University of Natural Resources and Life Sciences Vienna

Department for Biotechnology

Advisor:

Ao. Univ. Prof. DI Dr. Diethard Mattanovich

Towards *Pichia pastoris* Systems Biotechnology: Genome Sequence, Expression Microarrays and Genome-Scale Metabolic Model

Dissertation

For obtaining a doctorate degree at the

University of Natural Resources and Life Sciences Vienna

Submitted by

Mag. Dipl-Ing. (FH) Alexandra Graf

Vienna, September 2010

Abstract

The methylotrophic yeast *Pichia pastoris* is one of the major eukaryotic expression systems and extensively used for the production of recombinant proteins. Due to the possible humanization of its glycosylation pattern, it is especially well suited for pharmaceutical applications. Despite the ever growing interest in this cell factory, a genome sequence of *P. pastoris* was not publicly available at the time of this work. Without a genome sequence, the use of high throughput technologies is drastically hampered, and the analysis of system wide responses not feasible.

Previously, a commercial draft sequence of *P. pastoris* was used to develop gene expression microarrays. During this project, the microarray design was improved and an analysis pipeline was implemented to determine differential expression and gene set enrichment in the data set. The microarrays and the analysis pipeline were validated in a study comparing the effects of dithiothreitol (DTT) treatment and HAC1 overexpression on *P. pastoris*. Both factors are involved in the unfolded protein response, which plays an important role in the secretion of heterologous protein. The arrays and the pipeline were then used in an international, multi-organism study about the impact of environmental factors on protein production. During these studies, it soon became evident that a publicly available genome sequence of *P. pastoris* was needed to fully utilize the state of the art technologies.

Applying next generation sequencing technologies, the genome of the *P. pastoris* type strain DSMZ 70382 was sequenced, and subsequently annotated. Using the set of putative genes, and the homology based functional annotation, the secretome of *P. pastoris* could be computationally predicted. The *in-silico* secretome was then compared to the experimentally determined secretome from a glucose cultivation. Additionally, the growth characteristics of *P. pastoris* was elucidated by surveying the hexose transporters present in the genome. To visualize the genomic data, a Genome Browser was implemented and made publicly available [1].

Based on the genomic and functional data of *P. pastoris*, as well as experimental data from chemostat and fed-batch cultivations, the first genome-scale metabolic model of *P. pastoris* was created. In a validation study, the model was used to simulate the production of two heterologous proteins, namely the secreted human serum albumin (HSA) and the intracellular human superoxide dismutase (hSOD), under oxygen limiting conditions. Furthermore, the model is capable of predicting overexpression and knockout targets that have a beneficial effect on protein production. With a genome model at hand and established high throughput tools *P. pastoris* is now ready for Systems Biotechnology.

[1] `www.pichiagenome.org`

Zusammenfassung

Die methylotrophe Hefe *Pichia pastoris* ist ein bedeutendes Expressionssystem, welches vorrangig zur Produktion von rekombinanten Proteinen verwendet wird. Trotz der ständig wachsenden Beliebtheit dieser Zellfabrik, ist relativ wenig über ihr Genom bekannt, und zum Beginn dieser Arbeit, war keine öffentlich zugängliche Genomsequenz verfügbar. Ohne diese, ist jedoch die Anwendung von High-Throughput Methoden stark eingeschränkt, und eine systemweite Analyse nicht möglich. Nachdem das Ziel dieser Arbeit die Erstellung eines genomweiten metabolischen Modells von *P. pastoris* war, mussten zuerst die erforderlichen Methoden etabliert, und die für die Modellkonstruktion essentiellen Daten erzeugt werden.

Im Rahmen einer Weiterentwicklung von vorhandenen DNA Microarrays, wurde das Arraydesign überarbeitet, sowie eine Pipeline für die statistischen Auswertung der Daten implementiert. Das System konnte in einer Studie, die sich mit den Auswirkungen von Dithiothreitol (DTT) und von HAC1 Überexpression auf *P. pastoris* beschäftigte, erfolgreich validiert werden. Sowohl DTT wie auch HAC1 beeinflussen die 'Unfolded Protein Response', welche eine wichtige Rolle in der Proteinsekretion spielt. Die Arrays wurden daraufin in einer internationalen Studie, über die Auswirkung von Umweltfaktoren auf die Produktivität von verschiedenen Expressionssystemen, verwendet.

Der *P. pastoris* Stamm DSMZ 70382 wurde auf zwei 'Next-Generation-Sequencing' Platformen sequenziert, und danach funktionell annotiert. Da sowohl das Sekretom als auch die Wachstumseigenschaften wichtige Faktoren in der industriellen Biotechnologie sind, wurden die gewonnenen Genomdaten dazu verwendet um, einerseits das Sekretom von *P. pastoris* bioinformatisch und experimentell zu bestimmt, und andererseits die Hexosetransporter zu untersuchen. Die Resultate dieser Analyse ergaben neue und vielversprechende Erkenntnisse ber die Eigenschaften dieser Hefe. Um das Genom zu visualisieren und öffentlich zugänglich zu machen, wurde ein Genom Browser implementiert[2].

Basierend auf die funktionelle Annotation des Genoms, sowie experimentelle Daten, wurde das erste genomweite metabolischen Modell von *P. pastoris* entwickelt, mit dessen Hilfe die Produktion von zwei rekombinanten Proteinen (Humane Serum Albumin (HSA) und Humane Superoxid Dismutase (hSOD)) unter verschiedenen Sauerstoffaufnahmeraten, erfolgreich simuliert werden konnte. Durch die Implementierung der Proteinsynthese kann das Modell zur Vorhersage von Zielgenen für Knock-out oder Überexpressionsexperimenten verwendet werden. Mit der Entwicklung des genomweiten Modells, der Etablierung von Microarrays, und einer verfügbaren Genomsequenz, ist *P. pastoris* nun gut gerüstet für die Systembiotechnologie.

[2] `www.pichiagenome.org`

Acknowledgements

I would like to thank my supervisor Diethard Mattanovich for his support and trust during my thesis, and I want to thank him even more for his positive attitude under all circumstances and his innovative visions for the future of Biotechnology.

I want to thank all people who had a part in this work, especially Martin Dragosits, Gitti Gasser, and Kristin Baumann. Special thanks also goes to Andreas Redl for his great support and for reminding me that one can do good work and still relax and party. A big Thank You goes to the bioinformatics group (present and past) of Dr. Kreil for fruitful discussions, good advice and continuing friendship. I also would like to thank Brian for taking the time to read over this piece of work.

Last but not least, I want to thank my family for always being there for me, for making me laugh no matter how bad all seemed to be, in short - simply for being who they are. Special thanks belongs to Philipp, who did not complain about all the evenings and weekends I spend working. This is for you.

Contents

List of Figures

List of Tables

Aim of the study

The aim of the study was to construct a genome-scale metabolic model of the yeast *Pichia pastoris* and establish transcriptomics methods to facilitate a Systems Biotechnology approach in the analysis of this organism. A basic requirement for both objectives was the availability of a genome sequence of *P. pastoris* and its functional annotation. To be able to curate the annotation and gene prediction and to make the genome accessible to the community, it was necessary to implement a genome browser visualizing the genomic data. The transcriptome analysis of *P. pastoris* required the development of DNA expression microarrays and an analysis pipeline to calculate differential expression and higher level statistics.

1 Introduction

The methylotrophic yeast *Pichia pastoris* has been successfully used for heterologous protein production for several decades, and constantly gains popularity. For all its advantages, this production platform still has shortcomings, especially when trying to secrete complex proteins [Macauley-Patrick et al., 2005]. Several studies have targeted isolated processes involved in protein biosynthesis and secretion, such as the unfolded protein response or a number of helper factors [Inan et al., 2006; Marx et al., 2006; Zhang et al., 2006; Resina et al., 2009; Gasser et al., 2006, 2007a]. Although these studies could achieve a beneficial effect on the expression of recombinant proteins, they also showed the need to study recombinant protein production in the framework of the complete cellular system. This insight, which is reflected in the concept of Systems Biology, is not a new one. Yet, a system-level approach became feasible only in the last decade, after the development of new analysis tools and high throughput technologies [Graf et al., 2009]. In general, Systems Biology can be understood as an approach that tries to consider all components of a biological system, and attempts to represent and predict interactions between the components using mathematical-computational models. This approach also attracts attention from the biotechnological community, where the term Systems Biotechnology was recently introduced. Current applications of Systems Biotechnology in recombinant protein production are reviewed in Graf et al. [2009]. One aspect of Systems Biology, metabolic modelling, shifts more and more into focus as biology moves from a qualitative to a quantitative science. The ability to predict the reaction of a system to perturbation without doing laborious experiments represents a great advantage in any field, yet the complexity of biological systems constitutes a special challenge. Many gaps still exist in our knowledge about the function and regulation of cellular processes, but with the progress in genome sequencing, automated annotation, and high throughput analysis whole genome models of microorganisms are now feasible.

1.1 Yeasts and Pichia pastoris

Within the kingdom of fungi, the methylotrophic yeast *P. pastoris* belongs to the phylum of *Ascomycota* and the family of *Saccharomycotina*. After phylogenetic sequence analysis became available, it has been reclassified from the genus *Pichia* into the genus *Komagataella* and split into three species, namely *K. pastoris*, *K. pseudopastoris* and *K. phaffii* [Yamada et al., 1995]. Biotechnologically used strains belong either to the species *K. phaffi* or *K. pastoris* [Kurtzman, 2009], but the name *P. pastoris* is still widely used for both of these species.

Yeasts do not represent a homogeneous group that has a clear lineage. Instead, they constitute a group of unicellular organisms that developed several times during evolution. They are characterized by an unlimited clonal growth and sexual states that do not form within or on fruiting bodies [Dujon, 2010]. Yeast genomes are usually compact with a high gene density and a small number of intron-containing genes [Dujon et al., 2004]. The exact position of the *Komagataella* species in the phylogenetic tree is still unclear, but

according to Dujon [2010] they belong, together with *Yarrowia* and *Trichomonascus*, in a group called *Dipodascaceae*, which is characterized by 4-6 chromosomes and dispersed 5S RNA genes.

