Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition

Machtelt Braaksma,^{1,2} Sabina Bijlsma,¹ Leon Coulier,¹ Peter J. Punt^{1,2} and Mariët J. van der Werf^{1,2}

Correspondence

Machtelt Braaksma
machtelt.braaksma@tno.nl

¹TNO Quality of Life, PO Box 360, 3700 AJ Zeist, The Netherlands

²Kluyver Centre for Genomics of Industrial Fermentation, PO Box 5057, 2600 GA Delft, The Netherlands

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 highly relevant for classical improvement approaches, and even more so for modern systems biology approaches. In this study, the information content of a metabolomics dataset 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, also taking into account different time points of sampling during the fermentation. Both linear and nonlinear relationships 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 datasets 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 datasets. A detailed analysis of the metabolites that showed strong correlation with these quantitative phenotypes revealed that various sugar derivatives correlated with glucoamylase activity. For the reduction of protease activity, mainly as-yet-unidentified compounds correlated.

Received 4 May 2010

Revised 20 August 2010

Accepted 13 September 2010

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 enzymic 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 have been used. These include glucoamylase concentration (in g l^{-1}) (Withers *et al.*, 1998), activity (in U l^{-1}) (Wang *et al.*, 2008), yield [$\text{in mol product (mol substrate)}^{-1}$] (Melzer *et al.*, 2007), specific concentration or activity [$\text{in g (g cell dry weight)}^{-1}$ or $\text{U (g cell dry weight)}^{-1}$, respectively] (Swift *et al.*, 2000; Pedersen *et al.*, 2000; Schrickx *et al.*, 1993), and specific productivity [$\text{in mol, g or U (g cell dry weight)}^{-1} \text{ h}^{-1}$] (Melzer *et al.*, 2007; Withers *et al.*, 1998; Schrickx *et al.*, 1993).

The motivation for choosing a particular 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

Abbreviations: DWT, cell dry weight; IP-LC-MS, ion pair-LC-MS; LC-MS, liquid chromatography-MS; LN, natural logarithm; MVDA, multivariate data analysis; OS-GC-MS, oximation silylation-GC-MS; PLS, partial least-squares; RP-LC-MS, reverse phase-LC-MS.

Two supplementary data files, a supplementary figure and a supplementary table are available with the online version of this paper.

Kennedy & Krouse (1999) in their review of 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 important for classical optimization approaches, and perhaps even more so for modern, top-down systems biology approaches, particularly as the enormous quantity of data that arise from these systems biology studies may easily result in data overload (Braaksma *et al.*, 2010). 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 that 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 focusing on the yield itself during the strain improvement process. Focusing 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 analysed 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 dataset with respect to different quantitative phenotypes associated with the formation of either of the two different products. Subsequently, metabolites were identified that showed the strongest correlation with the phenotype studied.

METHODS

Strain and cultivation conditions. *A. 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 l working volume. Minimal medium (Bennett & Lasure, 1991) contained 7 mM KCl, 11 mM KH₂PO₄, 2 mM MgSO₄, 76 nM ZnSO₄, 178 nM H₃BO₃, 25 nM MnCl₂, 18 nM FeSO₄, 7.1 nM CoCl₂, 6.4 nM CuSO₄, 6.2 nM Na₂MoO₄ and 134 nM EDTA. This medium was supplemented with the appropriate carbon or nitrogen source at the concentrations 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 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 nine 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 *p*-nitrophenyl α -D-glucopyranoside (PNPG) (Sigma-Aldrich) as substrate (Withers *et al.*, 1998). The procedure was fully automated using a Cobas Mira Plus autoanalyser. Thirty microlitres of cleared culture supernatant was incubated with 90 μ l 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 μ l 0.1 M borate buffer, pH 9.3, and the A₄₀₅ was determined. One unit of glucoamylase activity was defined as the amount of enzyme that produced an A₄₀₅ value equivalent to 1 μ mol *p*-nitrophenol l⁻¹ min⁻¹ 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 with cell dry weight (DWT), the internal standards phenylalanine-d₅, leucine-d₃ (Spectral Stable Isotopes) and labelled [¹³C₁₀, ¹⁵N₅]GTP (Sigma-Aldrich) were added prior to extraction. The intracellular metabolites were extracted from the cell suspensions by chloroform extraction at -45 °C, as described by Ruijter & Visser (1996). The water/methanol phase was subsequently divided into two portions, one for GC- and one for liquid chromatography (LC)-MS analysis. The LC-MS sample was deproteinized by filtration using a Microcon YM-10 (Millipore) filter centrifuged at 18 000 g and -20 °C for 16 h. 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 ²D, ¹⁵N-labelled amino acids (Spectral Stable Isotopes) was added.

