	**0.0002**	**0.28**
[N]	**<.0001**	**0.17**	**<.0001**	**0.21**	**<.0001**	**0.21**	**<.0001**	**0.20**	**0.0007**	**0.20**	**0.0005**	**0.22**
Two-way interaction effects												
pH × C	**0.0045**	**0.00**	0.2348	0.00	0.0897	0.00	0.3673	0.00	0.7992	0.00	0.6387	0.00
pH × N	0.2362	0.00	**0.0171**	**0.01**	**0.0063**	**0.02**	**0.0273**	**0.02**	**0.0146**	**0.07**	**0.0176**	**0.07**
pH × [N]	**0.0034**	**0.01**	0.1037	0.00	0.1403	0.00	**0.0376**	**0.02**	0.0506	0.04	0.0663	0.03
C × N	0.5292	0.00	0.1953	0.00	0.7232	0.00	0.1155	0.01	0.1439	0.02	0.1419	0.02
C × [N]	**0.0175**	**0.00**	0.5587	0.00	**0.0151**	**0.01**	0.9660	0.00	0.6652	0.00	0.6661	0.00
N × [N]	**<.0001**	**0.14**	**<.0001**	**0.20**	**<.0001**	**0.19**	**<.0001**	**0.20**	**0.0003**	**0.27**	**0.0003**	**0.26**
Three-way interaction effects												
pH × C × N	**0.0035**	**0.01**	**0.0056**	**0.02**	**0.0094**	**0.01**	0.8227	0.00	0.8447	0.00	0.9268	0.00
pH × C × [N]	**0.0373**	**0.00**	0.8798	0.00	0.9339	0.00	0.2031	0.01	0.5394	0.00	0.6177	0.00
pH × N × [N]	**0.0358**	**0.00**	0.9171	0.00	0.6553	0.00	**0.0197**	**0.03**	**0.0171**	**0.07**	**0.0206**	**0.06**
C × N × [N]	0.1753	0.00	0.2583	0.00	0.9265	0.00	0.7193	0.00	0.8017	0.00	0.5989	0.00

* C, Carbon source; N, Nitrogen source ; [N], Nitrogen concentration

One of the most prominent interaction effects identified was between nitrogen source and nitrogen concentration, as the effect of concentration is dependent on the nitrogen source. Both individual factors have been reported to affect extracellular protease production in *A. niger* and other aspergilli. Several extracellular protease encoding genes in *A. nidulans* (Katz *et al.*, 1996; Katz *et al.*, 2008) and *A. niger* (van den Hombergh *et al.*, 1994; Jarai & Buxton, 1994) are derepressed when nitrogen source limitation occurs. In the presence of low-molecular-mass nitrogen sources, such as ammonium and nitrate, the induction of extracellular proteases is repressed (Cohen, 1972). In our research we found that in the presence of ammonium, extracellular protease activity only appeared as soon as the carbon source was depleted (results not shown). However, with nitrate as nitrogen source, protease activity levels started to increase before the carbon source depleted, as shown in Fig. 1. This is in agreement with the findings of van den Hombergh *et al.* (1997b), who found nitrate to be less repressive than ammonium and urea. Based on the above-mentioned findings, ammonium seems to be the preferred nitrogen source to repress protease activity. Excess ammonium has been suggested as a means to reduce proteolytic degradation of heterologous proteins (Wiebe *et al.*, 2001; Wiebe, 2003). However, the effect of ammonium as the nitrogen source was less advantageous as soon as derepression of extracellular proteases occurred, for instance due to carbon source depletion. We found that final protease activity levels were higher with ammonium than with nitrate. The existing interaction between nitrogen source and concentration was demonstrated by the increase in maximum protease activity with increasing initial ammonium concentrations, while proteases activities remained unchanged when nitrate concentrations were further elevated.

Less prominent interaction effects were observed involving culture pH. Culture pH itself often has been indicated as an important environmental parameter in controlling extracellular protease activity. It affects both the activity of the secreted proteases (O'Donnell *et al.*, 2001) as well as the expression of protease (Jarai & Buxton, 1994; van den Hombergh *et al.*, 1997b). Also, in our experiments we observed that extracellular protease activities were higher at a culture pH of 4 than at pH 5 or pH 6 (Table 1). Most of the extracellular proteases previously purified from culture filtrates of *A. niger* have acid pH optima (van den Hombergh *et al.*, 1997b; van Noort *et al.*, 1991; de Vries *et al.*, 2004), which is consistent with the acidifying properties of this fungus. Also, the recent sequencing of the *A. niger* genome revealed the presence of an abundance of genes encoding secreted proteases that are expected to be mostly active at low pH (such as aspartic proteases and carboxypeptidases) (Pel *et al.*, 2007).

Due to the observed interactions, the selection of culture conditions to reduce protease activity levels is not straightforward, as several factors are dependent on each other and may have unexpected antagonistic or synergistic effects. Moreover, the environmental parameters affect the biomass levels and protein secretion. For instance, in general protease activity levels are higher at pH 4 and with ammonium as nitrogen source, but so are biomass levels and, for example, glucoamylase levels (data not shown). The selection of the most optimal protein production conditions will therefore require a balance between reduction of protease activity on the one hand and optimization of growth and the level of production of the desired protein on the other. On top of this, a deliberate choice of the phenotype of interest is crucial before the start of an optimization route. In this research, we have illustrated that the effect of an environmental parameter on the six studied protease-related phenotypes is not always the same. It is likely that this is also the case for other fermentation products, both undesired - in this case protease - as well as desired products. When, for instance, a short process time is important, productivity can be the phenotype to be optimized, while in other cases time is less relevant and final product levels are crucial. One might also consider the optimal yield in relation to an expensive substrate or medium component.

ACKNOWLEDGEMENTS

The authors thank Wout Kroes for performing some of the cultivations, Carina M. de Jong-Rubingh for the smoothing algorithm analyses and calculating the six protease-related phenotypes, Peter M. van de Ven for the ANOVA analyses, and Robert A. van den Berg (Katholieke Universiteit Leuven) for fruitful discussions. This project was carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation, which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

CHAPTER 3

AN INVENTORY OF THE *ASPERGILLUS NIGER* SECRETOME BY COMBINING *IN SILICO* PREDICTIONS WITH SHOTGUN PROTEOMICS DATA

Machtelt Braaksma*, Elena Martens-Uzunova*, Peter J. Punt
and Peter J. Schaap

** These authors contributed equally to this study*

This chapter has been accepted for publication in:
BMC Genomics (2010), 11: doi:10.1186/1471-2164-11-584

Supplementary data are available with the online version of this paper at
http://www.biomedcentral.com/bmcgenomics

ABSTRACT

The ecological niche occupied by a fungal species, its pathogenicity and its usefulness as a microbial cell factory to a large degree depend on its secretome. Protein secretion usually requires the presence of a N-terminal signal peptide (SP) and by scanning for this feature using available highly accurate SP prediction tools, the fraction of potentially secreted proteins can be directly predicted. However, prediction of a SP does not guarantee that the protein is actually secreted and current *in silico* prediction methods suffer from gene-model errors introduced during genome annotation.

A majority rule-based classifier that also evaluates SP predictions from the best homologs of three neighbouring *Aspergillus* species was developed to create an improved list of potential SP containing proteins encoded by the *Aspergillus niger* genome. As a complement to these *in silico* predictions, the secretome associated with growth and upon carbon source depletion was determined using a shotgun proteomics approach. Overall, some 200 proteins with a predicted SP were identified to be secreted proteins. Concordant changes in the secretome state were observed as a response to changes in growth/culture conditions. Additionally, two proteins secreted via a non-classical route operating in *A. niger* were identified.

We were able to improve the *in silico* inventory of *A. niger* secretory proteins by combining different gene-model predictions from neighbouring aspergilli and thereby avoiding prediction conflicts associated with inaccurate gene-models. The expected accuracy of SP prediction for proteins that lack homologous sequences in the proteomes of related species is 85%. An experimental validation of the predicted proteome confirmed *in silico* predictions.

INTRODUCTION

Fungi are heterotrophic organisms that decompose and utilize a plethora of bio-organic carbon sources through secretion of biomass degrading enzymes. The fungal secretome is defined as the sub-proteome of soluble secreted proteins. A large part of this secretome consists of the many extracellular hydrolytic enzymes necessary to digest potential substrates. Other extracellular proteins play crucial roles in fungus-host interactions and in fungal pathogenicity. Therefore, gene classes expressed in the fungal secretome to a large degree define the ecological niche occupied by a fungal species, its impact on human health and agriculture and its usefulness as a production organism.

In the absence of direct experimental proof fungal secretomes are usually directly predicted from the genome sequence by analysing the deduced proteome for proteins with a putative N-terminal signal peptide (SP). Experimentally identified eukaryotic signal peptides on average have a sequence length between 17 to 30 amino acids and these SP are further characterized by a central hydrophobic core region of 6-15 amino acids flanked by hydrophilic N- and C- terminal regions. These features have been used to develop highly specific SP prediction tools, which all show very high prediction accuracies of 93% or higher when applied to benchmark data sets (Zhang *et al.*, 2003; Bendtsen *et al.*, 2004a; Käll *et al.*, 2004). However, the accuracy of a SP prediction for predicted proteins heavily relies on an accurate gene-model that provides correct N-terminal end translation of the encoded protein. Since signal peptides do not share an apparent sequence homology (Bendtsen *et al.*, 2004a), sequence variability between secreted homologous proteins of related species is usually significantly higher at the N-terminal end. This N-terminal heterogeneity proofs to be a serious problem for homology assisted gene-finding algorithms to create a reliable gene-model useful for accurate SP prediction. Therefore, the real problem in predicting an *in silico* proteome is not the accuracy of the present prediction tools, but are the inaccurate gene-models used as input for these tools. Furthermore, a number of proteins with a correctly predicted SP are in reality not secreted, for instance because they are resident ER proteins (Scott *et al.*, 2004). Thus, in the absence of direct experimental proof of secretion, an *in silico* predicted secretome does not correctly represent the actual secretome.

The genus *Aspergillus* represents an important group of filamentous fungi with significant impact on many facets of human welfare. Recently, genome-sequencing projects of at least 10 *Aspergillus* species have been completed or are nearing completion. The corresponding proteomes are usually inferred from gene-models derived with automated gene prediction tools. Consequently, the large majority of the predicted protein coding sequences are hypothetical and have a variable degree of accuracy. An encouraging exception is the extensively manually annotated genome sequence of *A. niger* (Pel *et al.*, 2007). Genome sequences are publicly available from two *A. niger* strains (Pel *et al.*, 2007; Baker, 2006), which allows for a direct cross-validation of genome data and for a direct comparison of most of the independently derived gene-models. *A. niger* is an excellent producer of a suite of extracellular enzymes and many of them have been granted a GRAS (Generally Recognized As Safe) status by the U.S. Food and Drug Administration (FDA). These properties have made this fungus a preferred production organism for a range of secreted commercial enzymes. Among the most important of them are amylases, asparaginases, β-galactosidases, glucose oxidase, glycosidases, lipases, phospholipases, proteases, phytases and several hemicellulases (Schafer *et al.*, 2007). Nevertheless, based on the recently elucidated genomic sequence of *A. niger*, it can be estimated that currently direct experimental proof of secretion of only a fraction of the potentially secreted proteins exists.

In this study we combined comparative *in silico* SP predictions for classically secreted proteins with an extensive set of experimental secretome data derived from mass spectrometry analysis of *A. niger* secretome enriched fractions. Cross-species validation of *in silico* SP predictions produced a more accurate list of potentially secreted proteins and an improved annotation of the underlying gene-models. The secretome of *A. niger* associated with growth on sorbitol and galacturonic acid and upon depletion of the carbon source was analyzed using a shotgun proteomics approach. This analysis provided insight into the dynamics of the *A. niger* secretome and direct experimental proof of secretion for known and unknown signal peptide directed proteins (SP proteins).

METHODS

Bioinformatics

Signal peptide predictions. SP predictions were done with a local implementation of the signalP3-NN (neural network) and the signalP3-HMM (hidden Markov model) algorithms (Bendtsen *et al.*, 2004a). Proteins were considered to be SP proteins if the signalP3-NN *D*-score was higher than 0.43. Additional signal anchor predictions were done with a local implementation of the signalP3-HMM algorithm.

Signal peptide cross-validation and construction of the classifier. The predicted proteomes of *A. niger* strains CBS 513.88 and ATCC 1015, *A. oryzae* RIB40, *A. fumigatus* AF293 and *A. nidulans* FGSC A4 were used as input. For each individual protein a SP prediction was done using the signalP algorithm. If the score was above the set threshold *D*-value the tag SP was added to the ordered locus name. Next, a bidirectional Blastp (Altschul *et al.*, 1997) was done between the *A. niger* CBS 513.88 proteome and the proteome of *A. niger*

ATCC 1015, and between the *A. niger* CBS 513.88 proteome and the proteome of the three other *Aspergillus* species. Each set of pairwise tabular outputs was stringently parsed for bidirectional best hit pairs using the following criteria (i) between the two sequences the percentage of identity must be above a set threshold level of 40%, (ii) the two aligned protein sequences must be of similar size (a difference in size of less than 20% was accepted) and (iii) the aligned region must include more than 70% of the smallest protein sequence. Next, these bidirectional best hits were used to form *A. niger* centred protein clusters. Protein clusters with at least one SP-tag added to an ordered locus name were selected for the construction of the classifier.

Implementation of the classifier. In comparison with single genome signalP3 predictions deviating classifier SP predictions were considered to be of better-quality when the two following criteria were met (i) a cluster-size of at least three species and (ii) between the non-*A. niger* classifier proteins a complete agreement in SP prediction. Muscle (Edgar, 2004) was used for protein multiple sequence alignments. The general Prosite consensus pattern was used to identify C-terminal ER retention motifs in predicted SP proteins.

Mass spectrometry data analysis. The 98.150 MS/MS spectra resulting from MS analysis of the *A. niger* secretome enriched samples (see below) were submitted to a local implementation of the OMSSA search engine (Geer *et al.*, 2004). MS/MS spectra were independently searched against peptide databases derived from the predicted proteomes of *A. niger* strain CBS 513.88 and of strain ATCC 1015 and against a database of randomized sequences constructed from the reverse of the CBS 513.88 proteome. All OMSSA searches used the following parameters: a precursor ion tolerance of 0.03 Da, fragment ion tolerance of 0.5 Da, a miss cleavage allowance of up to and including 2, all cysteines were considered to be carboxyamidomethylated, oxidation of methionine and deamination of glutamine and aspargine were treated as variable modifications.

The set *E*-value threshold was determined iteratively from the false discovery rate (FDR) and was set to 0.01. With this setting an FDR of <2% was obtained over all samples. FDR calculations were done as follows: for each identified spectrum with a threshold *E*-value <0.01 accepted peptide-spectrum matches (PSM) with each individual peptide database were ranked by their *E*-value and the top hit identified peptide sequence was selected. The FDR was calculated from top hit spectral matches to peptides in the reversed database as described by Elias & Gygi (2007).

The data is available in the PRIDE database[1] (Vizcaino *et al*., 2010) under accession numbers 13662, 13663, 13664 and 13665.

Culture conditions

The fungal strain used in this study was *A. niger* wild type N402, a *cspA1* (conferring short conidiophores) derivative of ATCC 9029 (alternative names NRRL 3, CBS 120.49, N400).

Conditions for growth on sorbitol and galacturonic acid. For pre-culture, 1.0 x 10^6 spores per millilitre were inoculated into 2.5-L fermentors (Applikon) containing 2.2 L of minimal medium (Pontecorvo *et al.*, 1953) with 0.01% yeast extract and either 50 mM D-sorbitol or 50 mM D-galacturonic acid as carbon source, at 30 °C and pH 3.5. Spore germination in bioreactors was as described previously (van der Veen *et al.*, 2009), with headspace aeration and a stirring speed of 300 r.p.m., and when dissolved oxygen levels were below 60%, stirring speed was changed to 750 r.p.m. and aeration was through sparger inlet. The amount of

[1] http://www.ebi.ac.uk/pride/; October 14, 2010

monomeric sugars remaining in the culture fluid was assessed by standard HPLC techniques. Culture supernatants were taken 24 h and 48 h after inoculation.

Conditions used for carbon source exhaustion. Cultures were grown in batch fermentations in a BioFlo 3000 (New Brunswick Scientific) bioreactor with a 5-L working volume. Cultivations were performed with varying carbon source (glucose or xylose), nitrogen source (ammonium chloride or sodium nitrate), nitrogen concentration (low (282.4 mM) or high (564.8 mM)), and pH (4 or 5) (see Table 1). The medium composition, cultivation conditions and operating procedure of the bioreactor have been described in detail previously (Braaksma *et al.*, 2009). Samples for analysis of the carbon source concentration were collected every six hours and analyzed as described previously (Braaksma *et al.*, 2009). From each growth condition culture supernatants were taken after carbon source exhaustion.

Analysis of total protein. The concentration of protein in cleared culture supernatants was measured by the Bio-Rad Protein Assay, using BSA as a standard. The procedure was fully automated using a COBAS MIRA Plus autoanalyzer.

Table 1. Overview of initial growth conditions used for carbon source exhaustion and time point of sampling

Experiment name	Carbon source	pH	Nitrogen source	Nitrogen source level (mM)	Sampling time (h)
Nitrate, pH 4					
4 G 4NO$_3$	glucose	4	NaNO$_3$	282.4	96
4 X 4NO$_3$	xylose	4	NaNO$_3$	282.4	95
Ammonium, pH 4					
4 G 8NH$_4$	glucose	4	NH$_4$Cl	564.8	91
4 X 4NH$_4$	xylose	4	NH$_4$Cl	282.4	85
Nitrate, pH 5					
5 G 8NO$_3$	glucose	5	NaNO$_3$	564.8	96
5 X 8NO$_3$	xylose	5	NaNO$_3$	564.8	156
Ammonium, pH 5					
5 G 8NH$_4$	glucose	5	NH$_4$Cl	564.8	84
5 X 4NH$_4$	xylose	5	NH$_4$Cl	282.4	90
5 X 8NH$_4$	xylose	5	NH$_4$Cl	564.8	96

Liquid chromatography tandem mass spectrometric analysis

For secretome enriched fractions obtained from growth on sorbitol and galacturonic acid equal amounts of protein sample (250 µg) were separated on 12% SDS polyacrylamide gels, and stained with Colloidal Blue Staining (Invitrogen, Carlsbad, CA, USA). Gel lanes were cut into five slices, and each slice was treated with 50 mM dithiothreitol (DTT) in 50 mM NH_4HCO_3 (pH 8.0) for 1 h at 60 °C. Next, slices were alkylated with 100 mM iodoacetamide in NH_4HCO_3 (pH 8.0) for 1 h at room temperature, washed with NH_4HCO_3 (pH 8.0). Slices were rehydrated in 10 ng/µl trypsin (Sequencing grade modified trypsin, Promega, Madison, WI, USA) and digested overnight at 37 °C.

LC-MS/MS conditions: samples were loaded on a preconcentration column and peptides were eluted to an analytical column with an acetonitrile gradient and a fixed concentration of formic acid. The resulting eluent was subjected to an electrospray potential *via* a coupled platinum electrode. MS spectra were measured on an LTQ-Orbitrap (Thermo Electron, San Jose, CA, USA) and MS scans of four most abundant peaks were recorded in data-dependent mode. To simplify the comparison between the two growth conditions the two galacturonic acid and the two sorbitol samples were pooled.

Secretome enriched samples obtained from carbon source exhaustion were analysed with LC-ESI-MS-MS performed by Eurogentec (Seraing, Belgium). From each sample a volume corresponding to 10-15 µg of total protein was digested with trypsin, without prior separation of the proteins. To simplify the comparison with growth, all samples were pooled.

RESULTS AND DISCUSSION

In silico prediction and validation of the secretome of *A. niger*

To estimate the prediction accuracy of the secretome of *A. niger*, we compared the SP predictions of both sequenced *A. niger* strains (*A. niger* CBS 513.88 *and A. niger* ATCC 1015), and further compared them with SP predictions of orthologous proteins from three closely related functionally annotated *Aspergillus* species (*A. oryzae* strain RIB40 [Machida *et al.*, 2005], *A. fumigatus* strain Af293 [Nierman *et al.*, 2005] and *A. nidulans* strain FGSC A4 [Galagan *et al.*, 2005]).

Cross-validation of SP predictions between *A. niger* CBS 513.88 and *A. niger* ATCC 1015

The genome of the industrial production strain *A. niger* CBS 513.88 was recently sequenced (Pel *et al.*, 2007). A total of 14,086 protein coding genes (CDS) were identified in its genome. Of these protein CDS 191 are known to be N-terminally truncated, because the corresponding loci are located at a contig border. When the signalP3 signal peptide prediction suite (Bendtsen *et al.*, 2004a) is used, a classical signal sequence for secretion is detected in at least 1831 predicted proteins (Table 2). For reasons argued above, this *in silico* prediction will not be very accurate, because it depends heavily on the correctness of the underlying gene-models.

The genome of *A. niger* strain ATCC 1015 was annotated independently and is predicted to encode approximately 11,200 genes (Baker, 2006). The signalP3-NN algorithm predicts that 1540 *A. niger* ATCC 1015 gene-models encode proteins with a SP (Table 1). In total 1257 of those gene-models orthologous to a single CBS 513.88 gene-model and undoubtedly derived from the equivalent locus. This subset was used to compare SP prediction results in the two strains. In as much as 30% of these predicted proteins conflicting signalP3-NN prediction results were obtained due to alternative start codon selection (Supplementary data file 1).

Cross-validation of SP predictions using other aspergilli as classifier species

To improve the precision of the *A. niger* whole proteome SP prediction, signalP3-NN prediction results of *the A. niger* CBS 513.88 proteome were also compared to those of the best homologous proteins of three closely related fully annotated *Aspergillus* species, i.e. *A. oryzae* strain RIB40 (Machida *et al.*, 2005), *A. fumigatus* strain Af293 (Nierman *et al.*, 2005) and *A. nidulans* strain FGSC A4 (Galagan *et al.*, 2005). A summary of the single genome signalP3 predictions of these aspergilli is presented in Table 2.

Table 2. Proteome size and single signalP3 signal peptide and signal anchor predictions of four selected *Aspergillus* species.

From left to right species are ranked by their phylogenetic distance to *A. niger* CBS513.88.

Species	*A. niger* CBS 513.88	*A. niger* ATCC 1015	*A. oryzae* RIB40	*A. fumigatus* Af293	*A. nidulans* FGSC A4
protein CDS	14086 †	11197 ‡	10406 ‡	9887 ‡	10665 ‡
signalP3-NN *	1831	1540	1751	1258	1469
signalP3-HMM SP *	2016	1687	1802	1067	1612
signalP3-HMM SA *	627	529	582	391	488

* NN, neural network method; proteins were considered to be SP proteins if the signalP3 *D*-score >0.43. HMM, hidden markov model; SP, signal peptide; SA, signal anchor.

† Number obtained from the Refseq section of GenBank.

‡ Numbers obtained from www.broad.mit.edu/annotation/genome/aspergillus_group/MultiHome.html.

At such a close phylogenetic distance, clusters of orthologous proteins not only are predicted to have the same molecular function in the different species, but also are expected to exert this molecular function at the equivalent location. If this is true, SP prediction results derived from individual signalP3 predictions for *Aspergillus* sp. proteins orthologous to an *A. niger* protein of interest can be used as an independent majority rule-based classifier. The classifier was constructed in the following way. For each genome the complete list of predicted SP proteins and their reciprocal top BlastP hits with *A. niger* CBS 513.88 proteins were sorted into *A. niger* centred orthologous protein clusters as is detailed in Methods. Subsequently, each cluster member was also screened for a possible signal anchor (SA). In this way 1527 *A. niger* centred orthologous clusters with at least one putative SP protein could be formed. Of these

1527 protein clusters, 1274 are spanning three to five genomes and 253 are formed by bidirectional "best hit" protein pairs (Supplementary data file 1). In total 669 thus formed protein pairs and clusters showed a pan-genomic cross validation of SP prediction.

It should be noted that not all of the cross-validated protein in these clusters are actually secreted. Proteins with a function in the secretion pathway or related compartments such as the vacuole may with this *in silico* approach be classified as (potentially) secreted proteins in. For instance, we have observed clustering of at least 12 resident ER proteins, which can be recognized by the presence of a C-terminal ER-retention motif (Scott *et al.*, 2004) (Supplementary data file 1). However, as for most of the classified proteins a molecular functional characterization is lacking, we have not taken this into account in our analysis. A further inspection of Supplementary data file 1 suggests that for all five analysed aspergilli the accuracy of a single genome *in silico* SP prediction is approximately 85%.

Improved annotation of *A. niger* gene models

In 33 protein clusters of the classifier an *A. niger* CBS 513.88 protein predicted to be a non-SP protein was clustered exclusively with classifier SP proteins being orthologous proteins from the other *Aspergillus* species. Four of those likely false negative signalP3 predictions were re-evaluated by aligning their N-terminal ends (Supplementary data file 2). In all cases selection of an alternative start codon in the most likely reading frame would (i) bring the predicted protein sequence length in better agreement with the predicted protein sequence length of the close by orthologs and (ii) add a predicted SP feature to the alternative N-terminal end of the predicted CBS 513.88 protein.

Vice versa, in 55 cases a SP prediction for an *A. niger* CBS 513.88 protein was not supported by predictions for the orthologous classifier proteins in the protein clusters. While the molecular function prediction of most of them clearly suggests an intracellular molecular function, in some cases the classifier also showed an ambiguous behaviour in separating SP and SA predictions. For instance, the protein sequences of An15g01200 (*A. niger* 513.88) and the equivalent protein 137591 (*A. niger* ATCC 1015) differ both in length and in their SP/SA prediction. However, compared to the best homologs of the other *Aspergillus* sp. both proteins appear to be N-terminally truncated and therefore both should be N-terminally extended. A screen of *A. niger* ATCC 1015 EST sequence data available at the Broad Institute[2] demonstrated the presence of an alternative start codon revealing a putative SA with a probability of 0.993 for the newly inferred protein (see Supplementary data file 2 for details).

Proteogenome analysis of secretome enriched fractions

Secretome enriched fractions of *A. niger* N402 cultured under controlled conditions in defined synthetic media were analyzed by high-throughput mass spectrometry (see Methods). The culture supernatants of three conditions were sampled. In the first two conditions, samples of secretome enriched fractions grown on sorbitol were compared with samples from secretome enriched fractions grown on galacturonic acid (GalA). For the induction of the pectinolytic system sorbitol is considered to be a neutral carbon source, while the carbon source GalA is the major constituent of pectin and a specific inducer (Martens-Uzunova & Schaap, 2008; Martens-Uzunova & Schaap, 2009). In the third condition, prolonged carbon source exhaustion was exploited. The Open Mass Spectrometry Search Algorithm (OMSSA) search engine (Geer *et al.*, 2004) was used for the analysis of these tandem mass spectra. One of the major causes for errors in protein identification is incompleteness of the peptide sequence database due to missed protein encoding genes and gene-models errors. Therefore, tandem mass spectra obtained by shotgun proteomics of the enriched secretome fractions were independently matched with two peptide databases derived from the predicted proteome sequences of both *A. niger* strains. To quantify false positive rates of peptide identification, all spectra were also independently searched against a reverse peptide database constructed from the reverse *A. niger* CBS 513.88 proteome (see Methods). At the selected *E*-value threshold <0.01 for acceptance of a PSM, the spectrum level FDR was limited to 2% or less under all conditions. The bioinformatics analysis workflow is presented in Fig. 1. The full list and functional annotation of thus identified proteins and the conditions under which they were detected are shown in Supplementary data file 3.

The genome of *A. niger* CBS 513.88 has been a subject of an extensive molecular function prediction, followed by thorough manual verification. As a result, the genome sequence of this strain encompasses a higher number of protein-coding genes compared to *A. niger* ATCC 1015. Therefore, the CBS 513.88 proteome was chosen as the primary database for further analysis. Overall, 7523 accepted PSMs identified 285 predicted *A. niger* CBS 513.88 proteins. Additionally, we detected 7 more *A. niger*

[2] http://www.broadinstitute.org/; September 5, 2010

ATCC 1015 proteins with no apparent matching locus in the genome of the other strain (Supplementary data file 3). Conversely, 25 identified *A. niger* CBS 513.88 proteins lacked an *A. niger* ATCC 1015 gene-model, even though in most cases the corresponding locus was present in the ATCC 1015 genome.

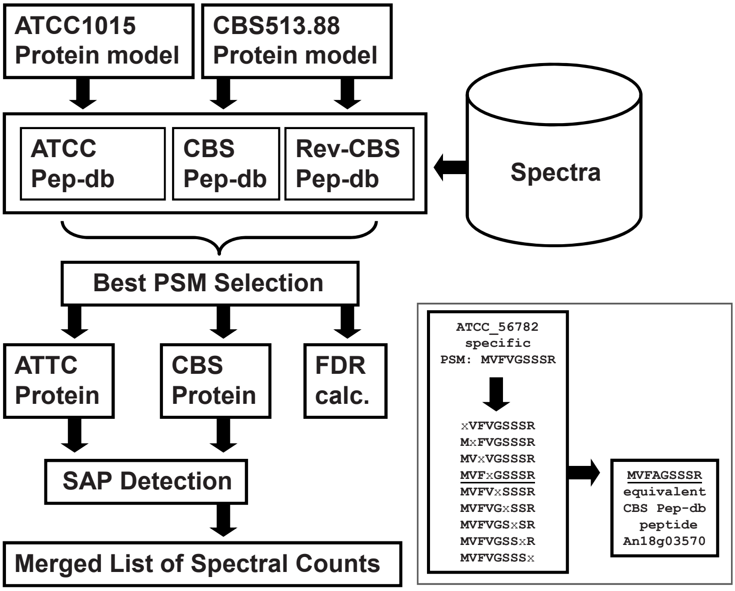


Fig. 1. Schematic of the Spectrum to Peptide matching pipeline. Forward and reversed (REV-CBS) databases were searched with local implementation of the OMSSA MS/MS search engine. Threshold Expect values for matching peptides were estimated from the false discovery rate (FDR). Best Accepted peptide-spectrum matches (PSM) selection was done by ranking for each MS/MS spectrum the output of each individual peptide database by *E*-value and selection of the top hit identified peptide sequence.
Insert: Detection of a single amino acid polymorphism (SAP). A wildcard character (x) is introduced at each position of a single proteome matching peptide, followed by a pattern search in the complementary proteome. In the given example using the ATCC single proteome matching peptide as a template, a single equivalent peptide, derived from An18g03570 is retrieved from the complementary proteome. An18g03570 is 99% identical to ATCC 56782.

Wright *et al.* (2009) and Tsang *et al.* (2009) used a similar shot-gun proteomics procedure to exploit the *A. niger* proteome. Very recently, also an *A. niger* proteome study based on 2D-gel electrophoresis was carried out by Lu *et al.* (2010). In the study of Wright *et al.* (2009), where frozen mycelium was used as study material, 214 different loci were identified. As expected by the differences in source material, the overlap with the present study is limited to only eight proteins. In the study by Lu *et al.* (2010) about 70 proteins were detected in the secretome of which the majority was also found in our data set. Frome these, only three SP directed proteins we did not identify in our data set. Similar to our results, the shotgun proteomics approach from Tsang *et al.* (2009) identified about 200 secretome-associated proteins, from which the large majority corresponds to *in silico* classified SP proteins, confirming the validity of our approach. About 40 of the proteins identified by Tsang *et al.* were not identified in our data set, whereas our experimental data set identified more than 80 SP proteins not identified by Tsang *et al.* (2009).

Peptide spectrum matching requires a high quality proteome. While by and large correct, gene-model predictions may suffer from exon-identification and exon-border errors, leading to a mismatch with identified peptide spectra. Another reason for not obtaining completely matching peptide spectra may be due to the presence of genetic variation, small strain differences leading to single amino acid polymorphisms between the investigated strain *A. niger* N402 and the two annotated *A. niger* genomes used for mass analysis. In a systematic analysis of matching peptides that are only present the peptide databases of the annotated genomes, 31 single proteome peptides were found to match with a single amino acid polymorphism in the equivalent protein of the other strain. The large majority of amino acid polymorphisms (29 out of 31) was observed between strain CBS 513.88 and strain N402, suggesting that strain N402 is more closely related to ATCC 1015.

Functional analysis of secretome enriched fractions

Fungal secretome enriched samples are expected to contain a complex mixture of possibly hundreds of SP proteins with a minimal contribution of proteins acquired through cell lysis.

A simple differential measure of relative protein abundance known as ‘spectral counting’ can be used to quantify the relative contribution of each protein to this mixture. It has been shown that the total number of spectra that identify peptides originating from a given protein shows good linear correlation with the abundance of that protein (Liu *et al.*, 2004; Zybailov *et al.*, 2005) and a good sensitivity for detecting changes in protein abundance (Old *et al.*, 2005, Fu *et al.*, 2008). The major analytical caveats to using this approach is that spectral count ratios can be biased by undersampling, the fact that different peptides have different physiochemical properties that affect MS detection, and that in complex mixtures for proteins with a low number of spectral counts this correlation is not very strong (Old *et al.*, 2005).

To overcome such limitations in interpreting relative presentation of proteins, functional annotation clustering was used to identify biological processes overrepresented among the proteins detected in the enriched secretome fractions. For this, detected proteins were clustered in nine groups. Seven groups were based on molecular function prediction by using the FunCat annotation scheme (Ruepp *et al.*, 2004) and the predicted molecular function as guidance. Functionally unclassified proteins with an SP prediction and a functionally diverse group of “non-SP proteins” formed two additional groups (Fig. 2). The group “C compound and carbohydrate metabolism” (CH) together with the enzymes of the pectinolytic system formed the largest functional annotation cluster. From Fig. 2 it is obvious that compared to growth on sorbitol the pectinolytic system is induced upon growth on GalA. Therefore “pectin-modifying proteins” were put in a cluster separate from the CH cluster. FunCat category “extracellular protein degradation” was used as a basis for the cluster “protein and peptide degradation”. Furthermore, we distinguished “cell wall components”, “oxidases”, “lipase-esterases” and “acid phosphatases”.

