A functional genomics study of extracellular protease production by Aspergillus niger

Braaksma, M.

CHAPTER 4

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

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

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

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

ABSTRACT

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

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

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

INTRODUCTION

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

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

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

In bioprocess optimization a high quantity, e.g. concentration, of a product is not automatically the most desired result. In the case the substrate is an expensive part of the total fermentation costs, a high yield may be more relevant. However, improvement of the product yield is not always achieved by focussing on the yield itself during the strain improvement process. Focussing on the productivity may require fewer strain improvement steps during a particular bioprocess optimization process, thus resulting in an improved yield more quickly. Reduction of the

fermentation time is another parameter to reduce production costs and can be realized by increasing the productivity. It is very likely that selection of either of these phenotypes for optimization will result in different targets to obtain the desired increase.

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

METHODS

Strain and cultivation conditions

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

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

Enzyme assays

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

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

Collection of samples, extraction and sample clean-up

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

Biomass determination

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

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

Analytical procedures

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

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

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

aliquots of the derivatized samples were injected splitless on a HP5-MS capillary column (30 m x 0.25 mm, 0.25 µm film thickness, Agilent). Detection was performed using MS detection in electron impact mode (70 eV).

Data preprocessing

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

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

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

Multivariate data analysis

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

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

RESULTS

Experimental setup

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

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

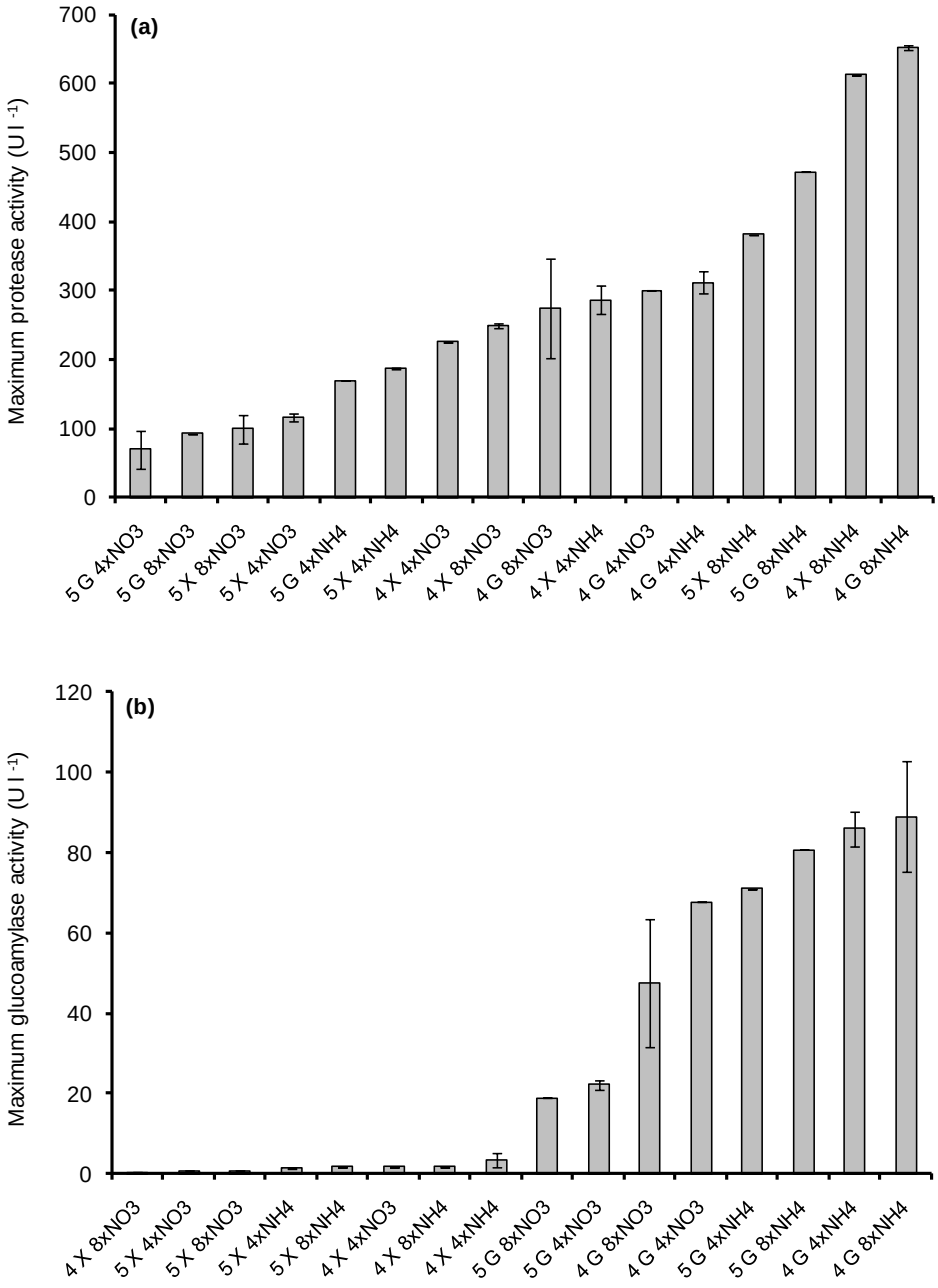

Fig. 1. (A) Maximum protease activity and (B) maximum glucoamylase activity in the different fermentations.

Quantitative phenotypes

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

Analysis of the information content of the data set

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

statistical models. For both products, cross validated PLS models were made for all different quantitative phenotypes (Table 1).

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

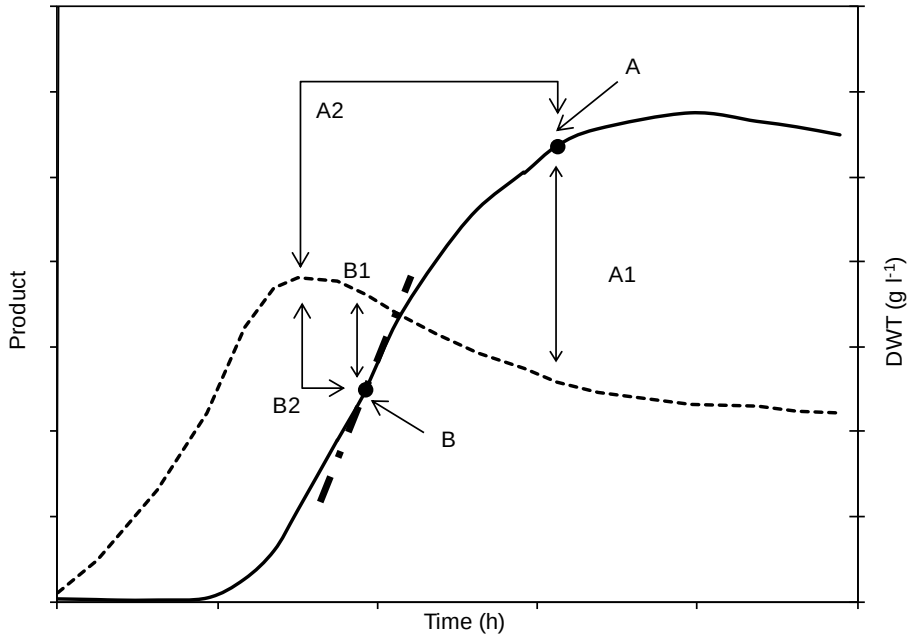

Fig. 2. A schematic representation of production in time to illustrate the various product-related phenotypes that can be defined. Solid line, product; dashed line, biomass concentration DWT. (A) activity at time point of sampling; (A1) specific activity – 1, based on the biomass at the time point of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

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

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

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

Table 1. Continued.