Biomass determination

Cell culture samples. For the quantification of DWT, a known volume of cell culture was filtered through a dried, pre-weighed filter paper, followed by washing with distilled water twice and then drying at 110 °C 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 (Ruijter & Visser, 1996). The metabolite concentrations in the extracts were correlated with dry weight by the use of the above-mentioned internal standards added prior to the extraction of the cell pellets.

Analytical procedures

Ion pair (IP)-LC-MS. Lyophilized metabolome samples were dissolved in 100 μ l methanol/water (1:3, v/v) and analysed as described by Coulier *et al.* (2006). Samples (10 or 20 μ l) were separated on a reversed-phase column (Chrompack Inertsil 5 mm ODS-3 100 $\times$ 3 mm) 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.). During data acquisition, the mass spectrometer probe voltage was maintained at 3–4 kV, and the heated capillary was kept at 250 °C.

Reverse phase (RP)-LC-MS. After analysis by IP-LC-MS, the redissolved metabolome samples were used for analysis by RP-LC-MS. Samples (10 or 20 µl) were separated on a reversed-phase column (Waters Sunfire C18, 150 × 3 mm, 3.5 µm) using a linear gradient from 100 % water + 0.1 % formic acid to 75 % methyl cyanide/water (80 %/20 %) + 0.1 % formic acid in 18 min followed by a linear gradient to 100 % methyl cyanide/water (80 %/20 %) + 0.1 % formic acid in 10 min 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.

Oximation silylation (OS)-GC-MS. 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.* (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 to 320 °C at a rate of 10 °C min⁻¹ on an Agilent 6890N GC and an Agilent 5973 mass selective detector. One microlitre aliquots of the derivatized samples were injected splitless on an HP5-MS capillary column (30 m × 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 into MATLAB [version 7.7.0.471 (R2008b); The Mathworks, Inc.]. 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 labelled amino acids. Finally, peaks were normalized with respect to the amount of extracted biomass per sample.

Preprocessed data files for both LC-MS and GC-MS 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 that were present in at least 25 % of the samples (Rubingh *et al.*, 2006; Bijlsma *et al.*, 2006). Next, all remaining zero values in the separate GC-MS, IP-LC-MS and RP-LC-MS datasets were replaced by a threshold value of half the lowest value in the dataset unequal to zero (Rubingh *et al.*, 2006). 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 datasets to be used as input for multivariate data analysis (MVDA).

MVDA. 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 dataset was mean-centred and, in order to compare the metabolites with 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). Partial least-squares (PLS) analysis was performed in the MATLAB environment using the PLS Toolbox (version 5.0.3, 2008; Eigenvector Research). The PLS results were cross-validated by using a 10-fold single cross-validation procedure. In addition to PLS analysis of the original metabolome and phenotype data, PLS analysis was also performed after either natural logarithm (LN) transformation of the phenotype data in combination with the original metabolome data or after LN 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 whether the number of latent variables (LVs) chosen by the algorithm seemed appropriate with respect to the root mean square error of cross validation (RMSECV) curve. In general, if more LVs are included in the PLS model, the given model will contain more noise. In the case that too many LVs were chosen by the algorithm, a new PLS model was generated by choosing a smaller number of LVs.

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 reanalysed using high-resolution and/or tandem MS (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 by *A. niger* of two industrially relevant products, i.e. glucoamylase and proteases, was studied. To this end, *A. niger* was grown under 16 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 DWT concentrations. One sample was collected at the middle of the exponential growth phase (mid-exponential), one at the end of the exponential growth phase (late-exponential) and one during the stationary growth phase. Samples were immediately quenched in a methanol solution to prevent alterations in the metabolite composition. Subsequently, the metabolites were extracted from the cells under quenched conditions, and the metabolites present were analysed using three analytical methods (see Methods).