Overall, 98% of the 2722 accepted PSMs obtained from the sorbitol samples could be traced back to a SP protein in one of the seven functional annotation clusters or the hypothetical SP protein cluster. Almost identical results were obtained for the GalA samples. For the carbon source starvation conditions this amounted to 88% of the accepted PSMs (Fig. 2 and Supplementary data file 3). These results suggest that the quantitative contribution of cell lysis to the secretome enriched fractions demonstrated by the detection of an array of functionally diverse non-SP proteins is indeed limited. The contribution of non-SP proteins is significantly higher in secretome samples derived from starvation conditions, but this difference is primarily caused by the specific expression of a single non-SP protein An01g09980, with a strong similarity to Asp-hemolysin from *A. fumigatus*. Asp-hemolysin has been purified from the culture filtrate of *A. fumigatus*, while no SP is detected (Kudo *et al.*, 2002). The fact that the *A. niger* homologous protein is detected in significant amounts in the culture filtrate as well, suggests that this is a non-classically secreted protein. If the Asp-hemolysin is indeed intentionally secreted, the relative contribution of cell lysis in secretome enriched fractions under starvation conditions is much more comparable to what is observed for sorbitol and GalA.

More than 98% of the here-identified secreted proteins are supported by signalP3 and majority-rule predictions. However, the list also includes the protein ATCC 1015 protein (128537), which is supported by the rule-based classifier prediction only. Others, such as An02g11390, are ambiguous in their signalP3 and classifier-based SP predictions, but are clearly present in secretome fractions. If we consider these proteins to be genuinely secreted proteins, the contribution of cell lysis in our data set is even lower than discussed above.

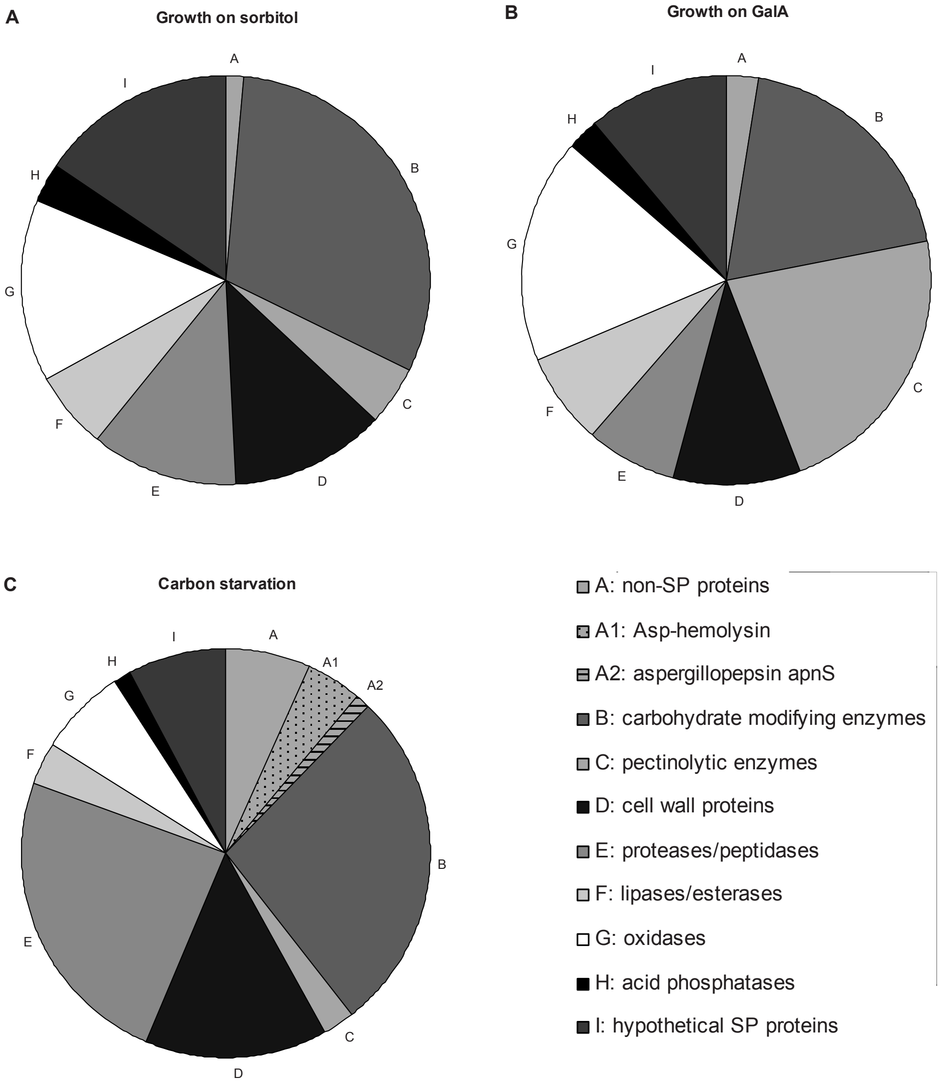


Fig. 2. Categorization of the *A. niger* secretome. Detected secretome when grown on sorbitol (A), on galacturonic acid (B), or under carbon starvation conditions (C). For each condition, the contribution of a protein to a category was normalized based on the total number of spectra.

Carbohydrate modifying enzymes

Three controlled fermentation conditions were chosen to study the relative contribution of various classes of carbohydrate modifying enzymes and proteases to the secretome. To minimize the effect of undersampling, sorbitol, GalA, and starvation-specific samples were pooled. Between the three conditions significant changes were observed for all functional annotation clusters, except for the cluster of acid phosphatases. Upon growth on GalA "pectinolytic enzymes" are overrepresented. In contrast, proteins present in the CH cluster are overrepresented upon growth on sorbitol.

Although some pectinolytic enzymes are found in all sampled secretome fractions, pinpointing to constitutive expression (Martens-Uzunova & Schaap, 2009), the pectinolytic system is strongly induced under GalA growth conditions. Compared to growth on sorbitol, not only the number of spectral counts related to the pectinolytic functions increase upon growth on GalA, but also the diversity of enzymatic functions. To identify *A. niger* genes potentially involved in galacturonic acid catabolism, we have previously compared N402 microarray data obtained upon growth of the fungus on various carbon sources. Fifteen highly correlating genes were found that were specifically induced on galacturonic acid (Martens-Uzunova & Schaap, 2008). GalA specifically activates the transcription rate of six extracellular enzymes. In the GalA derived secretome five of those, pectin lyase A, three exoPGs (PGAX, PGXB and PGXC) and An02g02540, a putative pectin acetylesterase, are detected, but only under this condition. An08g01710, a putative arabinofuranosidase with no apparent SP and part of this transcriptional cluster, is not found in any of the secretome fractions (Table 3).

Table 3. Pectinolytic enzymes with a correlating transcriptional profile in galacturonic acid transfer cultures in secretome enriched fractions

Locus tag	*r* *	Gene name	Molecular Function	Signal Peptide	Spectral counts		
					Sorbitol	GalA †	Starvation
An14g04370	0.999	*pelA*	Pectin lyase A	Yes	0	8	0
An12g07500	0.979	*pgaX*	Exopolygalacturonase X	Yes	0	18	0
An11g04040	0.978	*pgaA*	Exopolygalacturonase A	Yes		Not detected	
An03g06740	0.971	*pgxB*	Exopolygalacturonase B	Yes	0	40	0
An02g12450	0.964	*pgxC*	Exopolygalacturonase/ exoxylogalacturonan hydrolase	Yes	0	19	0
An08g01710	0.953		Putative arabino-furanosidase	No		Not detected	
An02g02540	0.963		Putative pectin acetylesterase	Yes	0	19	1

* *r*, correlation coefficient (data from Martens-Uzunova & Schaap, 2008).

† GalA, galacturonic acid.

Glucan is one of the major chemical components of the *Aspergillus* cell wall and 1,3-beta-glucanosyltransferases therefore play an active role in fungal cell wall biosynthesis (Mouyna *et al*., 2000). Overall, eight 1,3-beta-glucanosyltransferase genes from the GH72 family are present in the *A. niger* genomes. All eight encoded proteins are predicted to have a Glycosylphosphatidylinositol (GPI) anchor, that becomes linked to the C-terminal residue after a proteolytic cleavage occurring at the so-called ω-site (Pierleoni *et al*., 2008). A multiple alignment of the eight encoding protein sequences suggest that they can be assigned to three subgroups (Table 4). Four of those 1,3-beta glucanosyltransferases representing each of these subgroups are detected in the three main secretomes.

Table 4. Expression of 1,3-beta-glucanosyltransferase genes present in the *A. niger* genome

Group	Ordered locus name *	Signal peptide prediction		GPI-anchor Prediction ‡
		signalP3 †	classifier	
1	**ATCC 53033**	Yes	Yes	Highly probable
	An02g03070	Yes	Ambiguous	Highly probable
2	An02g09050	Yes	Yes	Probable
	An08g07350	Yes	Yes	Highly probable
	An10g00400	Yes	Yes	Highly probable
3	**An09g00670**	Yes	Yes	Highly probable
	An03g06220	No	Yes	Probable
	An16g06120	Yes	N.A. §	Highly probable

* Bold: proteins are detected by mass spectrometry in the secretome enriched fractions

† SignalP3 algorithm (Bendtsen *et al*., 2004a)

‡ Using the PredGPI algorithm (Pierleoni *et al*., 2008)

§ N.A.: Not available, classifier predictions are not valid, because the cluster-size is too small

Proteases

Carbon source starvation conditions were chosen to induce extracellular proteases. Indeed, where the fraction of spectral counts assigned to "extracellular protein degradation" is 7% and 12% for growth conditions GalA and sorbitol, respectively, under starvation conditions this is 24%. The extracellular aspartic proteinase aspergillopepsin I (PepA) is by far most abundant under starvation conditions. Other high abundant proteases are An08g04490 (Edens *et al.*, 2005), and the putative serine proteases An14g02470, An06g00190, and An03g05200, which together with PepA account for over 75% of the PSMs assigned to proteases under starvation conditions (Table 5). In addition, under all conditions tested, a protease (53364) was detected specific to *A. niger* ATCC 1015 locus only. This aspartic-type endopeptidase has a predicted SP and homologs are widespread in the genomes of other aspergilli.

Table 5. Proteases detected in secretome enriched fractions of *A. niger* N402 cultured under a set of controlled conditions.

MEROPS family	Locus tag	Spectral counts		
		sorbitol	GalA *	starvation
Peptidase family A1 (pepsin family)	An01g00370 †	0	0	27
	An02g07210	0	0	2
	An04g01440	0	0	2
	An12g03300	0	1	0
	An13g02130	1	1	0
	An14g04710	87	33	226
	An15g06280	4	0	0
	An18g01320	9	8	17
	ATCC 53364	2	3	4
Peptidase family M28	An03g01660	0	8	3
Peptidase family S10 (carboxypeptidase Y family)	An02g04690	9	5	3
	An03g05200	38	21	49
	An14g02150	0	1	4
Peptidase family S28 (lysosomal Pro-Xaa carboxypeptidase)	An08g04490	58	25	53
	An12g05960	42	20	29
Peptidase family S53 (sedolisin family)	An01g01750	14	5	24
	An03g01010	9	10	2
	An06g00190	7	16	68
	An08g04640	17	4	24
	An14g02470	18	18	67

* GalA, galacturonic acid.

† For An01g00370 no signal peptide could be predicted

Overall, 20 proteases were identified in this study (Table 5), of which all but An01g00370 have a SP prediction. An01g00370 is an aspartic protease with strong similarity to aspergillopepsin ApnS of *A. phoenicis*, and is only detected under starvation conditions. An01g00370 is not a protein directed by a classical signal peptide for secretion nor can such signal peptides be detected in orthologous (predicted) proteins from other aspergilli. Nevertheless, the number of spectra derived from this protein is relatively high (Table 5), making it unlikely that this protein was detected in these fractions due to lysis. Therefore, in addition to the highly expressed putative hemolysin homolog, this protease is the second likely candidate for non-classical secretion and indeed, when subjected to the Secretome P2.0 algorithm (Bendtsen *et al.*, 2004b), both protein sequences score above the set threshold value for non-classical secretion. Furthermore, disruption of this protease, results in a significant increase of the secreted level of heterologous laccase activity (Wang *et al.*, 2008), suggesting it is in fact a functional extracellular protease secreted by non-classical routing.

CONCLUSIONS

In this work we present an improved list of SP proteins encoded by the *A. niger* genome. The list of SP proteins as predicted by signalP3 was improved by the additional implementation of a rule-based classifier constructed from single genome signalP predictions of the best homologs combined with a simple decision rule. Conflicting predictions are mostly due to inaccurate gene-models. Re-evaluation of the CDS by N-terminal alignment showed that selection of an alternative start codon in the same reading frame is in most cases sufficient to obtain an agreement. For putative SP proteins that do not have clear homologs in the proteomes of the related species and thus depend on signalP predictions only, an accuracy of 85% can be expected. Proteogenome analysis of secretome enriched fractions subsequently provided evidence for secretion of at least 209 of these predicted SP proteins in our data set, whereas about 40 additional predicted SP proteins were identified in the data sets from Tsang *et al.* (2009) and Lu *et al.* (2010).

The *A. niger* secretome responds dynamically to changes of the carbon source. The majority of the detected carbohydrate modifying enzymes are present under both sorbitol and GalA growth conditions. However, the relative contribution of the individual enzymes significantly changed with the carbon source. As was already evident from transcriptome data (Martens-Uzunova & Schaap, 2008; Martens-Uzunova & Schaap, 2009), the pectinolytic system is most strongly induced under the GalA growth conditions, where 22 of 30 proteins are either solely present or significantly more abundant in samples from GalA cultures. The most prominent difference between the growth and starvation conditions is the relative contribution of a number of abundant proteases, which levels increase even further under starvation conditions. However, a few other proteases are exclusively detected upon growth on sorbitol or GalA.

Although a broad spectrum of non-SP proteins was identified in the secretome enriched fractions, the relative contribution of lysis was very limited, even under starvation conditions. Still, relative high concentrations of two non-SP proteins with a putative extracellular function, An01g09980 and An01g00370, were detected. Most probable, these proteins are exported outside the cell by active transport mechanisms, indicating that a non-classical secretion pathway operates in *A. niger*. Further experimental validation of this pathway will be required by more detailed analysis of trafficking of these proteins.

ACKNOWLEDGEMENTS

Part of this project was carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation, which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

CHAPTER 4

METABOLOMICS AS A TOOL FOR TARGET IDENTIFICATION IN STRAIN IMPROVEMENT: THE INFLUENCE OF PHENOTYPE DEFINITION

Machtelt Braaksma, Sabina Bijlsma, Leon Coulier, Peter J. Punt
and Mariët J. van der Werf

This chapter has been accepted for publication in:
Microbiology (2010), doi:10.1099/mic.0.041244-0

Supplementary data will be available with the online version of this paper at http://mic.sgmjournals.org/

ABSTRACT

For the optimization of microbial production processes, the choice of the quantitative phenotype to be optimized is crucial. For instance, for the optimization of product formation either product concentration or productivity can be pursued, potentially resulting in different targets for strain improvement. The choice of a quantitative phenotype is not only highly relevant for classical improvement approaches, but even more so for modern systems biology approaches.

In this study, the information content of a metabolomics data set was determined with respect to different quantitative phenotypes related to the formation of specific products. To this end, the production of two industrially relevant products by *Aspergillus niger* was evaluated; (i) the enzyme glucoamylase and (ii) the more complex product group of secreted proteases, consisting of multiple enzymes. For both products six quantitative phenotypes associated with activity and productivity were defined, taking also into account different time points of sampling during the fermentation. Both linear and non-linear relations between the metabolome data and the different quantitative phenotypes were considered.

The multivariate data analysis tool partial least squares (PLS) was used to evaluate the information content of the data sets for all the different quantitative phenotypes defined. Depending on the product studied, different quantitative phenotypes were found to have the highest information content in specific metabolomics data sets. A detailed analysis of the metabolites showing strong correlation with these quantitative phenotypes revealed that for glucoamylase activity various sugar-derivatives were found to be correlating. For the reduction of protease activity mainly as yet unidentified compounds were found to be correlating.

INTRODUCTION

The optimization of microbial production processes is an ongoing cycle of strain and/or process improvement. Traditionally, prior knowledge is the basis for identifying putative bottlenecks in the process. However, with the use of functional genomics technologies a more unbiased approach towards target selection for metabolic engineering or process optimization can be applied (van der Werf, 2005).

For optimization of the production process of a biological compound or enzymatic activity, a broad range of definitions of phenotypes can be selected for improvement. For instance, in studies reporting the production of glucoamylase by the filamentous fungus *Aspergillus niger* many different quantitative phenotypes for glucoamylase production were used. These included glucoamylase concentration (in g l^{-1}) (Withers *et al.*, 1998), activity (in U l^{-1}) (Wang *et al.*, 2008), yield (in mol product mol^{-1} substrate) (Melzer *et al.*, 2007), specific concentration or activity (in g g^{-1} DWT or U g^{-1} DWT, respectively) (Swift *et al.*, 2000, Pedersen *et al.*, 2000; Schrickx *et al.*, 1993), and specific productivity (in mol, gram or units g^{-1} DWT h^{-1}) (Melzer *et al.*, 2007; Withers *et al.*, 1998; Schrickx *et al.*, 1993).

The motivation for choosing a certain quantitative phenotype in bioprocess optimization is not always clear, and seems largely *ad libitum*. The choice of the quantitative phenotype to be pursued may have a major influence on the outcome of an optimization strategy. As stated by Kennedy & Krouse (1999) in their review on strategies for improving fermentation medium performance, some medium design studies flounder because the target variable to be improved is not clearly defined. Phenotype definition is not only important for classical optimization approaches, but perhaps even more so for modern, top-down systems biology approaches. In particular, as the enormous quantity of data that arise from these systems biology studies may easily result in a data overload (Braaksma *et al.*, 2010a). However, as far as we know, no systematic studies have been performed to study which quantitative phenotype is the most relevant in bioprocess optimization.

In bioprocess optimization a high quantity, e.g. concentration, of a product is not automatically the most desired result. In the case the substrate is an expensive part of the total fermentation costs, a high yield may be more relevant. However, improvement of the product yield is not always achieved by focussing on the yield itself during the strain improvement process. Focussing on the productivity may require fewer strain improvement steps during a particular bioprocess optimization process, thus resulting in an improved yield more quickly. Reduction of the fermentation time is another parameter to reduce production costs and can be realized by increasing the productivity. It is very likely that selection of either of these phenotypes for optimization will result in different targets to obtain the desired increase.

In this study, a metabolomics approach was used for target selection for process optimization and/or metabolic engineering of the host. Culture samples from *A. niger* fermentations were analyzed for the production of glucoamylase and protease. For both products different quantitative phenotypes associated with activity and productivity were defined. In a first step, we determined the information content of our metabolomics data set with respect to different quantitative phenotypes associated with the formation of either of the two different products. Subsequently, metabolites were identified showing the strongest correlation with the phenotype studied.

METHODS

Strain and cultivation conditions

Aspergillus niger N402, a *cspA1* (conferring short conidiophores) derivative of ATCC 9029 (Bos *et al.*, 1988), was used in this study.

Cultures were grown in batch fermentations in BioFlo 3000 (New Brunswick Scientific) bioreactors with a 5 litre working volume. Minimal medium (Bennett & Lasure, 1991) contained 7 mM KCl, 11 mM KH_2PO_4, 2 mM $MgSO_4$, 76 nM $ZnSO_4$, 178 nM H_3BO_3, 25 nM $MnCl_2$, 18 nM $FeSO_4$, 7.1 nM $CoCl_2$, 6.4 nM $CuSO_4$, 6.2 nM Na_2MoO_4 and 134 nM EDTA. This medium was supplemented with the appropriate carbon source or nitrogen source in concentrations as indicated below. To prevent foaming, 1 % (v/v) antifoam (Struktol J 673) was added to the medium and, when necessary, additional antifoam was added during the cultivation. The medium composition, cultivation conditions and operating procedure of the bioreactor have been described in detail previously (Braaksma *et al.*, 2009). Cultivations were performed according to a full factorial design (total 16 conditions, and 9 biological duplicates), varying the carbon source (277.5 mM glucose or 333.0 mM xylose), the nitrogen source (ammonium chloride or sodium nitrate), the nitrogen concentration (low (282.4 mM) or high (564.8 mM)), and the pH (4 or 5) (Braaksma *et al.*, 2009).

Enzyme assays

Protease activity. Extracellular proteolytic activities were measured at an assay pH of 4 as described previously (Braaksma *et al.*, 2009).

Glucoamylase activity. Glucoamylase activity was measured using PNPG (*p*-nitrophenyl α-D-glucopyranoside) (Sigma-Aldrich) as a substrate (Withers *et al.*, 1998). The procedure was fully automated using a COBAS MIRA Plus autoanalyser. 30 µl of cleared culture supernatant was incubated with 90 µl 0.1% (w/v) PNPG in 0.1 M sodium acetate buffer, pH 4.3, for 20 min. at 37 °C. The reaction was terminated by the addition of 135 µl 0.1 M borate buffer, pH 9.3, and the absorbance was read at 405 nm. One unit of glucoamylase activity was defined as the amount of enzyme that produces an absorbance at 405 nm equivalent to 1 µmol/l of *p*-nitrophenol in 1 minute under the given assay conditions.

Collection of samples, extraction and sample clean-up

Samples for metabolome analysis (25-100 ml, depending on the dry weight concentration) were taken rapidly from the bioreactor by closing the gas outlet and opening the sampling port. Cells were immediately quenched at -45 °C in methanol and collected as described previously (Pieterse *et al.*, 2006). Cell pellets were stored at -45 °C until use. To allow correlation of the metabolite concentrations to cell dry weight, the internal standards phenylalanine-d_5, leucine-d_3 (Spectral Stable Isotopes, Columbia, USA) and labelled $^{13}C_{10}$,$^{15}N_5$-GTP (Sigma-Aldrich, Zwijndrecht, the Netherlands) were added prior to extraction. The intracellular metabolites were extracted from the cell suspensions by chloroform extraction at -45 °C as described by Ruijter and Visser (Ruijter & Visser, 1996). The water/methanol phase was subsequently divided in two portions, one for GC- and one for LC-MS analysis. The LC-MS sample was deproteinized by filtration using a Microcon YM-10 (Millipore) filter centrifuged at 18000 g and -20 °C for 16 hours. Subsequently, all samples were lyophilized. To allow correction for the recovery of amino acids, the group of metabolites most susceptible to matrix effects (i.e. the effect that in complex samples the detection of some compounds is disturbed in the presence of other compounds), prior to lyophilizing the samples for GC-MS an internal standard mixture of ^{2}D,^{15}N-labeled amino acids (Spectra Stable Isotopes) was added.

Biomass determination

Cell culture samples. For the quantification of cell dry weight (DWT), a known volume of cell culture was filtered though a dried, pre-weighted filter paper, followed by washing with distilled water twice and then drying at 110 ºC for 24 h.

Metabolome samples. The extracted mycelium was collected and dried at 110 °C for 24 h to determine the dry weight of the sample (Ruiter & Visser, 1996). The metabolite concentrations in the extracts were correlated to dry weight by the use of the above mentioned internal standards added prior to the extraction of the cell pellets.

Analytical procedures

IP-LC-MS method. Lyophilized metabolome samples were dissolved in 100 µl methanol/water (1:3 v/v) and analyzed as described by Coulier *et al.* (Coulier *et al.*, 2006). Samples (10 or 20 µl) were separated on a reversed phase column (Chrompack Inertsil 5 mm ODS-3 100 x 3 mm, Middelburg, The Netherlands) using a 40 min linear gradient from 100% 5 mM hexylamine (pH 6.3) to 100% of 90% methanol-10 mM ammonium acetate (pH 8.5) at a flow rate of 0.4 ml min^{-1}. Compounds were detected by electrospray ionization (negative ion mode) in the range m/z 150/1000 using a Thermo Finnigan LTQ linear ion-trap system (Thermo Electron Corp. San Jose, USA). During data acquisition, the mass spectrometer probe voltage was maintained at 3–4 kV, the heated capillary was kept at 250 °C.

RP-LC-MS method. After analysis with the IP-LC-MS method, the redissolved metabolome samples were used for analysis with the RP-LC-MS method. Samples (10 or 20 µl) were separated on a reversed phase column (Waters Sunfire C18, 150 x 3 mm, 3.5 µm) using a linear gradient from 100% water + 0.1% formic acid to 75% MeCN/water (80%/20%) + 0.1% formic acid in 18 minutes followed by a linear gradient to 100% MeCN/water (80%/20%) + 0.1% formic acid in 10 minutes at a flow rate of 0.3 ml min^{-1}. Compounds were detected by electrospray ionization (ESI; positive ion mode) in the range m/z 150-2000.

OS-GC-MS method. Lyophilized metabolome samples were derivatized using a solution of ethoxyamine hydrochloride in pyridine as the oximation reagent followed by silylation with *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) as described by Koek *et al.* (Koek *et al.*, 2006). Before silylation, dicyclohexylphthalate (Sigma-Aldrich) was added as an internal standard for injection. GC-MS-analysis of the derivatized samples was performed using a temperature gradient from 70 °C to 320 °C at a rate of 10 °C min^{-1} on an Agilent 6890 N GC and an Agilent 5973 mass selective detector (Agilent, Palo Alto, USA). 1 µl aliquots of the derivatized samples were injected splitless on a HP5-MS capillary column (30 m x 0.25 mm, 0.25 µm film thickness, Agilent). Detection was performed using MS detection in electron impact mode (70 eV).

Data preprocessing

The LC-MS data were converted to .cdf-files and imported in Matlab (version 7.7.0.471 (R2008b), The Mathworks, Inc., Natick, MA). The homemade software packages Impress V1.2, Winlin V2.4 and Equest V2.3XP (Vogels *et al.*, 1996; van der Greef *et al.*, 2004) were used to align and peak-pick the LC-MS data. Following preprocessing, all peaks in the obtained target tables (in the form of peak identifiers [mass.retention time] and peak areas) were normalized with respect to the amount of extracted biomass per sample.

Also the data from the GC-MS analyses were converted into target tables, i.e. spreadsheets containing relative peak areas for all significant metabolite peaks in all samples. Peak areas were obtained by automated peak integration, followed by manual inspection. To several of the peaks a (partial) chemical identity could be assigned by comparing retention time and mass spectrum with an in-house database, otherwise a unique peak identifier [AN codes] was assigned. All peak areas were corrected for the recovery of the internal standard for injection. Subsequently, the amino acids were corrected for the recovery of the labeled amino acids. Finally, peaks were normalized with respect to the amount of extracted biomass per sample.

Both preprocessed LC-MS and GC-MS data files were combined in one data matrix. As the presence of values equal to zero can disturb the statistical analysis, prior to this, a so-called 25%-rule was applied: only those variables were retained which were present in at least 25% of the samples (Rubingh *et al.*, 2009; Bijlsma *et al.*, 2006). Next, all remaining zero values in the separate GC-MS, IP-LC-MS and RP-LC-MS data sets were replaced by a threshold value of half the lowest value in the data set unequal to zero (Rubingh *et al.*, 2009). In total 489 individual peaks, i.e. 131 GC-MS, 176 IP-LC-MS and 182 RP-LC-MS peaks, were retained in the final data sets to be used as input for multivariate data analysis MVDA.

Multivariate data analysis

Before data analysis, the curves with glucoamylase and protease activity were corrected for noise and possible outliers using a smoothing algorithm as described previously (Braaksma *et al.*, 2009). The phenotype data, e.g. protease or glucoamylase activity or productivity, were mean-centred [$(x - \bar{x})$] prior to MVDA in order to remove the overall offset from the data (van den Berg *et al.*, 2006). The metabolome data set was mean-centred and, in order to compare the metabolites relative to the biological response range, it was subsequently range scaled [$(x_i - \bar{x})/(x_{max} - x_{min})$] prior to MVDA (van den Berg *et al.*, 2006). PLS analysis were performed in the Matlab environment using the PLS Toolbox (version 5.0.3, 2008; Eigenvector Research, Manson, WA). The PLS results were cross-validated by using a tenfold single cross validation procedure. In addition to PLS analysis on the original metabolome and phenotype data, PLS analysis was also performed after either natural logarithm transformation of the phenotype data in combination with the original metabolome data or after natural logarithm transformation of the metabolome data in combination with the original phenotype data. An automatic procedure was written in Matlab code in order to run the many PLS models in a short time. Every generated PLS model was inspected manually to judge if the number of latent variables (LV's) chosen by the algorithm seemed appropriate with respect to the Root Mean Square Error of Cross Validation (RMSECV) curve. In general, if more LV's are included in the PLS model, the given model will contain more noise. In the case too many LV's were chosen by the algorithm, a new PLS model was generated by choosing a smaller number of LV's.

Compound identification

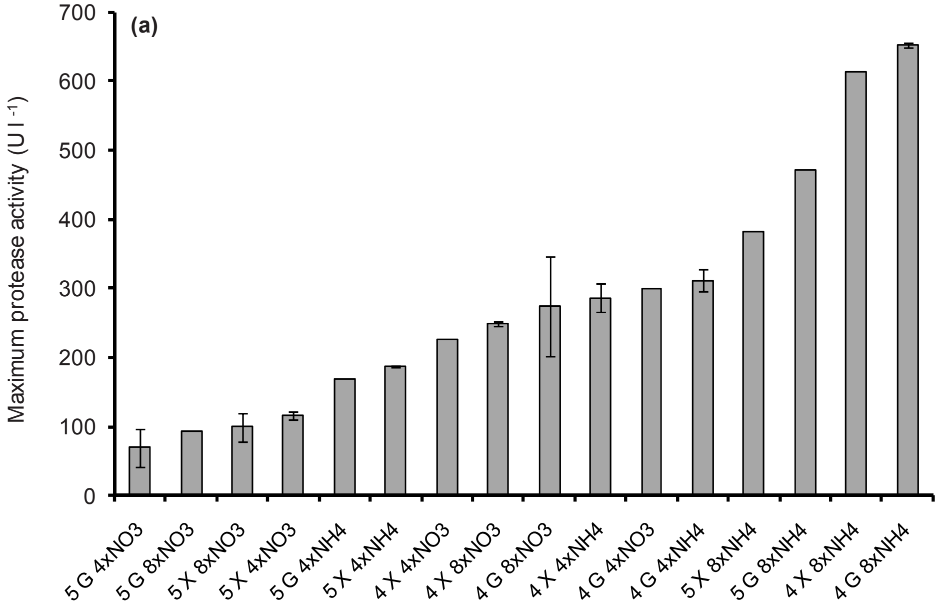
The identity of relevant peaks was established by verifying peak retention time and mass spectrum against in-house and public databases. If a peak could not be identified in this way, in several cases it was subsequently reanalyzed using high resolution and/or tandem mass spectrometry (MS/MS) analytical instruments (van der Werf *et al.*, 2007).

RESULTS

Experimental setup

In order to evaluate whether the definition of the phenotype used influences the outcome of a metabolomics study, or for that matter any optimization approach, the production of two industrially relevant products, i.e. glucoamylase and proteases, by *A. niger* was studied. To this end, *A. niger* was grown at sixteen different environmental conditions, with nine randomly selected biological duplicates (see also Braaksma *et al.*, 2009). Samples for metabolome analyses were taken at three different time points of the growth curve based on cell dry weight concentrations. One sample was collected at the middle of the logarithmic growth phase (mid log), one at the end of the logarithmic growth phase (late log) and one during the stationary growth phase. Samples were immediately quenched in a methanol solution to prevent alterations in the metabolite composition of the samples. Subsequently, the metabolites were extracted from the cells under quenched conditions, and the metabolites present were analyzed using three analytical methods (see Methods section).

The production of glucoamylase and protease was monitored during the course of the fermentation by analyzing culture samples every six hours. The variation in maximum protease and glucoamylase activities under the different experimental conditions is shown in Fig. 1. For protease activity the variation is evenly distributed over the different experimental conditions (Braaksma *et al.*, 2009). For glucoamylase the experiments can be clearly separated in two groups. One group with very low activities of conditions where the fungus was grown under non-induced conditions (on xylose) and another group with high activities of growth under induced conditions (on glucose).



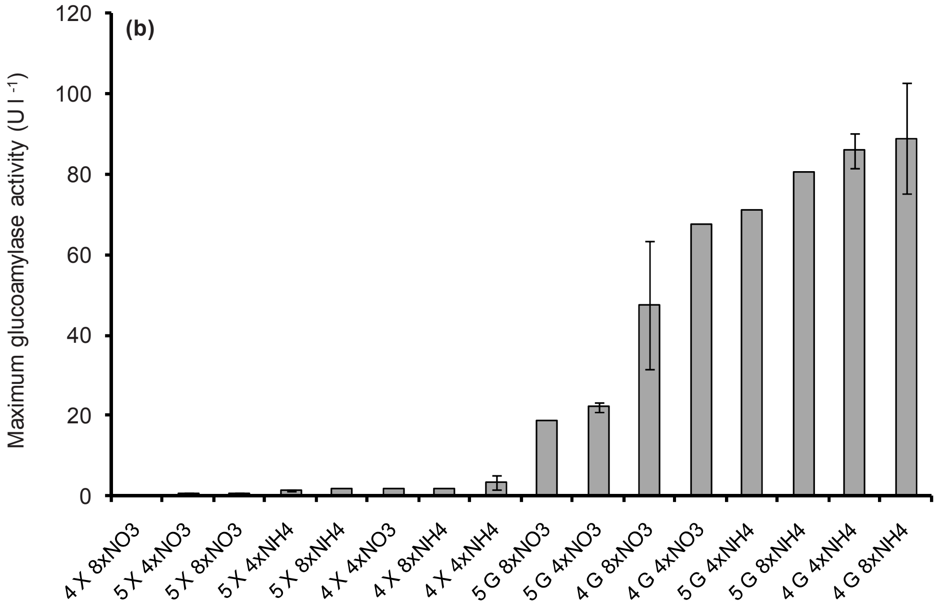


Fig. 1. (A) Maximum protease activity and (B) maximum glucoamylase activity in the different fermentations.