P. pastoris has several advantages as production host of recombinant proteins. Even though the diversity within the group *Saccharomycotina* is large, the similarity between *P. pastoris* and *Saccharomyces cerevisiae* is high enough to enable sharing of protocols and techniques for genetic manipulation. The growth conditions of *P. pastoris* are relatively simple, it has a high specific growth rate, and it can grow to a higher cell density than *S. cerevisiae* due to its preference of respirative growth. Moreover, *P. pastoris* has the ability to generate human-like post translational modifications (disulfide bridges and proteolytic processing), it produces N- as well as O-glycosylated proteins, and it was engineered to show a human-like glycosylation pattern, which is essential for the expression of many proteins that are destined for medical purposes [Wildt and Gerngross, 2005; Jacobs et al., 2009; De Pourcq et al., 2010]. Gasser and Mattanovich [2007] review the production of antibodies in yeasts and filamentous fungi and highlight the potential of fungal production systems that can be glycoengineered. Another advantage of *P. pastoris* is the secretion of very low levels of endogenous proteins, making downstream purification of heterologous protein easier [Cereghino and Cregg, 2000; Gellissen et al., 2005; Mattanovich et al., 2009a]. Therefore, it is not surprising that this methylotrophic yeast is widely used as production host for recombinant proteins. Lists of proteins successfully produced with *P. pastoris* can be found in Cregg et al. [2000] and Macauley-Patrick et al. [2005]. The remaining bottlenecks in the production of complex recombinant proteins were suspected to be in the folding machinery of the cell, and linked to the unfolded protein response and other stress related factors [Mattanovich et al., 2004; Hohenblum et al., 2004; Gasser et al., 2006]. A comprehensive analysis of the reaction of the whole cell to recombinant protein production was, however, not possible. Attempts to analyse the transcriptome of *P. pastoris* under protein producing conditions include, the use of heterologous *S. cerevisiae* microarrays [Gasser et al., 2007a; Sauer et al., 2004], or transcript analysis with aid of affinity capture (TRAC) [Gasser et al., 2007b]. While important lessons were learned in these studies, they were not able to capture the complete transcriptional response of the cell.

1.2 Next Generation Sequencing

When F. Sanger presented his DNA sequencing method in the 70s, it opened the door to a new understanding of life. In the 90s, not least because of the human genome project, the technique was improved to be faster and to produce longer reads. Automated sequencing machines were introduced, making large-scale production possible [Hall, 2007]. Still, the infrastructure costs for Sanger sequencing are too high for most small labs, and sequencing is therefore done primarily in large sequencing centres. To meet the challenges of the human genome project, it was essential to consider other sequencing approaches that could eliminate the time consuming step of cloning, and therewith increase the sequencing rate [Medini et al., 2008]. In the last few years, several new sequencing

Advantages	Disadvantages
Massively parallel throughput	Shorter read length
No time consuming cloning, no cloning bias	More complex data processing
Low volumes of DNA and reagents	Higher error rate, more coverage needed

Table 1: Advantages and Disadvantages of NGS

methods were brought to the market. These so-called Next Generation Sequencing (NGS) platforms reduced the reaction volume, while at the same time extending the number of sequencing reactions, making them faster and cheaper than Sanger sequencing [Schuster, 2008].

NGS methods do not use cloning to amplify the DNA fragments, but some form of Polymerase Chain Reaction (PCR), thus avoiding the bias introduced by cloning. On the other hand they do not reach the read length of the traditional approach and come with their own set of problems, not least the introduction of a new level of complexity into assembly and alignment. Apart from these features, the new techniques differ considerably, and depending on the biological question, one or another is more suited to find the answer. The advantages and disadvantages that are common to all new sequencing techniques are summarized in table 1 and figure 2 on page 11 shows which features of the new sequencing technologies are important for which type of application. Since the technical details of the NGS methods have been exhaustively reviewed [Mardis, 2008; MacLean et al., 2009; Metzker, 2010], they will not be repeated here. Many of the disadvantages of current sequencing methods could be solved by single molecule sequencing, a technology that determines the base composition of an individual DNA strand due to differences in physical, optical, electrical or magnetic properties of the four bases [Xu et al., 2009]. The first single molecule sequencing technology on the market was developed by Helicos BioSciences, and uses cyclic steps of labled nucleotide addition and fluorescence detection without the need of an amplification step[3]. With the omission of the amplification step, the error rate is lower than for NGS methods, and although the present read length of 35 bases places it at the lower end of the range, it is still comparable to the Illumina/Solexa approach. Another single molecule platform was introduced recently by Pacific Biosciences. It uses phospholinked nucleotides to measure incorperation during natural DNA synthesis in real time (SMRT technology)[4]. The read length promised by this platform lies above 1000 bases, but at the moment there are no independent studies concerning the performance of the SMRT. While a number of other commercial single molecule sequencing system are announced to be on the market within the next five years, there are still major issues to be tackled. Current single molecule sequencing methods are critically reviewed in Gupta [2008] and Xu et al. [2009]. The features of currently used sequencing methods are listed in figure 1.

[3] http://www.helicosbio.com(September2010)

[4] http://www.pacificbioscience.com(September2010)

	SOLiD	454 GS FLX	Pacific Biosciences	Solexa	Heliscope
Paired reads	✓	✓		✓	
Accurate reads	✓		✓		
Read length		✓	✓		
High number of reads	✓		✓	✓	✓
Sequencing single molecules			✓		✓
Read Length (bases)	50-75	400	>1,000	100	35
Primary error type	Substitutions	Homopolymer Indels	Not yet determined	Substitutions	Deletions

Figure 1: Features of currently used sequencing methods

1.2.1 Questions related to De Novo Assembly

In *de novo* assembly, the sequence reads are pieced together without any kind of additional information. This enables researchers to work with organisms for which no sequence data is available yet, but makes the analysis computationally more demanding. The main application for *de novo* sequencing is whole genome sequencing, but it is also attractive for other questions where a reference genome is not available and not needed. After the sequencing run, the reads are assembled into contigs, which are longer units of continuous DNA sequence, separated by gaps where assembly was not possible. In most cases, the gaps result from repetitive elements in the genome sequence [Griffiths et al., 2005]. Hence, the number of contigs and their length greatly depends on the complexity of the sequenced genome. To further assemble the contigs into scaffolds or supercontigs, it is often necessary to know the orientation and order of the contigs. This information can be obtained by using paired-end reads, a method where a genomic insert or DNA fragment of defined size is sequenced from opposite ends. Because of the cloning step in Sanger sequencing, paired-end data was readily available. With NGS a paired end methodology had to be specifically developed, and was therefore lacking at the beginning [Pop and Salzberg, 2008]. The remaining assembly gaps can be filled using PCR, primer walking, or Sanger sequencing. For this type of sequencing problem, longer reads are better suited as they span larger repeat regions and make the assembly easier to manage. At present, the 454 technique can reach a read length of 400 (454 GS FLX with Titanium reagents), and is therefore the best choice for *de novo* sequencing. Nevertheless, also the short read techniques like Illumina or SOLiD can be used for *de novo* sequencing, especially when the genome is small. Farrer et al. [2009] explored *de novo* sequence assembly of *Pseudomonas syringae*, using paired-end reads from Illumina's Genome Analyzer and a number of assembly programs. The result is promising even though they could not take full advantage of the paired-end data since the algorithms at that time were not yet optimized for this type of data. With a read length of over 1,000 bases, the SMRT technology of Pacific Bioscience is comparable to Sanger sequencing and would be especially well suited for *de novo* sequencing, but the strengths and weaknesses of this technology are not known yet.

	De-novo assembly	Gene expression	SNP detection	Protein binding sites	Re-sequencing	Non-coding RNAs	Sequence alignment
Paired reads	✓				✓		
Accurate reads	✓	✓	✓	✓	✓	✓	✓
Read length	✓						
High number of reads	✓	✓	✓	✓	✓	✓	✓
Sequencing single molecules					✓		

Figure 2: Requirements of Sequencing Applications [MacLean et al., 2009]

1.2.2 Questions related to Sequence Alignment

Whenever a genome sequence already exists, like for most model organisms, the NGS reads can be aligned to the reference genome. Sequence alignment is a much simpler task than *de novo* assembly since it is only necessary to find the matching position on the reference genome, but still poses certain challenges. Repetitive sequences and highly similar regions are very hard to align properly, especially when the sequence reads are short. It is not always easy to distinguish between sequencing errors and polymorphisms, and diverging sequences can result in wrong alignments. Apart from a well annotated high quality reference genome, read accuracy and sequencing depth (coverage) are the critical factors for most alignment problems (figure 2). Resequencing of strains or individuals can be used for mutational profiling, and is therefore of special interest for microbiological strain improvement, as well as the basis for personalized medicine. This application has similar requirements as *de novo* sequencing, but in combination with the analysis of single nucleotide polymorphisms (SNP), a large sequencing depth and good read accuracy is essential [MacLean et al., 2009]. Using sequencing technologies to analyse messenger RNA content (RNA-seq) has several advantages over expression profiling on microarrays. The analysis is not limited to known genes, and the precise location of transcript boundaries can be mapped to the genome. Technically, background signals are very low or nonexistent, and the results are highly reproducible and accurate when compared to qPCR. On the other hand, each sequencing method introduces a different bias, and splicing and splicing variants still present a problem [Wang et al., 2009]. Short read as well as long read platforms can be used for RNA-seq, but due to its lower per base cost the Illumina/Solexa platform has been used in the majority of studies up to date. NGS can also be used for the discovery of non-coding RNAs, a field that has made a great impact on genome research [Mardis, 2008]. Another application is the *in-vivo* examination of protein-DNA interaction using chromatin immunoprecipitation (ChIP) essays. Like with gene expression, the identification of resulting DNA fragments was done through hybridization to a microarray (ChIP-chip), but the comparative advantages of sequencing led to the establishment of a ChIP-seq protocol [Park, 2009]. The applications of sequencing will surely continue to increase in the future while the technology matures, and though cost is presently the main disadvantage of all sequencing

platforms, the per base cost of sequencing has been falling continually during the last few years and will probably keep on doing so in the near future.

1.2.3 Bioinformatics Challenges

Although bioinformatics software and algorithms were already available for the assembly and alignment of Sanger sequencing output, they could not be directly applied to the data generated by next generation sequencing machines. The new sequencing technologies have a higher coverage, which increases the data output drastically, and most of the formerly used algorithms did not scale well with the increasing data volume. Also, due to the shorter reads, the sequence overlaps are not as long as in Sanger sequencing, making the assembly problem more tricky. Lastly, because the sequencing methods differ, the data behave differently and the error models need to be changed for each new technology [MacLean et al., 2009; Pop and Salzberg, 2008]. The development of new assembly and alignment tools progresses fast, and already a whole range of options are available. The programs are updated continuously to improve the performance of the algorithms and to incorporate new features of NGS, like paired-end reads. It is hard to say what the advantages and disadvantages of the various tools are since they are all fairly new and quite sensitive to parameter values and the quality of the input data [Farrer et al., 2009]. MacLean et al. [2009] gives a good introduction into assembly and alignment tools available at the moment.

1.3 Gene Expression Microarray Analysis

Despite the promises of RNA-seq, many labs still rely on microarrays when measuring transcriptomics data. The main reason for this is an existing infrastructure and the cost, which is still lower than sequencing, especially when using high density chips. It is likely that expression microarrays will be replaced by sequencing in the future and the use of microarrays will shift to new applications [Graf et al., 2009], but for the time being they still have their place in transcriptomics. Generally, a gene expression microarray consists of nucleic acid sequences (probes) attached to a solid support like a glass slide. Depending on the type of microarray, the length of the probes vary, and so does the method, with which they are created and fixed on the support. A fluorescently labelled target sample is then hybridized to the microarray, and the fluorescence signal is scanned and transformed to an intensity value. Two-channel arrays are hybridized with two samples, each being labelled with a different fluorescent dye, usually Cy3 and Cy5 [Jaluria et al., 2007]. More details about microarrays can be found in Rogers and Cambrosio [2007], who review the development and the application of microarrays since their introduction, and Tarca et al. [2006], who give a concise introduction to the topic of gene expression profiling with microarrays.

