A functional genomics study of extracellular protease production by Aspergillus niger

Braaksma, M.

CHAPTER 6

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

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

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

INTRODUCTION

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

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

OPTIMIZATION OF FUNGAL PRODUCTION PROCESS

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

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

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

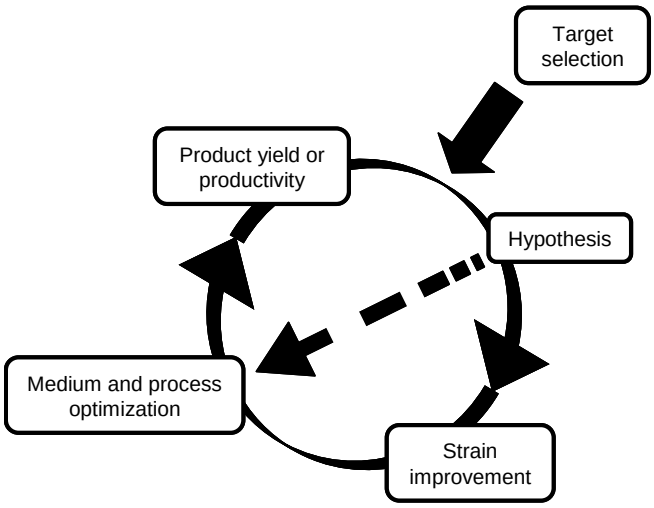


Fig 1. Iterative cycle of strain improvement and/or process optimization.

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

TOP-DOWN SYSTEMS BIOLOGY

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

In contrast, in top-down systems biology, data are used as the starting point and statistical data mining approaches are applied to come to a comprehensive understanding of the biological system. The principal behind top-down systems biology is that molecular components that respond similarly to changes in the experimental conditions are somehow functionally related. No other prior assumptions regarding the interactions of the studied molecular components are required. This allows the study of complex and relatively poorly characterized processes and strains, as extensive knowledge of the studied organism or process is not necessary. In this top-down systems biology approach there is also no *a priori* focus on specific biomolecules expected to relate to the biological question. Therefore, this approach also enables the discovery of previously unknown or unexpected

relations between specific biomolecules and the biological process studied. Despite the potential of top-down systems biology, the great majority of scientists applying systems biology use a bottom-up systems biology approach. The reluctances towards top-down systems biology might relate to the risk of being overwhelmed by the enormous quantity of data that arise from functional genomics technologies such as metabolomics and transcriptomics. The challenge is to be able to extract relevant information from these data sets. Principally, the success of this approach depends on balancing three interlinked key factors: (i) definition of the biological question, (ii) experimental design, and (iii) the data analysis tool (Fig. 2). These three factors are discussed in more detail below.

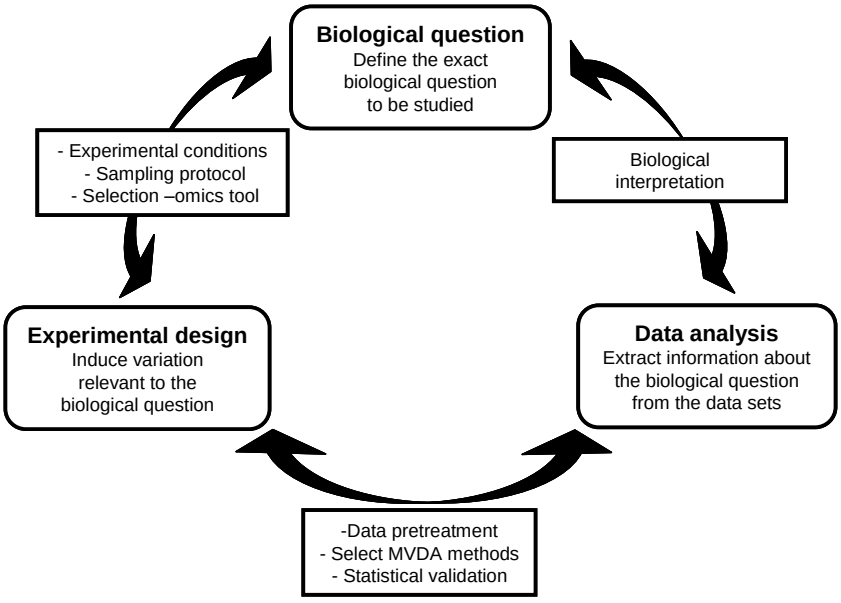


Fig. 2. Key conditions and their relation to a successful systems biology study. In top-down systems biology, three interlinked factors are crucial for success: (i) the biological question, (ii) the experimental design, and (iii) data analysis.