Quantitative phenotypes

Six different quantitative phenotype values for the three different products were determined. Glucoamylase and protease were expressed as activity (see A in Fig. 2), and for both products the rate of production, i.e. the productivity (see B in Fig. 2), was calculated. However, the amount of product formed also depends on the biomass concentration (DWT). Therefore, specific activity and specific productivity were also determined. These two specific phenotypes were calculated using the DWT at the time point of sampling (see A1 and B1, respectively, in Fig. 2). However, when a sample was collected during the stationary phase of the fermentation, the biomass concentration may already be declining due to autolysis of the fungal cells (White *et al.*, 2002), thus making specific activity and specific productivity dependent on the degree of lysis. Therefore, both phenotypes were also calculated in relation to the maximum biomass concentration (DWT_{max}) (see A2 and B2, respectively, in Fig. 2). By using DWT_{max}, the phenotypic value is not artificially increased when in certain fermentations severe cell lysis had occurred. In addition to the phenotypes described above, similar quantitative phenotypes values were also calculated using the *maximum* activity or productivity for these products (see also Braaksma *et al.*, 2009). Thus, in this latter case, for all three metabolome time samples the phenotypic value was identical. For a detailed description of how each phenotype was defined and a complete overview of the phenotypic values corresponding to each metabolome sample, see Supplementary data file 1.

Analysis of the information content of the data set

The multivariate data analysis (MVDA) tool partial least squares (PLS) was used to determine the information content of the metabolome data sets for all the different quantitative phenotypes defined. PLS is a regression tool that results in a model that describes a quantifiable phenotype of interest, such as protease activity or productivity, based on the concentrations of each of the metabolites determined. In MVDA analysis of metabolomics data it is important to realize that due to the relatively large number of variables and few number of samples, chance correlations are a serious issue. Therefore, the cross-validated correlation coefficient, R^2_{CV}, obtained from a PLS model after cross validation, is a better measure for the information content of a PLS model than the initial correlation coefficient R^2_{fit}, because R^2_{CV} also reflects the robustness of the model. A high R^2_{CV} indicates a high information content of the metabolome data in relation to the quantitative phenotype. In this study, cross validated PLS models with a R^2_{CV} of 0.6 or higher were considered good statistical models. For both products, cross validated PLS models were made for all different quantitative phenotypes (Table 1).

To investigate whether the information content of the metabolomics data set was growth phase specific, PLS models of these six quantitative phenotypes were calculated by including the metabolome data of different time samples in the PLS model. PLS models were determined using metabolome data of all three samples generated from the different fermentations as well as with the metabolome data of only the samples collected at one of the growth phases during the fermentation. In addition, also PLS models were generated evaluating non-linear relations between the quantitative phenotype and the metabolome data, in order to identify metabolites with a non-linear relation to the studied phenotype. An overview of the PLS models generated from the metabolome data of this study, including the R^2_{CV} of each model, is shown in Table 1.

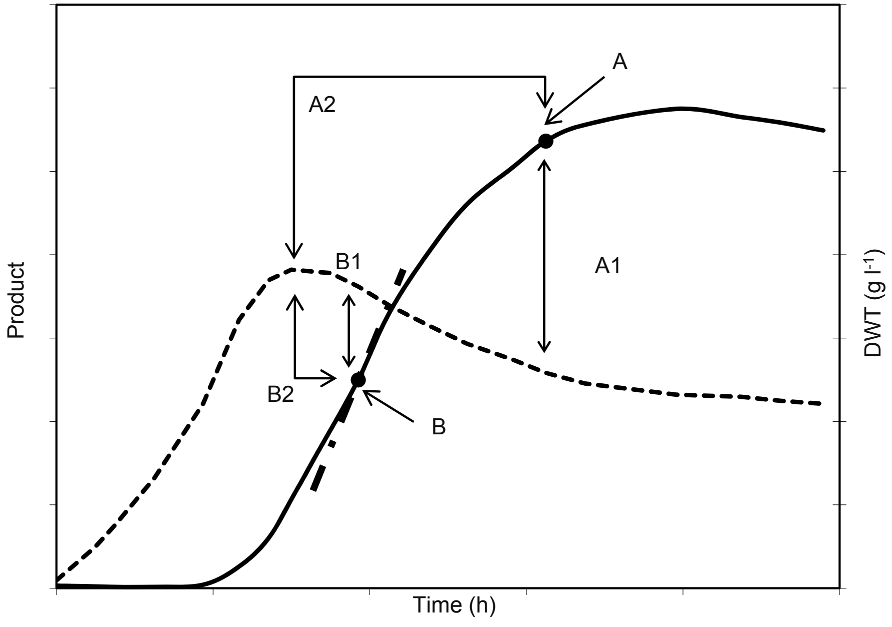


Fig. 2. A schematic representation of production in time to illustrate the various product-related phenotypes that can be defined. Solid line, product; dashed line, biomass concentration DWT. (A) activity at time point of sampling; (A1) specific activity – 1, based on the biomass at the time point of sampling; (A2) specific activity – 2, based on the maximal biomass concentration during the fermentation; (B) productivity at time point of sampling; (B1) specific productivity – 1, based on the biomass at the time point of sampling; (B2) specific productivity – 2, based on the maximal biomass concentration during the fermentation. (Adapted from Braaksma *et al.* (2009), Microbiology 155, 3430-3439.)

Table 1. Overview of the cross validation values (R^2_{CV}) of the PLS models made for glucoamylase (A) and protease (B).

Models with a R^2_{CV} of 0.6 or higher are considered good statistical models and are indicated in bold.

Glucoamylase							
Table 1A	**Phenotype ***	**P**	**R^2_{CV}**	**LN(P)**	**R^2_{CV}**	**LN(M)**	**R^2_{CV}**
Maximum phenotype, metabolome data of all samples	Max.Act.	G1	0.59	**G49**	**0.66**	**G97**	**0.75**
	Max.Spec.Act.-1	G2	0.47	**G50**	**0.64**	**G98**	**0.64**
	Max.Spec.Act.-2	G3	0.59	**G51**	**0.64**	**G99**	**0.77**
	Max.Prod.	G4	0.59	**G52**	**0.67**	**G100**	**0.73**
	Max.Spec.Prod.-1	**G5**	**0.60**	**G53**	**0.63**	**G101**	**0.78**
	Max.Spec.Prod.-2	G6	0.59	**G54**	**0.65**	**G102**	**0.74**
Maximum phenotype, metabolome data of mid log samples	Max.Act.	**G7**	**0.71**	**G55**	**0.76**	**G103**	**0.77**
	Max.Spec.Act.-1	G8	0.47	**G56**	**0.72**	**G104**	**0.67**
	Max.Spec.Act.-2	**G9**	**0.62**	**G57**	**0.74**	**G105**	**0.71**
	Max.Prod.	**G10**	**0.79**	**G58**	**0.75**	**G106**	**0.82**
	Max.Spec.Prod.-1	**G11**	**0.71**	**G59**	**0.74**	**G107**	**0.82**
	Max.Spec.Prod.-2	**G12**	**0.73**	**G60**	**0.74**	**G108**	**0.82**
Maximum phenotype, metabolome data of late log samples	Max.Act.	G13	0.43	**G61**	**0.63**	G109	0.50
	Max.Spec.Act.-1	G14	0.42	**G62**	**0.62**	G110	0.48
	Max.Spec.Act.-2	G15	0.43	**G63**	**0.61**	G111	0.49
	Max.Prod.	**G16**	**0.60**	**G64**	**0.72**	G112	0.57
	Max.Spec.Prod.-1	**G17**	**0.67**	**G65**	**0.71**	**G113**	**0.66**
	Max.Spec.Prod.-2	**G18**	**0.61**	**G66**	**0.70**	G114	0.58
Maximum phenotype, metabolome data of stationary samples	Max.Act.	G19	0.00	G67	0.01	G115	0.41
	Max.Spec.Act.-1	G20	0.01	G68	0.03	G116	0.45
	Max.Spec.Act.-2	G21	0.03	G69	0.04	G117	0.44
	Max.Prod.	G22	0.02	G70	0.01	G118	0.40
	Max.Spec.Prod.-1	G23	0.00	G71	0.03	G119	0.44
	Max.Spec.Prod.-2	G24	0.00	G72	0.02	G120	0.39
Phenotype at time point of sampling, metabolome data of all samples	Act.	G25	0.40	**G73**	**0.68**	G121	0.51
	Spec.Act.-1	G26	0.38	**G74**	**0.67**	G122	0.48
	Spec.Act.-2	G27	0.41	**G75**	**0.66**	G123	0.53
	Prod.	G28	0.55	G76	0.59	**G124**	**0.69**
	Spec.Prod.-1	G29	0.56	G77	0.57	**G125**	**0.66**
	Spec.Prod.-2	G30	0.59	G78	0.57	**G126**	**0.68**
Phenotype at time point of sampling, metabolome data of mid log samples	Act.	**G31**	**0.67**	**G79**	**0.69**	**G127**	**0.67**
	Spec.Act.-1	**G32**	**0.63**	**G80**	**0.69**	**G128**	**0.67**
	Spec.Act.-2 †	**G33**	**0.63**	**G81**	**0.69**	**G129**	**0.67**
	Prod.	**G34**	**0.78**	**G82**	**0.69**	**G130**	**0.78**
	Spec.Prod.-1	**G35**	**0.77**	**G83**	**0.70**	**G131**	**0.81**
	Spec.Prod.-2 †	**G36**	**0.77**	**G84**	**0.70**	**G132**	**0·81**
Phenotype at time point of sampling, metabolome data of late log samples	Act.	G37	0.22	G85	0.49	G133	0.30
	Spec.Act.-1	G38	0.23	G86	0.48	G134	0.33
	Spec.Act.-2 †	G39	0.23	G87	0.48	G135	0.33
	Prod.	G40	0.33	G88	0.28	G136	0.42
	Spec.Prod.-1	G41	0.29	G89	0.25	G137	0.38
	Spec.Prod.-2 †	G42	0.29	G90	0.25	G138	0.38
Phenotype at time point of sampling, metabolome data of stationary samples	Act.	G43	0.04	G91	0.00	G139	0.34
	Spec.Act.-1	G44	0.01	G92	0.01	G140	0.34
	Spec.Act.-2	G45	0.02	G93	0.01	G141	0.37
	Prod.	G46	0.05	G94	0.02	G142	0.40
	Spec.Prod.-1	G47	0.01	G95	0.02	G143	0.38
	Spec.Prod.-2	G48	0.01	G96	0.02	G144	0.40

Table 1. Continued.

Protease							
Table 1B	**Phenotype ***	**P**	**R^2_{CV}**	**LN(P)**	**R^2_{CV}**	**LN(M)**	**R^2_{CV}**
Maximum phenotype, metabolome data of all samples	Max.Act.	**P1**	**0.70**	**P49**	**0.75**	**P97**	**0.78**
	Max.Spec.Act.-1	**P2**	**0.66**	**P50**	**0.66**	**P98**	**0.72**
	Max.Spec.Act.-2	P3	0.57	**P51**	**0.60**	**P99**	**0.66**
	Max.Prod.	**P4**	**0.71**	**P52**	**0.69**	**P100**	**0.80**
	Max.Spec.Prod.-1	P5	0.58	P53	0.50	**P101**	**0.63**
	Max.Spec.Prod.-2	P6	0.58	P54	0.48	**P102**	**0.65**
Maximum phenotype, metabolome data of mid log samples	Max.Act.	P7	0.46	**P55**	**0.72**	P103	0.47
	Max.Spec.Act.-1	P8	0.38	P56	0.58	P104	0.32
	Max.Spec.Act.-2	P9	0.28	P57	0.55	P105	0.28
	Max.Prod.	P10	0.51	**P58**	**0.69**	P106	0.43
	Max.Spec.Prod.-1	P11	0.28	P59	0.44	P107	0.16
	Max.Spec.Prod.-2	P12	0.29	P60	0.45	P108	0.18
Maximum phenotype, metabolome data of late log samples	Max.Act.	P13	0.52	**P61**	**0.65**	**P109**	**0.62**
	Max.Spec.Act.-1	P14	0.37	P62	0.48	P110	0.47
	Max.Spec.Act.-2	P15	0.42	P63	0.49	P111	0.47
	Max.Prod.	P16	0.48	P64	0.59	P112	0.44
	Max.Spec.Prod.-1	P17	0.28	P65	0.30	P113	0.17
	Max.Spec.Prod.-2	P18	0.29	P66	0.34	P114	0.18
Maximum phenotype, metabolome data of stationary samples	Max.Act.	P19	0.11	P67	0.19	**P115**	**0.68**
	Max.Spec.Act.-1	P20	0.11	P68	0.25	P116	0.58
	Max.Spec.Act.-2	P21	0.11	P69	0.14	P117	0.59
	Max.Prod.	P22	0.17	P70	0.14	**P118**	**0.60**
	Max.Spec.Prod.-1	P23	0.14	P71	0.19	P119	0.44
	Max.Spec.Prod.-2	P24	0.18	P72	0.18	P120	0.47
Phenotype at time point of sampling, metabolome data of all samples	Act.	**P25**	**0.70**	P73	0.57	**P121**	**0.80**
	Spec.Act.-1	**P26**	**0.66**	P74	0.46	**P122**	**0.75**
	Spec.Act.-2	**P27**	**0.67**	P75	0.44	**P123**	**0.77**
	Prod.	P28	0.45	**P76**	**0.65**	**P124**	**0.61**
	Spec.Prod.-1	P29	0.32	P77	0.49	P125	0.45
	Spec.Prod.-2	P30	0.36	P78	0.48	P126	0.45
Phenotype at time point of sampling, metabolome data of mid log samples	Act.	P31	0.09	P79	0.01	P127	0.17
	Spec.Act.-1	P32	0.03	P80	0.01	P128	0.05
	Spec.Act.-2 †	P33	0.03	P81	0.01	P129	0.05
	Prod.	P34	0.21	P82	0.42	P130	0.18
	Spec.Prod.-1	P35	0.16	P83	0.24	P131	0.12
	Spec.Prod.-2 †	P36	0.16	P84	0.24	P132	0.12
Phenotype at time point of sampling, metabolome data of late log samples	Act.	P37	0.23	P85	0.29	P133	0.41
	Spec.Act.-1	P38	0.25	P86	0.09	P134	0.29
	Spec.Act.-2 †	P39	0.25	P87	0.09	P135	0.29
	Prod.	P40	0.49	P88	0.51	P136	0.51
	Spec.Prod.-1	P41	0.32	P89	0.23	P137	0.26
	Spec.Prod.-2 †	P42	0.32	P90	0.23	P138	0.26
Phenotype at time point of sampling, metabolome data of stationary samples	Act.	P43	0.18	P91	0.18	**P139**	**0.69**
	Spec.Act.-1	P44	0.20	P92	0.14	P140	0.57
	Spec.Act.-2	P45	0.19	P93	0.15	P141	0.59
	Prod.	P46	0.18	P94	0.39	P142	0.55
	Spec.Prod.-1	P47	0.05	P95	0.45	P143	0.38
	Spec.Prod.-2	P48	0.03	P96	0.45	P144	0.42

* For a detailed description of how each phenotype (P) was defined, see Supplementary data file 1. P is used to indicate models generated without LN transformation; LN(P) is used to indicate models generated after LN transformation of the phenotype; LN(M) is used to indicate models generated after LN transformation of the metabolome data.

† For these PLS models, the results for Spec.Act.-2 and Spec.Prod.-2 are identical to Spec.Act.-1 and Spec.Prod.-1, respectively. To calculate Spec.Act.-2 and Spec.Prod.-2 in principal DWT_{max} is used, except for samples collected before DWT_{max} was reached (as is the case for the mid log and late log samples). For these samples DWT at the time point of sampling was used, similar as for calculating Spec.Act.-1 and Spec.Prod.-1 (see also Supplementary data file 1).

Information content of the metabolomics data set with respect to the different quantitative phenotypes

About 44% of the PLS models generated for glucoamylase were considered good models ($R^2_{CV} \geq 0.6$); for protease, this was 19% (see Table 1). When comparing Tables 1A (glucoamylase) and 1B (protease) with each other, one thing is obvious: the highest information content of the metabolomics data set was obtained with different quantitative phenotypes for the different products. For glucoamylase good models were especially obtained when based on metabolome data of the samples from the mid log growth phase, while most good PLS models for protease were based on inclusion of metabolome data from all three time samples. Furthermore, LN transformation of either the metabolome data or the phenotype data resulted in general in an increased number of PLS models with $R^2_{CV} \geq 0.6$. In addition, more good PLS models were generated with the quantitative phenotype based on the *maximum* activity or productivity instead of the phenotype based at the activity or productivity *at the time point of sampling*. Moreover, for glucoamylase productivity resulted in more models with a R^2_{CV} above the cut-off of 0.6, while for protease on average the selection of activity (i.e., amount of product formed) as phenotype resulted in a somewhat higher number of good models.

Identification of metabolites that correlate with the phenotype studied

Metabolites contributing the most to, for instance, protease activity or productivity can be identified by ordering the (relative) statistical importance of the metabolites by virtue of the weight factors (regression factors) as determined in the PLS models for all metabolites. In other words, by applying PLS, metabolites important for a specific phenotype can be identified and ranked based on the strength of their correlation with the phenotype of interest. For both products, one good PLS model was chosen as starting point for analysing the strongest correlating metabolites in more detail. Based on this analysis subsequently lists of correlating metabolites from other good PLS models were compared.

Glucoamylase

For glucoamylase, most PLS models were above the threshold of $R^2_{CV} = 0.6$ when using metabolome data of the samples collected during the mid log growth phase. From this group of models, the PLS model in relation to maximum activity (PLS model G7), was selected as starting point for target identification and comparison to other good PLS models for glucoamylase. From this G7 PLS model, the 20 highest ranking metabolites are shown in Table 2. This top 20 included a relative high number of disaccharides and other sugar-derived compounds that were only present under glucoamylase inducing conditions (i.e. with glucose as carbon source). For all these dissacharides as well as some of the other compounds, such as DL-aminoadipic acid, 2,3-butanediol and xylitol the correlation is based on the absence of the compounds in all xylose samples and the presence in all glucose samples (Table 2). However, there is no clear correlation between their intracellular concentrations and maximum glucoamyase activity based on only the glucose samples (e.g. Fig. 3A). On the other hand, for putrescine, ornithine, glucose-6-phosphate, and fructose-6-phosphate there is a correlation between increasing intracellular levels of these compounds and maximum glucoamyase activity (e.g, see Fig. 3B).

When comparing the top 20's of other models with a R^2_{CV} ≥ 0.6 with each other, especially the use of metabolome samples from particular sampling times was of influence on the resulting top 20 (see Supplementary data file 2A). When either the metabolome data of all time samples was used (e.g. model G49), or only the metabolome data of the mid log or late log samples (models G55 and G61, respectively), only four metabolites are present in all three resulting top 20’s. These four metabolites were the compound tentatively identified as volemitol or perseitol, the compound tentatively identified as ribonic acid or xylonic acid, an unidentified disaccharide with a retention time of 42.02 min and another unidentified compound with ID AN 320-218 22.96 min (Supplementary data file 2A).

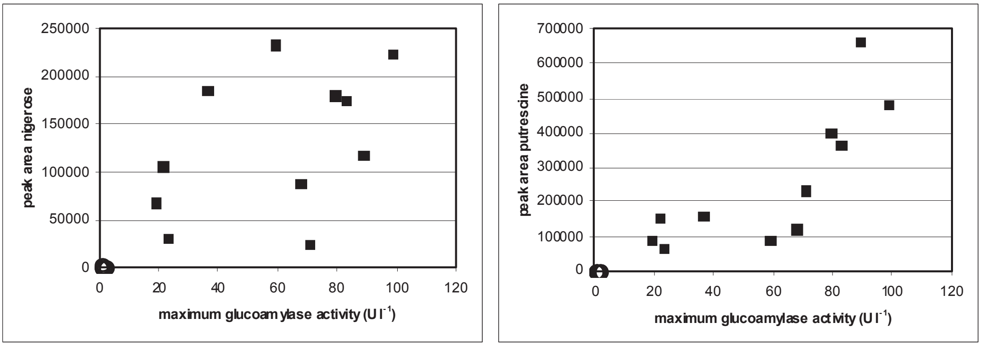


Fig. 3. Plot of the correlation between the metabolite tentatively identified as nigerose and maximum glucoamylase activity (A) and a similar plot for putrescine (B). ○, Metabolome samples from xylose fermentations (n=11); ■, metabolome samples from glucose fermentations (n=11).

Table 2. Twenty metabolites with the strongest correlation to glucoamylase as determined by PLS based on all mid log metabolome samples in relation to maximum activity (PLS model G7).

rank	metabolite ID *	tentative identity	regression factor	visual correlation to phenotype †
1	dissacharide 39.13 min	nigerose	+	+ ‡
2	C5 sugar alcohol	xylitol	+	+ ‡
3	DL-aminoadipic acid		+	+ ‡
4	putrescine		+	+
5	disaccharide 319-361	kojibiose	+	+ ‡
6	ornithine		+	+
7	disaccharide 40.41 min	isomaltose	+	+ ‡
8	disaccharide 40.89 min	isomaltose §	+	+ ‡
9	xylose		-	- | |
10	histidine		+	0
11	glucose-6-phosphate		+	+
12	glucose		+	+ ‡
13	fructose-6-phosphate		+	+
14	AN 292-333 24.26 min	ribonic acid or xylonic acid	-	- | |
15	AN 201 26.51 min	unknown	+	+ ‡
16	spermidine		+	0
17	tryptophan		+	0
18	glutamine		+	+
19	2,3-butanediol		+	+ ‡
20	uric acid		+	+ ‡

* All metabolites in this list were detected with the OS-GC-MS method.

†Visual correlation is indicated by + (positive correlation), – (negative correlation), or 0 (no apparent correlation); see also Supplementary data file 3A.

‡ Only or mainly high abundant on glucose, no apparent visual correlation within the glucose samples.

§ These are different mass fragments of the same compound.

| | Only high abundant on xylose.

The effect of LN transformation on the ranking of the potential targets was somewhat ambiguous. The effect of LN transformation of the phenotype or the metabolome data on the resulting top 20's was in several cases limited. For instance, for PLS models G7, G55 and G103 50% of the compounds were present in all three lists (for details, see Supplementary data file 2A). However, in other cases, i.e. PLS models G34, G82 and G130, this was only the case for 25% of the compounds (for details, see Supplementary data file 2A). The exact effect of LN transformation on the correlations of the metabolites with the phenotype was unclear; plotting of the peak areas of metabolites exclusively present in the top 20’s after LN transformation against the phenotype showed in some cases an improvement of the linear correlation, while in other cases the linear correlation deteriorated (data not shown).

Protease

For protease, most PLS models were above the threshold of $R^2_{CV} = 0.6$ when using the metabolome data of all three samples collected during the fermentation. The PLS model in relation to maximum activity, model P1, was selected from this group of models as starting point for target identification and comparison to other good PLS models for protease. From this PLS model, the 20 highest ranking metabolites are shown in Table 3. This top 20 mainly consisted of unidentified compounds detected by LC-MS, making interpretation of the results difficult. Two of the metabolites were tentatively identified as 2,3-dihydroxy-3-methylpentanoic acid and 2,3-dihydroxy-3-methylbutanoic acid, both known intermediates of the isoleucine and valine biosynthesis, respectively. A number of the compounds in the top 20 contained a phosphate-group; however, very little is known of a possible involvement of phosphorus sources on protease expression in aspergilli. In comparison to the glucoamylase results, the relative high contribution of compounds analyzed with the RP-LC-MS method was remarkable. Among others, RP-LC-MS is suitable for the detection of aromatic peptides and peptides larger than 4-5 amino acids, suggesting that at least some of the high ranked compounds could be peptide-derived. Unfortunately, for none of these compounds appropriate reference compounds are currently available to establish their exact identity.

When comparing the top 20's from good PLS models for protease with each other, the overall observations are in line with those for glucoamylase. Also for protease the largest differences between the top 20's were observed when comparing models which were based on different selections of the metabolome data, e.g. metabolome data of all time samples or only the metabolome data of mid log or late log samples (see Supplementary data file 2B for details). Furthermore, the influence on the resulting top 20's was very limited when using either activity or specific activity as phenotype. This is to be expected, given the strong correlation between activity and specific activity, or productivity and specific productivity. On the other hand, the effect of LN transformation of either the phenotype or the metabolome data was considerable, as the resulting top 20's showed 50% or less overlap with the top 20 without LN transformation (Supplementary data file 2B).

Table 3. Twenty metabolites with the strongest correlation to protease as determined by PLS based on all metabolome samples in relation to maximum activity (PLS model P1).

rank	metabolite ID *	tentative identity	regression factor	visual correlation to phenotype †
1	428.0417 (RP)	unknown	+	+
2	AN 110-336 13.53 min (GC)	unknown	+	+
3	phosphorylethanolamine related (GC)	unknown	+	0
4	712.1019 (RP)	unknown	+	+
5	AN 312 15.42 min (GC)	unknown	+	+
6	2,3-dihydroxy-3-methylpentanoic acid (GC)		+	+
7	223.0937 (IP)	monomethylphosphate	+	0
8	2,3-dihydroxy-3-methylbutanoic acid (GC)		+	+
9	AN 298-342 (GC)	unknown	+	+
10	AN 342-299 31.30 min (GC)	unknown	-	-
11	AN 211-283 20.80 min (GC)	unknown	+	0
12	446.0929 (IP)	monomethylphosphate ‡	+	0
13	monomethylphosphate (GC)		+	0
14	230.1734 (RP)	unknown	+	0
15	171.0420 (RP)	unknown	+	+
16	207.0929 (IP)	monomethylphosphate ‡	+	0
17	799.1182 (IP)	unknown	+	0
18	688.1035 (RP)	unknown	+	0
19	428.0743 (RP)	unknown	-	0
20	Adenosine (GC)		+	0

* The analytical method used to detect each metabolite is indicated in between brackets: GC, OS-GC-MS; IP, IP-LC-MS; and RP, RP-LC-MS.

† Visual correlation is indicated by + (positive correlation), – (negative correlation), or 0 (no apparent correlation); see also Supplementary data file 3B.

‡ These are different mass fragments of the same compound.

DISCUSSION

The choice for a certain quantitative phenotype in bioprocess optimization often seems rather random, but may have a major influence on the outcome of an optimization strategy. In this study, the information content of a metabolomics data set was determined with respect to different quantitative phenotypes related to the formation of two simple products, i.e. glucoamylase, and a more complex product, i.e. protease. When comparing the results of the two enzyme products glucoamylase and protease, it could be concluded that the information content of the metabolomics data set is higher for the simpler of these two products, i.e. glucoamylase. This is on the one hand remarkable, because the fermentation conditions from which the metabolome samples were collected in this study, were originally selected to result in large and evenly distributed variation in protease activity (Braaksma *et al.*, 2009).

Another important aspect influencing the information content of the metabolomics data set is the time point at which metabolome samples were collected. For instance, in this study the information content of the metabolome data from the mid log time samples was high in respect to glucoamylase (Table 1A), while it was low for protease (Table 1B). Based on this result, we conclude that data sets based on fewer experimental conditions but more metabolome samples in time may be more informative than a data set based on many experimental conditions and only one or a few time samples per condition. In addition, data sets based on more samples in time will allow the analysis of longitudinal effects in the data, i.e. metabolites whose correlation with product formation show a shift in time (Rubingh *et al.*, 2009).

Our results show that the effect of different ways to calculate the quantitative phenotype on the information content and resulting targets is much smaller than the effect of the time point of sampling. In general, the number of PLS models with a R^2_{CV} above the threshold value was higher when quantitative phenotypes were used that were based on the maximum activity or productivity instead of the activity or productivity at the time point of sampling (Table 1). A possible explanation for this is the more distinct variation in phenotypic values for the maximum phenotype. This may correlate better to the variation in the metabolome data present at a time point when phenotypic differences are perhaps not yet that clearly visible. Nevertheless, the effect of either maximum phenotype or phenotype at time point of sampling on the resulting top 20's is limited (Supplementary data file 2). This holds for the different description of the phenotype (e.g. activity versus productivity, or activity versus specific activity) as well. Conversely, the effect of LN transformation was considerable. Not only did the number of PLS models with a R^2_{CV} above the threshold value increase with LN transformation of the phenotype or the metabolome data, the resulting top 20's were often considerably different from the top 20 based on the data without LN transformation. However, it should be noted that it is difficult to interpret the effect of LN transformation, especially as it is not clear how LN transformation and data pretreatment methods (e.g. scaling methods such as range scaling) influence each other with regard to complex metabolome data (van den Berg *et al.*, 2006).

With the MVDA tool PLS the quantifiable phenotype of interest can be related to the metabolome data set as a whole and at the same time take into account the relationship between the metabolites (van der Werf *et al.*, 2007). Without this, it would be necessary to plot the metabolite concentrations of each metabolite against the phenotype in order to investigating the relation between individual metabolites and the quantifiable phenotype of interest. However, in case of a large number of metabolites, and as in our case a large number of phenotypes as well, this approach will result in an extremely large number of plots to analyze. Moreover, in such plots the intrinsic interdependency of the metabolites is neglected. However, despite these advantages of MVDA over a univariate approach, interpretation of the relation of the metabolites ranked by PLS to the quantifiable phenotype of interest is not straightforward. Several aspects, as listed below, have to be taken into account when interpreting the results of a PLS model.

(1) The positive or negative regression factors that are a measure for the contribution of a metabolite to the phenotype cannot be directly translated into how a metabolite actually correlates to the phenotype. These regression factors are not only a measure for the correlation of a single metabolite to the phenotype, but also for the correlation of this metabolite to other metabolites. Therefore, for a more detailed biological interpretation it is recommended to plot the concentrations of highly correlated metabolites against the quantifiable phenotype.

(2) Not all metabolites found to be correlating to the phenotype of interest are involved in the production of this product, either as inducer/inhibitor or precursor/side-product. With MVDA no distinction can be made between metabolites that correlate to the phenotype due to either a cause or an effect relation. For instance, one may conclude that the disaccharides found to be correlating to high glucoamylase activity (Table 2) induce glucoamylase secretion and thus cause the high activities. However, it is also possible that the identified disaccharides were formed from glucose by transglucosylation activity from glucoamylase (Nikolov *et al.*, 1989), and thus are an effect of glucoamylase activity (‘effect correlation’). For strain improvement in particular cause relations are of importance.

(3) Related to the previous subject is the occurrence of confounding effects, i.e. the situation that an extraneous factor correlates with both the phenotype and a metabolite. This can result in the false conclusion that there is a causal relationship between the phenotype and that specific metabolite. For example, there is only significant glucoamylase activity when *A. niger* is cultured on glucose instead of on xylose. Also several metabolites, such as uric acid and xylitol, are mainly present when *A. niger* is cultured on glucose. Therefore, one may conclude that there is a direct correlation between these metabolites and glucoamylase production. However, these

compounds may not be directly linked to glucoamylase production per se, but perhaps both glucoamylase and these metabolites independently correlate to growth on a specific carbon source.

(4) With the comprehensive analytical methods used in this study not only known compounds are analyzed, but also all peaks of unknown identity are included in the data set. One last aspect hampering the interpretation of the results of the data analysis is the correlation of these unidentified metabolites with the phenotype.

Taking into account the various aspects that influence the interpretation of the PLS results, as discussed above, specific metabolites identified as important to the question under study can be distilled from the initial list of potential targets that result from PLS. For optimization of glucoamylase production glucose-6-phosphate and fructose-6-phosphate are among the most likely targets. The enzyme glucose-6-phosphate isomerase catalyzes the conversion of glucose-6-phosphate into fructose-6-phosphate. The ratio between the concentrations of glucose-6-phosphate and fructose-6-phosphate is approximately a factor seven higher than expected based on the equilibrium constant for glucose-6-phosphate isomerase (data not shown). The relative accumulation of glucose-6-phosphate may on the one hand suggest that the activity of this enzyme is a bottleneck in the flux through the glycolysis. On the other hand, this aberration of the equilibrium may be required to obtain a sufficient flux in the direction of the pentose phosphate pathway (PPP), in order to generate sufficient NADPH. Melzer *et al.* (2007) also observed that under glucoamylase-producing conditions the flux of glucose through the PPP was higher than for non-producing conditions. However, in our study even under non-producing conditions the ratio between glucose-6-phosphate and fructose-6-phosphate concentrations is approximately a factor seven higher than expected. This weakens the hypothesis that the flux through the PPP may only be insufficient under glucoamylase-producing conditions, although when glucose was used as carbon source the absolute concentrations of both metabolites are higher. Alternatively, also absolute metabolite concentrations could be involved in regulation of metabolite fluxes (e.g. allosteric effects). All in all, in view of the crucial position of glucose-6-phosphate isomerase at the branch point between the glycolysis and the PPP, the regulation of the activity of this enzyme may be a means to regulate the fluxes through these two pathways and thus optimize glucoamylase production. Putrescine and ornithine are the two other most likely targets for optimization of glucoamylase production. Ornithine is the starting point for the synthesis of polyamines such as putrescine. Little is known about the actual function of putrescine and other polyamides in *A. niger*. In *A. nidulans*, there is an absolute requirement of polyamides in growth and development (Tabor &

Tabor, 1985; Jin *et al.*, 2002). The positive correlation between glucoamylase production and putrescine suggests that glucoamylase production may be stimulated by either addition of this polyamine to the medium or overexpression of the gene encoding ornithine decarboxylase, the enzyme responsible for the conversion of ornithine into putrescine.