There are several factors, which contribute to the fact that the task of analysing microarray data is not a trivial one:

- The measured fluorescence signal depends on many variables not related to the biological question, causing a certain amount of variation and possibly systematic error [Huber et al., 2005].
- Usually, many thousand genes are tested with only a very limited number of replicates.
- Microarray fluorescence data do not satisfy the assumptions of many classical statistical tests, like variance homogeneity, or normal distribution [Dabney and Storey, 2007].
- There is a nonlinear relationship between fluorescence measurement and abundance of mRNA transcripts in the sample [Huber et al., 2005].

The types of variation are classified into random variations (noise), and systematic deviations from the true signal (bias). Both can be introduced at various steps in the microarray analysis [Kreil and Russell, 2005; Mecham et al., 2010]. The non exhaustive list below describes some of the steps in the analysis that could lead to noise or bias in the data:

Dye bias: There can be variations in dye incorporation into the target sequence; when using two-colour arrays it is known that the two dyes (usually Cy3 and Cy5) show a different behaviour, especially in regard to stability in the presence of ozone [Fare et al., 2003].

Cross-hybridization: Depending on the specificity of the designed oligos, sequences may bind to unintended probes. The formation of secondary structure and fragmentation of the sample RNA are particularly problematic in this regard. Well designed oligos are a prerequisite to avoid too much cross-hybridization [Leparc et al., 2009].

Further causes of variation can include differences in the amount of reactants between samples, spatial variations in hybridization efficiency, and artefacts from contamination or washing steps. Also, variation in the production of the array, such as differences between print tips, can influence the fluorescence signal. Today, most of these technical factors can be controlled by sound oligo design, microarray production, hybridization, and normalization methods. However, purely biological signal variations exist, which reflect fluctuations between individuals or colonies, or random fluctuations within one individual or colony. Gene expression is very sensitive, reacting to rather minor changes in the environment of the organism, and leading to significantly biased results. The only solution to this dilemma is a sufficient amount of randomized biological replicates [Breitling, 2006]. How many replicates are sufficient is in practice mainly a function of the available resources and therefore often rather limited.

Microarray analysis starts with image analysis, where fluorescence signals are scanned and the pixel intensities are converted into probe-level data. In the second step, the quality of the arrays is assessed using key values like signal to noise ratio and data visualization. This step is very important in order to exclude bad arrays from the analysis or to flag outliers. Lower level analysis consists of background adjustment and normalization [Gentleman et al., 2005]. It is assumed that the intensity of a spot is comprised

of the fluorescence signal of the target (foreground signal) and non-specific fluorescence from other sources (background signal). To remove this bias, it has been common practice to measure and remove the background signal [Smyth et al., 2003]. However, this approach has also disadvantages. Subtracting the background will increase the variance of the expression values and distort the biological signal [Scharpf et al., 2007]. Additionally, it introduces processing problems, if the signal value falls below zero. The ultimate goal in microarray processing, and the main reason why raw array data is normalized, is to preserve the biological signal, while removing technical bias and random fluctuation. There exist a multitude of methods for microarray normalization, and the decision concerning which to apply depends on the biological question, and the sources of bias or noise in the data. A good summary of existing normalization procedures can be found in Steinhoff and Vingron [2006]. Many of the most commonly used microarray normalization methods are unsupervised methods, which means they do not utilize additional information about the study, but work with descriptive properties of the expression values. Some supervised procedures have also been developed, and apparently show better results than unsupervised methods. However, the results strongly depend on the knowledge of the factors influencing the study and the quality of the model that is created using these factors [Mecham et al., 2010]. Therefore, they depend on the close interaction between experimentalist and data analyst or an excellent documentation about the various factors influencing the study. The last step in the basic analysis of microarrays is the calculation of differentially expressed genes. The first indication of differential expression is the fold change between the two samples, though it is statistically not meaningful without the addition of a significance value which shows how reliable the fold changes are [Miron and Nadon, 2006]. Therefore, it is advisable to use a method that calculates some form of significance value like a t- or F-statistic. Gene expression data consists of many thousand genes that need to be tested with just a few replicates. Additionally, groups of co-expressed genes lead to correlation in the data. Under such conditions, not controlling the type I error rate would result in a large amount of false positives. While stringent multiple testing corrections work well to reduce the type I error, they also reduce the power of the test, increasing the false negative rate. Depending on the focus of the study, both can be of importance. For biotechnological questions, the less conservative approach of controlling the false discovery rate (FDR) is usually a better choice [Reiner et al., 2003]. A comparison of commonly used statistics to generate differentially expressed gene lists can be found in Jeffery et al. [2006].

Interpreting long lists of genes is not an easy task and often leads to a confusing and fragmented picture of the biological system. The introduction of a threshold value makes the gene lists shorter, though this arbitrarily chosen cutoff does not reflect the fact that genes act in functional groups and are not independent of each other. Hence, higher level analysis is applied to reduce the complexity of the data, and detect underlying patterns. The techniques for higher level analysis can be separated into supervised methods, which use information about a pre-existing classification, and unsupervised methods, for which only the expression data itself is needed. Class prediction studies, like for example finding a predictor to distinguish between healthy and sick patients,

primarily work with the first approach. Class discovery studies, like finding coexpressed genes, use the second approach [Raychaudhuri et al., 2001; Tarca et al., 2006]. The most common unsupervised method in microarray analysis is clustering. In clustering, genes with a similar expression profile are grouped together based on a distance metric (e. g. euclidean distance) [Chipman and Tibshirani, 2006]. It is important to use caution when interpreting clustering results, since there is no guarantee that the emerging pattern represents biologically relevant information. On the other hand, the data separation can be a good basis for the detection of overrepresented regulatory motifs. Another group of higher level analysis can be summarized as functional profiling, and deals with the enrichment of certain biological processes or pathways in a previously selected set of genes. The classical approach consists of a gene selection step, and a step, in which the enrichment is tested against a background to determine significance (e. g. Fisher's exact test). In the majority of cases, the gene selection is based on differential expression. Since this method is also based on an artificially selected gene list, it has the drawback that it misses functional relations in the rest of the genes below the cutoff. Gene set analysis (GSA) circumvents this disadvantage by using the whole gene set, and testing if certain annotation modules are grouped at the extreme of a ranked gene list [Dopazo, 2009]. The module annotation used in most studies is either Gene Ontology (GO) [Ashburner et al., 2000] or Kyoto Encyclopaedia of Genes and Genomes (KEGG) [Kanehisa and Goto, 2000], but can in principle be freely chosen.

Transcriptomics data are very useful but, when considered in isolation, they can not give a comprehensive picture of a system's status. Many of the functions in a cell are predominantly regulated through other processes, like post translational protein modification, or signalling cascades. The measurement of proteins, metabolites, and fluxes are necessary to understand the system in its totality [Graf et al., 2009]. If a metabolic network for the system in question or a closely related one exists, the measurements can be placed into an even better physiological context. Biocyc represents a tool to map transcriptomics or proteomics data on an existing metabolic network, facilitating a pathway oriented interpretation of the measurements [Caspi et al., 2007].

1.4 Metabolic Models

For many years, modelling and simulation techniques have been successfully used in various, but mainly technical fields. Typical applications are, the improvement of the performance of a system, and the prediction of its behaviour under specific conditions. Biotechnology pursues the same objectives with relation to biological systems, but the utilization of mathematical modelling in the biological sciences was hampered by technological shortcomings, and our lack of system-wide knowledge. Therefore, earlier models were solely based on biochemical characterization of enzymes, and available literature, limiting the models in scope, and confining them to well known pathways [Kim et al., 2008]. The development of high-throughput technologies enabled a system-wide analysis of genes, proteins and metabolites. Subsequently, whole genome models became the emphasis of the emerging field of Systems Biology. Today genome-scale models exist

for several organisms, of which the models for *S. cerevisiae* and *E. coli* are the most comprehensive [Nookaew et al., 2008; Feist et al., 2007; Thiele et al., 2009].

The first step towards a novel metabolic model is the identification of all the molecular constituents and their interactions. This information can be inferred from the genome sequence based on its functional annotation. After a genome sequence becomes available, coding sequences can be identified by gene-finding algorithms, and the putative genes can be annotated through sequence homology. The annotation relies heavily on comparison with information that is available in sequence, domain, and pathway databases. Lee et al. [2005] and Feist et al. [2009] list a good selection of databases and software tools, which are useful for model construction in microorganisms. There are some uncertainties involved in this approach, and it is beneficial to manually curate the functional annotation of the open reading frames (ORFs) [Feist et al., 2009]. Compartmentalization is an important feature of the cell. It creates distinct environments, with conditions adapted to the processes that take place inside such an organell. The functionality of proteins is therefore closely related to their subcellular context. Protein localization studies [Huh et al., 2003; Mattanovich et al., 2009a] can be used to determine possible interactions between proteins and to place reactions in the right context, as well as to show inconsistencies in the annotation. The biggest obstacle when building an *in silico* model capable of predictions, is the parameterization. Many of the required parameters can not be measured with state of the art technologies [Lee et al., 2005; Barrett et al., 2006]. Kinetic models have good predictive power, but are based on explicit enzyme-kinetic rate equations. They also rely on good quantitative *in-vivo* measurements, which are still scarce. Additionally, the computational complexity increases fast with the size of the model, making this approach unfeasible for larger models [Steuer, 2007]. To circumvent this drawback and still retain some predictive power, genome-scale models work with a stoichiometric matrix, and use assumptions and constraints to reduce the solution space. The stoichiometries for the matrix can be derived from the genome annotation, through the gene to protein to reaction relationship. Resources like the Enzyme Commission (EC) numbers classification[5], are an invaluable tool for the translation of function into biochemical reactions. Functions and pathways unique to the organism can only be elucidated through extensive experimental work, but indications as to the existence of such pathways can be drawn from gaps and inconsistencies in the model. When the basic model is in place, it needs to be checked for 'dead ends' or 'metabolic gaps', and validated with experimental data. Available omics data can be integrated to confirm concentrations or interactions. ^{13}C based flux measurements are especially valuable in this context. Fluxes represent time dependent movements of metabolites through the cell. Experimentally, they can not be measured directly, but have to be estimated using computer models. The models that are used in this approach, concentrate on the central carbon metabolism, and often exclude pathways due to interpretability issues [Sauer, 2006]. Usually, a ^{13}C labelled substrate is fed to the cell, and the enrichment of ^{13}C is measured with nuclear magnetic resonance (NMR) or mass spectrometry (MS). Several studies explore metabolic fluxes in *P. pastoris* using the ^{13}C method [Solà et al.,

[5] `http://www.chem.qmul.ac.uk/iubmb/enzyme/`

2004, 2007; Carnicer et al., 2009; Heyland et al., 2010]. Whole genome networks greatly benefit from the integration of experimentally validated flux rates, even if they can not show the underlying reactions that led to the metabolic system response.

