



Universiteit
Leiden
The Netherlands

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

Braaksma, M.

Citation

Braaksma, M. (2010, December 15). *A functional genomics study of extracellular protease production by Aspergillus niger*. Retrieved from <https://hdl.handle.net/1887/16246>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/16246>

Note: To cite this publication please use the final published version (if applicable).

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 proves 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 asparagine 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 database¹ (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, 1.0×10^6 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 NH₄HCO₃ (pH 8.0) for 1 h at 60 °C. Next, slices were alkylated with 100 mM iodoacetamide in NH₄HCO₃ (pH 8.0) for 1 h at room temperature, washed with NH₄HCO₃ (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	<i>A. niger</i> CBS 513.88	<i>A. niger</i> ATCC 1015	<i>A. oryzae</i> RIB40	<i>A. fumigatus</i> Af293	<i>A. nidulans</i> 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 Institute² 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.

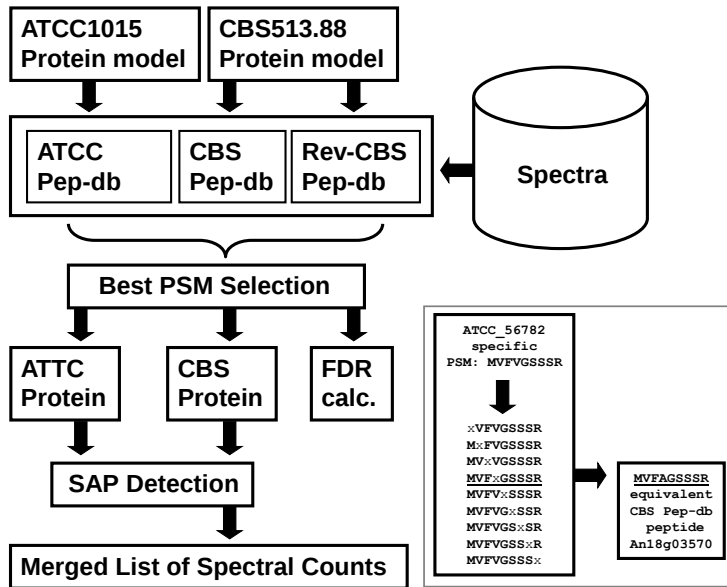


Fig. 1. Schematic of the Spectrum to Peptide matching pipeline. Forward and reversed (REV-CBS) databases were searched with local implementation of the OMSSA MS/MS search engine. Threshold Expect values for matching peptides were estimated from the false discovery rate (FDR). Best Accepted peptide-spectrum 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. From 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 in 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.

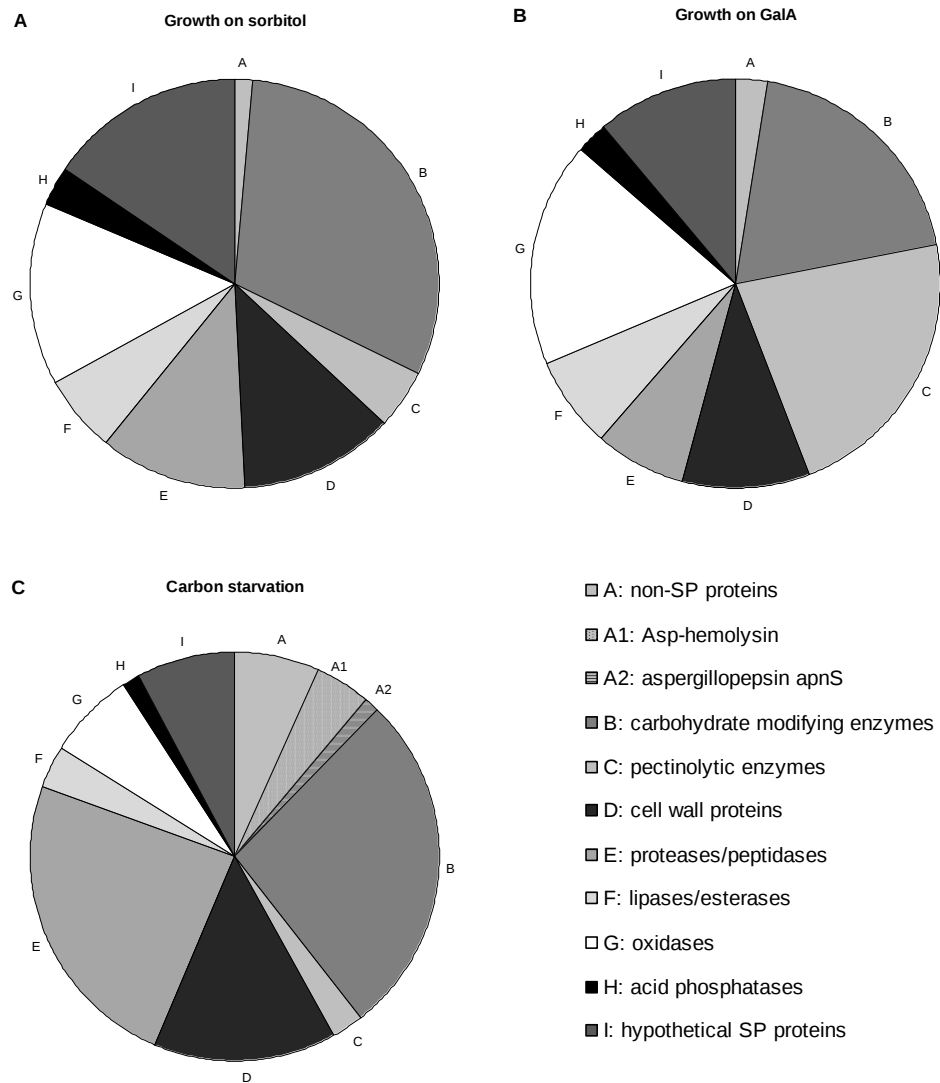


Fig. 2. Categorization of the *A. niger* secretome. Detected secretome when grown on sorbitol (A), on galacturonic acid (B), or under carbon starvation conditions (C). For each condition, the contribution of a protein to a category was normalized based on the total number of spectra.

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	<i>r</i> *	Gene name	Molecular Function	Signal Peptide	Spectral counts		
					Sorbitol	GalA †	Starvation
An14g04370	0.999	<i>pelA</i>	Pectin lyase A	Yes	0	8	0
An12g07500	0.979	<i>pgaX</i>	Exopolygalacturonase X	Yes	0	18	0
An11g04040	0.978	<i>pgaA</i>	Exopolygalacturonase A	Yes	Not detected		
An03g06740	0.971	<i>pgxB</i>	Exopolygalacturonase B	Yes	0	40	0
An02g12450	0.964	<i>pgxC</i>	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.