No obvious targets were found in relation to protease production. Moreover, the majority of the compounds correlating to protease activity are unidentified compounds (Table 3). The presence of several compounds analyzed with the RP-LC-MS method in Table 3 suggests the possible involvement of small peptides in protease induction. Unfortunately, identification of peptides with the RP-LC-MS method has proved to be quite difficult, also because of the lack of appropriate reference compounds. Therefore, in order to further investigate the possible role of peptides in protease induction, additional methods will have to be deployed that offer more detailed information on the (partial) identity of peptides.

It was anticipated that the relation between intracellular metabolite concentrations and extracellular protease activity would not be straightforward, because extracellular protease activity is a complex phenotype, consisting of multiple enzyme activities. Recent analysis of the secretome of *A. niger* has indicated the presence of up to 20 different secreted proteases in the medium (Tsang *et al.*, 2009; Braaksma *et al.*, 2010b). Possibly, an approach with metabolomics alone is not sufficient for identifying targets for such a complex phenotype and an integrated systems biology approach is required.

Besides glucoamylase and protease production, also citric acid production was analysed as a phenotype. Although the experimental design of our data set was not optimally suited for this product, resulting in very few reliable PLS models (Supplementary data file 4), several TCA cycle intermediates (isocitrate, α-ketoglutarate) were identified as correlating with citric acid production (results not shown). Altogether, this study illustrates that with a combined metabolomics/MVDA approach relevant targets for strain and process improvement can be identified, as the relevance of several of the identified leads seem confirmed by what already is known in literature (e.g. the role of glucose-6-phosphate isomerase in glucoamylase production). Moreover, this study demonstrates the importance of experimental design in top-down systems biology studies, not only with regard to the fermentation conditions, but also with respect to the time point of sampling and the selection and calculation of the quantitative phenotype to be pursued.

ACKNOWLEDGEMENTS

The authors thank Karin Overkamp for her helpful input on the sample work-up, Maud Koek and Maarten Hekman for the GC-MS and LC-MS analysis, Bas Muilwijk and Richard Bas for identification of the metabolites, and Carina de Jong-Rubingh for the smoothing algorithm analyses and calculating the quantitative phenotypes. This project was carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation, which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

CHAPTER 5

IDENTIFICATION OF MODULES IN *ASPERGILLUS NIGER* BY GENE CO-EXPRESSION NETWORK ANALYSIS

Robert A. van den Berg*, Machtelt Braaksma*, Douwe van der Veen*,
Mariët J. van der Werf, Peter J. Punt, John van der Oost
and Leo H. de Graaff

** These authors contributed equally to this study*

Fungal Genetics and Biology (2010), 47: 539-550

Supplementary data are available with the online version of this paper at
http://www.sciencedirect.com/science/journal/10871845.

ABSTRACT

The fungus *Aspergillus niger* has been studied in considerable detail with respect to various industrial applications. Although its central metabolic pathways are established relatively well, the mechanisms that control the adaptation of its metabolism are understood rather poorly. In this study, clustering of co-expressed genes has been performed on the basis of DNA microarray data sets from two experimental approaches. In one approach, low amounts of inducer caused a relatively mild perturbation, while in the other approach the imposed environmental conditions including carbon source starvation caused severe perturbed stress. A set of conserved genes was used to construct gene co-expression networks for both the individual and combined data sets. Comparative analysis revealed the existence of modules, some of which are present in all three networks. In addition, experimental condition-specific modules were identified. Module-derived consensus expression profiles enabled the integration of all protein-coding *A. niger* genes to the co-expression analysis, including hypothetical and poorly conserved genes. Conserved sequence motifs were detected in the upstream region of genes that cluster in some modules, e.g., the binding site for the amino acid metabolism-related transcription factor CpcA as well as for the fatty acid metabolism-related transcription factors, FarA and FarB. Moreover, not previously described putative transcription factor binding sites were discovered for two modules: the motif 5'-CGACAA is overrepresented in the module containing genes encoding cytosolic ribosomal proteins, while the motif 5'-GGCCGCG is overrepresented in genes related to 'gene expression', such as RNA helicases and translation initiation factors.

INTRODUCTION

Genome-wide gene expression levels as generated by DNA microarray technology give insight into the behavior of individual genes at the cellular level. Such expression levels can be considered as a reflection of the physiological state of an organism and can be used to reveal details of metabolic regulation. The first fungal microarray studies were reported for the model organism *Saccharomyces cerevisiae* (DeRisi *et al.*, 1997; Lashkari *et al.*, 1997), followed by application of this technology for over 20 filamentous fungi including *Aspergillus niger* (Breakspear & Momany, 2007). *A. niger* is of industrial importance as it is the major production organism for citric acid world-wide (Magnuson & Lasure, 2004) and an efficient producer of both homologous and heterologous proteins (Pel *et al.*, 2007; Punt *et al.*, 2002). So far, *A. niger* transcriptome studies have been used to characterize polysaccharide-degrading enzyme systems (Andersen *et al.*, 2008; Jørgensen *et al.*, 2009; Martens-Uzunova & Schaap, 2008; van der Veen *et al.*, 2009; Yuan *et al.*, 2008a,b), to describe spatial colony development (Levin *et al.*, 2007), and to study the *A. niger* response towards reductive stress (Guillemette *et al.*, 2007) or cell wall damage (Meyer *et al.*, 2007).

Most DNA microarray studies, including all above-mentioned *A. niger* studies, focus on differential gene expression between few experimental conditions. However, solely an observation of fold changes of expression of individual genes does not explain how biological processes work together to achieve the cell's objectives. Additional information regarding the activation and co-operation of biological processes can be obtained by comparing gene expression profiles over a range of conditions. For example, genes that encode subunits of a protein complex may have a consistently similar change of expression levels over many conditions. The similar expression of two or more genes over a range of conditions is referred to hereafter as gene co-expression.

Featherstone and Broadie deduced from a *S. cerevisiae* gene co-expression study (Featherstone and Broadie, 2002) that specific sets of genes interact extensively at the level of gene expression, and thus can be described in terms of an interconnected network. Such gene co-expression networks can provide a large-scale, global view of the transcriptional response of an organism.

A comparison of gene co-expression networks constructed from DNA microarray data of the evolutionary distinct organisms *S. cerevisiae*, *Homo sapiens*, *Escherichia coli*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* indicated that these networks share common structural, or topological, properties (Bergmann *et* *al.*, 2004). For example, the observed gene co-expression networks consist of co-expressed groups of genes, termed clusters or modules, which are associated with the same cellular function. However, while modules of genes involved in similar cellular functions were identified in all species analyzed (e.g., "glycolysis", "proteasome"), this study also indicated that the higher-order relations between modules differ significantly between the organisms (Bergmann *et al.*, 2004). For example, the average gene expression profiles for genes in the "secreted protein" and "proteasome" modules correlate positively in yeast and *A. thaliana*, negatively in *D. melanogaster*, and do not appear to correlate significantly in *H. sapiens*.

Before gene co-expression networks can be generated and analyzed, two problems need to be solved. First, genomes often encode thousands of proteins. Construction of a co-expression network of this amount of genes will in most cases results in a network that is difficult to interpret due the number of genes and their many connections. Second, the physiological role of a large proportion of proteins is unknown, or at best poorly understood (Hughes *et al.*, 2004). This lack of understanding further hampers the interpretation of any network generated. Different strategies to circumvent these problems can be employed in gene co-expression network analyses. The first strategy is to limit the analysis of co-expression networks to descriptive parameters only, such as the number of connections that a gene has with other genes (connectivity), e.g., see Jordan *et al.*, 2008; van Noort *et al.*, 2004. While this approach provides an understanding of the network at a higher abstraction level, such knowledge cannot be converted easily into understanding the actual underlying biological processes. The second strategy is to investigate only a subset of genes that is relevant for a certain research interest, e.g., see Bergmann *et al.*, 2004; Lee *et al.*, 2004; Neretti *et al.*, 2007. For example, Bergmann and co-workers selected genes participating in eight well-defined *S. cerevisiae* biological processes, and examined co-expression of their orthologous genes in five other organisms (Bergmann *et al.*, 2004). In a variation of this strategy, co-expression between all genes is calculated but the analysis is focused on a part of the network that is of particular biological interest, e.g., a certain oncogene (Basso *et al.*, 2005). These approaches, however, are biased towards already known biological processes. In addition, the selected biological processes might operate distinctly in organisms other than *S. cerevisiae*. To reduce bias of using only known biological processes, yet another approach limits the analysis to genes conserved in different species, for instance, in *S. cerevisiae*, *D. melanogaster* and *C. elegans* (Daub & Sonnhammer, 2008).

In this study, we extend the latter approach to the analysis of gene co-expression networks. For the construction of a gene co-expression network of *A. niger*, a subset of genes is selected that is based on evolutionary conservation of the proteins they encode among 19 different fungal species. Even when no defined function can be assigned to such proteins, their evolutionary conservation suggests a biological role. From expression data of these conserved protein-encoding genes, a gene co-expression network is generated. Subsequently, the topology of this network is used to extend the analysis to less conserved genes excluded from the initial analysis. This approach is followed for the analysis of two *A. niger* DNA microarray data sets cultivated under distinct experimental conditions.

MATERIALS AND METHODS

Culturing

Mildly perturbed conditions: A. niger 872.11 (*ΔargB pyrA6 prtF28 goxC17 cspA1*) is derived from CBS 120.49. All media were based on Pontecorvo's minimal medium (pMM) (Pontecorvo *et al.*, 1953), contained 100 mM sorbitol as carbon source and were supplemented with uridine and arginine. Glass 2.5-l fermentors (Applikon) with 2.2 l of pMM were kept at a constant temperature of 30 ± 0.5 °C while fermentor headplates were kept at 8 °C. A total of 1.0×10^6 of spores per ml were added to a fermentor. During germination, each fermentor was aerated through the headspace (50 l h^{-1}) and stirred at 300 r.p.m.. When dissolved oxygen tension levels dropped below 60% for over 5 min, the stirrer speed was set to 750 r.p.m. and aeration was switched to sparger inlet. In one experiment, fermentors were induced with either 0.1 mM sorbitol or D-xylose at *T* = 14 h as previously described (van der Veen *et al.*, 2009). In a second experiment, fermentors were induced with 1 mM of various oils at *T* = 14 hours and samples were taken before induction and up to 2 h after induction (Table 1).

Strongly perturbed conditions: A. niger N402 (*cspA1*) (van Hartingsveldt *et al.*, 1987) is derived from CBS 120.49. All media were based on Bennett's minimal medium (bMM) (Bennett & Lasure, 1991). Both shake flask and fermentor cultures were grown in bMM medium at a constant temperature of 30 ± 0.5 °C and with differing combinations of carbon source, nitrogen source and concentration, and pH of the medium (Table 1) (Braaksma *et al.*, 2009). Fermentor inoculum was pre-cultured in baffled 500 ml Erlenmeyer flasks containing 100 ml bMM (pH 6.5) supplemented with the carbon source and nitrogen source concentrations corresponding to fermentor conditions. These flasks were inoculated with 10^6 spores per liter and incubated in a rotary shaker at 125 r.p.m. until approximately half of the available carbon source was consumed. Cultivations were carried out in 6.6-l fermentors (New Brunswick Scientific) with 5.0 l of bMM. The fermentors were inoculated with 4% (w/v) pre-culture. To prevent foaming, 1% (v/vl) Struktol J-673 antifoam was added to the medium and additional antifoam was added during cultivation when necessary. Each fermentor was sparged with 75 l h^{-1} of air with the stirrer speed set at 400 r.p.m. at the start of the cultivation. When dissolved oxygen tension levels dropped below 20%, the stirrer speed was automatically increased to maintain oxygen tension at 20% or until the maximum of 1000 r.p.m. was reached. The pH was controlled by automatic addition of 8 M KOH or 1.5 M H_3PO_4.

RNA isolation

Culture samples from mildly perturbed conditions were filtrated and biomass was snap-frozen into liquid nitrogen and stored at –80 °C. Culture samples from strongly perturbed conditions were quenched immediately in methanol at –45 °C as described previously (Pieterse *et al.*, 2006) and centrifuged at –20 °C to remove supernatant. Biomass was frozen into liquid nitrogen and stored at –80 °C. A Trizol-chloroform extraction preceded total RNA extraction with RNeasy mini columns (Qiagen) according to the manufacturer's protocol for yeast. Concentration of total RNA was determined by spectrophotometry. RNA integrity was assessed on an Experion system (Biorad) for samples from mildly perturbed conditions by visual inspection of the electropherograms. Graphs depicting RNA integrity categories were used as visual aids (Schroeder *et al.*, 2006). Electropherograms approximating an RNA integrity number of 8 or lower or with a 28S/18S ratio below 1.8 were discarded. For samples from strongly perturbed conditions, RNA integrity was assessed on agarose gel, by its A260/A280 ratio, and on an Agilent 2100 Bioanalyzer.

Table 1. Fermentation conditions for DNA microarray samples used in this study

Sample name	pH	Carbon source	Nitrogen source	Initial nitrogen source level (mM)	Inducer compound *	Sampling time (hr)	Growth Phase †
Mildly perturbed conditions							
29 ‡	3.5	sorbitol	$NaNO_3$	70.5	D-xylose	14	E
44 ‡	3.5	sorbitol	$NaNO_3$	70.5	D-xylose	14	E
52 ‡	3.5	sorbitol	$NaNO_3$	70.5	sorbitol	14	E
76 ‡	3.5	sorbitol	$NaNO_3$	70.5	D-xylose	14	E
86-1 ‡	3.5	sorbitol	$NaNO_3$	70.5	D-xylose	14	E
96 ‡	3.5	sorbitol	$NaNO_3$	70.5	sorbitol	14	E
Triton-0 ‡	3.5	sorbitol	$NaNO_3$	70.5	---	14	E
Triton-0.5	3.5	sorbitol	$NaNO_3$	70.5	Triton-X-100	14.5	E
Triton-1	3.5	sorbitol	$NaNO_3$	70.5	Triton-X-100	15	E
Triton-2	3.5	sorbitol	$NaNO_3$	70.5	Triton-X-100	16	E
Olive-0 ‡	3.5	sorbitol	$NaNO_3$	70.5	---	14	E
Olive-0.5	3.5	sorbitol	$NaNO_3$	70.5	olive oil	14.5	E
Olive-1	3.5	sorbitol	$NaNO_3$	70.5	olive oil	15	E
Olive-2	3.5	sorbitol	$NaNO_3$	70.5	olive oil	16	E
DGDG-0 ‡	3.5	sorbitol	$NaNO_3$	70.5	---	14	E
DGDG-0.5	3.5	sorbitol	$NaNO_3$	70.5	DGDG oil	14.5	E
DGDG-1	3.5	sorbitol	$NaNO_3$	70.5	DGDG oil	15	E
DGDG-2	3.5	sorbitol	$NaNO_3$	70.5	DGDG oil	16	E
Wheat-0 ‡	3.5	sorbitol	$NaNO_3$	70.5	---	14	E
Wheat-0.5	3.5	sorbitol	$NaNO_3$	70.5	wheat oil	14.5	E
Wheat-1	3.5	sorbitol	$NaNO_3$	70.5	wheat oil	15	E
Wheat-2	3.5	sorbitol	$NaNO_3$	70.5	wheat oil	16	E

Table 1. Continued

Sample name	pH	Carbon source	Nitrogen source	Initial nitrogen source level (mM)	Inducer compound	Sampling time (hr)	Growth Phase †
Strongly perturbed conditions							
4 G $4NO_3$-1	4	glucose	$NaNO_3$	282.4	---	66	LS
4 G $4NO_3$-2	4	glucose	$NaNO_3$	282.4	---	96	LS
4 G $8NO_3$	4	glucose	$NaNO_3$	564.8	---	53	LE
4 G $4NH_4$	4	glucose	NH_4Cl	282.4	---	57	LS
4 G $8NH_4$-1a §	4	glucose	NH_4Cl	564.8	---	36	LE
4 G $8NH_4$-2a §	4	glucose	NH_4Cl	564.8	---	36	LE
4 G $8NH_4$-1b §	4	glucose	NH_4Cl	564.8	---	60	LS
4 G $8NH_4$-2b §	4	glucose	NH_4Cl	564.8	---	60	LS
4 X $4NO_3$	4	xylose	$NaNO_3$	282.4	---	66	LS
4 X $8NO_3$	4	xylose	$NaNO_3$	564.8	---	91	LS
4 X $4NH_4$	4	xylose	NH_4Cl	282.4	---	60	LS
4 X $8NH_4$	4	xylose	NH_4Cl	564.8	---	66	LS
5 G $4NO_3$	5	glucose	$NaNO_3$	282.4	---	48	S
5 G $8NO_3$	5	glucose	$NaNO_3$	564.8	---	49	LE
5 G $4NH_4$	5	glucose	NH_4Cl	282.4	---	35.25	LE
5 G $8NH_4$	5	glucose	NH_4Cl	564.8	---	35	LE
5 X $4NO_3$	5	xylose	$NaNO_3$	282.4	---	93.5	S
5 X $8NO_3$	5	xylose	$NaNO_3$	564.8	---	112	S
5 X $4NH_4$	5	xylose	NH_4Cl	282.4	---	41	LE
5 X $8NH_4$	5	xylose	NH_4Cl	564.8	---	47.5	LE

* The concentration of inducer compound added was 0.1 mM for D-xylose and sorbitol; 0.002% for Triton-X-100; and 1 mM for olive oil, digalactoside-diglyceride (DGDG) and wheat oil.

† E, exponential growth phase; LE, late exponential growth phase; S, stationary phase, carbon source will become depleted within 1 hour; LS, late stationary growth, over 10 hours of carbon depletion.

‡ These DNA microarray samples are independent biological replicates, i.e., all thus labeled samples are grown in different fermentor vessels but with identical media composition and are sampled at identical time point, even though they are part of different experiments and grown at different dates.

§ These DNA microarray samples are technical replicates. The thus labeled microarray samples are derived from one fermentor sample from which the extracted RNA was processed further in duplicate.

Microarray processing

cDNA and cRNA synthesis and labeling, and array hybridization were performed following the Affymetrix users' manual (Affymetrix, 2004) using the One-cycle Target Labeling and Control Reagents Kit to synthesize 15 µg of cRNA from 5 µg of total RNA as template material for mildly perturbed conditions samples. For strongly perturbed conditions samples, the Bioarray High Yield RNA Transcript Labeling Kit (Enzo) was used to synthesize at least 30 µg of cRNA from 10 µg of total RNA as template material. Fifteen microgram of fragmented and labeled cRNA was hybridized to custom-made *A. niger* arrays at 45 °C for 16h.

Washing and staining was done using the Hybridization, Wash and Stain Kit (Affymetrix) using a GeneChip FS-450 Fluidics station and an Agilent G2500A Gene Array scanner. Scanned images were converted into .CEL files using MicroArray Suite software (Affymetrix).

Microarray data accession number

Raw and RMA-normalized array data were deposited at the NCBI Gene Expression Omnibus database (Edgar *et al.*, 2002) under series entries **GSE11405** and **GSE14285** for the mildly perturbed conditions and under series entry **GSE17329** for the strongly perturbed conditions.

Data preprocessing

DNA microarrays were normalized using Affymetrix' MicroArray Suite Software version 5 (MAS5) with the target value set at 100 (Affymetrix, 2001). MAS5 was preferred over another often-used normalization strategy, Robust Multichip Average, or RMA (Irizarry *et al.*, 2003), as MAS5 normalization is calculated over each individual array alone, thus excluding a potential influence of normalization to the correlation structure of the whole data set. Some probe sets have a signal above background in only few of the experimental conditions examined. Since their limited number of observations hampers the calculation of reliable correlations, for each of the three data sets, probe sets flagged "absent" in more than 80% of the microarrays per data set were discarded from that specific data set. Under mildly perturbed conditions, 7,955 probe sets (55%) were discarded while under strongly perturbed conditions 5,084 probe sets (35%) were discarded. Probe set values were normalized per microarray by dividing each probe set value by the mean signal over the whole microarray. Signals that are close to the detection limit are more influenced by random noise signal and thus yield varying values on different arrays. This variation hampers the calculation of reliable correlations as well and therefore the mean "absent" call value divided by two was taken as uniform "lowest in the data set" value. The remaining signals flagged as "absent" (those not included in the removed probe sets) as well as all other probe set signals with a value below this lowest uniform value were replaced by this uniform value. Probe sets were not filtered for a certain fold change threshold as the magnitude of fold change is not necessarily a measure for biological relevance (van den Berg *et al.*, 2006). In addition, the correlation analysis of the data has its own selection criterion, namely the ρ threshold value. No artifacts or outliers in the signals distribution for the microarrays within the data sets were observed, and the per-array signal distributions were similar.

Correlation analysis

The correlations between genes were determined by the Spearman correlation coefficient ρ. The correlation coefficient ranges from 0 (no correlation) till either 1 (full positive correlation between expression levels) or –1 (full negative correlation between expression levels, i.e. perfect antagonists). The Spearman ρ is a non-parametric correlation measure based on the rank of the expression values instead of the detected values, and is robust against outliers and mild non-linear behavior (Zar, 1996). The Spearman correlation measure was recently shown to be slightly more robust in the analysis of gene co-expression compared to other correlation methods including the Pearson correlation, Euclidian distance, and the mutual information measures (Daub & Sonnhammer, 2008). The *p*-value for the Spearman correlation for the mildly perturbed conditions data set, which is the smallest data set, was 4.12×10^{-6} for the lowest cut-off value for ρ (ρ = 0.85).

Correlation networks were drawn in Cytoscape (Shannon *et al.*, 2003). Initial networks were constructed using the "spring embedded" layout function, and individual genes within the resulting networks were manually re-positioned for improved interpretation (Fig. 1). Manual arrangement was based on the ρ values associated with each gene pair, by the sign of ρ, and by the number of connections per gene that were visible at a certain ρ threshold value. In this iterative process, while switching back and forth between ρ threshold values, only genes visible at a certain ρ threshold value were relocated. The length of the connecting lines does not represent the degree of correlation between the two connected genes. The term "module" is used for a group of genes that has a core of interconnected genes at high ρ threshold values, and to which group genes appear to attach preferentially upon lowering of the ρ threshold. A coloring scheme was deduced from the combined data network, where at ρ 0.95, eight modules can be identified. These modules were labeled A–H, and genes within each module were assigned a color to assist localization of these genes within the networks.

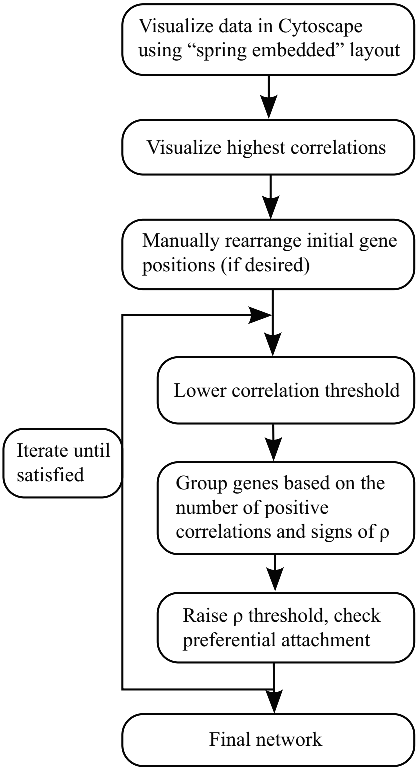

Visualize data in Cytoscape using “spring embedded” layout
Visualize highest correlations
Manually rearrange initial gene positions (if desired)
Lower correlation threshold
Iterate until satisfied
Group genes based on the number of positive correlations and signs of ρ
Raise ρ threshold, check preferential attachment
Final network

Fig. 1. Procedure for construction of gene co-expression networks. Schematic representation of the process to construct gene co-expression networks.

Validation

The influence of individual microarrays on the correlation analysis was evaluated by a “leave *N* samples out” validation. This validation procedure tests whether the strongest co-expressed gene pairs also remain the strongest co-expressed gene pairs in case DNA microarray samples are removed from the complete set of DNA microarrays. The following procedure was used: (i) random selection of two (for each single data set) or five (for the combined data sets) microarrays and removal from the data set; (ii) calculation of new correlation coefficients; (iii) continuation until all microarrays were excluded whilst ensuring that previously removed arrays were not removed again; (iv) repetition of this procedure for 20 times; (v) calculation of the mean correlation coefficient per gene pair; (vi) selection of gene pairs matching the 2.5 and 97.5 percentiles of the mean correlation coefficient; (vii) comparison of the selected genes with the genes present in the correlation networks. For the mildly perturbed data set, 80% of its strongest correlating gene pairs fell within the 2.5th and 97.5th percentiles in this validation procedure. All of the actually observed strongest correlating gene pairs for both the strongly perturbed data set and the combined data set belonged to the 5% strongest correlating gene pairs of the validation.

Consensus expression profile analysis

The "combined data set" network was used for consensus expression profile analysis. Genes with three or more connections with other genes within a module were selected. Their expression profiles were converted in a rank order analogous to the rank order used for the Spearman correlation. Following (Horvath & Dong, 2008), the consensus expression profile was defined as the profile obtained from the first principal component score vector of a principal component analysis (PCA) (Jackson, 1991; Joliffe, 2002) of the converted expression profiles. Subsequently, the correlation between the obtained consensus expression profile and the expression profile of all measured genes was calculated. All calculations were performed on a Pentium 4 personal computer with 1 GB internal memory using Matlab (The Mathworks), the Statistics Toolbox (The Mathworks), and homemade scripts.

Promoter analysis

Promoter analysis was done in GeneSpring, version 7.2 (Agilent), using the "find potential regulatory sequences" tool. The promoter region from 10 to 800 bases upstream of a gene was searched for oligonucleotides ranging from 5 to 10 bases, with at maximum one single point discrepancy allowed, and correcting for local nucleotide density. The likelihood of random occurrence of identified sequences was compared relative to the upstream region of all 14,165 genes in the *A. niger* genome.

KEGG pathway analysis

For the combined data set network, all genes present within each module A–H at ρ 0.90 were exported to a tab-delimited file and imported into the KegArray program (Wheelock *et al.*, 2009). The “PathwayMap” tool was applied to extract *A. niger* genes linked to a KEGG pathway from the KEGG database (Kanehisa & Goto, 2000) by using the “Ang” organism abbreviation.

Gene ontology

The genomes of the fungi *Aspergillus niger*, *Aspergillus fumigatus*, *Aspergillus nidulans*, *Penicillium chrysogenum*, *Neurospora crassa*, *Magnaporthe grisea*, *Stagonospora nodorum*, *Ustilago maydis*, and *Trichoderma reesei*, and the yeasts *Ashbya gossypii*, *Candida albicans*, *Candida neoformans*, *Debaromyces hansenii*, *Giberella zeae*, *Kluyveromyces lactis*, *Phanerochaete chrysosporium*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica* were used to construct an in-house built database of orthologous protein sequences (S. Basmagi & P. Schaap, unpublished data). Protein sequences were placed into an orthology cluster based on bi-directional first-hit BLAST alignment of protein sequences of other species. Conserved proteins were defined as having an ortholog in at minimum 15 of 19 species; the absence of a gene in a species while present in over 15 other genomes is due mostly to mis-annotation or incorrect intron-predictions in our experience. A total of 2,749 genes fulfilled this criterion. All 455 genes that have a *S. cerevisiae* ortholog but did not meet the criterion were added because of the extensive body of knowledge that is available for genes of this model organism. On average, these latter 455 genes have an ortholog in 11 species. Gene ontology terms, available from *S. cerevisiae* orthologous genes per module, were browsed at the *Saccharomyces* Genome Database website (Hong *et al.*, 2008).

RESULTS

Gene co-expression networks based on a subset of genes have been generated and some of their generic properties will be described. Two series of *A. niger* DNA microarray data sets were used, that were analyzed as separate data sets as well as in combination. Groups of co-expressed genes, termed modules, were observed within the visualized networks. Next, the biological properties of these modules were analyzed using the combined data set network as our reference network. Subsequently, the module structure found for the combined data set was compared with the co-expression networks based on the two individual data sets. We conclude our results by extending our network analysis from a subset of genes to all genes for which a probe set is available on the *A. niger* DNA microarray.

Construction of gene co-expression networks

It is expected that gene co-expression networks will be influenced by the experimental setups of the microarray data sets used to generate these networks. Therefore, a total of 42 microarrays that originated from *A. niger* strains grown in batch fermentation under two different experimental setups were used (Table 1). The microarrays used in this study were obtained from different experimental perspectives (e.g., investigate D-xylose metabolism [van der Veen *et al.*, 2009] or lipid metabolism [van der Veen, 2009], or extracellular protease activity [Braaksma *et al.*, 2009]), but were selected on the basis of covering diversity, especially in relation to cell culture perturbations. We expect that these perturbations will have a larger impact on the physiology than the differences between the closely related strains used. It was decided to not include further *A. niger* microarray data that are available in public repositories, as at the time of this study only shake flask-cultivated experiments were deposited. Shake flask cultivation generally introduces more culture heterogeneity as pH and the transfer of oxygen, nutrients, and heat are not controlled (van der Veen *et al.*, 2009).

Twenty microarrays were obtained from fungal cells growing exponentially with 100 mM sorbitol as primary carbon source, to which either 0.1 mM sorbitol or D-xylose or 1 mM vegetable oils were applied. Under these growth conditions, the cells do not experience any nutrient limitation and grow at maximum growth rate. We reasoned that the applied pulses would provoke only a minor disturbance in global gene expression levels, and hence labeled these cultivation conditions "mildly perturbed". In contrast, the other 22 microarrays were obtained from fungal cells growing in much more perturbed conditions at the time of sampling (e.g., carbon source deprivation). These cells were expected to yield more drastic changes in gene expression levels, and therefore these conditions were labeled "strongly perturbed". From a biological point of view, the differing experimental conditions are expected to yield both condition-specific gene expression (e.g., induction with D-xylose leads to increased expression of the xylan–metabolic system) as well as expression of genes involved in general metabolic processes required for both conditions (e.g., growth in Minimal Medium broth requires de novo amino acid biosynthesis under both conditions).

Construction of a co-expression network using the data of the over 14 thousand predicted *A. niger* genes will result in a network that is difficult to interpret due to the many resulting gene–gene interactions. Therefore, a subset of genes was selected according to their evolutionary conservation among fungal species, and their signal value. The evolutionary conserved subset consisted of only those protein-encoding genes for which an ortholog is present in 15 or more of the 19 fungal species analyzed, or for which a *S. cerevisiae* ortholog is identified (see Materials and Methods). Even in case no clear biological function has been assigned to such protein, its evolutionary conservation suggested a functional role. In addition, a present signal for a gene in at least 20% of the arrays ensures that enough relevant data points are available to calculate an expression profile for that gene. The selected gene list comprised 2,773 genes.

The similarity in expression of two genes was expressed in the correlation coefficient ρ and was calculated for all pair-wise combinations of the 2,773 genes for each of the three data sets. The ρ distribution detailed the strength of pair-wise correlations and gave an impression on the nature of the three gene co-expression networks (Fig. 2). For the mildly perturbed conditions data set, the ρ values were centered around zero (Fig. 2, left), which suggests that most gene pairs in this data set were weakly co-expressed with only relatively few genes being strongly co-expressed. In contrast, the histogram for the strongly perturbed conditions data showed a much broader base. This broader base translated into a tendency of gene pairs to be more strongly correlated or anti-correlated (i.e., two genes that have antagonistic expression patterns) (Fig. 2, middle). The histogram of the combined data set resembled the histogram of the strongly perturbed conditions in shape, although less strongly correlating ρ values were observed for this network (Fig. 2, right). The different ρ value distributions suggested that co-expression was different between the data sets.

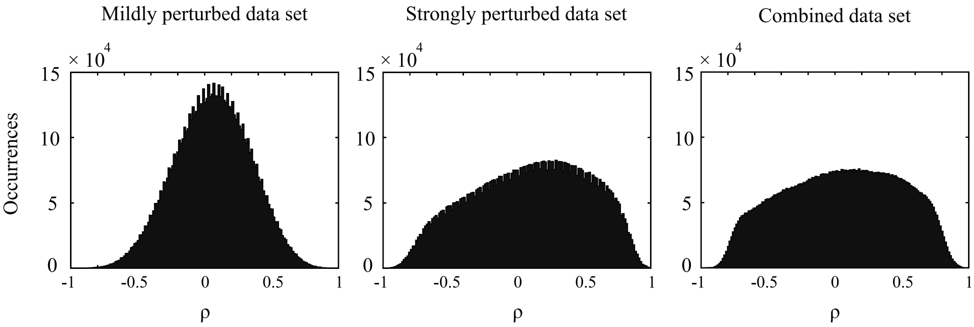


Fig. 2. Correlation coefficient distribution per data set. Histogram of the values of *ρ* as calculated for all possible gene pair combinations in each data set. The distribution of *ρ* for all gene pairs possible is visualized for the subset of 2,773 genes by dividing the range of *ρ* values in equally spaced bins (e.g., one bin would range from 0-0.1, the next from 0.1-0.2, and so on), followed by counting the number of occurrences for a range of *ρ* values per bin.