The prediction of growth phenotypes is one of the major uses of genome-scale metabolic models. It is realized by constraints-based flux balance analysis (FBA), which is based on a steady-state assumption around each metabolite, and an objective function that defines the purpose of the system [Kim et al., 2008]. The optimization of growth rate is an appropriate objective function for cells grown under energy limited conditions, but it has also been shown that, under other conditions, other objective functions work as well [Nielsen, 2007]. The fluxes through the stoichiometric model, under the selected constraints, are then simulated by linear programming algorithms. With the addition of a layer of Gene Protein Reaction (GPR), the effect of gene knockout or overexpression on the phenotype can be explored. Algorithms, like minimization of metabolic adjustment (MoMA) and regulatory on/off minimization (ROOM), assume that the metabolism in the mutant is as similar as possible to the wild-type. This approach has been shown to improve predictions in comparison to the optimal growth assumption used by FBA [Durot et al., 2009].

Regulatory networks can not be represented in the same way as metabolic networks due to their non-linear multivariate relationships. One approach to model regulatory behaviour nonetheless, is to use control logic functions [Price and Shmulevich, 2007]. The algorithms to implement the control logic functions range from Boolean formalism, wherein each component has an on/off state and the relation between the components is described by a Boolean function, to Bayesian networks, which are based on probabilistic graphical models [Schlitt and Brazma, 2005; Price and Shmulevich, 2007]. At the moment, great effort is being made to implement transcriptional and translational processes into the genome-scale model, but since these processes are different for each expressed protein, and because protein biosynthesis is highly interconnected with metabolic as well as regulatory processes, the implementation of transcription and translation into a whole genome model is a challenging task. At the present time, the only model that incorporates protein biosynthesis is the model of *E. coli* [Thiele et al., 2009].

2 Materials and Methods

This section only deals with materials and methods used in the microarray analysis pipeline since it is unpublished work. For the other topics, materials and methods are described in the respective publication.

The *R* programming language[6], provides a free open source software environment for statistical computing and graphics. *R* is based on the commercial programming language *S* but its source code is freely available under the GNU General Public License [Dudoit et al., 2003].

Bioconductor is a development software project for the analysis and comprehension of genomic data, using the computing environment *R*. The broad goals of the project are to provide access to a wide range of statistical and graphical methods for the analysis of genomic data, the integration of biological meta data and the fast development of software [Gentleman et al., 2004].

The microarray analysis pipeline was written as an in-house solution for Agilent 2-colour expression microarrays but can be adapted to other platforms by changing the input function. It consists of a set of R-scripts which can be run by calling a wrapper function with a targets file list and the experiment name as parameters. Additionally, it must be specified if the experiment was a reference or a direct design.

The following packages were used in the microarray analysis pipeline:

limma: The package is designed for differential expression analysis of microarray data. It deals with complex experimental setups, contains a variety of pre-processing and normalization methods and uses linear models and empirical Bayes methods to calculate differentially expressed genes [Smyth, 2004].

marray: Class definitions and associated methods to process pre- and post-normalization intensity data from microarrays.

vsn: A package to pre-process microarray intensity data. It calibrates the data and performs variance stabilization [Huber et al., 2002]. The package compatible with the *limma* framework.

gsa: Performs gene set analysis, using pre-defined gene groups and a gene expression value.

[6] `http://www.r-project.org/`

3 Results

To be able to use a Systems Biology approach in the work with *Pichia pastoris*, it was essential to develop, and establish the necessary omics tools. Unfortunately, in the beginning of this work, no public genome sequence of *P. pastoris* was available, and the NCBI contained only 173 nucleotide and 245 amino acid sequences. A commercial draft sequence of *P. pastoris* provided by Integrated Genomics[7] (IG), which consisted of 474 contigs and 5,425 putative genes, could be obtained. For the development of the first whole-genome expression microarrays, the IG contigs were used to predict additional open reading frames (ORFs), and design oligos for the microarray platform Agilent [Graf, 2007]. Through this process, it became clear that a publicly available *P. pastoris* sequence was needed to be able to fully explore the possibilities the omics methods could offer. Therefore, the sequencing project of *P. pastoris* strain DSMZ 70382 was started.

The sections one to three of this chaper shortly describe the main results for the different topics of this work. The last section describes my contribution to those parts of the project that are still unpublished. For all published work my contribution is mentioned before listing the respective papers (see page 42, 57 and 82).

3.1 Pichia pastoris DSMZ 70382 Genome Sequence

Sequencing and assembly of *P. pastoris* strain DSMZ 70382 was conducted by GATC Biotech AG[8]. The selected platform was the new 454 system (Roche GS FLX-Titanium Series) with a read length of 400 bases. To examine the sequencing quality, all *P. pastoris* genes available from the NCBI were blasted against the assembly. Unfortunately, it was found that about one third of the genes had frameshifts, mostly due to homopolymer repeats. This is a known problem of the 454 technique, but was reported to be manageable in literature [Margulies et al., 2005; Droege and Hill, 2008]. As it was found to be beneficial for *de novo* assemblies to combining different NGS techniques [Metzker, 2010], an additional Illumina paired-end run was used to improve the sequence quality. The draft genome consisted of 326 contigs, with the largest contig being 419 475 bases, and the smallest 128 bases. Using the paired-end data, the number of contigs could be reduced to 317, and 125 of these could be aggregated into 38 supercontigs. At the same time, the genome of the *P. pastoris* strain GS115 was sequenced by the group of Nico Callewaert in Ghent [De Schutter et al., 2009]. An alignment of the strain DSMZ 70382 and the strain GS115, visualized in figure 3, showed a chromosome rearrangement, with parts of chromosome III being transferred to chromosome I, and parts of chromosome IV being transferred to chromosome II (unpublished data). This confirms the chromosome polymorphism that was published by Ohi et al. [1998] between the *P. pastoris* strain DSMZ 70382[9] and the strain GS115. With a genome size of 9.4 Mb, *P. pastoris* belongs

[7] `http://www.integratedgenomics.com`

[8] `http://www.gatc.com`

[9] In the work of Ohi et al. [1998] the *Komagataella pastoris* type strain DSMZ 70382 is named ATCC 28485 and corresponds to the NRRL Y-1603 strain.

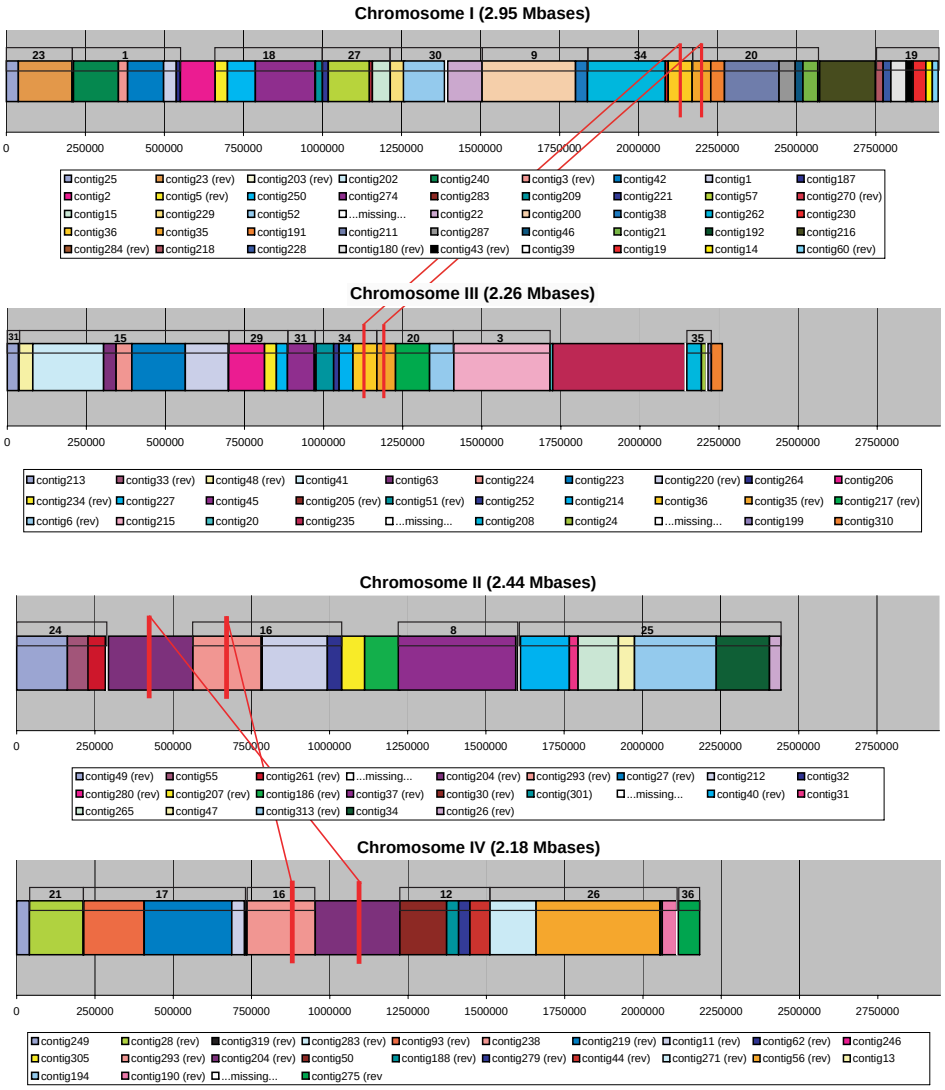


Figure 3: Mapping of the contigs of *P. pastoris* strain DSMZ 70382 to the chromosomes of strain GS115. The boxes indicate supercontigs (scaffolds) and the red lines chromosome rearrangements between the two strains

to the yeast organisms with smaller genome size. The GC content in the whole genome was 41.34% and in the coding regions it was slightly higher with 41.90%. In comparison, *S. cerevisiae* has an overall GC content of 38%[10]. An analysis of rDNA showed that the 35S rDNA cluster is present on each of the four chromosomes, and that the 5S rDNA is dispersed, which has already been noted in Dujon [2010] for the GS115 strain. The *P. pastoris* strain from IG and the strain GS115 belong to the species *K. phaffi* [Kurtzman, 2009] and the strain DSMZ 70382 is the type strain of the species *K. pastoris*.