The production of glucoamylase and protease was monitored during the course of the fermentation by analysing

culture samples every 6 h. The variation in maximum protease and glucoamylase activities under the different experimental conditions is shown in Fig. 1. For protease activity, the variation was evenly distributed over the different experimental conditions (Braaksma *et al.*, 2009). For glucoamylase, the experiments could be clearly separated in two groups: one group with very low activities representing conditions where the fungus was grown under non-induced conditions (on xylose), and another group with high activities representing growth under induced conditions (on glucose).

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

Fig. 1. (a) Maximum protease activity and (b) maximum glucoamylase activity in the different fermentations.

Fig. 2. Schematic representation of production over 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).

point of sampling (see A1 and B1, respectively, in Fig. 2). However, when a sample is collected during the stationary phase of 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 has occurred. In addition to the phenotypes described above, similar quantitative phenotype values were also calculated using the maximum activity or productivity for these products (see also Braaksma *et al.*, 2009). Thus, in this last 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 File 1.

Analysis of the information content of the dataset

The MVDA tool PLS was used to determine the information content of the metabolome datasets 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 of metabolomics data it is important to realize that due to the relatively large number of variables and small 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 of 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 an 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 dataset 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 the samples collected at only one of the growth phases of the fermentation. In addition, PLS models were generated to evaluate nonlinear relationships between the quantitative phenotype and the metabolome data, in order to identify metabolites with a nonlinear relationship 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.

Information content of the metabolomics dataset 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 Table 1(a) (glucoamylase) and Table 1(b) (protease) with each other, one thing is obvious: the highest information content of the metabolomics dataset was obtained with different quantitative phenotypes for the different products. For glucoamylase, good models were particularly obtained when based on metabolome data of the samples from the mid-exponential growth phase, while most good PLS models for protease were based on the inclusion of metabolome data from samples at all three time points.

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 type.