Three gene co-expression networks were constructed using the calculated *ρ* values. In these networks, a connecting line was drawn between each pair of genes for which their expression profile correlated stronger than by the set *ρ* threshold. The networks were visualized at different *ρ* threshold values. When the *ρ* threshold value was lowered, more connecting gene pairs appeared in the network. The network based on the combined data set at different *ρ* threshold values is visualized in Fig. 3 (panels A–D), while for the mildly and strongly perturbed conditions-derived networks only the lowest *ρ* threshold value with a meaningful clustering was visualized in panels E and F of Fig. 3 (full networks are accessible in Supplementary material file 1). Upon lowering the *ρ* threshold, additional connecting gene pairs preferred attachment to genes already present, instead of being randomly placed within the network (Fig. 3, panels A–D). This preferential attachment of new genes to genes already in the network is a common observation for biological networks (Almaas, 2007; Barabási & Albert, 1999; Barabási & Oltvai, 2004). A result of preferential attachment was the presence of a small number of genes that correlated strongly with many other genes within a network, while many genes only correlated strongly with few other genes. The distribution of the number of correlations per gene, or the gene connectivity, is given in Fig. 4. For the three networks described here, the gene connectivity distribution could be described by a power-law distribution with a connectivity exponent *γ* around 1.2 (Fig. 3). Similar values were found for other gene co-expression networks: a 4077-genes network of *S. cerevisiae* had *γ* around 1.0 (van Noort *et al.*, 2004), whereas this value ranged between 1.1 and 1.8 for gene co-expression networks for six distinct organisms (Bergmann *et al.*, 2004).

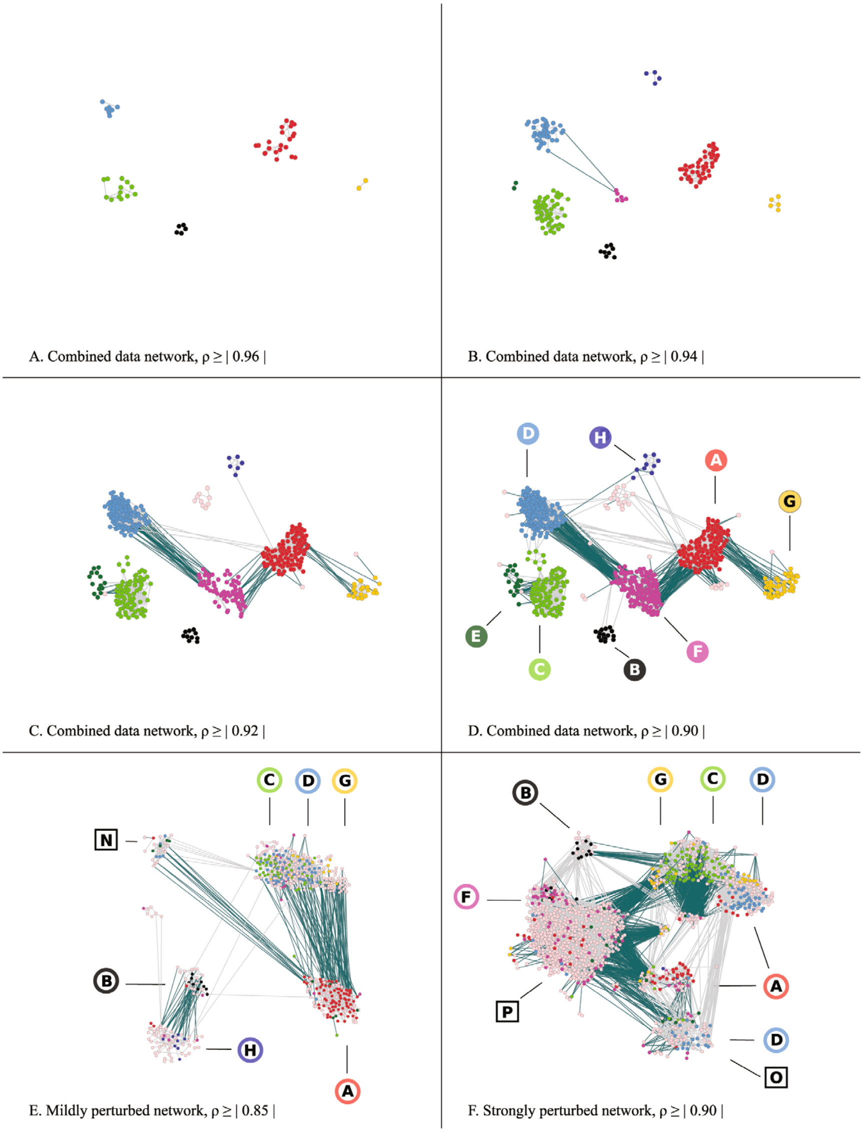

A. Combined data network, ρ ≥ | 0.96 |
B. Combined data network, ρ ≥ | 0.94 |
C. Combined data network, ρ ≥ | 0.92 |
D
H
A
G
E
C
B
F
D. Combined data network, ρ ≥ | 0.90 |
C
D
G
N
B
H
A
E. Mildly perturbed network, ρ ≥ | 0.85 |
B
G
C
D
F
A
P
D
O
F. Strongly perturbed network, ρ ≥ | 0.90 |

Fig. 3. Gene co-expression networks. Panels A-D: the gene co-expression network constructed from all 42 microarrays for 4 threshold settings of *ρ* as indicated in each lower left corner. Circles represent genes, while lines represent a *ρ* value above the set threshold. Positive *ρ* values are shown as solid gray lines while negative *ρ* values are represented as green lines. The networks constructed from mildly perturbed conditions data set (panel E) and strongly perturbed conditions data set (panel F) are given at their lowest *ρ* threshold value only. Colouring is based on the modules identified in the combined data sets network (panel D), with module labels indicated in solid coloured circles. This colouring is superimposed on the networks, and groups of genes with identical colour are indicated by open coloured circles (panels E and F). Boxed letters indicate modules that are not present in the combined data sets network.

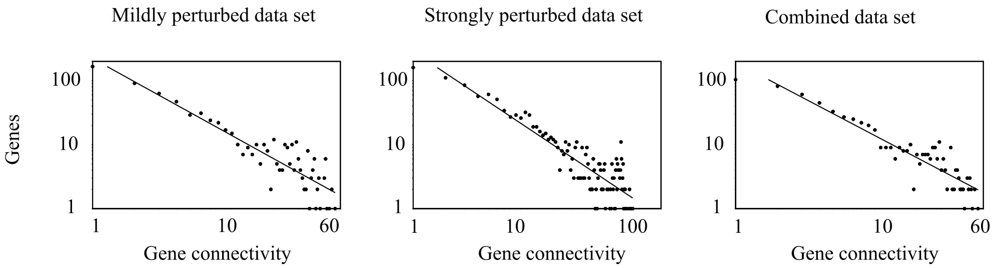


Fig. 4. Gene connectivity distribution per data set. For each gene within the networks at a *ρ* threshold of 0.90, the number of genes it partners with (horizontal axis) is plotted against the number of genes with identical number of gene pairs (vertical axis). The fitted line is for a power-law distribution, $P(k) \sim k^{-\gamma}$, which describes the probability that a gene has *k* gene pairs. For all networks, *γ* is around 1.2.

Modules relate to biological functions

As the combined data network was derived from expression data obtained from both mildly and strongly perturbed conditions, we expected that co-expressed genes within this network are less prone to condition-specific peculiarities. Therefore, the combined data network was analyzed with respect to biological processes. Eight modules were observed in the combined data set network. These were coloured and labelled A–H (Fig. 3, panel D). As genes with similar expression level profiles often encode proteins that are involved in a similar biological process (Walker *et al.*, 1999; Wolfe *et al.*, 2005), we searched for indications for biological processes that were overrepresented in the modules using the *S. cerevisiae* Gene Ontology vocabulary for genes in these modules that had a *S. cerevisiae* ortholog and the annotation of these genes. In addition, we examined whether genes within each module were assigned to metabolic pathways using the KEGG database. Lastly, we examined the upstream regions of genes within these modules for conserved upstream elements that hint to co-regulation by a common transcription factor. Indeed, when using the Gene Ontology annotations, an overrepresentation of similar ontology terms was found for genes that were present in some modules (Fig. 5; Supplementary material file 2). Also, conserved sequences were identified for genes of some modules.

Module A contains an overrepresentation of genes predicted to encode proteins involved in amino acid metabolic processes, including amino acid biosynthetic enzymes and tRNA-ligases. For 62 of 137 genes (45%), the conserved sequence 5'-TGA-(C/G)-TCA was identified (*p*-value 4.6×10^{-15} and 6.7×10^{-14}, respectively), which is a known binding site of the DNA-binding protein CpcA (Wanke *et al.*, 1997). This transcription factor is a global regulator in *A. niger*. Upon amino acid starvation, CpcA co-ordinates a transcriptional response by derepressing transcription of many genes encoding enzymes involved in amino acid biosynthetic pathways, as well as enzymes involved in nucleotide biosynthesis.

Module B consists of genes encoding proteins involved in fatty acid metabolism or peroxisome organization. Genes encoding the peroxins Pex6, Pex10, and Pex11, as well as the FoxA bifunctional enzyme that catalyzes the second and third step of fatty acid β-oxidation are in this module. We identified the conserved sequence 5'-CCTCGG or its reverse complement sequence within the upstream region of 16 of 20 genes of this module (*p*-value 2.6 × 10^{-5}). This sequence has been shown to be present upstream of a large number of genes predicted to encode proteins involved in fatty acid metabolism and peroxisome proliferation in filamentous fungi (Hynes *et al.*, 2006). Hynes and co-workers showed experimentally that two transcription factors involved in fatty acid utilization, FarA and FarB, bind to this sequence in *A. nidulans* (Hynes *et al.*, 2006).

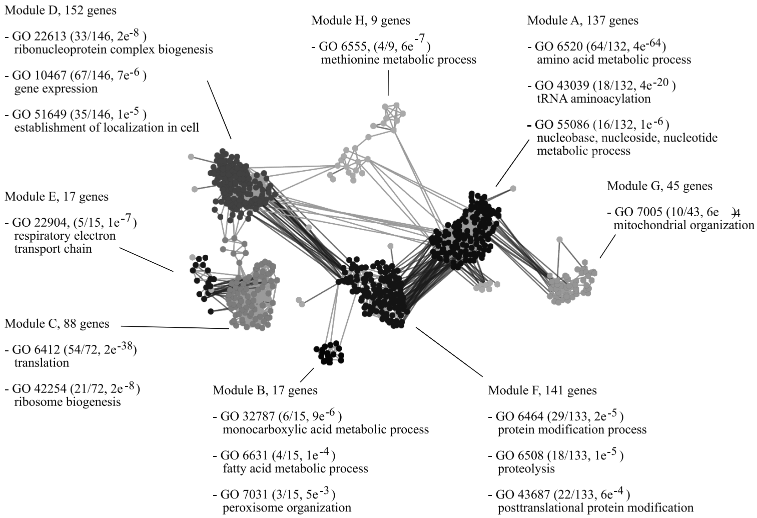

Module D, 152 genes
- GO 22613 (33/146, 2e^{-8})
ribonucleoprotein complex biogenesis
- GO 10467 (67/146, 7e^{-6})
gene expression
- GO 51649 (35/146, 1e^{-5})
establishment of localization in cell
Module H, 9 genes
- GO 6555, (4/9, 6e^{-7})
methionine metabolic process
Module A, 137 genes
- GO 6520 (64/132, 4e^{-64})
amino acid metabolic process
- GO 43039 (18/132, 4e^{-20})
tRNA aminoacylation
- GO 55086 (16/132, 1e^{-6})
nucleobase, nucleoside, nucleotide metabolic process
Module G, 45 genes
- GO 7005 (10/43, 6e)4
mitochondrial organization
Module E, 17 genes
- GO 22904, (5/15, 1e^{-7})
respiratory electron transport chain
Module C, 88 genes
- GO 6412 (54/72, 2e^{-38})
translation
- GO 42254 (21/72, 2e^{-8})
ribosome biogenesis
Module B, 17 genes
- GO 32787 (6/15, 9e^{-6})
monocarboxylic acid metabolic process
- GO 6631 (4/15, 1e^{-4})
fatty acid metabolic process
- GO 7031 (3/15, 5e^{-3})
peroxisome organization
Module F, 141 genes
- GO 6464 (29/133, 2e^{-5})
protein modification process
- GO 6508 (18/133, 1e^{-5})
proteolysis
- GO 43687 (22/133, 6e^{-4})
posttranslational protein modification

Fig. 5. Assignment of biological functions to modules. For the combined data sets network at a *ρ* threshold of 0.90 (as shown in Fig. 3, panel D), enriched biological processes are indicated within modules using the Gene Ontology terms of genes with a *S. cerevisiae* ortholog. The number of *A. niger* genes per module is indicated after the module code. The GO-number points to the observed Gene Ontology process. Between brackets, the number of genes with that annotated Gene Ontology process relative to the total number of genes queried is given, followed by a p-value that gives the likelihood that the identified GO process is found by chance alone. Genes that encode conserved proteins but have no *S. cerevisiae* ortholog make up the difference in the total number of *A. niger* genes per module and the number of genes queried for Gene Ontology enrichment.

Module C contains mostly cytosolic ribosomal protein-encoding genes. In the upstream region of 47% of the 88 genes within this module, the conserved sequence 5'-CGACAA was identified, while the core sequence 5'-CGAC was found upstream 80% of the genes. The probability of observing these upstream sequences for these genes by chance alone is very low (*p*-value 4 × 10^{-6} and 2 × 10^{-4}, respectively). This sequence does not resemble any of the known binding sites associated with ribosomal proteins in *S. cerevisiae* or *S. pombe* (Tanay *et al.*, 2005). The presence of such conserved sequence hints to the existence of a yet unidentified DNA-binding transcription factor that is involved in the regulation of genes encoding cytosolic ribosomal proteins in a fungal system.

In the D-labelled module, genes categorized by the generic Gene Ontology term "gene expression" are overrepresented. For example, this module contains genes that encode putative RNA helicases, spliceosome assembly proteins, and 16 putative translation initiation factors. We identified the sequence 5'-GGCCGCG for 111 of 152 genes (*p*-value 8 × 10^{-4}). This upstream element is located 400 base pairs or more away from the gene's start site for 60% of these 111 genes. Also for this module, the presence of a specific conserved upstream sequence suggested regulation by a yet unidentified DNA-binding transcription factor involved in the regulation of genes whose products appear involved in "gene expression" processes. Next to the above-mentioned sequence motif, we identified an overrepresentation of pyrimidine-rich sequences upstream for 80% of the genes of module D. However, it should be noted that CT-rich regions are relatively common in upstream regions of filamentous fungi and that they are mostly related to the position of the transcription start site (Punt & van den Hondel, 1992).

Overrepresented Gene Ontology processes were also found for modules E–H (Fig. 5; Supplementary material file 2). However, no conserved upstream sequences were found. Module E pertains to energy metabolism related processes like “electron transport chain”, “oxidative phosphorylation”, and “ATP synthesis coupled electron transport”. Module F contains the following overrepresented processes: “cellular catabolic process”, “proteolysis”, and “protein modification process”. Module G is related to “organelle organization” and “mitochondrion organization”. Module H pertains to processes related to different amino acid metabolic processes.

For each module, we assessed whether their genes could be related to metabolic pathways (Supplementary material file 2). We observed a good agreement between *A. niger* genes within a module that can be linked to biological pathways, and the observed overrepresentation of Gene Ontology processes. For example, module A, which contains an overrepresentation of genes encoding proteins involved in amino acid metabolic processes, contains 71 genes that encode proteins of amino acid related biochemical pathways (Supplementary material file 2). For the 16-gene containing module E, which has overrepresenting Gene Ontology terms related to energy metabolism, the four genes that encode proteins that relate to KEGG biochemical pathways are within the “oxidative phosphorylation” pathway.

Modular structure is retained in the two other networks

Seventy-five percent of the genes that were present in the combined data sets network were found in at least one other network (Fig. 6). The localization of these genes within each network was examined by colouring of each module identified in the combined data sets network (Fig. 3, panel D), and superimposition of these colours to both other networks (Fig. 3, panels E and F).

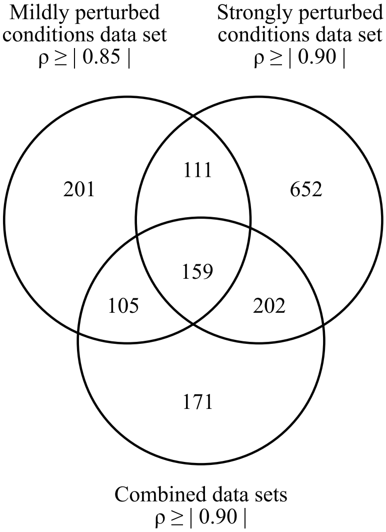


Fig. 6. Overlap of genes between networks. Venn-diagram showing the overlap of genes present in any of the three networks analyzed.

Both the mildly perturbed and strongly perturbed conditions networks appeared less structured compared to the combined data sets network, but modules can be recognized nevertheless.

The modules labelled C, D, and G, were relatively well separated in the combined data sets network while these modules overlapped or were closely connected in the two other networks. In the combined data sets network, these modules were enriched for genes encoding proteins involved in “ribosome biogenesis” and “translation”, “gene expression” and “ribonucleoprotein complex biogenesis”, and “mitochondrial organization” respectively. In the mildly perturbed conditions network, 323 genes are in the module that contains many of the C, D, and G-coloured genes; 138 of these genes (43%) were present as well in the combined data sets network. A Gene Ontology terms search on all 323 genes yielded similar GO terms for this C–D–G module (Supplementary material file 3). Likewise, in the strongly perturbed conditions network, the 284 genes that included many of the genes of the C, D, and G modules in the combined data sets network yielded the same GO terms (Supplementary material file 3).

Modules A and B found in the network based on the combined data set are present in the mildly and strongly perturbed networks as well. Other genes are also associated to these modules, and these genes have the same GO terms associated to them as in the combined data set network (Supplementary material file 3), namely “cellular amino acid biosynthetic process” and related GO terms for module A and “fatty acid β-oxidation” and “carboxylic acid metabolic process” for module B.

In addition, network-specific modules appeared that were not visible in the combined data sets network. The mildly perturbed conditions network gave one such module, labelled N (Fig. 3, panel E). The N module consisted of 24 genes, of which half are related to the respiratory electron transport chain. Indeed, this module contains subunits of the ubiquinol–cytochrome C oxidase complex. The expression of these genes could be a specific adaptation to the exponential growth phase these cells were in at the time of sampling.

For the strongly perturbed network, two specific modules were observed and were labelled O and P in Fig. 3, panel F. The 84 genes present in module O were enriched for “metabolic processes”, to which term 62 of 84 genes are assigned. Twenty-four percent of the 84 genes were involved in the generation of precursor metabolites and energy.

The second module P (Fig. 3, panel F) was large and contained half of the genes present in the strongly perturbed data set. No biological processes were overrepresented for this module although the consensus expression profile (see below) seemed to correlate strongly to the presence or absence of a carbon source in the medium. No overrepresented motifs were detected in the upstream region of genes in this module. Ten percent of the genes located in this module were also located in module F in the combined data sets network. These observations supported our choice to use the combined data set as a basis as indeed the individual data sets seemed to contain condition-specific modules.

Extending the subset of genes by means of a consensus expression profile

The thus far constructed gene co-expression networks were based on a subset of 2,773 genes that encode evolutionary conserved proteins. However, these genes made up only 43% of the total of 6,416 *A. niger* protein-encoding genes that were evaluated as present in more than 20% of the arrays on the combined data set microarrays. Genes that did not take part in our initial selection were examined using the modules identified in the network. Genes within a module have similar gene expression profiles, and this similarity was used to calculate a consensus expression profile (Horvath & Dong., 2008) for each module in the combined data set. The correlation between each module's consensus expression profile and all 6416 genes' expression profile was calculated and expressed as an associated consensus expression profile correlation coefficient ρ_{cons}. Associated ρ_{cons} values for all modules are given in Supplementary material file 4. Here, the results of this approach are exemplified by description of module B. This module contained 19 genes at ρ of 0.90, with most genes being related to peroxisome proliferation and fatty acid metabolism. As genes in this module are relatively well characterized, interpretation of the resulting data and analysis of this proof-of-concept is made easier.

Table 2 presents the genes with a ρ_{cons} to the module B consensus expression profile of 0.90 or higher. Half of the 20 genes that correlate most strongly with the module B consensus expression profile did fall outside our initial selection criteria (Table 2). Most of these genes could be associated with fatty acid metabolic activity or peroxisome functioning based on inspection of their gene annotation. Interestingly, the ortholog of the *A. nidulans* FarA fatty acid-related transcription factor, encoded by gene An14g00920, also correlated strongly with the module B consensus expression profile. A motif sequence for the FarA and FarB transcription factors could be identified in the upstream region of all but three genes listed in Table 2, including the FarA-encoding gene itself.

Similar results were obtained when the consensus expression profiles derived from the other modules were analyzed (Supplementary material file 4). For example, module F in the combined data sets network had an overrepresentation of "protein modification process" and related Gene Ontology terms. These Gene Ontology terms are also overrepresented when the genes with a *S. cerevisiae* ortholog that have an expression profile ρ_{cons} of 0.80 or more were analyzed. For instance, the Gene Ontology term "protein modification process" is found for 22% of these genes (*p*-value 4.4×10^{-11}). In addition, of the 86 genes with an expression profile similar to the consensus expression profile by over ρ_{cons} 0.90, 40 genes are annotated as "hypothetical protein" (Supplementary material file 4).

DISCUSSION

Variations in the timing and levels of gene transcription, mRNA translation, and protein maturation have considerable consequences for a cell. For understanding the dynamics of the physiological processes in *A. niger*, insight into the interactions and combined activity of these processes or events is required, in addition to knowledge of individual components of the cellular system. This study queried transcriptomes obtained from cultures grown under different experimental conditions, with the aim to gain insight into the relations between genes, and, at a higher hierarchical level, into relations between modules. For this, initial analysis of gene co-expression was performed on an evolutionary highly conserved subset of the *A. niger* genes, and the analysis was subsequently extended to the whole genome.

Our approach reveals that the gene co-expression networks consist of modules of co-expressed genes (Fig. 3). Subsequent analysis of the discovered modules provides evidence for their biological relevance: (i) modules are enriched for Gene Ontology terms, (ii) genes within the modules relate to biochemical pathways, and (iii) conserved motifs are present in the upstream region of many genes in several of the modules. Experimentally confirmed upstream sequences corresponding to the DNA-binding sites of the transcription factors CpcA (involved in amino acid related processes) (Wanke *et al.*, 1997) and FarA/FarB (involved in β-oxidation and perosixome biosynthesis) (Hynes *et al.*, 2006) are found in modules A and B, respectively, that have an overrepresentation of related Gene Ontology terms. These findings indicate that our approach is able to infer "true" biological processes.

Table 2. Genes strongly correlating with consensus expression profile of module B

Rank	*A. niger* probe ID (corresponding gene ID)	ρ_{cons} *	Description	*S. cerevisiae* ID	No. of orthologous species containing this protein	Gene present in Combined Data network?	Upstream motif †
1	An00g06872_at An13g01920)	0.982	strong similarity to acetyl-CoA C-acetyltransferase precursor - *R. norvegicus*	YPL028W	17	yes	(61), 437
2 ‡	An00g11070_at (An16g07150)	0.979	strong similarity to soluble cytoplasmic fumarate reductase FRDS1 - *S. cerevisiae*	YEL047C	17	yes	(208)
3	An00g06382_at (An16g05340)	0.977	similarity to trans-2-enoyl-ACP reductase II fabK - *S. pneumoniae*	YJR149W	15	yes	(165), 192
4 ‡	An00g11070_s_at (An16g07150)	0.970	strong similarity to soluble cytoplasmic fumarate reductase FRDS1 - *S. cerevisiae*	YEL047C	17	yes	(208)
5	An00g11023_at (An01g12960)	0.970	strong similarity to short/branched chain specific acyl-CoA dehydrogenase precursor ACADSB - *H. sapiens*		12	no	145, (161)
6	An00g09716_at (An15g01920)	0.968	strong similarity to methylcitrate synthase mcsA - *A. nidulans*	YPR001W	12	yes	(152), 368
7	An00g06734_at (An14g00430)	0.958	strong similarity to 3-hydroxybutyryl-CoA dehydrogenase BHBD - *C. acetobutylicum*		11	no	109
8	An00g06380_at (An07g03290)	0.953	similarity to trans-2-enoyl-ACP reductase II fabK - *S. pneumoniae*		11	no	(250)
9	An00g09575_at (An08g10110)	0.940	strong similarity to lipid transfer protein POX18 - *C. tropicalis*		14	no	(612)
10	An00g09583_at (An04g03290)	0.939	strong similarity to long-chain acyl-CoA dehydrogenase - *R. norvegicus*		8	no	784
11	An00g06703_at (An12g07630)	0.938	strong similarity to 2-methylisocitrate lyase ICL2 - *S. cerevisiae*	YPR006C	14	yes	246

Table 2. Continued

Rank	*A. niger* probe ID (corresponding gene ID)	ρ_{cons} *	Description	*S. cerevisiae* ID	No. of orthologous species containing this protein	Gene present in Combined Data network?	Upstream motif †
12	An00g08462_at (An01g09830)	0.927	strong similarity to glutathione S-transferase GTT1 - *S. cerevisiae*	YIR038C	15	yes	170
13	An00g09552_at (An08g07520)	0.921	strong similarity to levodione reductase lvr - *C. aquaticum*		11	no	(259)
14	An00g05624_at (An02g05230)	0.921	similarity to protein fragment SEQ ID NO: 65270 of patent EP1033405-A2 - *A. thaliana*			no	(298)
15	An00g09578_at (An12g08270)	0.919	strong similarity to L-lactate 2-monooxygenase LA2M - *M. smegmati*		9	no	152
16	An00g10969_at (An07g00440)	0.918	strong similarity to secretory lipase LIP2 - *C. albicans*		9	no	Not present
17	An00g11952_at (An14g00990)	0.915	strong similarity to trifunctional protein of the β-oxidation fox-2 - *N. crassa*	YKR009C	17	yes	226
18	An00g12118_at (An07g09190)	0.906	strong similarity to very long-chain fatty acyl-CoA synthetase FAT1 - *S. cerevisiae*			no	318
19	An00g13622_at (An02g04350)	0.906	weak similarity to the helix-loop-helix transcription factor Max - *M. musculus*		5	no	Not present
20	An00g10237_at (An01g03680)	0.901	strong similarity to peroxisomal ABC transporter ALDR - *M. musculus*	YPL147W	17	yes	315
21	An00g07656_at (An14g00920)	0.900	strong similarity to FarA transcription factor - *A. nidulans*		11	no	Not present

* The correlation coefficient of a gene's expression profile with the consensus expression profile constructed from genes present in module B.

† Number indicates the position of the upstream motif 5'-CCGAGG relative to the gene's start codon in base pairs; number in brackets indicates the position of the reverse complement motif relative to the start codon.

‡ Gene An16g07150 is represented on the DNA microarray by two probe sets, both of which are in this list.

In addition to observations that can be related to experimentally verified data, our approach yielded novel targets for experimental validation, such as the upstream sequences observed in genes of modules C and D. The sequence 5'-CGACAA in the upstream region of many ribosomal protein-encoding genes in module C appears of special interest, as this sequence does not resemble the conserved upstream sequences found genes encoding ribosomal proteins in the yeasts *S. cerevisiae* and *S. pombe* (Tanay *et al.*, 2005).

Previous gene co-expression network studies used a subset of genes already known (Bergmann *et al.*, 2004; Herrgård *et al.*, 2003) or suspected (Neretti *et al.*, 2007) to be involved in specific biological processes, or discussed network characteristics without zooming into biological details (Jordan *et al.*, 2008; van Noort *et al.*, 2004). In this study, however, an approach similar to the approach of Daub and Sonnhammer (2008) was followed. A subset of genes was selected based on their highly conserved nature among fungal species, without taking into account their role in biological processes. An advantage of this approach is that also protein-encoding genes for which no function is assigned are analyzed, while a potential pitfall of this selection criterion is that evolutionary less well conserved co-expressed genes (e.g., species-specific genes that encode biopolymer-degrading enzymes) will not be examined in this initial selection. Genes within the observed modules are related to essential cellular processes; for instance, ribosomes are required for protein synthesis (Fig. 5, module C), genes must be transcribed (Fig. 5, module D), and amino acids must be synthesized *de novo* when not supplied in the medium (Fig. 5, Module A).

The advantage of selecting evolutionary conserved genes likely extends to the consensus expression profile analyses. As the modules identified are based on evolutionary conserved sequences, it is likely that the consensus expression profiles of these modules are more robust than consensus expression profiles based on subsets of known genes. The evolutionary conservation suggests that a large time span has past in which the regulation of such a module was tuned, while for organism-specific modules the regulation could be more variable. Thus, using consensus expression profiles based on an evolutionary conserved subset will probably result in more accurate lists of genes with similar expression profiles to the consensus expression profile.

Research towards a better understanding of the higher-order structures that play a role in *A. niger* cellular functioning did only start recently, after high throughput technologies such as DNA microarray platforms became available for this organism. The usefulness of studying these higher-order structures is illustrated in this paper;

the networks of evolutionary conserved genes of *A. niger* resulted in the identification of biologically relevant gene co-expression modules. In addition, the use of consensus-profiles extended the analysis to include the full gamut of genes of *A. niger*.

ACKNOWLEDGEMENTS

This project was in part carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation (RvdB, MB, PP, and MvdW) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research. In addition, RvdB was supported by the Research Fund of Katholieke Universiteit Leuven (EF/05/007 SymBioSys) and by IWT-Flanders (IWT/060045/SBO Bioframe). The authors would like to thank DSM Food Specialties for access to the DSM *A. niger* Affymetrix DNA microarrays, and for providing the microarrays to MB.

CHAPTER 6

A TOP-DOWN SYSTEMS BIOLOGY APPROACH FOR THE IDENTIFICATION OF TARGETS FOR FUNGAL STRAIN AND PROCESS DEVELOPMENT

Machtelt Braaksma, Robert A. van den Berg, Mariët J. van der Werf and Peter J. Punt

This chapter has been published with minor modifications in:
Cellular and Molecular Biology of Filamentous Fungi (2010), pp. 25-35.
Edited by K. A. Borkovich & D. J. Ebbole. Washington, DC: ASM Press

INTRODUCTION

For many years, filamentous fungi have been used for the industrial production of a large variety of metabolites and proteins. A well-known example of a fungal bioprocess is the production of the secondary metabolite penicillin by *Penicillium chrysogenum*, developed about 60 years ago (Ligon, 2004). Fungal production processes of other β-lactam antibiotics as well as drugs such as hypolipidemic agents (e.g., lovastatin by *Aspergillus terreus*) (Tobert, 2003), have been developed since. Furthermore, many of the commercial biological production processes for organic acids are fungal bioprocesses, including the production of citric, gluconic, and itaconic acid by *Aspergillus* species or lactic acid by Rhizopus oryzae (Magnuson & Lasure, 2004). Filamentous fungi also play an important role in the industrial production of proteins and enzymes. In particular, *Trichoderma* and *Aspergillus* species, but also *Penicillium* and *Rhizopus* species, are used to produce a large number of different enzymes, e.g., (hemi)cellulases, xylanases, chitinases, amylases, proteases, and many more (see the list of commercial enzymes from the Association of Manufacturers and Formulators of Enzyme Products[1]). The first industrial fungal bioprocess for proteins dates back even further than that for penicillin. For instance, the product takadiastase appeared on the market in 1894 and is in fact fungal amylase produced by *Aspergillus oryzae* (Gwynne & Devchand, 1992).

Some of the above-mentioned production processes have been developed and optimized over a period of decades, like penicillin, citric acid and amylase; others have been developed more recently and are still being optimized to reach commercial viable production levels. This is particularly true for production of non-native proteins by use of genetically engineered fungal strains. This chapter discusses approaches to select targets for improvement of production processes, with special focus on the application of functional genomics technologies as an unbiased approach towards target selection.

OPTIMIZATION OF FUNGAL PRODUCTION PROCESS

The development of a fungal production process starts with the selection of a strain that produces the compound of interest or with the construction of such a strain. Once this strain is available, production levels need to be increased in order for the process to become economically viable. Optimization of the fungal production process, or any bioprocess for that matter, can be achieved by an iterative cycle of strain

[1] http://www.amfep.org/list.html; August 24, 2010

improvement and/or process optimization (Fig. 1). Process optimization includes improving medium performance as well as identifying optimal environmental process parameters, such as pH, temperature, and aeration. Many techniques are available for process optimization: straightforward methods like the change-one-factor-at-the-time approach or more advanced methods using the experimental design approach, for which various design and optimization techniques are available (Kennedy & Krouse, 1999; Weuster-Botz, 2000). Many of these techniques rely on prior knowledge of components and environmental parameters likely to affect product yields. This obviously means that many more components and parameters are overlooked that could be beneficial to bioprocess performance, but about which no prior knowledge is available. Similarly, strain optimizations until now mainly include alleviating bottlenecks identified in case-by-case studies. Often only the obvious targets for metabolic engineering are addressed (van der Werf, 2005). In the case of protein production, targeting known putative bottlenecks at the post-transcriptional stage is a commonly applied approach of optimizing production levels, for instance by alleviating blockages along the secretion pathway (Conesa *et al.*, 2001) or by eliminating extracellular proteases (Braaksma & Punt, 2008). From the almost infinite number of genetic changes that can be introduced by overexpression or knocking out of genes, only those that are known from the current and generally limited knowledge of the metabolic pathway are selected to optimize product formation. Biological processes or interactions that are not currently known to be important for bioproduct formation or that are not yet known to exist are not taken into account.