[10] `http://www.ncbi.nlm.nih.gov/sites/entrez` (September 2010)

3.1.1 Functional Annotation

Gene finding in yeast is not as complex as in higher eukaryotes, but due to the compact genome, most gene finders overpredict the number of intron containing genes as well as the number of introns per gene. For this project, the prokaryotic gene finder Glimmer3 [Delcher et al., 2007] and the eukaryotic gene finder Augustus [Stanke et al., 2008] were used to predict open reading frames. The resulting gene set was then annotated, using a reciprocal protein BLAST [Altschul et al., 1990], and Interproscan [Zdobnov and Apweiler, 2001]. Protein families that share a high similarity can not be properly annotated with the reciprocal best hit strategy. Especially proteins involved in transport and gene regulation (transcription factors) were found to pose a problem. Francke et al. [2005] report a similar problem for the functional classes transport and signalling, as well as processes related to carbohydrate and amino acid conversion in bacteria. To improve the quality of *in-silico* gene prediction and annotation, the putative genes were manually curated. From the comparison to *S. cerevisiae*, it was clear that the *Komagataella* group was separated before the gene duplication event occured in the subphylum of *Saccharomycotina*. Nonetheless, the number of genes is comparable to *S. cerevisiae*, pointing to metabolic differences between the two species, of which one example is of course the methanol pathway that does not exist in the model yeast.

The amino acid distribution (figure 4) was compared between *P. pastoris* strain DSMZ 70382, strain GS115, *S. cerevisiae*, and *E. coli*. As expected, no differences can be seen between the *P. pastoris* strains and the distribution follows the same pattern as *S. cerevisiae*, but differs considerably from *E. coli*. Figure 5 on page 23 shows the codon usage of both *P. pastoris* strains, listing the total number the particular codon occurs in the gene set (N), the codon frequency per thousand (fpt), and the relative synonymous codon usage (RSCU), which describes the number of times a particular codon is observed relative to the number of times that the codon would be observed in the absence of any codon usage bias [Sharp et al., 1986]. It is clearly visible that the ftp and RSCU are quite similar between the two strains. They are, however, different from the data in the codon usage database on http://www.kazusa.or.jp/codon/[11]. The small number of coding sequences (137), from which the values in the database were calculated, could be an explanation for the deviance. The additional information gained by analysing the putative genes with regard to their codon content can, for example, help to identify dubious ORFs or be correlated to expression patterns.

For heterologous protein production, it is beneficial if an organism secretes only low amounts of endogenous protein into the medium. Proteases in particular are problematic because of their potential proteolytic activity on the product. Using motif and homology based prediction programs, 88 of the putative genes of the strain DSMZ 70382 were identified as secreted. The experimentally determined secretome of *P. pastoris* on glucose consisted of only 28 genes, none of which were proteases. This discrepancy is partly due to the specific growth conditions, which could not be taken into account in the *in silico* prediction. However, it was also evident that the programs have difficulties

[11]Last viewed September 2010

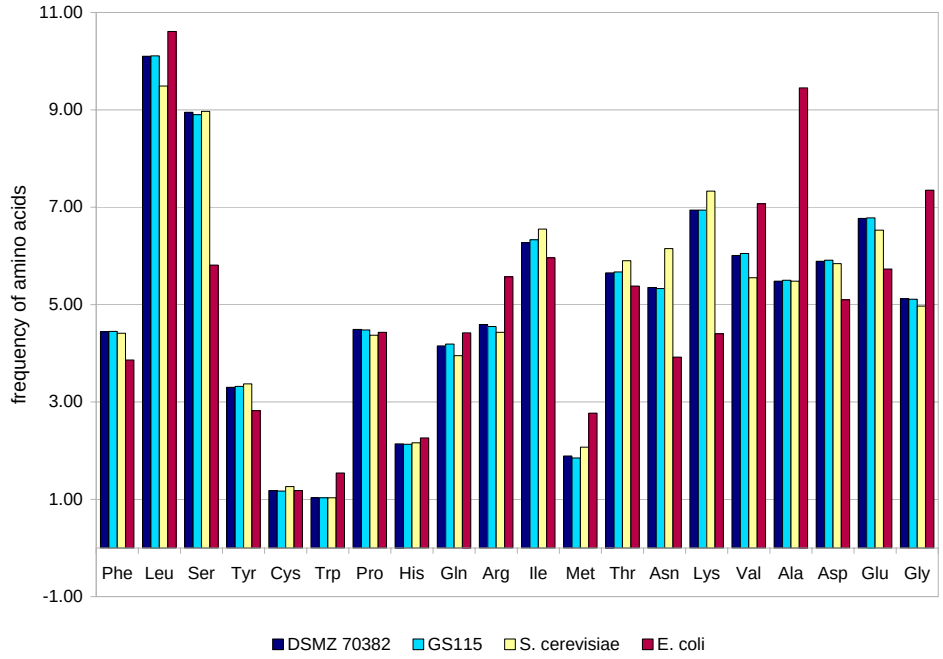
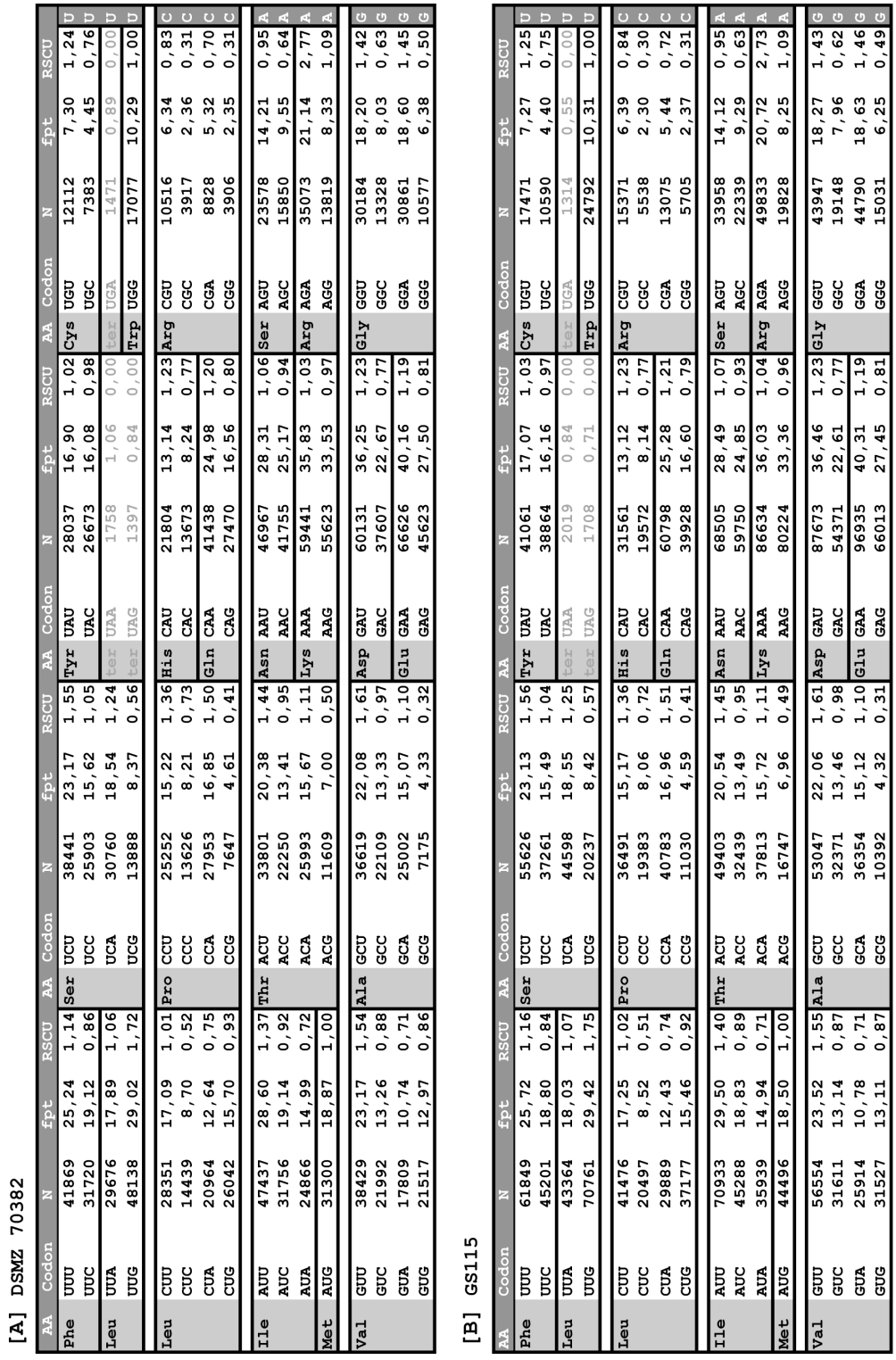


Figure 4: Average frequency of amino acids per protein for each *P. pastoris* strain (DSMZ 70382 and GS115), *S. cerevisiae*, and *E. coli*

distinguishing between secreted and ER/Golgi resident proteins, emphasising the need for experimental validation of *in silico* predictions. The comparatively low amount of naturally secreted proteins, and the absence of proteases in the culture media, highlight the advantages of the *P. pastoris* expression system in a glucose cultivation. *P. pastoris* is known to be a Crabtree-negative yeast, giving it the advantage of reaching high cell densities in cultivations. An investigation of sugar transporters confirmed that, in contrast with the 20 hexose transporters in *S. cerevisiae*, only 2 low affinity and 2 high affinity glucose transporters are present in *P. pastoris*. Interestingly, 4 genes exhibit similarity to glycerol transporters from *K. lactis* and *Y. lipolytica*. The large number of glycerol transporters result in a high glycerol uptake, and accordingly the specific growth rate of *P. pastoris* on glycerol is analogous to its growth rate on glucose.

3.1.2 Genome Browser

The genome browser for *P. pastoris* is based on the Generic Genome Browser (GBrowse) [Stein et al., 2002] and is available under `http://www.pichiagenome.org`. A postgres database was implemented, containing location information of all genes mapped on the respective contigs. The browser includes an overview of each contig, with zoom and search functions. Genomic, transcript and exon sequence as well as gene annotation information are available on a detail page for each gene. A WU-BLAST search and the domain search CDART as well as precalculated Interproscan pages were implemented [Redl, 2008; Mattanovich et al., 2009b]. The data of strain DSMZ 70382 as well as the strain GS115 are available in the genome browser.