Metabolome data for:		Maximum phenotype						Phenotype at time point of sampling						
		Phenotype*	P	R ² _{CV}	LN(P)	R ² _{CV}	LN(M)	R ² _{CV}	Phenotype*	P	R ² _{CV}	LN(P)	R ² _{CV}	LN(M)
(a) Glucoamylase														
All samples	Max.Act.	G1	0.59	G49	0.66	G97	0.75	Act.	G25	0.40	G73	0.68	G121	0.51
	Max.Spec.Act.-1	G2	0.47	G50	0.64	G98	0.64	Spec.Act.-1	G26	0.38	G74	0.67	G122	0.48
	Max.Spec.Act.-2	G3	0.59	G51	0.64	G99	0.77	Spec.Act.-2	G27	0.41	G75	0.66	G123	0.53
	Max.Prod.	G4	0.59	G52	0.67	G100	0.73	Prod.	G28	0.55	G76	0.59	G124	0.69
	Max.Spec.Prod.-1	G5	0.60	G53	0.63	G101	0.78	Spec.Prod.-1	G29	0.56	G77	0.57	G125	0.66
	Max.Spec.Prod.-2	G6	0.59	G54	0.65	G102	0.74	Spec.Prod.-2	G30	0.59	G78	0.57	G126	0.68
Mid-exponential samples only	Max.Act.	G7	0.71	G55	0.76	G103	0.77	Act.	G31	0.67	G79	0.69	G127	0.67
	Max.Spec.Act.-1	G8	0.47	G56	0.72	G104	0.67	Spec.Act.-1	G32	0.63	G80	0.69	G128	0.67
	Max.Spec.Act.-2	G9	0.62	G57	0.74	G105	0.71	Spec.Act.-2†	G33	0.63	G81	0.69	G129	0.67
	Max.Prod.	G10	0.79	G58	0.75	G106	0.82	Prod.	G34	0.78	G82	0.69	G130	0.78
	Max.Spec.Prod.-1	G11	0.71	G59	0.74	G107	0.82	Spec.Prod.-1	G35	0.77	G83	0.70	G131	0.81
	Max.Spec.Prod.-2	G12	0.73	G60	0.74	G108	0.82	Spec.Prod.-2†	G36	0.77	G84	0.70	G132	0.81
Late-exponential samples only	Max.Act.	G13	0.43	G61	0.63	G109	0.50	Act.	G37	0.22	G85	0.49	G133	0.30
	Max.Spec.Act.-1	G14	0.42	G62	0.62	G110	0.48	Spec.Act.-1	G38	0.23	G86	0.48	G134	0.33
	Max.Spec.Act.-2	G15	0.43	G63	0.61	G111	0.49	Spec.Act.-2†	G39	0.23	G87	0.48	G135	0.33
	Max.Prod.	G16	0.60	G64	0.72	G112	0.57	Prod.	G40	0.33	G88	0.28	G136	0.42
	Max.Spec.Prod.-1	G17	0.67	G65	0.71	G113	0.66	Spec.Prod.-1	G41	0.29	G89	0.25	G137	0.38
	Max.Spec.Prod.-2	G18	0.61	G66	0.70	G114	0.58	Spec.Prod.-2†	G42	0.29	G90	0.25	G138	0.38
Stationary samples only	Max.Act.	G19	0.00	G67	0.01	G115	0.41	Act.	G43	0.04	G91	0.00	G139	0.34
	Max.Spec.Act.-1	G20	0.01	G68	0.03	G116	0.45	Spec.Act.-1	G44	0.01	G92	0.01	G140	0.34
	Max.Spec.Act.-2	G21	0.03	G69	0.04	G117	0.44	Spec.Act.-2	G45	0.02	G93	0.01	G141	0.37
	Max.Prod.	G22	0.02	G70	0.01	G118	0.40	Prod.	G46	0.05	G94	0.02	G142	0.40
	Max.Spec.Prod.-1	G23	0.00	G71	0.03	G119	0.44	Spec.Prod.-1	G47	0.01	G95	0.02	G143	0.38
	Max.Spec.Prod.-2	G24	0.00	G72	0.02	G120	0.39	Spec.Prod.-2	G48	0.01	G96	0.02	G144	0.40
(b) Protease														
All samples	Max.Act.	P1	0.70	P49	0.75	P97	0.78	Act.	P25	0.70	P73	0.57	P121	0.80
	Max.Spec.Act.-1	P2	0.66	P50	0.66	P98	0.72	Spec.Act.-1	P26	0.66	P74	0.46	P122	0.75
	Max.Spec.Act.-2	P3	0.57	P51	0.60	P99	0.66	Spec.Act.-2	P27	0.67	P75	0.44	P123	0.77
	Max.Prod.	P4	0.71	P52	0.69	P100	0.80	Prod.	P28	0.45	P76	0.65	P124	0.61
	Max.Spec.Prod.-1	P5	0.58	P53	0.50	P101	0.63	Spec.Prod.-1	P29	0.32	P77	0.49	P125	0.45
	Max.Spec.Prod.-2	P6	0.58	P54	0.48	P102	0.65	Spec.Prod.-2	P30	0.36	P78	0.48	P126	0.45
Mid-exponential samples only	Max.Act.	P7	0.46	P55	0.72	P103	0.47	Act.	P31	0.09	P79	0.01	P127	0.17
	Max.Spec.Act.-1	P8	0.38	P56	0.58	P104	0.32	Spec.Act.-1	P32	0.03	P80	0.01	P128	0.05
	Max.Spec.Act.-2	P9	0.28	P57	0.55	P105	0.28	Spec.Act.-2†	P33	0.03	P81	0.01	P129	0.05
	Max.Prod.	P10	0.51	P58	0.69	P106	0.43	Prod.	P34	0.21	P82	0.42	P130	0.18
	Max.Spec.Prod.-1	P11	0.28	P59	0.44	P107	0.16	Spec.Prod.-1	P35	0.16	P83	0.24	P131	0.12
	Max.Spec.Prod.-2	P12	0.29	P60	0.45	P108	0.18	Spec.Prod.-2†	P36	0.16	P84	0.24	P132	0.12
Late-exponential samples only	Max.Act.	P13	0.52	P61	0.65	P109	0.62	Act.	P37	0.23	P85	0.29	P133	0.41
	Max.Spec.Act.-1	P14	0.37	P62	0.48	P110	0.47	Spec.Act.-1	P38	0.25	P86	0.09	P134	0.29
	Max.Spec.Act.-2	P15	0.42	P63	0.49	P111	0.47	Spec.Act.-2†	P39	0.25	P87	0.09	P135	0.29
	Max.Prod.	P16	0.48	P64	0.59	P112	0.44	Prod.	P40	0.49	P88	0.51	P136	0.51
	Max.Spec.Prod.-1	P17	0.28	P65	0.30	P113	0.17	Spec.Prod.-1	P41	0.32	P89	0.23	P137	0.26
	Max.Spec.Prod.-2	P18	0.29	P66	0.34	P114	0.18	Spec.Prod.-2†	P42	0.32	P90	0.23	P138	0.26
Stationary samples only	Max.Act.	P19	0.11	P67	0.19	P115	0.68	Act.	P43	0.18	P91	0.18	P139	0.69
	Max.Spec.Act.-1	P20	0.11	P68	0.25	P116	0.58	Spec.Act.-1	P44	0.20	P92	0.14	P140	0.57
	Max.Spec.Act.-2	P21	0.11	P69	0.14	P117	0.59	Spec.Act.-2	P45	0.19	P93	0.15	P141	0.59