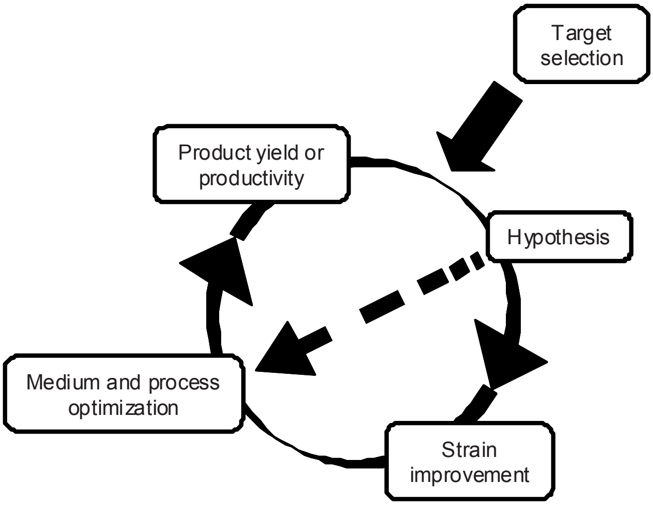

Target selection
Product yield or productivity
Hypothesis
Medium and process optimization
Strain improvement

Fig 1. Iterative cycle of strain improvement and/or process optimization.

In our research we have aimed at using a strain and process development approach which is not *a priori* hypothesis driven but relies on first acquiring data sets rich in information with regard to the bioprocess under study from functional genomics technologies and using these for target selection from the broadest possible ranges of expressed genes (transcriptomics), proteins (proteomics), or metabolites (metabolomics). In this chapter such a systems biology approach, based on the information gathered with functional genomics technologies and in combination with multivariate data analysis tools, is discussed as a method to achieve unbiased selection and ranking of targets for both strain improvement and bioprocess optimization.

TOP-DOWN SYSTEMS BIOLOGY

In systems biology the organism is studied as an integrated and interacting network of genes, proteins, and biochemical reactions. Principally, at its extreme, two approaches are recognized within systems biology: top-down and bottom-up systems biology (Bruggeman & Westerhoff, 2007). In bottom-up systems biology, biological knowledge is used as the starting point and a comprehensive mathematical model of the biological system under study is built. In fungal research metabolic stoichiometric or kinetic models and metabolic network topology models have been used for a systems-level investigation of mainly *P. chrysogenum* and *Aspergillus* species (David *et al.*, 2006; Andersen *et al.*, 2008a; Melzer *et al.*, 2007; Gheshlaghi *et al.*, 2007; Nasution *et al.*, 2008). Similar to the more classical approaches for target selection, these methods require prior knowledge about the studied system. The models are built from known components only and demand an extensive knowledge of the individual parts of the model, and they exclude all components and reactions whose functions are not yet (fully) known.

In contrast, in top-down systems biology, data are used as the starting point and statistical data mining approaches are applied to come to a comprehensive understanding of the biological system. The principal behind top-down systems biology is that molecular components that respond similarly to changes in the experimental conditions are somehow functionally related. No other prior assumptions regarding the interactions of the studied molecular components are required. This allows the study of complex and relatively poorly characterized processes and strains, as extensive knowledge of the studied organism or process is not necessary. In this top-down systems biology approach there is also no *a priori* focus on specific biomolecules expected to relate to the biological question. Therefore, this approach also enables the discovery of previously unknown or unexpected relations between specific biomolecules and the biological process studied. Despite the potential of top-down systems biology, the great majority of scientists applying systems biology use a bottom-up systems biology approach. The reluctances towards top-down systems biology might relate to the risk of being overwhelmed by the enormous quantity of data that arise from functional genomics technologies such as metabolomics and transcriptomics. The challenge is to be able to extract relevant information from these data sets. Principally, the success of this approach depends on balancing three interlinked key factors: (i) definition of the biological question, (ii) experimental design, and (iii) the data analysis tool (Fig. 2). These three factors are discussed in more detail below.

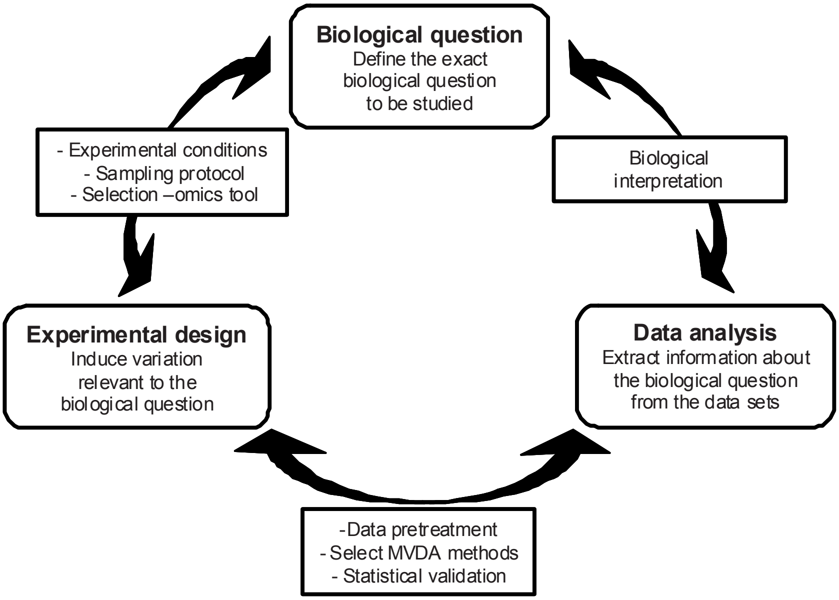

Biological question
Define the exact biological question to be studied
- Experimental conditions
- Sampling protocol
- Selection –omics tool
Biological interpretation
Experimental design
Induce variation relevant to the biological question
Data analysis
Extract information about the biological question from the data sets
-Data pretreatment
- Select MVDA methods
- Statistical validation

Fig. 2. Key conditions and their relation to a successful systems biology study. In top-down systems biology, three interlinked factors are crucial for success: (i) the biological question, (ii) the experimental design, and (iii) data analysis.

THE BIOLOGICAL QUESTION

A clear definition of the biological question to be answered is the crucial starting point in any top-down systems biology research project, because only then can a suitable experimental setup and data analysis strategy be selected (van der Werf *et al.*, 2005; Trygg *et al.*, 2007). To explain this in more practical terms, two examples are given of ways to define the biological question in a study to gain more insight in the regulation of the proteolytic system of *Aspergillus niger*. First, when this problem is approached on a metabolic level, the biological question could be, “Which metabolites induce protease activity in *A. niger*?” On the other hand, when this problem is approached on a genetic level the biological question could be stated as, “Which transcriptional regulators are associated with protease activity in *A. niger*?” In the first case metabolite levels are the relevant biomolecules to be measured, in the second case transcript levels are to be determined, and in both cases protease activities will have to be determined. What is important is that the biological question be translated into a quantifiable biomolecule level, which can be measured at different biochemical levels (i.e., at the transcriptome, metabolome, proteome level). In addition, it is often possible to specify a quantifiable phenotype that is relevant for the biological question, such as protease activity in this case. It is also very important to clearly define this phenotype. For instance, in the production of a biological compound or activity, among others, the following definitions of phenotypes could be chosen for improvement: concentration (in grams per litre) or activity (in units per litre); specific concentration or activity (in grams per gram dry cell weight or in units per gram dry cell weight); productivity (in grams per litre per hour or in units per litre per hour); specific productivity (in grams per gram dry cell weight per hour or in units per gram dry cell weight per hour). When reducing costs of nutrients is the key goal, one could also think of defining the phenotype as cost of nutrients per unit product (in U.S. dollars per gram of product) or cost of nutrients per unit productivity (in U.S. dollars per gram of product formed per litre per hour) (Kennedy & Krouse, 1999). The biological question and its translation into a practical format strongly influence the other key factors of a top-down systems biology study, i.e., experimental design and data analysis. The experimental setup should ensure that experimental conditions that induce variation relevant for the biological question are selected and that data analysis is able to extract the information relevant to the biological question from functional genomics data set.

EXPERIMENTAL DESIGN

Based on the biological question, the experimental design of the top-down systems biology study should be aimed at generating large information-rich data sets in order for data analysis to extract relevant biological information from the data set. Not only experimental conditions for the experimental design should be considered, but also sampling, sample work-up, and the functional genomics tool to be used to analyze the samples.

Experimental conditions

The first step in establishing how to plan and conduct the experiments is to identify those parameters affecting the response of the phenotype. These parameters can be process type (batch, fed-batch, or continuous), environmental conditions such as pH and nutrients, or selected strains. In the case of using various mutant strains to induce variation in the data set (for an example, see Askenazi *et al.*, 2003), one should keep in mind that each strain may have its own bottleneck, making identification of specific targets for a general improvement more complex. When a phenotype relevant to the biological question is available, the experimental conditions should be targeted to induce variation in this phenotype. When it is unclear what experimental factors are involved in the induction of biological variation relevant to the biological problem, screening experiments need to be conducted to obtain more information regarding these experimental factors.

Traditionally, one of the most frequently used approaches to study which parameters affect biological responses is the change-one-factor-at-a-time approach, in which one independent variable is studied while all others are fixed at a specific level. An advantage of this simple and easy method is that any change in response can be attributed to a specific change. On the other hand, this change-one-factor-at-a-time approach has some serious drawbacks, perhaps the most important being that possible interactions between components are ignored. As a result, this approach frequently fails to find optimal conditions for experiments. Another disadvantage is the unnecessarily large number of experiments that are required when testing more than a few variables. Therefore, the change-one-factor-at-a-time method is acknowledged to have severe shortcomings and is more and more being replaced by statistics-based experimental designs, also called “Design of Experiments”. For an initial screening of factors possibly related to the biological question, different types of experimental designs, so-called screening designs, are available, including the full factorial design (Lundstedt *et al.*, 1998). In a full factorial design, every level of a factor is investigated at all levels of all other factors. Often the factors are investigated at two levels, requiring a number of runs equal to 2^k for k factors, which results in a large number of experiments when many factors are investigated (Fig. 3). When the factors are investigated at three or more levels, requiring 3^k runs in the case of three levels and n^k runs for n levels, the number of experiments rapidly becomes impracticable. To reduce the number of experiments without the loss of too much information, several experimental designs derived from the full factorial design are available. The most commonly used one is the fractional factorial design (Lundstedt *et al.*, 1998; Trygg *et al.*, 2006), which requires only n^{k-p} number of runs, with k as the number of investigated factors at *n* different levels, and *p* describing the size of the fraction of the full factorial used. With this type of design, three-way and higher interactions are ignored. Another useful screening tool is the Plackett-Burman design (Plackett & Burman, 1946; Weuster-Botz, 2000). This experimental design is a variation on the fractional factorial design, but instead of ignoring only higher interactions it considers all interactions between factors negligible. The downside of these two last designs is that when interactions between factors are not negligible, they are confounded with the estimated effects. This means that the estimated effects and those interaction effects cannot be distinguished from one another.

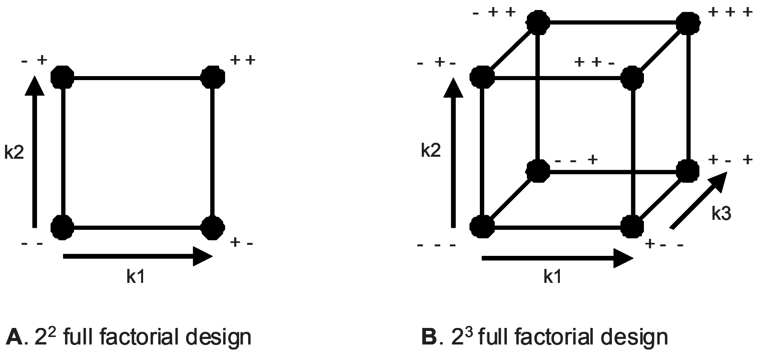


Fig. 3. Full factorial designs, with two factors (A) or three factors (B) investigated at two different levels.

Based on this first phase, the main factors relevant to the biological question under study are selected for the final setup of experiments for the top-down systems biology study. In principal, statistical experimental designs for this phase can be any of the methods as described above. While in the screening phase the goal was to find out a little about many factors, in this phase the goals is to extract the maximum amount of information from the experiments, preferably in the fewest number of runs. Types of statistical experimental designs suitable for this phase of the study include central composite designs and Box-Behnken designs, which are both based on (fractional) factorial designs, or D-optimal designs, a computer-aided design method (Kennedy & Krouse, 1999; Trygg *et al.*, 2006; Lundstedt *et al.*, 1998). On top of that, response surface methodology can be applied to generate a data set with an evenly distributed variation. Response surface methodology is commonly used in industry for process optimization (Dobrev *et al.*, 2007; Li *et al.*, 2007). Based on a set of designed experiments, e.g., from a factorial design, a model that predicts the biological response to different levels of the various factors included in the study is built. In contrast, from such a model, conditions that will result in various levels of the relevant biological response for the top-down systems biology study can be selected.

Selection of a functional genomics tool

Selection of the functional genomics tool to be used in a top-down systems biology study depends on the level at which the biological phenomena relevant for the biological question occur. With transcriptomics the expression levels of mRNA under a given condition are examined. The transcriptome reacts very fast, within in a few minutes, to environmental changes. This makes transcriptomics a very suitable tool to study the cell exposed to changing environmental conditions, such as the addition of toxic or chemical compounds (Arvas *et al.*, 2006; Guillemette *et al.*, 2007) or transfer from one medium to another (Yuan *et al.*, 2006). However, mRNA levels do not directly correlate to the levels of the encoded protein, due to post-transcriptional regulation steps at the level of mRNA stability, processing, and translation. Therefore, transcriptomics is only an indirect approach to study the function of a cell. On the other hand, the proteome and the metabolome together determine the actual function of the cell (the phenotype) (Oliver, 2000).

The proteome, meaning all proteins present at a given moment under defined environmental conditions, gives an indication of which metabolic pathways occur under those conditions (Kim *et al.*, 2007a), as for many proteins are enzymes that catalyze biochemical reactions. In contrast to transcriptomics, quantitative proteomics is still far from being a comprehensive analysis tool, mainly due to the limited dynamic detection range and poor reproducibility of proteomic analysis. Because of this there is a very strong bias towards identifying only the more abundant proteins in a complex proteome sample. Nonetheless, to study post-translational modifications of proteins, such as phosporylation and glycosylation, proteomics is the most obvious tool of choice (Fryksdale *et al.*, 2002; Kim *et al.*, 2007b).

The metabolome of the cell, i.e., all metabolites present in a cell at a certain moment, provides valuable information about the regulatory or catalytic properties of either mRNA or enzyme, as metabolites are downstream of all genome and proteome regulatory structures (Oldiges *et al.*, 2007). As the metabolome is closest to the phenotype of a cell, it will be most relevant in order to understand biological functioning. Similar to what was noted above for proteomics, full coverage of the complete metabolome is not (yet) accomplished by the available analytical platforms, although some metabolomics platform are approaching the ultimate goal of providing a universal platform for the comprehensive and quantitative analysis of microbial metabolomes (van der Werf *et al.*, 2007).

Sampling strategy

The sampling strategy is part of the experimental setup and describes when and how samples for the functional genomics analysis are collected. It embraces two main issues, namely, collecting the sample at a time point where the biological response relevant to the biological question is present and ensuring that levels of biomolecules remain unchanged from the moment of sampling. Concerning this first issue, if it is unknown beforehand which phases during the cultivation contain information related to the biological question, the sampling protocol should cover all possibly relevant growth phases and phase transitions (Trygg *et al.*, 2007). At the same time, practical matters have to be considered as well. For instance, the sampling volumes can limit the number of obtainable samples, or the costs of sample analysis can influence the sampling strategy. In the case of continuous cultures, time issues are of no importance, but due to technical difficulties this fermentation technique is not as commonly applied in fungal research as it is in research involving other microorganisms. Besides, with the application of continuous cultures the approach is quite different, as time is no longer a factor, excluding longitudinal effects (e.g., induction or other perturbations during the fermentation process). In addition, it should be noted that although the process conditions are fixed during continuous cultures, changes in the production organism are frequently observed (Swift *et al.*, 1998; Withers *et al.*, 1995), making continuous cultures prone to transitions, albeit of a different kind.

The second issue relates to the high turnover of mRNA and metabolites (for proteins this is not so much of an issue), risking the introduction of unwanted changes in RNA or metabolite levels during sample harvesting or work-up. In order to obtain samples that reflect the state of the cell under the environmental conditions at the time of harvesting, rapid sampling (Nasution *et al.*, 2006) and immediate inactivation (quenching) of the cellular metabolism are a necessity. In the literature, the quenching methods used for filamentous fungi mainly include rapid filtration followed by immediate freezing of the cells (mostly used for transcriptomics samples) (David *et al.*, 2006) or dilution of the cells in a methanol solution of -45 °C (more often used for metabolomics samples) (Ruijter & Visser, 1996; Nasution *et al.*, 2006; Kouskoumvekaki *et al.*, 2008).

After quenching the cells, conditions should be maintained during sample work-up in order to prevent changes in the metabolite composition of RNA levels due to residual enzymatic activity present in the samples. Extraction of RNA from mycelium is often accomplished by disruption of the cells by either grinding under liquid nitrogen using a mortar and pestle (Kimura *et al.*, 2008; Foreman *et al.*, 2003) or bead-milling at temperatures of approximately 4 °C (Andersen *et al.*, 2008b), followed by a standard RNA isolation protocol. Extraction of proteins is done in a similar way, without the stringent control of temperature (Carberry *et al.*, 2006). For fungal metabolomics samples, two methods in particular have been described for extracting metabolites from the cells. The first is boiling the cells in an ethanol-buffer solution and subsequent reduction of the volume by evaporation in a rotavapor (Nasution *et al.*, 2006). The second is chloroform extraction at -45 °C (Ruijter & Visser, 1996).

A final issue to consider as part of the sampling strategy is replicates. As the total variation in data set is the sum of technical, uninduced biological, and induced biological variation, repeated measurements may be necessary to estimate the individual contributions of these various parts. However, in general the biological variation is much larger than variation induced by sample work-up or variation in the analytical method (van den Berg *et al.*, 2006). This makes repeating the experimental procedure with identical samples not very worthwhile in most cases. Some biological replicates will have to be included in the experimental design to estimate the overall uninduced biological variation due to small differences between biological conditions or biological variability. In this way, the induced biological variation can be established, as calculated on the basis of the differences between the experimental conditions.

Based on the various aspects of the experimental setup discussed above, it becomes clear that it is necessary to balance the demands from the biological question and the data analysis on one side with practical considerations on the other.

DATA ANALYSIS

After having generated data sets under several different conditions with hundreds or thousands of proteins, mRNAs, or metabolites, the remaining challenge is to extract information about the biological question from these enormous data sets. Multivariate data analysis (MVDA) tools are preferably used, as those tools take into consideration the intrinsic interdependency of the biomolecules. But before the data sets can be analyzed by MVDA tools, the data output from the various functional genomics methods often requires data pretreatment.

Data pretreatment methods

In addition to the specific preprocessing steps of the data output from the various genomic methods, such as deconvolution of data files generated by gas chromatography-mass spectrometry for metabolomics (van der Werf *et al.*, 2005) or normalization of cDNA microarrays (Leung & Cavalieri, 2003), another critical step before applying MVDA tools is data pretreatment of the data sets. Data pretreatment procedures correct for the influence of factors such as the abundance of a biomolecule or the magnitude of the change, which are generally not a reflection for the importance of a biomolecule (van den Berg *et al.*, 2006). Appropriate data pretreatment methods will articulate the *biological* information content and will consequently allow more relevant biological interpretation of the data set. Three classes of data pretreatment methods can be distinguished: centring, scaling, and transformation. The last two methods are always applied in combination with centring. In MVDA, mean-centring and autoscaling are the two most commonly used data pretreatment methods. With mean centring, the average level of a biomolecule is subtracted from each individual experiment, thereby adjusting for differences in the offset between high-abundance and low-abundance biomolecules. With autoscaling, the values are subsequently divided by the standard deviation of each biomolecule, adjusting for disparities in increase/decrease differences between the various biomolecules. In addition to these two methods, range scaling holds great promise, as the mean centred values are not divided by a statistical measure for data spread, as is the case with autoscaling, but by a biological measure, namely, the biological range. The biological range is the difference between the minimal and maximal levels reached by a certain biomolecule in a set of experiments. In Fig. 4 the effect of data pretreatment on principal component analysis (PCA) results of a metabolomics data set of *Trichoderma reesei* is shown (van der Werf *et al.*, unpublished data). With data pretreatment the biological information content in the data set is accentuated. In this particular case, it is range scaling that especially emphasizes the biological variation among the different biological groups. This data pretreatment method allows a clear separation of these different groups, whereas no grouping or a less obvious grouping is observed in the data sets when the other two methods are used.

MVDA tools

Choices in data analysis strategy are influenced by the biological question, the characteristics of the experimental design, the behaviour of the relevant biomolecules, and the dimensions of the data set. There are various MVDA methods that address different biological questions. In general, these methods can be divided in two main groups, namely, unsupervised methods and supervised methods. Unsupervised methods include PCA (Jackson, 1991; Jolliffe, 2002) or hierarchical clustering analysis (Eisen *et al.*, 1998) that visualize relations/patterns in data sets without prior knowledge.

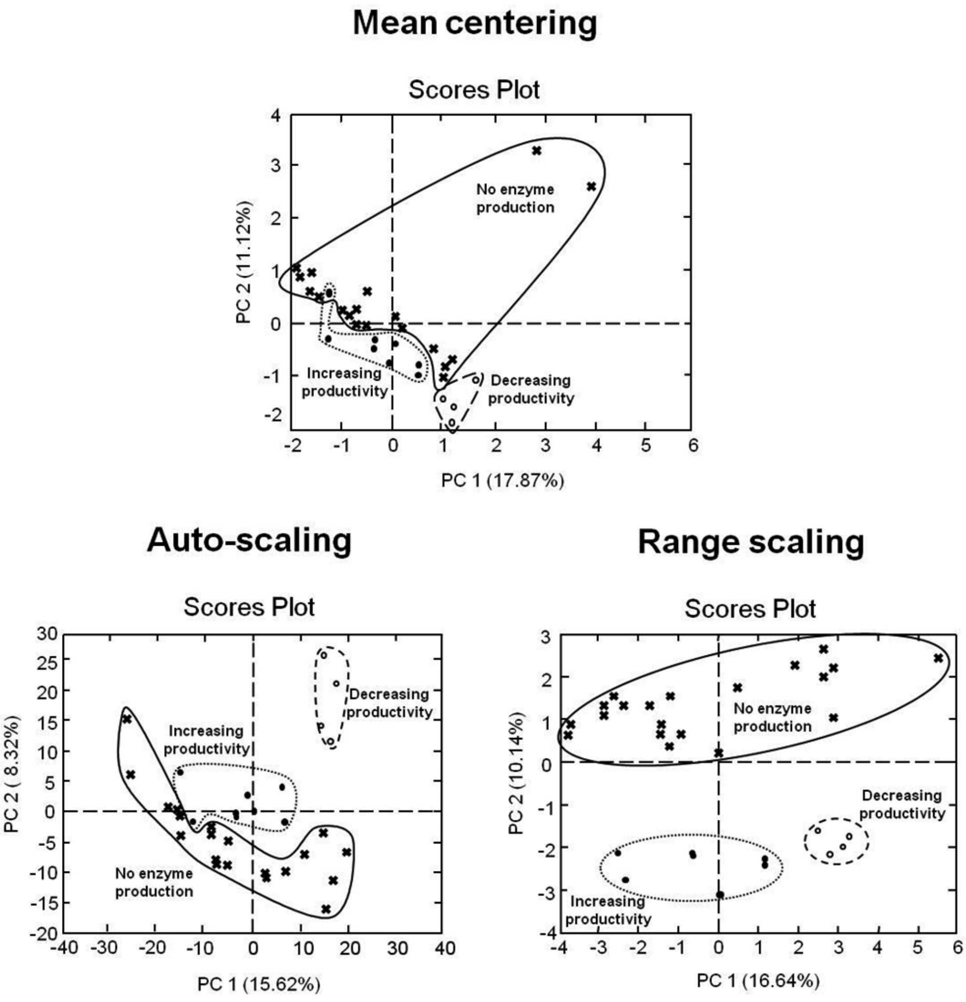


Fig. 4. The effect of mean scaling, autoscaling, or range scaling of metabolomics data sets on PCA data results. The data sets are derived from research related to induction of cellulase activity in *T. reesei* (van der Werf *et al.*, unpublished data). The metabolomes of three groups of samples (no enzyme production, increasing productivity, and decreasing productivity) were analyzed and pretreated with these three different approaches and subsequently analyzed by PCA.

Supervised methods, which include regression methods such as partial least squares (Geladi & Kowalski, 1986) and principal component regression (Mardia *et al.*, 1979) or classification methods such as partial least squares-discriminant analysis (Barker & Rayens, 2003) and principal component discriminant analysis (Hoogerbrugge *et al.*, 1983), do the same as unsupervised methods while at the same time prior knowledge about one or more biological properties of the data set are taken into consideration. Discriminant methods are particularly suitable for samples with no quantifiable phenotype other than the presence or absence of a certain biological characteristic, e.g., morphological traits such as colour or hyperbranching or certain environmental conditions or perturbations. For discriminant methods, this means that the samples are divided in (biological) groups, e.g., a group of samples from the wild-type strain and a group of samples from a mutant. Although each sample within such a biological group is designated as equal, there will always be biomolecules correlating to specific groups that are irrelevant to the biological question under study (so-called chance correlations). Therefore, when it is possible to express the phenotype as a numerical figure, this is preferred as the risk of chance correlations is reduced when analyzing such data with regression methods. Regression methods find correlations between a numerical phenotype response and the biomolecule composition for the different samples in the data set. Regression methods are preferably applied to a set of experiments with large and evenly distributed variation in the biological response of interest.

In addition, validation of the data analysis results is of crucial importance, as it will provide an indication for the risk that correlations were found by chance due to the relatively low number of samples in relation to the large number of measured biomolecules. As multivariate statistical methods were developed for data sets containing many samples and few variables, this is a serious risk. Frequently applied data analysis validation strategies in top-down systems biology are cross validation, permutation, jackknifing, and bootstrapping (Rubingh *et al.*, 2006; Westerhuis *et al.*, 2008; Efron & Tibshirani, 1993). Based on the results of these validation steps, the reliability of the obtained models is established. Finally, a list of biomolecules can be obtained with the largest contribution to the model, i.e., those with the highest absolute regression factor. The biomolecules with the highest ranking are considered to be most relevant to the studied biological phenomenon.

Biological interpretation

Based on the list of biomolecules identified by the MVDA tools as being important in relation to the question under study, targets for improvement of the production process have to be selected. There is a possibility with MVDA tools that biomolecules that do not show an unambiguous interaction with the specific biological question will be identified. Therefore, one of the first steps is to go back to the original data sets and examine fluctuation of the concentration of the biomolecule in relation to the studied phenotype. Moreover, not all biomolecules that exhibit an apparently strong interaction with the studied phenotype are biologically related to it. For that reason, as much information as possible should be acquired about the biological function of these biomolecules in the context of the biological question under study. From this knowledge, biological hypotheses will have to be formulated and new experiments will have to be setup to test them. For targets from transcriptomics studies, this can be quite straightforward, by either overexpression or deletion of the designated relevant genes, depending on a positive or negative correlation to the phenotype. On the other hand, several options for the ultimate improvement of the process are possible for targets identified in metabolomics studies. An easy way to increase product levels might be the addition or omission to the growth medium of a relevant metabolite identified by data analysis. This approach bears the risk that the transport of the compound into the cell will limit its suitability. More complex is the segue from a relevant metabolite identified by using metabolomics relevant to a gene target for metabolic engineering. This requires knowledge about the metabolic pathway(s) involving the metabolite and its putative (allosteric) regulatory effects. Even then, it is not straightforward to translate this knowledge into a gene target. For instance, when a positive correlation between the product of interest and an intermediate in the biosynthesis route for the product is observed (increase in the concentration of this intermediate correlates with elevated product levels), the enzyme converting the intermediate is not active enough and the corresponding gene should therefore be overexpressed. In another example elevated product levels correlate with increased levels of an intermediate via a side reaction. Elimination of this competitive pathway by deletion of the corresponding gene should result in an increased flux through the biosynthetic pathway of interest and thus elevated levels of the desired product.

CONCLUSIONS

The available selection methods for relevant targets for fungal strain and process development, or for that matter any microbial production process, have been very successful in numerous cases. However, the exclusion of all biological processes or interactions that are not currently known to exist has been shown to hamper further improvement while using these approaches. Recently introduced functional genomics technologies in combination with MVDA tools enable an open and comprehensive top-down systems biology approach towards target selection. Nevertheless, the success of such an approach depends heavily on a systematic study covering all aspects, from a clear description of the biological question up to statistical data analysis. As this involves knowledge beyond the biologist expertise (e.g., biostatistics), the assistance of experts in those fields will be indispensable. Due to its unbiased nature, a successful top-down system biology approach will provide a new boost in the ongoing cycle of bioprocess optimization.

SUMMARY

The filamentous fungus *Aspergillus niger* is widely used in industrial biotechnology for the production of many substances, including citric acid and a broad range of enzymes. Because of its natural ability to secrete large amounts of proteins, it also has been explored as a host organism for the commercial production of enzymes from fungal as well as non-fungal origin. However, although heterologous fungal proteins are efficiently expressed, so far the levels of most non-fungal proteins produced are too low to be commercially interesting. Extracellular degradation of non-fungal proteins by the fungus' native protein-degrading enzymes, so-called proteases, has long been recognized as one of the bottlenecks that reduce yield.

Currently, approaches for the development and use of *A. niger* as a production host are rapidly changing by recent advances in fungal genomics and related functional genomics tools such as microarrays, proteomics and metabolomics. The aim of the work presented in this thesis was to use functional genomics tools to study the production of extracellular proteases in *A. niger*.

Chapter 1 reviews the role of proteases in strain and process development of *A. niger* and other aspergilli and provides an outlook on how functional genomics techniques may play a role in further understanding the proteolytic system of aspergilli. To this day, classical mutagenesis and molecular genetic methods both are successfully applied to generate strains with reduced protease activity. With the resulting mutants and disruptants a significant improvement of heterologous protein production levels can be reached. Disruption strategies have not only focused on individual protease-encoding genes, but also on specific regulators of protease genes as well as wide-domain regulators. However, the latter approach seems unsuitable to generate protease-deficient fungal host strains for protein production due to pleiotropic growth defects of wide-domain regulatory mutants. In addition to strain development, the selection and development of fermentation conditions that repress protease production can improve heterologous protein production, but this has never been investigated systematically. The few available studies on this subject have mainly focused on the environmental parameters pH, carbon and nitrogen source.

A systematic study of the influence of several environmental factors on the production of extracellular proteases of *A. niger* in controlled batch cultivations is described in **Chapter 2**. In the first part of the study, a change-one-factor-at-a-time approach was used to establish the factors and their levels that affect protease activity. Subsequently, four selected factors, i.e. carbon source, nitrogen source, nitrogen concentration and pH, were investigated further using a full two-level factorial experimental design. Amongst others, the results showed a clear interaction effect between nitrogen source and nitrogen concentration. Interaction effects between environmental factors in relation to protease secretion of *A. niger* have not been reported before. Due to the occurrence of interaction effects, the selection of environmental factors to reduce protease activity is not straightforward, as unexpected antagonistic or synergistic effects may occur. Furthermore, in addition to maximum protease activity the effects of the process parameters on five other protease-related phenotypes, such as maximum specific protease activity and maximum protease productivity, were investigated. The results indicated significant differences in the effect of the environmental parameters on the various protease-related phenotypes. For instance, pH significantly affected final levels of protease activity, but not protease productivity. The experiments of the full factorial design showed large and evenly distributed variation of protease activity, and were therefore a suitable starting point for a full-scale systems biology approach (Chapters 3-5).

The availability of the sequenced genome of *A. niger* has allowed the prediction of the fraction of potentially secreted proteins by scanning for the presence of a N-terminal signal peptide (SP). Due to gene-model errors, the *in silico* predicted secretome is not an accurate description of the real secretome. Moreover, not all proteins with a SP are actually secreted; some are resident endoplasmic reticulum proteins. In **Chapter 3** an improved list of potential SP directed proteins encoded by the *A. niger* genome is presented. When compiling this list, in addition to SP predictions of *A. niger* CBS 513.88, SP predictions of the best homologs of *A. niger* ATCC 1015 and three neighbouring *Aspergillus* species were taken into account as well. We propose that the SP prediction of an *A. niger* non-SP protein was likely to be incorrect, when a SP prediction was present for the majority of the homologous proteins and vice versa. Four of those likely false negative SP predictions and one likely false positive SP prediction were re-evaluated by aligning N-terminal ends. In all cases of the false negative SP predictions, selection of an alternative start codon in the most likely reading frame would add a predicted SP feature to the alternative N-terminal end of the predicted CBS protein.

As a complement to the *in silico* data, shotgun proteomics approach was used to determine the secretome associated with *A. niger* growth and upon carbon source depletion. In our study, more than 200 proteins with a predicted SP were identified. Additionally, at least two secreted proteins, including an aspartic protease (An01g00370) with a strong similarity to aspergillopepsin apnS of *A. phoenicis*, were identified that apparently use a non-classical route for secretion. Of the other 19

proteases identified in this study and that did have a predicted SP, one aspartic protease (ATCC 53364) lacked a CBS gene-model.