THE BIOLOGICAL QUESTION

A clear definition of the biological question to be answered is the crucial starting point in any top-down systems biology research project, because only then can a suitable experimental setup and data analysis strategy be selected (van der Werf *et al.*, 2005; Trygg *et al.*, 2007). To explain this in more practical terms, two examples are given of

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

EXPERIMENTAL DESIGN

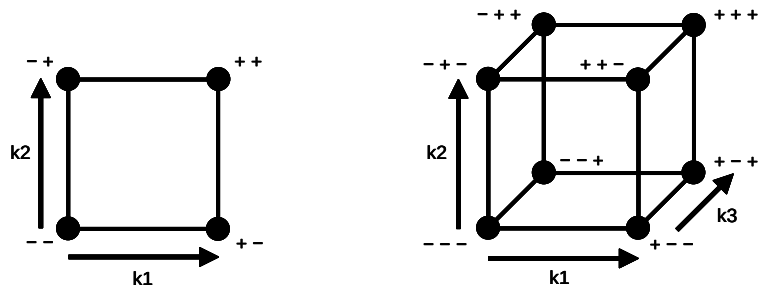
Based on the biological question, the experimental design of the top-down systems biology study should be aimed at generating large information-rich data sets in order for data analysis to extract relevant biological information from the data set. Not only experimental conditions for the experimental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

Experimental conditions

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

Traditionally, one of the most frequently used approaches to study which parameters affect biological responses is the change-one-factor-at-a-time approach, in which one independent variable is studied while all others are fixed at a specific level. An advantage of this simple and easy method is that any change in response can be attributed to a specific change. On the other hand, this change-one-factor-at-a-time approach has some serious drawbacks, perhaps the most important being that possible interactions between components are ignored. As a result, this approach frequently fails to find optimal conditions for experiments. Another disadvantage is the unnecessarily large number of experiments that are required when testing more than a few variables. Therefore, the change-one-factor-at-a-time method is acknowledged to have severe shortcomings and is more and more being replaced by statistics-based experimental designs, also called "Design of Experiments". For an initial screening of factors possibly related to the biological question, different types of experimental designs, so-called screening designs, are available, including the full factorial design (Lundstedt *et al.*, 1998). In a full factorial design, every level of a factor is investigated at all levels of all other factors. Often the factors are investigated at two levels, requiring a number of runs equal to 2^k for k factors, which results in a large number of experiments when many factors are investigated (Fig. 3). When the factors are investigated at three or more levels, requiring 3^k runs in the case of three levels and n^k runs for n levels, the number of experiments rapidly becomes impracticable. To reduce the number of experiments without the loss of too much information, several experimental designs derived from the full factorial design are available. The most commonly used one is the fractional factorial design (Lundstedt *et al.*, 1998; Trygg *et al.*, 2006), which requires only n^{k-p} number of runs, with k as the number of

investigated factors at *n* different levels, and *p* describing the size of the fraction of the full factorial used. With this type of design, three-way and higher interactions are ignored. Another useful screening tool is the Plackett-Burman design (Plackett & Burman, 1946; Weuster-Botz, 2000). This experimental design is a variation on the fractional factorial design, but instead of ignoring only higher interactions it considers all interactions between factors negligible. The downside of these two last designs is that when interactions between factors are not negligible, they are confounded with the estimated effects. This means that the estimated effects and those interaction effects cannot be distinguished from one another.



A. 2^2 full factorial design **B**. 2^3 full factorial design

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

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

Selection of a functional genomics tool

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

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

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

Sampling strategy

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

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

After quenching the cells, conditions should be maintained during sample work-up in order to prevent changes in the metabolite composition of RNA levels due to residual enzymatic activity present in the samples. Extraction of RNA from mycelium is often accomplished by disruption of the cells by either grinding under liquid nitrogen using a mortar and pestle (Kimura *et al.*, 2008; Foreman *et al.*, 2003) or bead-milling at

temperatures of approximately 4 °C (Andersen *et al.*, 2008b), followed by a standard RNA isolation protocol. Extraction of proteins is done in a similar way, without the stringent control of temperature (Carberry *et al.*, 2006). For fungal metabolomics samples, two methods in particular have been described for extracting metabolites from the cells. The first is boiling the cells in an ethanol-buffer solution and subsequent reduction of the volume by evaporation in a rotavapor (Nasution *et al.*, 2006). The second is chloroform extraction at -45 °C (Ruijter & Visser, 1996).