[A] DSMZ 70382

AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	
Phe	UUU	41869	25,24	1,14	Ser	UCU	38441	23,17	1,55	Tyr	UAU	28037	16,90	1,02	Cys	UGU	12112	7,30	1,24	U
	UUC	31720	19,12	0,86		UCC	25903	15,62	1,05		UAC	26673	16,08	0,98		UGC	7383	4,45	0,76	U
Leu	UUA	29676	17,89	1,06		UCA	30760	18,54	1,24	ter	UAA	1758	1,06	0,00	ter	UGA	1471	0,89	0,00	U
	UUG	48138	29,02	1,72		UCG	13888	8,37	0,56	ter	UAG	1397	0,84	0,00	Trp	UGG	17077	10,29	1,00	U
Leu	CUU	28351	17,09	1,01	Pro	CCU	25252	15,22	1,36	His	CAU	21804	13,14	1,23	Arg	CGU	10516	6,34	0,83	C
	CUC	14439	8,70	0,52		CCC	13626	8,21	0,73		CAC	13673	8,24	0,77		CGC	3917	2,36	0,31	C
	CUA	20964	12,64	0,75		CCA	27953	16,85	1,50	Gln	CAA	41438	24,98	1,20		CGA	8828	5,32	0,70	C
	CUG	26042	15,70	0,93		CCG	7647	4,61	0,41		CAG	27470	16,56	0,80		CGG	3906	2,35	0,31	C
Ile	AUU	47437	28,60	1,37	Thr	ACU	33801	20,38	1,44	Asn	AAU	46967	28,31	1,06	Ser	AGU	23578	14,21	0,95	A
	AUC	31756	19,14	0,92		ACC	22250	13,41	0,95		AAC	41755	25,17	0,94		AGC	15850	9,55	0,64	A
	AUA	24866	14,99	0,72		ACA	25993	15,67	1,11	Lys	AAA	59441	35,83	1,03	Arg	AGA	35073	21,14	2,77	A
Met	AUG	31300	18,87	1,00		ACG	11609	7,00	0,50		AAG	55623	33,53	0,97		AGG	13819	8,33	1,09	A
Val	GUU	38429	23,17	1,54	Ala	GCU	36619	22,08	1,61	Asp	GAU	60131	36,25	1,23	Gly	GGU	30184	18,20	1,42	G
	GUC	21992	13,26	0,88		GCC	22109	13,33	0,97		GAC	37607	22,67	0,77		GGC	13328	8,03	0,63	G
	GUA	17809	10,74	0,71		GCA	25002	15,07	1,10	Glu	GAA	66626	40,16	1,19		GGA	30861	18,60	1,45	G
	GUG	21517	12,97	0,86		GCG	7175	4,33	0,32		GAG	45623	27,50	0,81		GGG	10577	6,38	0,50	G

[B] GS115

AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	AA	Codon	N	fpt	RSCU	
Phe	UUU	61849	25,72	1,16	Ser	UCU	55626	23,13	1,56	Tyr	UAU	41061	17,07	1,03	Cys	UGU	17471	7,27	1,25	U
	UUC	45201	18,80	0,84		UCC	37261	15,49	1,04		UAC	38864	16,16	0,97		UGC	10590	4,40	0,75	U
Leu	UUA	43364	18,03	1,07		UCA	44598	18,55	1,25	ter	UAA	2019	0,84	0,00	ter	UGA	1314	0,55	0,00	U
	UUG	70761	29,42	1,75		UCG	20237	8,42	0,57	ter	UAG	1708	0,71	0,00	Trp	UGG	24792	10,31	1,00	U
Leu	CUU	41476	17,25	1,02	Pro	CCU	36491	15,17	1,36	His	CAU	31561	13,12	1,23	Arg	CGU	15371	6,39	0,84	C
	CUC	20497	8,52	0,51		CCC	19383	8,06	0,72		CAC	19572	8,14	0,77		CGC	5538	2,30	0,30	C
	CUA	29889	12,43	0,74		CCA	40783	16,96	1,51	Gln	CAA	60798	25,28	1,21		CGA	13075	5,44	0,72	C
	CUG	37177	15,46	0,92		CCG	11030	4,59	0,41		CAG	39928	16,60	0,79		CGG	5705	2,37	0,31	C
Ile	AUU	70933	29,50	1,40	Thr	ACU	49403	20,54	1,45	Asn	AAU	68505	28,49	1,07	Ser	AGU	33958	14,12	0,95	A
	AUC	45288	18,83	0,89		ACC	32439	13,49	0,95		AAC	59750	24,85	0,93		AGC	22339	9,29	0,63	A
	AUA	35939	14,94	0,71		ACA	37813	15,72	1,11	Lys	AAA	86634	36,03	1,04	Arg	AGA	49833	20,72	2,73	A
Met	AUG	44496	18,50	1,00		ACG	16747	6,96	0,49		AAG	80224	33,36	0,96		AGG	19828	8,25	1,09	A
Val	GUU	56554	23,52	1,55	Ala	GCU	53047	22,06	1,61	Asp	GAU	87673	36,46	1,23	Gly	GGU	43947	18,27	1,43	G
	GUC	31611	13,14	0,87		GCC	32371	13,46	0,98		GAC	54371	22,61	0,77		GGC	19148	7,96	0,62	G
	GUA	25914	10,78	0,71		GCA	36354	15,12	1,10	Glu	GAA	96935	40,31	1,19		GGA	44790	18,63	1,46	G
	GUG	31527	13,11	0,87		GCG	10392	4,32	0,31		GAG	66013	27,45	0,81		GGG	15031	6,25	0,49	G

Figure 5: [**A**] Codon Usage of strain DSMZ 70382 and, [**B**] strain GS115 ; N = occurrences in total genes, fpt = frequency per thousand, RSCU = Relative Synonymous Codon Usage

3.2 Establishment of Pichia pastoris Expression Microarrays

The first *P. pastoris* expression microarrays were designed based on the commercial genome sequence by IG [Graf, 2007]. The existing gene prediction was validated, and additional ORFs were predicted and functionally annotated. The thresholds of the gene finder were deliberately set to be low, so that the filtering procedure could be better controlled. Oligos were designed for 17,161 ORFs and evaluated in a self hybridization experiment. Based on the results from this experiment, the prediction score, and the gene annotation, a final gene set was selected. Oligomere probes for 15,035 ORFs could be designed. After preliminary hybridizations, a new generation of microarrays was developed, and validated in a study about the effects of dithiothreitol (DTT) treatment and HAC1 overexpression on protein production [Graf et al., 2008]. The study was a direct comparison of wild type against mutant in the case of HAC1, and untreated against treated state in the case of DTT. Both factors are known to influence the unfolded protein response (UPR), a cellular process with an important role in the secretion of recombinant protein. For the analysis of differential expression a microarray analysis pipeline was developed. The analysis pipeline was designed for Agilent Feature Extraction (FE) output files, but can be easily adapted to other 2-colour platforms. Agilents FE also performs preprocessing of the data, though it has been reported that the variability of these values is significantly large,r as compared to loess normalized raw data [Zahurak et al., 2007; Kerr, 2007]. For higher level analysis, a Fisher's exact test and a gene set analysis (GSA) were implemented. As already mentioned in the introduction, both methods have their advantages and disadvantages that need to be taken into account when interpreting the results. The software can be used for direct (mutant vs. wild type) designs as well as reference designs, and consists of three parts (figure 6). The aim of the first step is to check the quality of the microarrays, and analyze the effects of data preprocessing and normalization steps. The quality plots available are, false colour plots, scatterplots, MA-, density-, QQ- and volcano-plots. The background correction and normalization procedures that are compared in this step are listed in Figure 7. Most normalization methods assume that the majority of genes is not differentially expressed, and react more or less sensitive to the violation of this assumption. Therefore, the percentage of differentially expressed genes is calculated from the raw data and additionally estimated using the *R* function *convest*.

Based on the first step, a background correction and normalization procedure can be selected for the second step, in which gene expression fold changes and p-values are calculated using the *eBayes* method from the *limma* package. Multiple testing correction is done using the FDR controlling method of Benjamini and Yekutieli (BY). The third step analyses the enrichment of differentially expressed data for a Gene Ontology (GO) group. The GO categorizes biological function, molecular process, and cellular component attributed to proteins into groups. The classes and their relations are represented as a directed acyclic graph (DAG), where the most general terms are at the root growing more specific with each subsequent level. The decision which level of the DAG to use in the analysis has an impact on the results of the GSA [Dopazo, 2009]. For the gene set

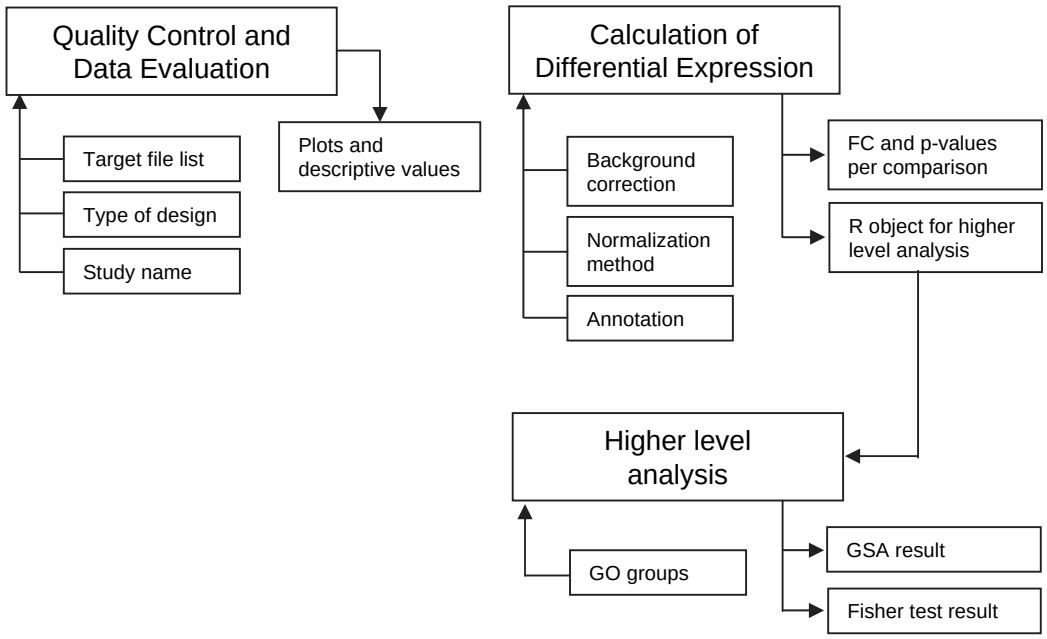


Figure 6: Steps in the microarray analysis pipeline with input and output data

analysis used in the microarray pipeline, the GOSlim Terms for biological process from the Saccharomyces Genome Database (SGD) were used, with the addition of medium level groups for terms related to protein secretion and stress response pathways[12] This represnets a compromise between very general GO terms, containing a large number of proteins, which increases the statistical power, and more specific terms, to better determine the particular effect of the study.

The *P. pastoris* microarrays and the microarray analysis pipeline were utilized in the GENOPHYS project. This international research cooperation studies the production of a heterodimeric protein in different host organisms, under several physiological conditions. For the GENOPHYS study, the antibody fragment Fab 3H6 was chosen for recombinant production. Fab fragments are of considerable medical interest, and consist of a full antibody light chain and a shorter antibody heavy chain, linked by a disulfide bond. The factors tested in *P. pastoris* were, temperature, osmolarity [Dragosits et al., 2010] and oxygen [Baumann et al., 2010]. For each factor, three conditions were tested (low, normal and high) and compared to each other as well as between producing and non-producing strain. For each condition and strain, the growth rate and the specific production was measured. Samples were analysed regarding transcriptome and proteome changes. For the expression microarrays six replicates were used, including dye swap. After examination of the data, it was decided to use no background correction, as it increased the noise in the data, and Zahurak et al. [2007] reports a slightly better AUC value without background correction for Agilent two-color data. The quality plots

[12]A full list of the used GO terms can be found in the appendix.