The secretome state was observed to change with the growth condition. Growth on sorbitol lead to the secretion of a range of carbohydrate active enzymes, but a pectinolytic subset was specifically induced during growth on galacturonic acid. Carbon source exhaustion induced the expression of proteases, as was already observed in Chapter 2. However, it was not the number of different proteases, but the relative contribution of specific proteases that increased upon carbon source starvation. These included the most prominent protease in *A. niger*, the aspartic protease aspergillopepsin A, for which the number of spectral counts was almost upto seven times higher under carbon source exhaustion compared to the growth conditions. Also data sets from other studies confirmed the relevance of the SP-classifier approach.

One of the application fields of functional genomics tools is the optimization of microbial production processes. In **Chapter 4** we investigated the influence of the choice of the quantitative phenotype to be optimized on the outcome of a optimization strategy using metabolomics. To this end, we evaluated the production by *A. niger* of the industrially relevant products glucoamylase and protease. For both products different quantitative phenotypes associated with production were defined, taking the different time points of sampling into account as well. The information content of the metabolomics data set in relation to all these different quantitative phenotypes defined was evaluated using the multivariate data analysis tool partial least squares (PLS). The results showed that the effect of different ways to define the quantitative phenotype on the information content and resulting targets for production process optimization is much smaller than the effect of the time point of sampling.

A detailed analysis of specific metabolites identified by PLS as important to the two products under study revealed that for glucoamylase activity various sugar-derivatives were correlating. However, the identified disaccharides can be either inducers of glucoamylase secretion or formed from glucose by transglucosylation activity from glucoamylase. Especially in the first case, the disaccharides are relevant in relation to strain improvement strategies. Glucose-6-phosphate isomerase was another potential target identified by PLS for optimization of glucoamlyase in *A. niger*. For the reduction of protease activity no obvious targets were found. It was anticipated that the relation between intracellular metabolite concentrations and extracellular protease activity would not be straightforward, because multiple enzyme activities are involved in proteolytic degradation, as was also found in Chapter 3.

Chapter 5 provides a large-scale, global view of the transcriptional response of *A. niger* by the clustering of co-expressed genes. Gene co-expression networks were constructed on the basis of DNA microarray data sets from two experimental approaches. In one approach *A. niger* cultures were relatively mild perturbated by the addition of low amounts of inducer. In the other approach *A. niger* cultures were severely pertubated by the imposed environmental conditions, which included carbon source starvation.

Initially, gene co-expression networks for both the individual and combined data sets were constructed using only a set of conserved genes. Comparative analysis revealed the existence of modules, some of which were present in all three networks, while others were condition-specific. Next, all protein-coding *A. niger* genes, including hypothetical and poorly conserved genes, were integrated to the co-expression analysis by the application of module-derived consensus expression profiles. Evidence for the biological relevance of the discovered modules was provided by the overrepresentation of specific Gene Ontology terms within the modules, the overrepresentation of genes which related to specific biochemical pathways, and the presence of conserved motifs in the upstream region of many genes in several of the modules. Some of the conserved sequence motifs detected represented known binding sites for transcription factors. These included the amino acid metabolism-related transcription factor CpcA and the fatty acid metabolism-related transcription factors, FarA and FarB. In addition, not previously described putative transcription factor binding sites were discovered for the module containing genes encoding cytosolic ribosomal proteins and the module with genes related to 'gene expression'.

Finally, in **Chapter 6** a top-down systems biology approach, based on the information gathered with functional genomics technologies and in combination with multivariate data analysis tools, is discussed as a method to achieve unbiased selection and ranking of targets for both strain improvement and bioprocess optimization.

SAMENVATTING

De schimmel *Aspergillus niger* wordt veel gebruikt in de industriële biotechnologie voor de productie van verschillende metabolieten, waaronder citroenzuur, en een grote verscheidenheid aan enzymen. Van nature is *A. niger* in staat grote hoeveelheden eiwit uit te scheiden. Door deze eigenschap is bij de opkomst van genetische modificatietechnieken ook onderzocht of deze schimmel kon worden gebruikt als productieorganisme voor de commerciële productie van enzymen die van nature in andere schimmels of andere organismen voorkomen. Deze soortvreemde (heterologe) eiwitten worden vaak efficiënt tot expressie gebracht als het eiwitten van andere schimmels betreft. Wanneer het echter om heterologe eiwitten met een andere oorsprong dan schimmels gaat, is de productie tot op heden meestal onvoldoende om rendabel te zijn. Afbraak van deze heterologe eiwitten door eiwitafbrekende enzymen, zogenaamde proteases die de schimmel van nature uitscheidt, wordt als één van de oorzaken gezien van deze tegenvallende opbrengsten.

De recente opmars van systeembiologie, waaronder ook de onderzoeksgebieden genomics, transcriptomics, proteomics en metabolomics vallen, heeft een nieuwe impuls gegeven aan de ontwikkeling en toepassing van *A. niger* als een productieorganisme. Het werk dat in dit proefschrift beschreven is, had tot doel de productie van extracellulaire proteases in *A. niger* te onderzoeken, waarbij gebruik is gemaakt van een systeembiologie aanpak.

In **Hoofdstuk 1** wordt de rol van proteases in stam- en procesverbetering van *A. niger* en andere aspergilli beschreven. Ook wordt vooruitgeblikt op de rol die systeembiologie zou kunnen spelen om het begrip van het proteolytische systeem van aspergilli te vergroten. Zowel klassieke mutagenese als moleculair genetische methoden worden tot op de dag van vandaag succesvol toegepast om stammen te genereren met een verlaagde protease activiteit. Het gebruik van de resulterende stammen voor de productie van heterologe eiwitten kan tot een significante toename van de geproduceerde hoeveelheden eiwit leiden. Strategieën waarbij een gen wordt uitgeschakeld hebben zich niet alleen gericht op individuele protease genen. Ook is gekeken naar het effect van het uitschakelen van specifieke regulatoren van protease genen. Een andere mogelijkheid is het uitschakelen van regulatiegenen die de expressie sturen van een breder scala aan genen, waaronder bepaalde protease genen, als reactie op externe factoren zoals pH of koolstofbron. Deze laatste aanpak lijkt echter ongeschikt voor het genereren van protease-deficiënte gastheerschimmels voor eiwitproductie. Het uitschakelen van zulke regulatoren resulteert namelijk ook in pleiotrope groeidefecten.

Ook de selectie en ontwikkeling van fermentatiecondities waarbij de productie van proteases wordt onderdrukt kan, naast stamontwikkeling, helpen de productie van heterologe eiwitten te verbeteren. Hier is echter nooit systematisch onderzoek naar verricht. De enkele studies die over dit onderwerp zijn verschenen hebben zich met name gericht op de kweekparameters pH, koolstof- en stikstofbron.

Hoofdstuk 2 beschrijft een systematische studie naar de invloed van verschillende omgevingsfactoren op de productie van extracellulaire proteases door *A. niger*. Hierbij is gebruik gemaakt van gecontroleerde batch fermentaties. In het eerste deel van de studie is door steeds één parameter per experiment te veranderen geprobeerd vast te stellen welke factoren met name protease activiteit beïnvloeden. Vervolgens zijn vier geselecteerde factoren, namelijk pH, koolstofbron, stikstofbron en concentratie van de stikstofbron, verder onderzocht. Hiervoor werd een aanpak gebruikt waarbij elke factor op twee niveaus werd getest en in alle mogelijke combinaties met de andere factoren. De resultaten toonden onder andere een duidelijk interactie-effect aan tussen stikstofbron en concentratie van de stikstofbron. Dergelijke interactie-effecten tussen omgevingsfactoren in relatie tot extracellulaire protease productie door *A. niger* zijn nog niet eerder beschreven. Het selecteren van omgevingsfactoren om protease activiteit te reduceren is door deze interactie-effecten niet eenvoudig, omdat onverwachte antagonistische of synergetische effecten kunnen optreden.

Hiernaast is in deze studie niet alleen gekeken naar het effect van procesparameters op de maximale protease activiteit, maar ook op nog vijf andere fenotypen gerelateerd aan protease. De resultaten duidden op significante verschillen in het effect van de omgevingsfactoren op de verschillende protease-gerelateerde fenotypen. De pH beïnvloedde bijvoorbeeld duidelijk de eindwaarden van protease activiteit, maar niet die van protease productiviteit. De uitgevoerde experimenten resulteerden in een grote en evenredig verdeelde variatie van protease activiteit. Dit maakt de resultaten van deze experimenten een geschikt startpunt voor een systeembiologie aanpak (Hoofdstukken 3-5).

Met de beschikbaarheid van de genoomsequentie van *A. niger* is het mogelijk geworden *in silico* het secretoom, oftewel de subset van het proteoom dat alle uitgescheiden eiwitten omvat, te voorspellen. Door alle voorspelde eiwitten te analyseren op de aanwezigheid van een signaalpeptide (SP) aan de N-terminus kan een lijst worden opgesteld met potentieel uitgescheiden eiwitten. Echter, de genmodellen die voor deze voorspellingen worden gebruikt bevatten soms fouten, waardoor het *in silico* voorspelde secretoom niet een accurate weergave is van het daadwerkelijke secretoom. In **Hoofdstuk 3** wordt een verbeterde lijst gepresenteerd met potentiële SP eiwitten gecodeerd door het *A. niger* genoom. Bij het opstellen van deze lijst zijn, naast de SP voorspellingen van *A. niger* CBS 513.88, ook de SP voorspellingen van de beste homologen van *A. niger* ATCC 1015 en drie naburige *Aspergillus* soorten meegenomen. Ons uitgangspunt was dat de SP voorspelling van een *A. niger* non-SP eiwit zeer waarschijnlijk onjuist is, wanneer voor de meerderheid van de homologe eiwitten wel een SP voorspelt is en vice versa.

Als aanvulling op de *in silico* data is een shotgun proteomics aanpak gebruikt om het secretoom van *A. niger* te onderzoeken bij zowel groei- als koolstoflimiterende condities. In onze studie hebben we in het kweekmedium meer dan 200 eiwitten met een SP voorspelling gevonden. Daarnaast zijn ook tenminste twee uitgescheiden eiwitten gevonden waarvoor geen SP voorspeld was. Eén van deze twee eiwitten is een aspartylprotease (An01g00370) dat sterk overeenkomt met aspergillopepsin apnS van *A. phoenicis*. Blijkbaar wordt de export van dit eiwit niet door een SP gedirigeerd. Behalve dit aspartylprotease werden negentien andere proteases in deze studie gevonden die allen wel een SP hadden.

Het secretoom verandert met de groeicondities. Bij groei op sorbitol werd een breed scala aan koolhydraatacterende enzymen gevonden, terwijl bij groei op galacturonzuur vooral een specifieke subset van pectinolytische enzymen werd geïnduceerd. Koolstoflimitatie induceerde de expressie van proteases, zoals ook al was waargenomen in Hoofdstuk 2. Echter, niet zozeer het aantal verschillende proteases veranderde bij koolstoflimitatie, maar de relatieve bijdrage van specifieke proteases nam toe. Hieronder ook het belangrijkste protease in *A. niger*, het aspartylprotease aspergillopepsin.

Eén van de toepassingsgebieden van systeembiologie is het optimaliseren van microbiële productieprocessen. In **Hoofdstuk 4** hebben we onderzocht hoe de keus van een kwantitatief fenotype voor optimalisatie de uitkomst beïnvloedt van een optimalisatiestrategie gebaseerd op een metabolomics aanpak. Hiervoor hebben we de productie van de industrieel belangrijke *A. niger* producten glucoamylase en protease bestudeerd. Voor beide producten zijn verschillende kwantitatieve fenotypen verwant aan productie gedefinieerd. Daarbij is ook rekening gehouden met de verschillende tijdstippen waarop monsters genomen zijn. Met behulp van de multivariate data-analyse techniek partial least squares (PLS) is het informatiegehalte van de metabolomics dataset in relatie tot al deze verschillende kwantitatieve fenotypen geëvalueerd. De resultaten lieten zien dat het tijdstip van bemonstering een veel grotere invloed heeft op het informatiegehalte van de data dan de manier waarop een kwantitatief fenotype is gedefinieerd.

Uit analyse van de metabolieten die door de PLS analyse als belangrijk zijn aangemerkt voor de twee bestudeerde producten bleken verschillende suikers met glucoamylase activiteit te correleren. Echter, de geïdentificeerde disaccharides kunnen zowel mogelijk glucoamylase productie induceren, of producten zijn die door glycosyltransferase activiteit van glucoamylase uit glucose worden gevormd. Vooral in het eerste geval zijn de disaccharides van belang voor stamoptimalisatie. Voor optimalisatie van glucoamylase productie in *A. niger* werd ook glucose-6-fosfaat isomerase als één van de andere interessante aanknopingspunten voor verder onderzoek geïdentificeerd.

Om protease activiteit in *A. niger* te reduceren werden geen direct voor de hand liggende aanknopingspunten voor verder onderzoek gevonden. Er was van te voren al rekening gehouden met het feit dat de relatie tussen intracellulaire metabolieten en extracellulaire protease activiteit gecompliceerd zou zijn, omdat bij proteolytische afbraak verschillende enzymactiviteiten betrokken zijn, zoals ook al in Hoofdstuk 3 was gevonden.

Hoofdstuk 5 geeft een genoom-brede blik op de transcriptierespons van *A. niger* door genen op basis van co-expressie te groeperen. Co-expressie netwerken van genen werden gemaakt op basis van twee verschillende DNA microarray datasets. De ene dataset is gebaseerd op experimenten waarbij *A. niger* relatief mild verstoord werd door kleine hoeveelheden van een inducerende component toe te voegen. Voor de andere dataset werd *A. niger* veel heviger verstoord door verschillende groeicondities op te leggen, waaronder koolstoflimitatie.

In eerste instantie werden voor zowel de individuele als de gecombineerde datasets gen co-expressie netwerken gemaakt op basis van een selectie van geconserveerde genen. Bij het vergelijken van de verschillende netwerken bleken sommige clusters van genen, zogenaamde modules, aanwezig in alle drie netwerken, terwijl andere conditiespecifiek waren. Vervolgens zijn alle *A. niger* genen, waaronder ook hypothetische en slecht geconserveerde genen, geïntegreerd in de co-expressie analyse. Dit werd gedaan door voor elk oorspronkelijke module een representatief expressieprofiel op te stellen en daarmee de expressieprofielen van alle overige genen te vergelijken. De biologische relevantie van de ontdekte modules blijkt uit de oververtegenwoordiging binnen deze modules van genen die gerelateerd zijn aan specifieke biochemische routes. Voor een aantal modules wordt dit tevens ondersteund door de aanwezigheid van geconserveerde motieven in de promotorregio van veel van de genen binnen zo’n module. Enkele van de gevonden geconserveerde sequentiemotieven zijn bekende bindingsites van DNA-bindende transcriptiefactoren. Voorbeelden hiervan zijn de bindingsites van de transcriptie-factor CpcA, die gerelateerd is aan het aminozuurmetabolisme, en die van de transcriptiefactoren FarA and FarB, die gerelateerd zijn aan het verzuurmetabolisme. Daarnaast zijn enkele nog niet eerder beschreven mogelijke transcriptiefactor bindingsites ontdekt voor de module met genen die coderende voor ribosomale eiwitten en de module met genen gerelateerd aan algemene gen expressie processen.

Tenslotte wordt in **Hoofdstuk 6** een ‘top-down’ systeembiologie aanpak, gebaseerd op de informatie verkregen middels functionele genomics technologieën en in combinatie met multivariate data analyse, bediscussieerd als een methode om onbevooroordeeld aanknopingspunten voor stam- en procesverbetering te selecteren en op hun belangrijkheid te rangschikken.

REFERENCES

Affymetrix (2001). Statistical algorithms reference guide, technical report. Santa Clara, CA: Affymetrix.

Affymetrix (2004). GeneChip expression analysis technical manual. Santa Clara, CA: Affymetrix.

Almaas, E. (2007). Biological impacts and context of network theory. *J Exp Biol* **210**, 1548-1558.

Altschul, S.F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

Andersen, M. R. & Nielsen, J. (2009). Current status of systems biology in Aspergilli. *Fungal Genet Biol* **46**, S180-S190.

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., Salazar, M. P., Lehmann, L. & Nielsen, J. (2008b). A trispecies *Aspergillus* microarray: Comparative transcriptomics of three *Aspergillus* species. *Proc Natl Acad Sci U S A* **105,** 4387-4392.

Archer, D. B., MacKenzie, D. A., Jeenes, D. J. & Roberts, I. N. (1992). Proteolytic degradation of heterologous proteins expressed in *Aspergillus niger*. *Biotechnol Lett* **14**, 357-362.

Arvas, M., Pakula, T., Lanthaler, K., Saloheimo, M., Valkonen, M., Suortti, T., Robson, G. & Penttilä, M. (2006). Common features and interesting differences in transcriptional responses to secretion stress in the fungi *Trichoderma reesei* and *Saccharomyces cerevisiae*. *BMC Genomics* **7:**32.

Askenazi, M., Driggers, E. M., Holtzman, D. A., Norman, T. C., Iverson, S., Zimmer, D. P., Boers, M. E., Blomquist, P. R., Martinez, E. J. & other authors (2003). Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat Biotechnol* **21,** 150-156.

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

Barabási, A. & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-512.

Barker, M. & Rayens, W. (2003). Partial least squares for discrimination. *J Chemometr* **17,** 166-173.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390.

Bendtsen, J. D., Nielsen, H., von Heijne & G., Brunak, S. (2004a). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Bendtsen, J. D., Jensen, L. J., Blom, N., von Heijne, G. & Brunak, S. (2004b). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* **17**, 349-356.

Bennett, J. W. & Lasure, L. L. (1991). Growth media. In *More Gene Manipulations in Fungi*, pp. 441-447. Edited by J. W. Bennett & L. L. Lasure. San Diego, CA: Academic Press.

Bergmann, S., Ihmels, J. & Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* **2:**E9.

Berka, R. M., Carmona, C. L., Hayenga, K. J., Thompson, S. A. & Ward, M. (1993). Isolation and characterization of the *Aspergillus oryzae* gene encoding aspergillopepsin O. *Gene* **125**, 195-198.

Berka, R. M., Ward, M., Wilson, L. J., Hayenga, K. J., Kodama, K. H., Carlomagno, L. P. & Thompson, S. A. (1990). Molecular cloning and deletion of the gene encoding aspergillopepsin A from *Aspergillus awamori*. *Gene* **86**, 153-162.

Bignell, E., Negrete-Urtasun, S., Calcagno, A. M., Haynes, K., Arst, H. N., Jr. & Rogers, T. (2005). The *Aspergillus* pH-responsive transcription factor PacC regulates virulence. *Mol Microbiol* **55**, 1072-1084.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., van Ommen, B. & Smilde, A. K. (2006). Large scale human metabolomics studies. A strategy for data (pre-) processing and validation. *Anal Chem* **78**, 567-574.

Bos, C. J., Debets, A. J. M., Swart, K., Huybers, A., Kobus, G. & Slakhorst, S. M. (1988). Genetic analysis and the construction of master strains for assignment of genes to six linkage groups in *Aspergillus niger*. *Curr Genet* **14**, 437-443.

Braaksma, M. & Punt, P. J. (2008). *Aspergillus* as a cell factory for protein production: controlling protease activity in fungal production. In *The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods*, pp. 441-455. Edited by G. H. Goldman & S. A. Osmani. Boca Raton, FL: CRC Press.

Braaksma, M., Smilde, A. K., van der Werf, M. J. & Punt, P. J. (2009). The effect of environmental conditions on extracellular protease activity in controlled fermentations of *Aspergillus niger*. *Microbiol* **155**, 3430-3439.

Braaksma, M., van den Berg, R. A., van der Werf, M. J. & Punt, P. J. (2010a). A top-down systems biology approach for the identification of targets for fungal strain and process development. In *Cellular and Molecular Biology of Filamentous Fungi*, pp. 25-35. Edited by K. A. Borkovich and D. J. Ebbole. Washington, DC: ASM Press.

Braaksma, M., Martens-Uzunova, E., Punt, P. J. & Schaap, P.J. (2010b). An inventory of the *Aspergillus niger* secretome by combining in silico predictions with shotgun proteomics data. *BMC Genomics* **11**, doi:10.1186/1471-2164-11-584.

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

Bruggeman, F. J. & Westerhoff, H. V. (2007). The nature of systems biology. *Trends Microbiol.* **15,** 45-50.

Buxton, F. P., Gwynne, D. I. & Davies, R. W. (1985). Transformation of *Aspergillus niger* using the *argB* gene of *Aspergillus nidulans*. *Gene* **37**, 207-214.

Buxton, F., Gwynne, D. & Davies, R. (1989). Cloning of a new bidirectionally selectable marker for *Aspergillus* strains. *Gene* **84**, 329-334.

Carberry, S. & Doyle, S. (2007). Proteomic studies in biomedically and industrially relevant fungi. *Cytotechnology* **53**, 95-100.

Carberry, S., Neville, C. M., Kavanagh, K. A. & Doyle, S. (2006). Analysis of major intracellular proteins of *Aspergillus fumigatus* by MALDI mass spectrometry: Identification and characterisation of an elongation factor 1B protein with glutathione transferase activity. *Biochem Biophys Res Commun* **341,** 1096-1104.

Christensen, T. & Hynes, M. J. (2000). Fungus wherein the *areA* gene has been modified and an *areA* gene from *Aspergillus oryzae*. US patent 6025185.

Christensen, T. & Lehmbeck, J. (2000). Fungus wherein the *areA*, *pepC* and/or *pepE* genes have been inactivated. US patent 6013452.

Christensen, T., Hynes, M. J. & Davis, M. A. (1998). Role of the regulatory gene *areA* of *Aspergillus oryzae* in nitrogen metabolism. *Appl Environ Microbiol* **64**, 3232-3237.

Cohen, B. L. (1972). Ammonium repression of extracellular protease in *Aspergillus nidulans*. *J Gen Microbiol* **71**, 293-299.

Cohen, B. L. (1973). Regulation of intracellular and extracellular neutral and alkaline proteases in *Aspergillus nidulans*. *J Gen Microbiol* **79**, 311-320.

Cohen, B. L. (1981). Regulation of protease production in *Aspergillus*. *Trans Br Mycol Soc* **76**, 447-450.

Conesa, A., Punt, P. J., van Luijk, N. & van den Hondel, C. A. M. J. J. (2001). The secretion pathway in filamentous fungi: A biotechnological view. *Fungal Genet Biol* **33,** 155-171.

Connelly, M. and Brody, H. (2004). Methods for producing biological substances in enzyme-deficient mutants of *Aspergillus niger.* World patent WO/2004/090155.

Coulier, L., Bas, R., Jespersen, S., Verheij, E., van der Werf, M. J. & Hankemeier, T. (2006). Simultaneous quantitative analysis of metabolites using ion-pair liquid chromatography-electrospray ionization mass spectrometry. *Anal Chem* **78**, 6573-6582.

Coutinho P. M., Andersen M. R., Kolenova K., vanKuyk P. A., Benoit I., Gruben B. S., Trejo-Aguilar B., Visser H., van Solingen P. & others authors (2009). Post-genomic insights into the plant polysaccharide degradation potential of *Aspergillus nidulans* and comparison to *Aspergillus niger* and *Aspergillus oryzae*. *Fungal Genet Biol* **46**, S161-S169.

Daub, C. O. & Sonnhammer, E. L. (2008). Employing conservation of co-expression to improve functional inference. *BMC Syst Biol* **2:**81.

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:**R108.

de Vries, R. P., Burgers, K., van de Vondervoort, P. J., Frisvad, J. C., Samson, R. A., & Visser, J. (2004). A new black *Aspergillus* species, *A. vadensis*, is a promising host for homologous and heterologous protein production. *Appl Environ Microbiol* **70**, 3954-3959.

DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.

Dobrev, G. T., Pishtiyski, I. G., Stanchev, V. S. & Mircheva, R. (2007). Optimization of nutrient medium containing agricultural wastes for xylanase production by *Aspergillus niger* B03 using optimal composite experimental design. *Bioresour Technol* **98,** 2671-2678.

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

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

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

Edens, L., Dekker, P., van der Hoeven, R., Deen, F., de Roos, A. & Floris, R. (2005). Extracellular prolyl endoprotease *from Aspergillus niger* and its use in the debittering of protein hydrolysates. *J Agric Food Chem* **53**, 7950-7957.

Edgar, R., Domrachev, M. & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32**, 1792-1797.

Efron, B. & R. J. Tibshirani (1993). An introduction to the bootstrap. New York, NY: Chapman & Hall.

Eilers, P. H. C. (2003). A Perfect Smoother. *Anal Chem* **75**, 3631-3636.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95,** 14863-14868.

Elias, J. E. & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Meth* **4**, 207-214.

Featherstone, D. E. & Broadie, K. (2002). Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays* **24**, 267-274.

Foreman, P. K., Brown, D., Dankmeyer, L., Dean, R., Diener, S., Dunn-Coleman, N. S., Goedegebuur, F. Houfek, T. D., England, G. J. & other authors (2003). Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus *Trichoderma reesei*. *J Biol Chem* **278,** 31988-31997.

Fraissinet-Tachet, L., van den Hombergh, J. P. T. W., van de Vondervoort, P. J. I., Jarai, G. & Visser, J. (1996). Complex regulation of extracellular proteases in *Aspergillus niger*; an analysis of wide domain regulatory mutants demonstrates CREA, AREA and PACC control. In *An analysis of the proteolytic system in Aspergillus in order to improve protein production*, pp. 229-250. Edited by J. P. T. W. van den Hombergh. PhD thesis, Wageningen Agriculture University, Wageningen.

Frisvad, J. C., Rank, C., Nielsen, K. F. & Larsen, T. O. (2009). Metabolomics of *Aspergillus fumigatus*. *Med Mycol* **47**, S53-S71.

Fryksdale, B. G., Jedrzejewski, P. T., Wong, D. L., Gaertner, A. L. & Miller, B. S. (2002). Impact of deglycosylation methods on two-dimensional gel electrophoresis and matrix assisted laser desorption/ionization-time of flight-mass spectrometry for proteomic analysis. *Electrophoresis* **23,** 2184-2193.

Fu, X., Gharib, S. A., Green, P. S., Aitken, M. L., Frazer, D. A., Park, D. R., Vaisar, T. & Heinecke, J. W. (2008). Spectral index for assessment of differential protein expression in shotgun proteomics. *J Proteome Res* **7**, 845-854.

Fujinaga, M., Cherney, M. M., Oyama, H., Oda, K. & James, M. N. G. (2004). The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *PNAS* **101**, 3364-3369.

Fukushima, Y., Itoh, H., Fukase, T. & Motai, H. (1989). Continuous protease production in a carbon-limited chemostat culture by salt tolerant *Aspergillus oryzae*. *Appl Microbiol Biotechnol* **30**, 604-608.

Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L. J., Wortman, J. R., Batzoglou, S., Lee, S. I., Basturkmen, M., Spevak, C. C. & other authors (2005). Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105-1115.

Geer, L.Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W. & Bryant, S. H. (2004). Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964.

Geladi, P. & Kowalski, B. R. (1986). Partial least squares regression: A tutorial. *Anal Chim Acta* **185,** 1-17.

Gheshlaghi, R., Scharer, J. M., Moo-Young, M. & Douglas, P. L. (2007). Metabolic flux analysis for optimizing the specific growth rate of recombinant *Aspergillus niger*. *Bioprocess Biosyst Eng* **30,** 397-418.

Gordon, C. L., Archer, D. B., Jeenes, D. J., Doonan, J. H., Wells, B., Trinci, A. P. J. & Robson, G. D. (2000). A glucoamylase::GFP gene fusion to study protein secretion by individual hyphae of *Aspergillus niger*. *J Microbiol Methods* **42**, 39-48.

Grimm, L., Kelly, S., Krull, R. & Hempel, D. (2005). Morphology and productivity of filamentous fungi. *Appl Microbiol Biotechnol* **69**, 375-384.

Guillemette, T., van Peij, N. N. M. E., Goosen, T., Lanthaler, K., Robson, G. D., van den Hondel, C. A. M. J. J., Stam, H. & Archer, D. B. (2007). Genomic analysis of the secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. *BMC Genomics* **8:**158.

Gwynne, D. I. & Devchand, M. (1992). Expression of foreign proteins in the genus *Aspergillus*. In *Aspergillus: Biology and industrial applications*, pp. 203-214. Edited by J. W. Bennett & M. A. Klich. Stoneham, MA: Butterworth-Heinemann.

Hara, S., Kitamoto, K. & Gomi, K. (1992). New developments in fermented beverages and foods with *Aspergillus*. In *Aspergillus: Biology and industrial applications*, pp. 133-153. Edited by J. W. Bennett & M. A. Klich. Stoneham, MA: Butterworth-Heinemann.

Hartingsveldt, W. v., Mattern, I. E., Zeijl, C. M. J., Pouwels, P. H. & Hondel, C. A. M. J. (1987). Development of a homologous transformation system for *Aspergillus niger* based on the *pyrG* gene. *Mol Gen Genet* **206**, 71-75.

Heerikhuisen, M., van den Hondel, C. A. M. J. J. & Punt, P. J. (2005). *Aspergillus sojae*. In *roduction of Recombinant Proteins. Novel Microbial and Eucaryotic Expression Systems*, pp. 191-214. Edited by G. Gelissen. Weinheim, DE: Wiley-VCH.

Herrgård, M. J., Covert, M. W. & Palsson, B. Ø. (2003). Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res* **13**, 2423-2434.

Holm, K. A. (1980). Automated colorimetric determination of acid proteinase activity in fermentation samples using a trinitrobenzenesulphonic acid reagent. *Analyst* **105**, 18-24.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R. & other authors (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res* **36**, D577-581.

Hoogerbrugge, R., Willig, S. J. & Kistemaker, P. G. (1983). Discriminant analysis by double stage principal component analysis. *Anal Chem* **55,** 1710-1712.

Horvath, S. & Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol* **4**:e1000117.

Hughes, T. R., Robinson, M. D., Mitsakakis, N. & Johnston, M. (2004). The promise of functional genomics: completing the encyclopedia of a cell. *Curr Opin Microbiol* **7**, 546-554.

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*. *Eukaryot Cell* **5**, 794-805.

Iimura, Y., Gomi, K., Uzu, H. & Hara, S. (1987). Transformation of *Aspergillus oryzae* through plasmid-mediated complementation of the methionine-auxotrophic mutation. *Agric Biol Chem* **51**, 323-328.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B. & Speed, T. (2003). Summaries of Affymetrix Genechip probe level data. *Nucleic Acids Res* **31**:e15.

Jackson, J. E. (1991). A user's guide to principal components. New York, NY: John Wiley & Sons.

Jarai, G. & Buxton, F. (1994). Nitrogen, carbon, and pH regulation of extracellular acidic proteases of *Aspergillus niger*. *Curr Genet* **26**, 238-244.

Jaton-Ogay, K., Paris, S., Huerre, M., Quadroni, M., Falchetto, R., Togni, G., Latge, J. & Monod, M. (1994). Cloning and disruption of the gene encoding an extracellular metalloprotease of *Aspergillus fumigatus*. *Mol Microbiol* **14**, 917-928.

Jin, Y., Bok, J. W., Guzman-de-Peña, D. & Keller, N. P. (2002). Requirement of spermidine for developmental transitions in *Aspergillus nidulans*. *Mol Microbiol* **46**, 801-812.

Jolliffe, I. T. (2002). Principal Component Analysis. New York, NY: Springer-Verlag.

Jordan, I. K., Katz, L. S., Denver, D. R. & Streelman, J. T. (2008). Natural selection governs local, but not global, evolutionary gene coexpression networks in *Caenorhabditis elegans*. *BMC Syst Biol* **2**:96.

Jørgensen, T. R., Goosen, T., van den Hondel, C. A. M. J. J., Ram, A. F. J. & Iversen, J. J. L. (2009). Transcriptomic comparison of *Aspergillus niger* growing on two different sugars reveals coordinated regulation of the secretory pathway. *BMC Genomics* **10**:44.

Käll, L., Krogh, A. & Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036.

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30.

Katz, M. E., Bernardo, S. M. & Cheetham, B. F. (2008). The interaction of induction, repression and starvation in the regulation of extracellular proteases in *Aspergillus nidulans*: Evidence for a role for CreA in the response to carbon starvation. *Curr Genet* **54**, 47-55.

Katz, M. E., Flynn, P. K., vanKuyk, P. A. & Cheetham, B. F. (1996). Mutations affecting extracellular protease production in the filamentous fungus *Aspergillus nidulans*. *Mol Gen Genet* **250**, 715-724.

Katz, M. E., Gray, K. A. & Cheetham, B. F. (2006). The *Aspergillus nidulans xprG* (*phoG*) gene encodes a putative transcriptional activator involved in the response to nutrient limitation. *Fungal Genet Biol* **43**, 190-199.

Katz, M. E., Masoumi, A., Burrows, S. R., Shirtliff, C. G. & Cheetham, B. F. (2000). The *Aspergillus nidulans xprF* gene encodes a hexokinase-like protein involved in the regulation of extracellular proteases. *Genetics* **156**, 1559-1571.

Katz, M. E., Rice, R. N. & Cheetham, B. F. (1994). Isolation and characterization of an *Aspergillus nidulans* gene encoding an alkaline protease. *Gene* **150,** 287-292.

Kennedy, M. & Krouse, D. (1999). Strategies for improving fermentation medium performance: A review. *J Ind Microbiol Biotechnol* **23**, 456-475.

Kim, S., Kim, J. H., Whang, S. S. & Chae, K. S. (2001). Isolation and the nucleotide sequence of the *creA* gene for a carbon catabolite repressor of *Aspergillus oryzae*. *Food Sci Biotechnol* **10**, 90-93.