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

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

DATA ANALYSIS

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

Data pretreatment methods

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

MVDA tools

Choices in data analysis strategy are influenced by the biological question, the characteristics of the experimental design, the behaviour of the relevant biomolecules, and the dimensions of the data set. There are various MVDA methods that address different biological questions. In general, these methods can be divided in two main

groups, namely, unsupervised methods and supervised methods. Unsupervised methods include PCA (Jackson, 1991; Jolliffe, 2002) or hierarchical clustering analysis (Eisen *et al.*, 1998) that visualize relations/patterns in data sets without prior knowledge.

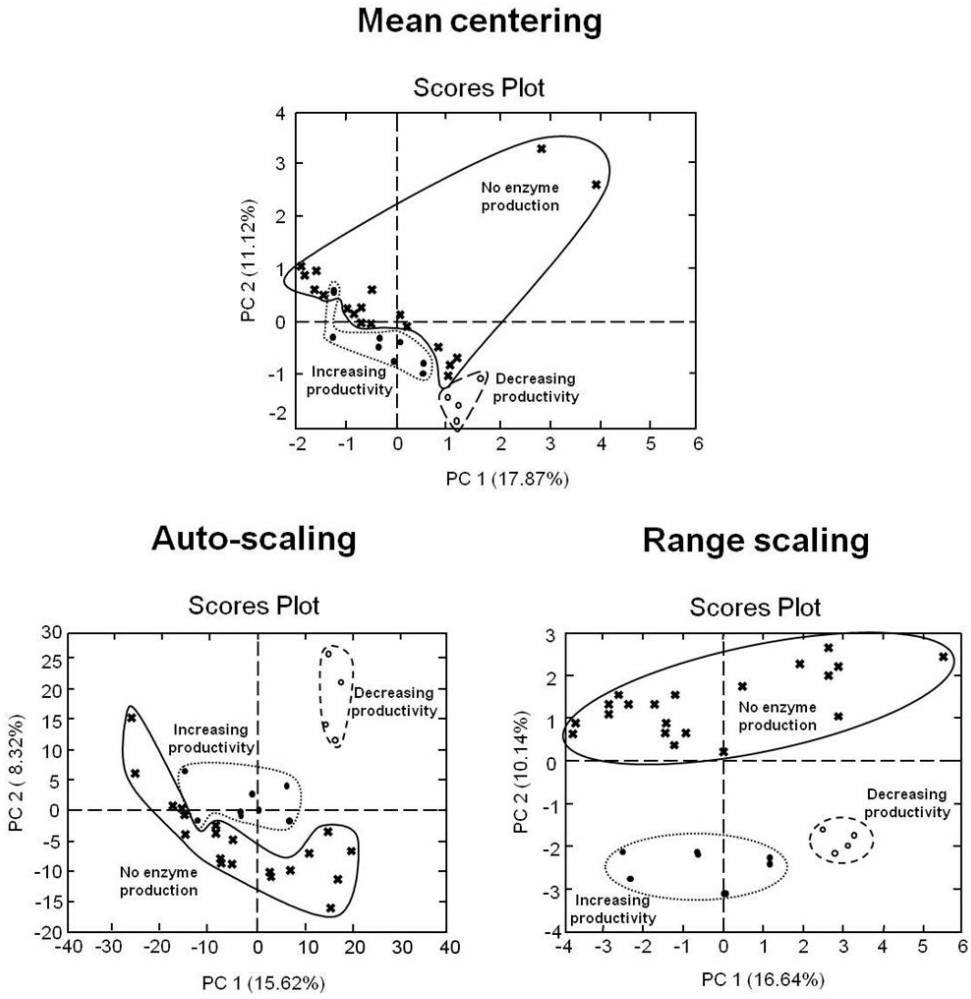


Fig. 4. The effect of mean scaling, autoscaling, or range scaling of metabolomics data sets on PCA data results. The data sets are derived from research related to induction of cellulase activity in *T. reesei* (van der Werf *et al.*, unpublished data). The metabolomes of three groups of samples (no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

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

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

Biological interpretation

Based on the list of biomolecules identified by the MVDA tools as being important in relation to the question under study, targets for improvement of the production

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

CONCLUSIONS

The available selection methods for relevant targets for fungal strain and process development, or for that matter any microbial production process, have been very successful in numerous cases. However, the exclusion of all biological processes or interactions that are not currently known to exist has been shown to hamper further improvement while using these approaches. Recently introduced functional genomics technologies in combination with MVDA tools enable an open and comprehensive top-down systems biology approach towards target selection. Nevertheless, the success of

such an approach depends heavily on a systematic study covering all aspects, from a clear description of the biological question up to statistical data analysis. As this involves knowledge beyond the biologist expertise (e.g., biostatistics), the assistance of experts in those fields will be indispensable. Due to its unbiased nature, a successful top-down system biology approach will provide a new boost in the ongoing cycle of bioprocess optimization.