Table 1. cont.

Metabolome data for:	Maximum phenotype							Phenotype at time point of sampling						
	Phenotype*	P	R ² _{CV}	LN(P)	R ² _{CV}	LN(M)	R ² _{CV}	Phenotype*	P	R ² _{CV}	LN(P)	R ² _{CV}	LN(M)	R ² _{CV}
	Max.Prod.	P22	0.17	P70	0.14	P118	0.60	Prod.	P46	0.18	P94	0.39	P142	0.55
	Max.Spec.Prod.-1	P23	0.14	P71	0.19	P119	0.44	Spec.Prod.-1	P47	0.05	P95	0.45	P143	0.38
	Max.Spec.Prod.-2	P24	0.18	P72	0.18	P120	0.47	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 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 those of Spec.Act.-1 and Spec.Prod.-1, respectively. To calculate Spec.Act.-2 and Spec.Prod.-2, in principle DWT_{max} is used, except for samples collected before DWT_{max} was reached (as is the case for the mid-exponential and late-exponential samples). For these samples, DWT at the time point of sampling was used, in a manner similar to the calculation of Spec.Act.-1 and Spec.Prod.-1 (see also Supplementary File 1).

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 an 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 a starting point for analysing the most strongly correlating metabolites in more detail. Based on this analysis, lists of correlating metabolites from other good PLS models were subsequently 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-exponential growth phase. From this group of models, the PLS model in relation to maximum activity (PLS model G7) was selected as a starting point for target identification and comparison with 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 relatively 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 disaccharides as well as some of the other compounds, such as DL-amino adipic acid, 2,3-butanediol and xylitol, the correlation was based on the absence of the compounds in all xylose samples and their presence in all glucose samples (Table 2). However, there is no clear correlation between their intracellular concentrations and maximum glucoamylase activity based on the glucose samples alone (e.g. Fig. 3a). On the other hand, for putrescine, ornithine, glucose 6-phosphate and fructose 6-phosphate there was a correlation between increasing intracellular levels of these compounds and maximum glucoamylase activity (e.g. see Fig. 3b).

When comparing the top 20s of other models with an $R^2_{CV} \geq 0.6$ with each other, the use of metabolome samples from particular sampling times was especially of influence on the resulting top 20 (see Supplementary File 2A). When either the metabolome data of all time samples (e.g. model G49) or only the metabolome data of the mid- or late-exponential samples (models G55 and G61, respectively) was used, only four metabolites were present in all three resulting top 20s. 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 the ID AN 320-218 22.96 min (Supplementary File 2A).

The effect of LN transformation on the ranking of the potential targets was somewhat ambiguous. The effect of LN transformation of the phenotype or metabolome data on the resulting top 20s 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 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 File 2A).

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

Rank	Metabolite*	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 by OS-GC-MS.