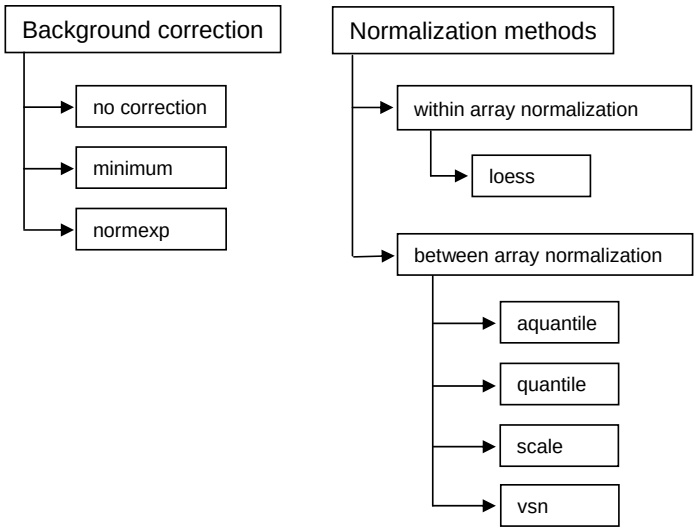


Figure 7: Background correction and normalization methods in the microarray analysis pipeline

showed a clear dye effect, which made loess normalization the appropriate choice. Additonally the aquantile method was used to normalize between arrays. The proteome analysis consisted of 2D-DIGE and LC-ESI-MS/MS for protein identification. The study was complemented by metabolic flux calculations performed by the group of Pau Ferrer in Barcelona. The results showed a beneficial effect of low temperature and low oxygen levels on the specific production of Fab 3H6 in *P. pastoris*. When trying to determine the underlying cause of this advantage, differences between transcriptome and proteome analysis highlighted the influence of post-transcriptional regulation in certain pathways, like the TCA cycle [Dragosits et al., 2009; Baumann et al., 2010]. The effects of osmolarity on *P. pastoris* were more pronounced in the control strain, but no clear effect on protein production was found [Dragosits et al., 2010].

3.3 Genome-Scale Metabolic Model

With a sequenced and annotated genome as well as established transcriptomics [Graf et al., 2008] and proteomics [Dragosits et al., 2009] tools, the next step towards Systems Biotechnology could be taken. The construction of the initial genome-scale metabolic model was done based on the gene to protein to reaction relationship, but to create a model capable of sensible predictions about the behaviour of the system, it was essential to include as much information about the physiological characteristics of the organism as possible. In the model for *P. pastoris*, the annotation was complemented with data from literature and public database, but also with experimental data from fed batch and chemostat cultivations under various conditions. The basis used for model construction was the genome and annotation of the *P. pastoris* strain DSMZ 70382, with 5,407 putative genes [Mattanovich et al., 2009a]. Additionally, the published genome and annotation of the strain GS115, with 5,313 putative genes [De Schutter et al., 2009]

was included. In both strains, about 75 % of the genes could be functionally annotated, and 96.5 % of the predicted ORFs are present in both genomes. The average identity of homologous proteins between the GS115 and the DSMZ 70382 strain was 93.7% on amino acid level. The genome-scale metabolic model PpaMBEL1254 comprises 1,254 metabolic reactions, of which 9.4 % are essential, and 1,147 metabolites. PpaMBEL1254 is compartmentalized into 7 cellular compartments and the cell exterior, which is also treated as a compartment. The gene coverage of 540 genes represents 9.9 % of the gene set. In comparison with existing models of *S. cerevisiae*, the *P. pastoris* model has a similar number of reactions and metabolites. The gene coverage in *S. cerevisiae* is higher, reflecting the large number of isogenes present in *S. cerevisiae* due to the gene duplication event [Herrero, 2005]. Also, the present information about *P. pastoris* can not match the exhaustive knowledge base that has been generated for the model organism over the years. For the analysis of heterologous protein production, protein synthesis was included into the model. Subsequently, the production of two heterologous proteins, the secreted HSA and the intracellular hSOD, was simulated and analysed under oxygen limitation. The maximal oxygen transfer rate is an important and limiting factor in large scale high density cultivations. Also, studies indicated beneficial effects of hypoxic conditions on the fermentation process and an increase in productivity for *P. pastoris* grown on glucose [Baumann et al., 2008, 2010]. To simulate the effect of oxygen limitation with the model, a flux variability analysis was carried out. The increase in productivity under oxygen limitation could be confirmed, but additionally, the simulation showed that HSA and hSOD had different optima and reacted with different sensitivity to the changes in oxygen level and growth rate. Shortly after the publication of Sohn et al. [2010], another model of the same organism was presented by Chung et al. [2010]. This model (iPP668) contains 1,361 reactions, 1,177 metabolites and 668 genes. While the work with the PpaMBEL1254 concentrated on the enhancement of protein production in *P. pastoris* when grown on glucose, Chung et al. [2010] focused on the production of metabolites on methanol and examined the biocatalytic potential of *P. pastoris*. It will be interesting in the future to compare these models moving towards an improved representation of the cellular factory *P. pastoris*.

3.4 Authors' Contribution

Much of the work presented here was made possible through the jont effort of many people. All experimental microarray work was performed by Martin Dragosits and Kristin Baumann. Several people from the Mattanovich lab had a part in the manual curation of the putative genes of *P. pastoris*. Brigitte Gasser contributed in many ways to this work, especially with regard to the biological interpretation of the data. The genome browser for *P. pastoris* was implemented by Andreas Redl in the frame of his diploma thesis and the metabolic model was created together with the group of Sang Yup Lee from the KAIST in Korea.

Of all the work presented here, the development of the microarray pipeline and the comparison of the two *P. pastoris* strains has not been published yet. The microarray

analysis pipeline, as described in section 3.2 on page 24 (and following), was conceived and implemented by me as an in-house solution for the analysis of transcriptomics data. The transcriptomics platform and the microarray analysis were validated in Graf et al. [2008], and subsequently improved to make the analysis more flexible user friendly. The microarray analysis pipeline was then applied in the transcriptomics part of the GENOPHYS project for the organisms *S. cerevisiae* and *P. pastoris*, as documented in Dragosits et al. [2010] and Baumann et al. [2010]. For the comparison of the two sequenced *P. pastoris* strains, I mapped the genomes against each other using various alignment methods, resulting in a chromosome map, which is visualized on page 20. As shown in figure 4 on page 22 and figure 5 on page 23, I also compared the codon usage and the amino acid composition of the predicted proteins between the two strains. It will be interesting to do a more detailed genome comparison in the future, to elucidate the differences and similarities of these two organisms.

4 Conclusion

The availability of expression microarrays facilitated a comprehensive study of gene expression in *P. pastoris*, revealing changes in to date unexpected pathways. Furthermore, comparisons of the proteomics and transcriptomics analysis clearly support the need for a Systems Biology approach. Each method on its own has certain strengths and weaknesses and is biased towards certain pathways. Integrative analysis of omics data increases the power of a study and improves our knowledge about biological systems considerably. With the availability of the genome sequence of by now several *P. pastoris* strains, a basis for many high-throughput technologies is provided. Furthermore, it is now possible to use the genome as a reference for mutational profiling, RNA-seq or ChIP-seq studies. Apart from the purely scientific interest of understanding how this organism works, the genome annotation will help to elucidate functions and pathways that can be targeted in metabolic engineering. Without the need to rely on the comparison to *S. cerevisiae*, it will be easier to discover and characterize specific features of this highly interesting cell factory. Genome-scale models have already been utilized with great success to improve production in microorganisms and broaden the product range. The availability of a metabolic model enables researchers to save resources by focusing on the most likely targets, using a primarily hypotheses driven approach. *P. pastoris* is now one of a few eukaryotic organisms for which such a genome-scale model exists. The first simulation to investigate the effect of oxygen limiting conditions on the production of two heterologous proteins was already successful and will help in the design of optimized fermentation strategies. Metabolic engineering will benefit greatly from the capability of the model to predict knockout or overexpression targets for the improvement of protein or metabolite production. Though the work on the model is not completed, each round of prediction and validation steps will increase our knowledge about the organism and its behaviour and in turn improve the quality of the genome-scale metabolic model of *P. pastoris*.

References

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, and et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.

C. L. Barrett, T. Y. Kim, H. U. Kim, B. O. Palsson, and S. Y. Lee. Systems biology as a foundation for genome-scale synthetic biology. *Nucleic Acids Research*, 17:488–492, 2006.

K. Baumann, M. Maurer, M. Dragosits, O. Cos, P. Ferrer, and D. Mattanovich. Hypoxic fed-batch cultivation of Pichia pastoris increases specific and volumetric productivity of recombinant proteins. *Biotechnol Bioeng.*, 100(1):177–83, 2008.

K. Baumann, M. Carnicer, M. Dragosits, A.B. Graf, J. Stadlmann, P. Jouhten, H. Maaheimo, B. Gasser, J. Albiol, D. Mattanovich, and P. Ferrer. A multi-level study of recombinant Pichia pastoris in different oxygen conditions as knowledge base for strain improvement. 2010.

R. Breitling. Biological microarray interpretation: the rules of engagement. *Biochim Biophys Acta*, 1759(7):319–27, 2006.

M. Carnicer, K. Baumann, I. Töplitz, F. Sánchez-Ferrando, D. Mattanovich, P. Ferrer, and J. Albiol. Macromolecular and elemental composition analysis and extracellular metabolite balances of Pichia pastoris growing at different oxygen levels. *Microbial Cell Factories*, 8(65), 2009.

R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, Tissier C., T. C. Walk, P. Zhang, and P. D. Karp. The Metacyc database of metabolic pathways and enzymes and the Biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 36:623–631, 2007.

J.L. Cereghino and J.M. Cregg. Heterologous protein expression in the methylotrophic yeast Pichia pastoris. *FEMS Microbiology Reviews*, 24:45–66, 2000.

H. Chipman and R. Tibshirani. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7(2):286–301, 2006.

B. K. S. Chung, S. Selvarasu, C. Andrea, J. Ryu, H. Lee, J. Ahn, H. Lee, and D.-Y. Lee. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast Pichia pastoris for strain improvement. *Microbial Cell Factories*, 9(50), 2010.

J.M. Cregg, J.L. Cereghino, J. Shi, and D.R. Higgins. Recombinant protein expression in Pichia pastoris. *Molecular Biotechnology*, 16, 2000.