Kim, Y., Nandakumar, M. P. & Marten, M. R. (2007a). Proteome map of *Aspergillus nidulans* during osmoadaptation. *Fungal Genet Biol* **44,** 886-895.

Kim, Y., Nandakumar, M. P. & Marten, M. R. (2007b). Proteomics of filamentous fungi. *Trends Biotechnol* **25**, 395-400.

Kim, Y., Nandakumar, M. P. & Marten, M. R. (2008). The state of proteome profiling in the fungal genus *Aspergillus*. *Brief Funct Genomic Proteomic* **7**, 87-94.

Kimura, S., Maruyama, J. I., Takeuchi, M. & Kitamoto, K. (2008). Monitoring global gene expression of proteases and improvement of human lysozyme production in the nptB gene disruptant of *Aspergillus oryzae*. *Biosci Biotechnol Biochem* **72,** 499-505.

Koek, M. M., Muilwijk, B., van der Werf, M. J. & Hankemeier, T. (2006). Microbial metabolomics with gas chromatography/mass spectrometry. *Anal Chem* **78**, 1272-1281.

Kolattukudy, P. E., Lee, J. D., Rogers, L. M., Zimmerman, P., Ceselski, S., Fox, B., Stein, B. & Copelan, E. A. (1993). Evidence for possible involvement of an elastolytic serine protease in aspergillosis. *Infect Immun* **61**, 2357-2368.

Kouskoumvekaki, I., Yang, Z., Jónsdóttir, S. O., Olsson, L. & Panagiotou, G. (2008). Identification of biomarkers for genotyping *Aspergilli* using non-linear methods for clustering and classification. *BMC Bioinformatics* **9:**59.

Kouskoumvekaki, I., Yang, Z., Jónsdóttir, S. Ó., Olsson, L. & Panagiotou, G. (2008). Identification of biomarkers for genotyping *Aspergilli* using non-linear methods for clustering and classification. *BMC Bioinformatics* **9:**59.

Kudla, B., Caddick, M. X., Langdon, T., Martinez-Rossi, N. M., Bennett, C. F., Sibley, S., Davies, R. W. & Arst, H. N., Jr. (1990). The regulatory gene *areA* mediating nitrogen metabolite repression in *Aspergillus nidulans*. Mutations affecting specificity of gene activation alter a loop residue of a putative zinc finger. *EMBO J* **9**, 1355-1364.

Kudo, Y., Ootani, T., Kumagai, T., Fukuchi, Y., Ebina, K. & Yokota, K. (2002). A novel oxidized low-density lipoprotein-binding protein, Asp-hemolysin, recognizes lysophosphatidylcholine. *Biol Pharm Bull* **25**, 787-790.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. & Davis, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A* **94**, 13057-13062.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res* **14**, 1085-1094.

Lee, J. D. & Kolattukudy, P. E. (1995). Molecular cloning of the cDNA and gene for an elastinolytic aspartic proteinase from *Aspergillus fumigatus* and evidence of its secretion by the fungus during invasion of the host lung. *Infect Immun* **63**, 3796-3803.

Lehmbeck, J. (1999). Host cell expressing reduced levels of a metalloprotease and methods using the host cell in protein production. US patent 5968774.

Lehmbeck, J. (2001). Alkaline protease deficient filamentaous fungi. US patent 6291209.

Lenouvel, F., Fraissinet-Tachet, L., van de Vondervoort, P. J. I. & Visser, J. (2001). Isolation of UV-induced mutations in the *areA* nitrogen regulatory gene of *Aspergillus niger*, and construction of a disruption mutant. *Mol Genet Genomics* **266**, 42-47.

Leung, Y. F. & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends Genet.* **19,** 649-659.

Levin, A. M., de Vries, R. P., Conesa, A., de Bekker, C., Talon, M., Menke, H. H., van Peij, N. N. M. E. & Wösten, H. A. B. (2007). Spatial differentiation in the vegetative mycelium of *Aspergillus niger*. *Eukaryot Cell* **6**, 2311-2322.

Li, Q., Harvey, L. M. & McNeil, B. (2008). The effects of bioprocess parameters on extracellular proteases in a recombinant *Aspergillus niger* B1-D. *Appl Microbiol Biotechnol* **78**, 333-341.

Li, Y., Liu, Z., Cui, F., Xu, Y., Zhao, H. & Liu, Z. (2007). Application of statistical experimental design to optimize culture requirements of *Aspergillus* sp. Zh-26 producing xylanase for degradation of arabinoxylans in mashing. *J Food Sci* **72,** e320-e329.

Ligon, B. L. (2004). Penicillin: Its discovery and early development. *Semin Pediatr Infect Dis* **15,** 52-57.

Lin, Y., Means, G. E. & Feeney, R. E. (1969). The action of proteolytic enzymes on N,N-dimethyl proteins. Basis for a microassay for proteolytic enzymes. *J Biol Chem* **244**, 789-793.

Liu, F., Li, W., Ridgway, D., Gu, T. & Moo-Young, M. (1998). Inhibition of extracellular protease secretion by *Aspergillus niger* using cell immobilization. *Biotechnol Lett* **20**, 539-542.

Liu, H., Sadygov, R. G. & Yates, J. R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201.

Lu, X., Sun, J., Nimtz, M., Wissing, J., Zeng, A.-P. & Rinas, U. (2010). The intra- and extracellular proteome of *Aspergillus niger* growing on defined medium with xylose or maltose as carbon substrate. Microb Cell Fact **9**, doi:10.1186/1475-2859-9-23.

Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, Å., Pettersen, J. & Bergman, R. (1998). Experimental design and optimization. *Chemometr Intell Lab Syst* **42**, 3-40.

MacCabe, A. P., van den Hombergh, J. P. T. W., Tilburn, J., Arst, H. N., Jr. & Visser, J. (1996). Identification, cloning and analysis of the *Aspergillus niger* gene *pacC*, a wide domain regulatory gene responsive to ambient pH. *Mol Gen Genet* **250**, 367-374.

MacCabe, A. P., Vanhanen, S., Sollewign, G., van, d., V, Arst, H. N., Jr. & Visser, J. (1998). Identification, cloning and sequence of the Aspergillus niger areA wide domain regulatory gene controlling nitrogen utilisation. *Biochim Biophys Acta* **1396**, 163-168.

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K. I., Arima, T., Akita, O. & 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. In *Advances in fungal biotechnology for industry, agriculture, and medicine*, pp. 307-339. Edited by J. S. Tkacz & L. Lange. New York, NY: Kluwer Academic/Plenum Publishers.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). Multivariate analysis. London; New York, NY: Academic Press.

Martens-Uzunova, E. S. & Schaap, P. J. (2008). An evolutionary conserved d-galacturonic acid metabolic pathway operates across filamentous fungi capable of pectin degradation. *Fungal Genet Biol* **45**, 1449-1457.

Martens-Uzunova, E. S. & Schaap, P. J. (2009). Assessment of the pectin degrading enzyme network of *Aspergillus niger* by functional genomics. *Fungal Genet Biol* **46**, S170-S179.

Mattern, I. E., van Noort, J. M., van den Berg, P., Archer, D. A., Roberts, I. N. & van den Hondel, C. A. M. J. J. (1992). Isolation and characterization of mutants of *Aspergillus niger* deficient in extracellular proteases. *Mol Gen Genet* **2**, 332-336.

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

Meyer, V., Damveld, R. A., Arentshorst, M., Stahl, U., van den Hondel, C. A. M. J. J., & Ram, A. F. J. (2007). Survival in the presence of antifungals: genome-wide expression profiling of *Aspergillus niger* in response to sublethal concentrations of caspofungin and fenpropimorph. J *Biol Chem* **282**, 32935-32948.

Monod, M., Paris, S., Sarfati, J., Jaton-Ogay, K., Ave, P. & Latge, J. (1993). Virulence of alkaline protease-deficient mutants of *Aspergillus fumigatus*. *FEMS Microbiol Lett* **106**, 39-46.

Moralejo, F. J., Cardoza, R. E., Gutiérrez, S., Lombraňa, M., Fierro, F. & Martín, J. F. (2002). Silencing of the aspergillopepsin B (*pepB*) gene of *Aspergillus awamori* by antisense RNA expression or protease removal by gene disruption results in a large increase in thaumatin production. *Appl Environ Microbiol* **68**, 3550-3559.

Moralejo, F. J., Cardoza, R. E., Gutiérrez, S., Sisniega, H., Faus, I. & Martín, J. F. (2000). Overexpression and lack of degradation of thaumatin in an aspergillopepsin A-defective mutant of *Aspergillus awamori* containing an insertion in the *pepA* gene. *Appl Microbiol Biotechnol* **54**, 772-777.

Mouyna, I., Fontaine, T., Vai, M., Monod, M., Fonzi, W. A., Diaquin, M., Popolo, L., Hartland, R. P. & Latgé, J. P. (2000). Glycosylphosphatidylinositol-anchored glucanosyltransferases play an active role in the biosynthesis of the fungal cell wall. *J Biol Chem* **275**, 14882-14889.

Nasution, U., van Gulik, W. M., Kleijn, R. J., van Winden, W. A., Proell, A. & Heijnen, J. J. (2006). Measurement of intracellular metabolites of primary metabolism and adenine nucleotides in chemostat cultivated *Penicillium chrysogenum*. *Biotechnol Bioeng* **94,** 159-166.

Nasution, U., van Gulik, W. M., Ras, C., Proell, A. & Heijnen, J. J. (2008). A metabolome study of the steady-state relation between central metabolism, amino acid biosynthesis and penicillin production in *Penicillium chrysogenum*. *Metab Eng* **10,** 10-23.

Natorff, R., Balinska, M. & Paszewski, A. (1993). At least four regulatory genes control sulphur metabolite repression in Aspergillus nidulans. *Mol Gen Genet* **238**, 185-192.

Natorff, R., Sienko, M., Brzywczy, J. & Paszewski, A. (2003). The *Aspergillus nidulans metR* gene encodes a bZIP protein which activates transcription of sulphur metabolism genes. *Mol Microbiol* **49**, 1081-1094.

Neretti, N., Remondini, D., Tatar, M., Sedivy, J. M., Pierini, M., Mazzatti, D., Powell, J., Franceschi, C., & Castellani, G. C. (2007). Correlation analysis reveals the emergence of coherence in the gene expression dynamics following system perturbation. *BMC Bioinformatics* **8**:S16.

Nevalainen, K. M. H. & Te'-o, V. S. J. (2003). Enzyme production in industrial fungi - molecular genetic strategies for integrated strain improvement. In *Applied Mycology and Biotechnology* **3**, pp. 241-259. Edited by D. K. Arora & G. G. Khachatourians. Amsterdam: Elsevier Science.

Nierman, W. C., Pain, A., Anderson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., Berriman, M., Abe, K., Archer, D. B. & other authors (2005). Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438**, 1151-1156.

Nikolov, Z. L., Meagher, M. M. & Reilly, P. J. (1989). Kinetics, equilibria, and modeling of the formation of oligosaccharides from D-glucose with *Aspergillus niger* glucoamylases I and II. *Biotechnol Bioeng* **34**, 694-704.

O'Donnell D., Wang L., Xu J., Ridgway D., Gu T. & Moo-Young M. (2001). Enhanced heterologous protein production in *Aspergillus niger* through pH control of extracellular protease activity. *Biochem Eng J* **8**, 187-193.

Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A. & Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **4**, 1487-1502.

Oldiges, M., Lütz, S., Pflug, S., Schroer, K., Stein, N. & Wiendahl, C. (2007). Metabolomics: Current state and evolving methodologies and tools. *Appl Microbiol Biotechnol* **76,** 495-511.

Oldiges, M., Lütz, S., Pflug, S., Schroer, K., Stein, N. & Wiendahl, C. (2007). Metabolomics: Current state and evolving methodologies and tools. *Appl Microbiol Biotechnol* **76**, 495-511.

Oliver, S. (2000). Guilt-by-association goes global. *Nature* **403,** 601-603.

Papagianni, M. & Moo-Young, M. (2002). Protease secretion in glucoamylase producer *Aspergillus niger* cultures: fungal morphology and inoculum effects. *Process Biochem* **37**, 1271-1278.

Papagianni, M., Joshi, N. & Moo-Young, M. (2002). Comparative studies on extracellular protease secretion and glucoamylase production by free and immobilized *Aspergillus niger* cultures. *J Ind Microbiol Biotechnol* **29**, 259-263.

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., Dyer, P. S., Hofmann, G., Schaap, P. J., Turner, G., de Vries, R. P., Albang, R. & other authors (2007). Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat Biotechnol* **25**, 221-231.

Peñalva, M. A. & Arst, H. N., Jr. (2002). Regulation of gene expression by ambient pH in filamentous fungi and yeasts. *Microbiol Mol Biol Rev* **66**, 426-446.

Pierleoni, A., Martelli, P. L. & Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**:392.

Pieterse, B., Jellema, R. H. & van der Werf, M. J. (2006). Quenching of microbial samples for increased reliability of microarray data. *J Microbiol Methods* **64**, 207-216.

Plackett, R. L. & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika* **33,** 305-325.

Pontecorvo, G., Roper, J. A., Hemmons, L. M., Macdonald, K. D., & Bufton, A. W. (1953). The genetics of *Aspergillus nidulans*. *Adv Genet **5***, 141-238.

Punt, P. J., Schuren, F. H. J., Lehmbeck, J., Christensen, T., Hjort, C. & van den Hondel, C. A. M. J. J. (2008). Characterization of the *Aspergillus niger prtT*, a unique regulator of extracellular protease encoding genes. *Fungal Genet Biol* **45**, 1591-1599.

Punt, P. J., van Biezen, N. A., Conesa, A., Albers, A., Mangnus, J. & van den Hondel, C. A. M. J. J. (2002). Filamentous fungi as cell factories for heterologous protein production. *Trends Biotechnol* **20**, 200-206.

Punt, P. J., van den Hondel, C. A. M. J. J. (1992). Analysis of transcription control sequences in filamentous fungi. In *EMBO Workshop on Molecular Biology of Filamentous Fungi*, pp. 177-187. Edited by U. Stahl & P. Tudzynscki. Weinheim: VCH.

Ramesh, M. V. & Kolattukudy, P. E. (1996). Disruption of the serine proteinase gene (*sep*) in *Aspergillus flavus* leads to a compensatory increase in the expression of a metalloproteinase gene (*mep20*). *J Bacteriol* **178**, 3899-3907.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G., Myers, C., Parsons, A., Friesen, H., Oughtred, R. & other authors (2006). Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol* **5**, 11.

Reichard, U., Monod, M., Odds, F. & Ruchel, R. (1997). Virulence of an aspergillopepsin-deficient mutant of *Aspergillus fumigatus* and evidence for another aspartic proteinase linked to the fungal cell wall. *J Med Vet Mycol* **35**, 189-196.

Rubingh, C. M., Bijlsma, S., Derks, E. P. P. A., Bobeldijk, I., Verheij, E. R., Kochhar, S. & Smilde, A. K. (2009). Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2,** 53-61.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G. & other authors (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucl Acids Res* **32**, 5539-5545.

Ruijter, G. J. G. & Visser, J. (1996). Determination of intermediary metabolites in *Aspergillus niger*. *J Microbiol Methods* **3**, 295-302.

Ruijter, G. J. G. & Visser, J. (1997). Carbon repression in *Aspergilli*. *FEMS Microbiol Lett* **151**, 103-114.

Salamov, A. A. & Solovyev, V. V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* **10**, 516-522.

Schäfer, T., Borchert, T. W., Nielsen, V. S., Skagerlind, P., Gibson, K., Wenger, K., Hatzack, F., Nilsson, L. D., Salmon, S. & other authors (2007). Industrial enzymes. *Adv Biochem Eng Biotechnol* **105**, 59-131.

Schrickx, J. M., Krave, A. S., Verdoes, J. C., van den Hondel, C. A. M. J. J., Stouthamer, A. H. & van Verseveld, 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. *J Gen Microbiol* **139**, 2801-2810.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M. & Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* **7**:3.

Scott, M., Lu, G., Hallett, M. & Thomas, D. Y. (2004). The Hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics* **20**, 937-944.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504.

Shintani, T. & Ichishima, E. (1994). Primary structure of aspergillopepsin I deduced from nucleotide sequence of the gene and aspartic acid-76 is an essential active site of the enzyme for trypsinogen activation. *Biochim Biophys Acta* **1204**, 257-264.

Shroff, R. A., O'Connor, S. M., Hynes, M. J., Lockington, R. A. & Kelly, J. M. (1997). Null alleles of *creA*, the regulator of carbon catabolite repression in *Aspergillus nidulans*. *Fungal Genet Biol* **22**, 28-38.

Singh, A., Ghosh, V. & Ghosh, P. (1994). Production of thermostable acid protease by *Aspergillus niger*. *Lett Appl Microbiol* **18**, 177-180.

Singh, D. P. & Vyas, S. R. (1977). Standardization of cultural conditions for maximum acid protease production by fungi. *Haryana Agric Univ J Res* **7**, 37-42.

Sokal, R. R. & Rohlf, F. J. (1995). Biometry: the principles and practice of statistics in biological research. New York, NY: W. H. Freeman.

Srinivasan, M. & Dhar, S. C. (1990). Factors influencing extracellular protease synthesis in an *Aspergillus flavus* isolate. *Acta Microbiol Hung* **37**, 15-23.

Sumner, J. B. & Somers, G. F. (1949). Dinitrosalicylic method for glucose. In *Laboratory experiments in biological chemistry*, pp. 38-39. Edited by J.B. Sumner & G. F. Somers. New York, NY: Academic Press.

Swift, R. J., Karandikar, A., Griffen, A. M., Punt, P. J., van den Hondel, C. A. M. J. J., Robson, G. D., Trinci, A. P. J. & Wiebe, M. G. (2000). The effect of organic nitrogen sources on recombinant glucoamylase production by *Aspergillus niger* in chemostat culture. *Fungal Genet Biol* **31**, 125-133.

Swift, R. J., Wiebe, M. G., Robson, G. D. & Trinci, A. P. J. (1998). Recombinant glucoamylase production by *Aspergillus niger* B1 in chemostat and pH auxostat cultures. *Fungal Genet Biol* **25,** 100-109.

Tabor, C. W. & Tabor, H. (1985). Polyamines in microorganisms. *Microbiol Rev* **49**, 81-99.

Tanay, A., Regev, A. & Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* **102**, 7203-7208.

Tang, C. M., Cohen, J. & Holden, D. W. (1992). An *Aspergillus fumigatus* alkaline protease mutant constructed by gene disruption is deficient in extracellular elastase activity. *Mol Microbiol* **6**, 1663-1671.

Tilburn, J., Sarkar, S., Widdick, D. A., Espeso, E. A., Orejas, M., Mungroo, J., Peñalva, M. A. & Arst, H. N., Jr. (1995). The *Aspergillus PacC* zinc finger transcription factor mediates regulation of both acid- and alkaline-expressed genes by ambient pH. *EMBO J* **14**, 779-790.

Tobert, J. A. (2003). Lovastatin and beyond: The history of the HMG-CoA reductase inhibitors. *Nat Rev Drug Discov* **2,** 517-526.

Tomonaga, G., Õhama, H. & Yanagita, T. (1964). Effect of sulfur compounds on the protease formation by *Aspergillus niger*, *J Gen Appl Microbiol* **10**, 373-386.

Trygg, J., Gullberg, J., Johansson, A., Jonsson, P. & Moritz, T. (2006). Chemometrics in metabolomics - An introduction. In *Plant Metabolomics*, p. 117-128. Edited by K. Saito, R. A. Dixon & L. Willmitzer. Berlin Heidelberg: Springer-Verlag.

Trygg, J., Holmes, E. & Lundstedt, T. (2007). Chemometrics in metabonomics. *J Proteome Res* **6,** 469-479.

Tsang, A., Butler, G., Powlowski, J., Panisko, E .A. & Baker, S.E. (2009). Analytical and computational approaches to define the *Aspergillus niger* secretome. *Fungal Genet Biol* **46**, S153-S160.

Unkles, S. E., Campbell, E. I., de Ruiter-Jacobs, Y. M. J. T., Broekhuijsen, M., Macro, J. A., Carrez, D., Contreras, R., van den Hondel, C. A. M. J. J. & Kinghorn, J. R. (1989). The development of a homologous transformation system for *Aspergillus oryzae* based on the nitrate assimilation pathway: a convenient and general selection system for filamentous fungal transformation. *Mol Gen Genet* **218**, 99-104.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142.

van den Hazel, H. B., Kielland-Brandt, M. C. & Winther, J. R. (1996). Review: biosynthesis and function of yeast vacuolar proteases. *Yeast* **12**, 1-16.

van den Hombergh, J. P. T. W. (1996). An analysis of the proteolytic system in *Aspergillus* in order to improve protein production. PhD thesis, Wageningen Agriculture University, Wageningen.

van den Hombergh, J. P. T. W. & Visser, J. (1997). Fungal metallo protease genes. World patent WO9746689.

van den Hombergh, J. P. T. W. , MacCabe, A. P., van de Vondervoort, P. J. I. & Visser, J. (1996). Regulation of acid phosphatases in an *Aspergillus niger pacC* disruption strain. *Mol Gen Genet* **251**, 542-550.

van den Hombergh, J. P. T. W., Jarai, G., Buxton, F. P. & Visser, J. (1994). Cloning, characterization and expression of *pepF*, a gene encoding a serine carboxypeptidase from *Aspergillus niger*. *Gene* **151**, 73-79.

van den Hombergh, J. P. T. W., Sollewijn-Gelpke, M. D., van de Vondervoort, P. J. I., Buxton, F. P. & Visser, J. (1997a). Disruption of three acid proteases in *Aspergillus niger*: effects on protease spectrum, intracellular proteolysis, and degradation of target proteins. *Eur J Biochem* **247**, 605-613.

van den Hombergh, J. P. T. W., van de Vondervoort, P. J. I., van der Heijden, N. C. B. A. & Visser, J. (1995). New protease mutants in *Aspergillus niger* result in strongly reduced in vitro degradation of target proteins; genetical and biochemical characterization of seven complementation groups. *Curr Genet* **28**, 299-308.

van den Hombergh, J. P. T. W., van de Vondervoort, P. J., Fraissinet-Tachet, L. & Visser, J. (1997b). *Aspergillus* as a host for heterologous production: the problem of proteases. *Trends Biotechnol* **7**, 256-263.

van der Greef, J., Vogels, J. T. W. E., Wulfert, F. & Tas, A. C. (2004). Method and system for identifying and quantifying chemical components of a mixture. US Patent 267459.

van der Veen, D. (2009). Transcriptional profiling of *Aspergillus niger*. PhD thesis, Wageningen University, Wageningen.

van der Veen, D., Oliveira, J. M., van den Berg, W. A. M. & de Graaff, L. H. (2009). Variance components analysis reveals contribution of sample processing to transcript variation. *Appl Environ Microbiol* **75**, 2414-2422.

van der Werf, M. J. (2005). Towards replacing closed with open target selection strategies. *Trends Biotechnol* **23**, *11-16.*

van der Werf, M. J., Jellema, R. H. & Hankemeier, T. (2005). Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *J Ind Microbiol Biotechnol* **32**, 234-252.

van der Werf, M. J., Overkamp, K. M., Muilwijk, B., Coulier, L. & Hankemeier, T. (2007). Microbial metabolomics: Toward a platform with full metabolome coverage. *Anal Biochem* **370**, 17-25.

van Esse, H. P., van't Klooster, J. W., Bolton, M. D., Yadeta, K. A., van Baarlen, P., Boeren, S., Vervoort, J., de Wit, P. J. G. M. & Thomma, B. P. H. J. (2008). The *Cladosporium fulvum* virulence protein Avr2 inhibits host proteases required for basal defense. *Plant Cell* **20**, 1948-63.

van Hartingsveldt, W., Mattern, I. E., van Zeijl, C. M. J., Pouwels, P. H. & van den Hondel, C. A. M. J. J. (1987). Development of a homologous transformation system for *Aspergillus niger* based on the *pyrG* gene. *Mol Gen Genet* **206**, 71-75.

van Noort, J. M., van den Berg, P. & Mattern, I. E. (1991). Visualization of proteases within a complex sample following their selective retention on immobilized bacitracin, a peptide antibiotic. *Anal Biochem* **198**, 385-390.

van Noort, V., Snel, B. & Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep* **5**, 280-284.

VanKuyk, P. A., Cheetham, B. F. & Katz, M. E. (2000). Analysis of two *Aspergillus nidulans* genes encoding extracellular proteases. *Fungal Genet Biol* **29**, 201-210.

Vizcaino, J. A., Cote, R., Reisinger, F., Barsnes, H., Foster, J. M., Rameseder, J., Hermjakob, H. & Martens, L. (2010). The Proteomics Identifications database: 2010 update. *Nucl Acids Res* **38**, D736-742.

Vogels, J. T. W. E., Tas, A. C., Venekamp, J. & van der Greef, J. (1996). Partial linear fit: a new NMR spectroscopy preprocessing tool for pattern recognition applications. *J Chemom* **10**, 425-438.

Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* **9**, 1198-1203.

Wang, L. P., Ridgway, D., Gu, T. Y. & Moo-Young, M. (2003). Effects of process parameters on heterologous protein production in *Aspergillus niger* fermentation. *J Chem Technol Biotechnol* **78**, 1259-1266.

Wang, Y., Xue, W., Sims, A. H., Zhao, C., Wang, A., Tang, G., Qin, J. & Wang, H. (2008). Isolation of four pepsin-like protease genes from *Aspergillus niger* and analysis of the effect of disruptions on heterologous laccase expression. *Fungal Genet Biol* **45**, 17-27.

Wanke, C., Eckert, S., Albrecht, G., van Hartingsveldt, W., Punt, P. J., van den Hondel, C. A. M. J. J., & Braus, G. H. (1997). The *Aspergillus niger* GCN4 homologue, *cpcA*, is transcriptionally regulated and encodes an unusual leucine zipper. *Mol Microbiol* **23**, 23-33.

Ward, O. P., Qin, W. M., Dhanjoon, J., Ye, J. & Singh, A. (2005). Physiology and biotechnology of *Aspergillus*. In *Advances in Applied Microbiology* **58**, pp. 1-75. Edited by A. I. Laskin, J. W. Bennett, G. M. Gadd & S. Sariaslani. Amsterdam: Acadamic Press.

Weatherburn, M. W. (1967). Phenol-hypochlorite reaction for determination of ammonia. *Anal Chem* **39**, 971-974.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., Velzen, E. J. J., Duijnhoven, J. P. M. & Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics* **4,** 81-89.

Weuster-Botz, D. (2000). Experimental design for fermentation media development: Statistical design or global random search? *J Biosci Bioeng* **90,** 473-483.

Wheelock, C. E., Wheelock, A. M., Kawashima, S., Diez, D., Kanehisa, M., van Erk, M., Kleemann, R., Haeggström, J. Z. & Goto, S. (2009). Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol Biosyst **5***, 588-602.

White, S., McIntyre, M., Berry, D. R. & McNeil, B. (2002). The autolysis of industrial filamentous fungi. *Crit Rev Biotechnol* **22**, 1-14.

Wiebe, M. G. (2003). Stable production of recombinant proteins in filamentous fungi - problems and improvements. *Mycologist* **17**, 140-144.

Wiebe, M. G., Karandikar, A., Robson, G. D., Trinci, A. P. J., Candia, J.-L. F., Trappe, S., Wallis, G., Rinas, U., Derkx, P. M. F. & other authors (2001). Production of tissue plasminogen activator (t-PA) in *Aspergillus niger*. *Biotechnol Bioeng* **76**, 164-174.

Withers, J. M., Swift, R. J., Wiebe, M. G., Robson, G. D., Punt, P. J., van den Hondel, C. A. M. J. J. & Trinci, A. P. J. (1998). Optimization and stability of glucoamylase production by recombinant strains of *Aspergillus niger* in chemostat culture. *Biotechnol Bioeng* **59**, 407-418.

Withers, J. M., Wiebe, M. G., Robson, G. D. , Osborne, D., Turner, G. & Trinci, A. P. J. (1995). Stability of recombinant protein production by *Penicillium chrysogenum* in prolonged chemostat culture. *FEMS Microbiol Lett* **133,** 245-251.

Wolfe, C. J., Kohane, I. S. & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* **6**:227.

Wright, J. C., Sugden, D., Francis-McIntyre, S., Riba-Garcia, I., Gaskell, S. J., Grigoriev, I. V., Baker, S. E., Beynon, R. J. & Hubbard, S. J. (2009). Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. *BMC Genomics* **10:**61.

Xu, J. F., Wang, L. P., Ridgway, D., Gu, T. Y. & Moo-Young, M. (2000). Increased heterologous protein production in *Aspergillus niger* fermentation through extracellular proteases inhibition by pelleted growth. *Biotechnol Progr* **16**, 222-227.

Yoder, W. T. & Lembeck, J. (2004). Heterologous expression and protein secretion in filamentous fungi. In *Advances in fungal biotechnology for industry, agriculture, and medicine*, pp. 201-219. Edited by J. S. Tkacz & L. Lange. New York, NY: Kluwer Academic/Plenum Publishers.

Yuan, X. L., Goosen, C., Kools, H., van der Maarel, M. J. E. C., van den Hondel, C. A. M. J. J., Dijkhuizen, L. & Ram, A. F. J. (2006). Database mining and transcriptional analysis of genes encoding inulin-modifying enzymes of *Aspergillus niger*. *Microbiology* **152,** 3061-3073.

Yuan, X., Roubos, J. A., van den Hondel, C. A. M. J. J. & Ram, A. F. J. (2008a). Identification of InuR, a new Zn(II)2Cys6 transcriptional activator involved in the regulation of inulinolytic genes in *Aspergillus niger*. *Mol Genet Genomics* **279**, 11-26.

Yuan, X., van der Kaaij, R. M., van den Hondel, C. A. M. J. J., Punt, P. J., van der Maarel, M. J. E. C., Dijkhuizen, L. & Ram, A. F. J. (2008b). *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.

Zar, J. H. (1996). Biostatistical analysis. Upper Saddle River, NJ: Prentice-Hall.

Zhang, Z. & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19**, 307-308.

Zheng, X. F., Kobayashi, Y. & Takeuchi, M. (1998). Construction of a low-serine-type-carboxypeptidase-producing mutant of *Aspergillus oryzae* by the expression of antisense RNA and its use as a host for heterologous protein secretion. *Appl Microbiol Biotechnol* **49**, 39-44.

Zybailov, B., Colemanm M. K., Florens, L. & Washburn, M.P. (2005). Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* **77**, 6218-6224.

LIST OF PUBLICATIONS

R. te Biesebeke, A. Boussier, N.A. van Biezen, M. Braaksma, C.A.M.J.J. van den Hondel, W.M. de Vos & P.J. Punt (2006). Expression of *Aspergillus* hemoglobin domain activities in *Aspergillus oryzae* grown on solid substrates improves growth rate and enzyme production. Biotechnology Journal 1, 822-827.

M. Braaksma & P.J. Punt (2008). *Aspergillus* as a cell factory for protein production: controlling protease activity in fungal production. In *The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research Methods*, pp. 441-455. Edited by G.H. Goldman & S.A. Osmani. Boca Raton, FL: CRC Press.

M. Braaksma, A.K. Smilde, M.J. van der Werf & P.J. Punt (2009). The effect of environmental conditions on extracellular protease activity in controlled fermentations of *Aspergillus niger*. Microbiology 155, 3430-3439.

M. Braaksma, R.A. van den Berg, M.J. van der Werf & P.J. Punt (2010). A top-down systems biology approach for the identification of targets for fungal strain process development. In *Cellular and Molecular Biology of Filamentous Fungi*, pp. 25-35. Edited by K.A Borkovich & D.J. Ebbole. Washington, DC: ASM Press.

R.A. van den Berg, M. Braaksma, D. van der Veen, M.J. van der Werf, P.J. Punt, J. van der Oost & L.H. de Graaff (2010). Identification of modules in *Aspergillus niger* by gene co-expression network analysis. Fungal Genetics and Biology 47, 539-550.

M. Braaksma, E. Martens-Uzunova, P.J. Punt & P.J. Schaap (2010). An inventory of the *Aspergillus niger* secretome by combining *in silico* predictions with shotgun proteomics data. BMC Genomics 11, doi:10.1186/1471-2164-11-584.

M. Braaksma, S. Bijlsma, L. Coulier, P.J. Punt & M.J. van der Werf (2010). Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition. Microbiology, doi:10.1099/mic.0.041244-0.

CURRICULUM VITAE

Machtelt Braaksma werd geboren op 30 oktober 1977 te Stadskanaal en groeide op in Lollum. In 1996 behaalde zij haar VWO diploma aan het Marne College te Bolsward. In datzelfde jaar begon zij met de opleiding Biotechnologie aan het Van Hall Instituut / de Noordelijke Hogeschool Leeuwarden, met als specialisatie voeding en bioprocestechnologie. Tijdens haar studie heeft ze stage gelopen bij de Slowaakse bierbrouwerij Zlatý Bažant, een dochter van Heineken, en bij de afdeling Koolhydraattechnologie van TNO Voeding te Groningen.

Na het behalen van haar hbo-diploma in augustus 2000 is Machtelt als analist bij TNO Voeding (het huidige TNO Kwaliteit van Leven) in Zeist gaan werken. Vanaf juli 2004 heeft ze bij TNO en binnen programma 2 van het Kluyver Centre for Genomics of Industrial Fermentation het promotieonderzoek uitgevoerd wat heeft geleid tot dit proefschrift. Met ingang van juli 2008 heeft Machtelt haar werkzaamheden bij TNO Kwaliteit van Leven voortgezet, eerst als assistent projectleider, nu als scientist / projectleider.