†Visual correlation is indicated by + (positive correlation), — (negative correlation) or 0 (no apparent correlation); see also Supplementary Fig S1(a).

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

§These are different mass fragments of the same compound.

||Only highly abundant on xylose.

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 20s 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 a starting point for target identification and comparison with 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 isoleucine and valine biosynthesis, respectively. A number of the compounds in the top 20 contained a phosphate group; however, very little is known about the possible involvement of

phosphorus sources on protease expression in *aspergilli*. In comparison with the glucoamylase results, the relatively high contribution of compounds analysed by RP-LC-MS was remarkable. Among other applications, 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, appropriate reference compounds to establish their exact identity are not currently available for any of these compounds.

When comparing the top 20s from good PLS models for protease with each other, the overall observations were in line with those for glucoamylase. Also, for protease, the largest differences between the top 20s 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- or late-exponential samples (see Supplementary File 2B for details). Furthermore, the influence on the resulting top 20s 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

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 ($n=11$).

20s showed 50 % or less overlap with the top 20 without LN transformation (Supplementary Table File 2B).

DISCUSSION

The choice of a particular 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 dataset was determined with respect to different quantitative phenotypes related to the formation of one simple product, 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 dataset was higher for the simpler of these two products, i.e. glucoamylase. This is in one sense remarkable, because the fermentation conditions from which the metabolome samples were collected in this study were originally selected to result in a large and evenly distributed variation in protease activity (Braaksma *et al.*, 2009).

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

Our results show that the effect of different ways of calculating the quantitative phenotype on the information content and the resulting targets is much smaller than the effect of the time point of sampling. In general, the number of PLS models with an R^2_{CV} above the threshold value was higher when quantitative phenotypes were used that were

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*	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 parentheses: GC, OS-GC-MS; IP, IP-LC-MS; RP, RP-LC-MS.

†Visual correlation is indicated by + (positive correlation), – (negative correlation) or 0 (no apparent correlation); see also Supplementary Fig S1(b).

‡These are different mass fragments of the same compound.

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 with 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 the maximum phenotype or the phenotype at the time point of sampling on the resulting top 20s is limited (Supplementary File 2). This holds for the different descriptions 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 an R^2_{CV} above the threshold value increase with LN transformation of the phenotype or the metabolome data, but also the resulting top 20s 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 dataset 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 concentration of each metabolite against the phenotype in order to investigate the relationship between individual metabolites and the quantifiable phenotype of interest. However, in the case of a large number of metabolites, and in our case a large number of phenotypes as well, this approach will result in an extremely large number of plots to analyse. Moreover, in such plots, the intrinsic interdependence of the metabolites is neglected. However, despite these advantages of MVDA over a univariate approach, interpretation of the relationship 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 of the contribution of a metabolite to the phenotype cannot be directly translated into how a metabolite actually correlates with the phenotype. These regression factors are a measure not only of the correlation of a particular metabolite with the phenotype but also of the correlation of the metabolite with 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 correlate with the phenotype of interest are involved in the production of the product, either as inducer/inhibitor or precursor/side-product. With MVDA, no distinction can be made between metabolites that correlate with the phenotype due to a cause and those that correlate due to an effect relationship. For instance, one might conclude that the disaccharides found to correlate with 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 the transglucosylation activity of glucoamylase (Nikolov *et al.*, 1989), and thus are an effect of glucoamylase activity ('effect correlation'). For strain improvement in particular, cause relationships are of importance.

(3) Related to the previous subject is the occurrence of confounding effects, i.e. the situation in which 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 particular metabolite. For example, significant glucoamylase activity is only observed when *A. niger* is cultured on glucose instead of xylose. Also, several metabolites, such as uric acid and xylitol, are mainly present when *A. niger* is cultured on glucose. Therefore, one might 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; instead, both glucoamylase and these metabolites may be independently correlated with growth on a specific carbon source.

(4) With the comprehensive analytical methods used in this study, not only known compounds are analysed but also all peaks of unknown identity are included in the dataset. One last aspect hampering the interpretation of the data is the correlation of these unidentified metabolites with the phenotype. To further investigate the possible role of these unknown compounds, an additional effort will have to be made to (partially) identify these metabolites.