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

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

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

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

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

Identification of metabolites that correlate with the phenotype studied

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

Glucoamylase

For glucoamylase, most PLS models were above the threshold of $R^2_{CV} = 0.6$ when using metabolome data of the samples collected during the mid log growth phase. From this group of models, the PLS model in relation to maximum activity (PLS model G7), was selected as starting point for target identification and comparison to other good PLS

models for glucoamylase. From this G7 PLS model, the 20 highest ranking metabolites are shown in Table 2. This top 20 included a relative high number of disaccharides and other sugar-derived compounds that were only present under glucoamylase inducing conditions (i.e. with glucose as carbon source). For all these dissacharides as well as some of the other compounds, such as DL-aminoadipic acid, 2,3-butanediol and xylitol the correlation is based on the absence of the compounds in all xylose samples and the presence in all glucose samples (Table 2). However, there is no clear correlation between their intracellular concentrations and maximum glucoamyase activity based on only the glucose samples (e.g. Fig. 3A). On the other hand, for putrescine, ornithine, glucose-6-phosphate, and fructose-6-phosphate there is a correlation between increasing intracellular levels of these compounds and maximum glucoamyase activity (e.g, see Fig. 3B).

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

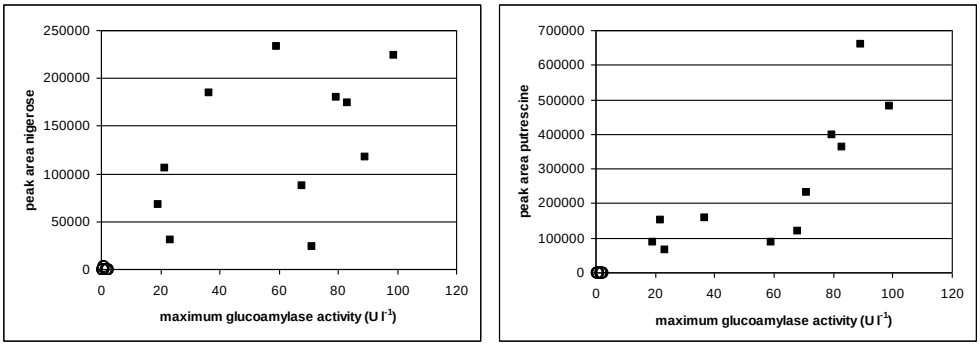

Fig. 3. Plot of the correlation between the metabolite tentatively identified as nigerose and maximum glucoamylase activity (A) and a similar plot for putrescine (B). ○, Metabolome samples from xylose fermentations (n=11); ■, metabolome samples from glucose fermentations (n=11).

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

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

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

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

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

§ These are different mass fragments of the same compound.

|| Only high abundant on xylose.

The effect of LN transformation on the ranking of the potential targets was somewhat ambiguous. The effect of LN transformation of the phenotype or the metabolome data on the resulting top 20's was in several cases limited. For instance, for PLS models G7, G55 and G103 50% of the compounds were present in all three lists (for details, see Supplementary data file 2A). However, in other cases, i.e. PLS models G34, G82 and G130, this was only the case for 25% of the compounds (for details, see Supplementary data file 2A). The exact effect of LN transformation on the correlations of the metabolites with the phenotype was unclear; plotting of the peak areas of metabolites exclusively present in the top 20's after LN transformation against the

phenotype showed in some cases an improvement of the linear correlation, while in other cases the linear correlation deteriorated (data not shown).

Protease

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

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

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

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

* The analytical method used to detect each metabolite is indicated in between brackets: GC, OS-GC-MS; IP, IP-LC-MS; and RP, RP-LC-MS.
† Visual correlation is indicated by + (positive correlation), – (negative correlation), or 0 (no apparent correlation); see also Supplementary data file 3B.
‡ These are different mass fragments of the same compound.

DISCUSSION

The choice for a certain quantitative phenotype in bioprocess optimization often seems rather random, but may have a major influence on the outcome of an optimization strategy. In this study, the information content of a metabolomics data set was determined with respect to different quantitative phenotypes related to the formation of two simple products, i.e. glucoamylase, and a more complex product, i.e. protease. When comparing the results of the two enzyme products glucoamylase and protease, it could be concluded that the information content of the metabolomics data

set is higher for the simpler of these two products, i.e. glucoamylase. This is on the one hand remarkable, because the fermentation conditions from which the metabolome samples were collected in this study, were originally selected to result in large and evenly distributed variation in protease activity (Braaksma *et al.*, 2009).

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

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

With the MVDA tool PLS the quantifiable phenotype of interest can be related to the metabolome data set as a whole and at the same time take into account the

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

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

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENTS

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