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 investigation can be distilled from the initial list of potential targets that results 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 catalyses the conversion of glucose 6-phosphate into fructose 6-phosphate. The ratio between the concentration of glucose 6-phosphate and that of fructose 6-phosphate is approximately sevenfold 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 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 under non-producing conditions. However, in our study, even under non-producing conditions, the ratio between glucose 6-phosphate and fructose 6-phosphate concentrations was approximately sevenfold higher than expected. This weakens the hypothesis that the flux through the PPP is only insufficient under glucoamylase-producing conditions, even though when glucose was used as the carbon source the absolute concentrations of both metabolites were higher. Alternatively, absolute metabolite concentrations could also be involved in the 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 glycolysis and the PPP, the regulation of the activity of this enzyme may be a means of regulating the fluxes through these two pathways and thus optimizing glucoamylase production.

Putrescine and ornithine are the other two 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 *Aspergillus nidulans*, there is an absolute requirement for polyamides in growth and development (Tabor & Tabor, 1985; Jin *et al.*, 2002). The positive correlation between glucoamylase production and putrescine suggests that glucoamylase production is stimulated either by addition of this polyamine to the medium or by overexpression of the gene that encodes 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 that correlate with protease activity are unidentified compounds (Table 3). The presence of several compounds analysed by RP-LC-MS in Table 3 suggests the involvement of small peptides in protease induction. Unfortunately, the identification of peptides by RP-LC-MS has proved to be quite difficult, 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 relationship 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.*, 2010). Possibly, an approach with metabolomics alone is not sufficient to identify targets for

such a complex phenotype, and an integrated systems biology approach is required.

Besides glucoamylase and protease production, citric acid production was also analysed as a phenotype. Although the experimental design of our dataset was not optimally suited for this product, resulting in very few reliable PLS models (Supplementary Table S1), several tricarboxylic acid (TCA) cycle intermediates (isocitrate, α -ketoglutarate) were identified as correlating with citric acid production (results not shown). Altogether, this 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 seems confirmed by what already is known in the 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, with regard to not only the fermentation conditions, but also 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 Mulwijk and Richard Bas for identification of the metabolites, and Carina de Jong-Rubingh for the smoothing algorithm analyses and the calculation of 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.

REFERENCES

- 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.
- 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., 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*. *Microbiology* **155**, 3430–3439.
- Braaksma, M., van den Berg, R. A., van der Werf, M. J. & Punt, P. J. (2010). 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 & D. J. Ebbel. Washington, DC: American Society for Microbiology.
- Braaksma, M., Martens-Uzunova, E. S., Punt, P. J. & Schaap, P. J. (2010). An inventory of the *Aspergillus niger* secretome by combining *in silico* predictions with shotgun proteomics data. *BMC Genomics* **11**, 584.
- 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.
- 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.
- Kennedy, M. & Krouse, D. (1999). Strategies for improving fermentation medium performance: a review. *J Ind Microbiol Biotechnol* **23**, 456–475.
- Koek, M. M., Mulwijk, B., van der Werf, M. J. & Hankemeier, T. (2006). Microbial metabolomics with gas chromatography/mass spectrometry. *Anal Chem* **78**, 1272–1281.
- 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.
- 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.
- 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.
- 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.
- Rubingh, C. M., Bijlsma, S., Jellema, R. H., Overkamp, K. M., van der Werf, M. J. & Smilde, A. K. (2009). Analyzing longitudinal microbial metabolomics data. *J Proteome Res* **8**, 4319–4327.
- Ruijter, G. J. G. & Visser, J. (1996). Determination of intermediary metabolites in *Aspergillus niger*. *J Microbiol Methods* **3**, 295–302.
- 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 *glcA* gene. *J Gen Microbiol* **139**, 2801–2810.
- 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.
- Tabor, C. W. & Tabor, H. (1985). Polyamines in microorganisms. *Microbiol Rev* **49**, 81–99.
- 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.
- 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 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 Werf, M. J. (2005). Towards replacing closed with open target selection strategies. *Trends Biotechnol* **23**, 11–16.

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

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 Chemometr* 10, 425–438.

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.

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

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.

Edited by: R. P. Oliver