A. R. Dabney and J. D. Storey. Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biology*, 8(3), 2007.

K. De Pourcq, K. De Schutter, and N. Callewaert. Enineering of glycosylation in yeast and other fungi: current state and perspectives. *Appl Microbiol Biotechnol*, 87:1617–1631, 2010.

K. De Schutter, Y.-C. Lin, P. Tiels, A. Van Hecke, S. Glinka, J. Weber-Lehmann, P. Rouzé, Y. Van de Peer, and N. Callewaert. Genome sequence of the recombinant protein production host Pichia pastoris. *Nature Biotechnology*, 27(6):561–566, 2009.

A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–9, 2007.

J. Dopazo. Formulating and testing hypotheses in functional genomics. *Artificial Intelligence in Medicine*, 45:97–107, 2009.

M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, M. Sauer, F. Altmann, P. Ferrer, and D. Mattanovich. The effect of temperature on the proteome of recombinant Pichia pastoris. *Journal of Proteom Research*, 8:1380–1392, 2009.

M. Dragosits, A. Stadlmann, J.and Graf, B. Gasser, M. Maurer, M. Sauer, D. P. Kreil, F. Altmann, and D. Mattanovich. The response to unfolded protein is involved in osmotolerance of Pichia pastoris. *BMC Genomics*, (11):13, 2010.

M. Droege and B. Hill. The genome sequencer FLX system - longer reads, more applications, straight forward bioinformatics and more complete data set. *Journal of Biotechnology*, 136:3–10, 2008.

S. Dudoit, R. C. Gentleman, and J. Quackenbush. Open source software for the analysis of microarray data. *Bio Techniques*, 34:45–51, 2003.

B. Dujon. Yeast evolutionary genomics. *Nature Reviews Genetics*, 11:512–524, 2010.

B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lfontaine, and J. et al. de Montigny. Genome evolution in yeasts. *Nature*, 430:35–44, 2004.

M. Durot, P.-Y. Bourguignon, and V. Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev*, 33:164–190, 2009.

T. L. Fare, E. M. Coffey, H. Dai, Y. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, R. B. Meyer, M. R. Stoughton, G. Y. Tokiwa, and Y. Wang. Effects of atmospheric ozone on microarray data quality. *Analytical Chemistry*, 75(17):4672–4675, 2003.

R. A. Farrer, E. Kemen, J. D. G. Jones, and D. J. Studholme. De novo assembly of the Pseudomonas syringae pv. syringae B728A genome using illumina/solexa short sequence reads. *FEMS Microbiol Lett.*, 291(1):103–111, 2009.

A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. . Palsson. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *PLoS Computational Biology*, 3(121), 2007.

A. M. Feist, M. J. Herrgard, I. Thiele, J. L. Reed, and B. . Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7:129–143, 2009.

C. Francke, R. J. Siezen, and B. Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.

B. Gasser and D. Mattanovich. Antibody production with yeasts and filamentous fungi: on the road to large scale? *Biotechnol Lett*, 29(2):201–12, 2007.

B. Gasser, M. Maurer, J. Gach, and D. Mattanovich. Engineering of Pichia pastoris for improved production of antibody fragments. *Biotechnology and Bioengineering*, 94(2), 2006.

B. Gasser, M. Maurer, J. Rautio, M. Sauer, A. Bhattacharyya, M. Saloheimo, M. Penttila, and D. Mattanovich. Monitoring of transcriptional regulation in Pichia pastoris under protein production conditions. *BMC Genomics*, 8(179), 2007b.

B. Gasser, M. Sauer, M. Maurer, G. Stadlmayr, and D. Mattanovich. Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts. *Appl Environ Microbiol*, 73(20):6499–507, 2007a.

G. Gellissen, G. Kunze, J. M. Gaillardin, C. Cregg, E. Berardi, M. Veenhuis, and I. van der Klei. New yeast expression platforms based on methylotrophic Hansenula polymorpha and Pichia pastoris and on dimorphic Arxula adeninivorans and Yarrowia lipolytica - a comparison. *FEMS Yeast Research*, 5:1079–1096, 2005.

R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

A. Graf. Building a bioinformatics pipeline for the design of Pichia pastoris whole genome microarrays. Master's thesis, FH-Campus Wien, 2007.

A. Graf, B. Gasser, M. Dragosits, M. Sauer, G. G. Leparc, T. Tüchler, D. P. Kreil, and D. Mattanovich. Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays. *BMC Genomics*, (9):13, 2008.

A. Graf, B. Gasser, M. Dragosits, and D. Mattanovich. Yeast systems biotechnology for the production of heterologous proteins. *FEMS Yeast Research*, 9(3):335–348, 2009.

A. J. F. Griffiths, S. R. Wessler, R. C. Lewontin, W. M. Gelbart, D. T. Suzuki, and J. H. Miller. *Introduction to Genetic Analysis*. W.H. Freeman and Company, 2005.

P. K. Gupta. Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology*, 26(11):602–611, 2008.

N. Hall. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology*, 209:1518–1525, 2007.

E. Herrero. Evolutionary relationships between Saccharomyces cerevisiae and other fungal species as determined from genome comparisons. *Rev Iberoam Micol*, 22(4): 217–22, 2005.

J. Heyland, J. Fu, L. M. Blank, and A. Schmid. Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production. *Biotechnol Bioeng.*, 107(2):357–68, 2010.

H. Hohenblum, B. Gasser, M. Maurer, N. Borth, and D. Mattanovich. Effects of gene dosage, promoters, and substrates on unfolded protein stress of recombinant Pichia pastoris. *Biotechnol Bioeng.*, 85(4):367–75, 2004.

W. Huber, Sültmann H. von Heydebreck, A., A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):96–104, 2002.

W. Huber, von A. Heydebreck, and M. Vingron. An introduction to low-level analysis methods of DNA microarray data. The Berkeley Electronic Press, 2005. Bioconductor Project Working Papers - Paper 9.

W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425: 686–691, 2003.

M. Inan, D. Aryasomayajula, J. Sinha, and M. M. Meagher. Enhancement of protein secretion in Pichia pastoris by overexpression of protein disulfide isomerase. *Biotechnol Bioeng*, 93(4):771–8, 2006.

P.S. Jacobs, S. Geysens, W. Vervecken, R. Contreras, and N. Callewaert. Enineering complex-type N-glycosylation in Pichia pastoris using glycoswitch technology. *Nature Protocols*, 4:58–70, 2009.

P. Jaluria, K. Konstantopoulos, M. Betenbaugh, and J. Shiloach. A perspective on microarrays: current applications, pitfalls, and potential uses. *Microb Cell Fact*, 6(4), 2007.

I. B. Jeffery, D. G. Higgins, and A. C. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7, 2006.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

K. F. Kerr. Extended analysis of benchmark datasets for agilent two-color microarrays. *BMC Bioinformatics*, 8(371), 2007.

T. Y. Kim, S. B. Sohn, H. U. Kim, and S. Y. Lee. Strategies for systems-level metabolic engineering. *Nucleic Acids Research*, 3:612–623, 2008.

D. P. Kreil and R. R. Russell. There is no silver bullet – a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics*, 6 (1):86–97, 2005.

C. P. Kurtzman. Biotechnological strains of Komagataella (Pichia) pastoris are Komagataella phaffi as determined from multigene sequence analysis. *J. Ind. Microbiol. Biotechnol.*, 2009.

S. Y. Lee, D.-Y. Lee, and T. Y. Kim. Systems biotechnology for strain improvement. *Trends in Biotechnology*, 23(7):349–358, 2005.

G. G. Leparc, T. Tüchler, G. Striedner, K. Bayer, P. Sykacek, I. L. Hofacker, and D. P. Kreil. Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Research*, 37(3), 2009.

S. Macauley-Patrick, M.L. Fazenda, B. McNeil, and L.M. Harvey. Heterologous protein production using the Pichia pastoris expression system. *Yeast*, 2:249–270, 2005.

D. MacLean, D. G. Jonathan, and D. J. Studholme. Application of next-generation sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7:287–296, 2009.

E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 2008.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Goodwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant,

B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, 437(7075):376–380, 2005.

H. Marx, M. Sauer, D. Resina, M. Vai, D. Porro, F. Valero, P. Ferrer, and D. Mattanovich. Cloning, disruption and protein secretory phenotype of the GAS1 homologue of Pichia pastoris. *FEMS Microbiol Lett*, 264(1):40–7, 2006.

D. Mattanovich, B. Gasser, H. Hohenblum, and M. Sauer. Stress in recombinant protein producing yeasts. *Journal of Biotechnology*, 113:121–135, 2004.

D. Mattanovich, N. Callewaert, P. Rouzé, Y. C. Lin, A. Graf, A. Redl, P. Thiels, B. Gasser, and K. Schutter. Open access to sequence: browsing the Pichia pastoris genome. *Microb. Cell Fact.*, 8(13), 2009b.

D. Mattanovich, A. Graf, J. Stadlmann, M. Dragosits, A. Redl, M. Kleinheinz, M. Sauer, F. Altmann, and B. Gasser. Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris. *Microb. Cell Fact.*, 8(1), 2009a.

B. H. Mecham, P. S. Nelson, and J. D. Storey. Supervised normalization of microarrays. *Bioinformatics*, 26(10):1308–1315, 2010.

D. Medini, D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli. Microbiology in the post-genomic era. *Nature Reviews Microbiology*, 6: 419–430, 2008.

M. L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11:31–46, 2010.

M. Miron and R. Nadon. Inferential literacy for experimental high-throughput biology. *Trends in Genetics*, 22(2), 2006.

J. Nielsen. Principles of optimal metabolic network operation. *Molecular Systems Biology*, 3(126), 2007.

I. Nookaew, M. C. Jewett, A. Meechai, C. Thammarongtham, K. Laoteng, S. Cheevadhanarak, J. Nielsen, and S. Bhumiratana. The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. *BMC Systems Biology*, 2, 2008.

H. Ohi, N. Okazaki, S. Uno, M. Miura, and R. Hiramatsu. Chromosomal DNA patterns and gene stability of Pichia pastoris. *Yeast*, 14(10):895–903, 1998.

P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

M. Pop and L. Salzberg. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 2008.

N. D. Price and I. Shmulevich. Biochemical and statistical network models for systems biology. *Curr Opin Biotechnol*, 18(4):365–70, 2007.

S. Raychaudhuri, P. D. Sutphin, J. T. Chang, and R. B. Altman. Basic microarray analysis: grouping and feature reduction. *Trends in Biotechnology*, 19(5):189–193, 2001.

A. Redl. Implementing a genome browser for the yeast Pichia pastoris. Master's thesis, FH-Campus Wien, 2008.

A. Reiner, D. Yekutieli, and Y. Benjamini. Identifiying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

D. Resina, M. Maurer, O. Cos, C. Arnau, M. Carnicer, H. Marx, B. Gasser, F. Valero, D. Mattanovich, and P. Ferrer. Engineering of bottlenecks in Rhizopus oryzae lipase production in Pichia pastoris using the nitrogen source-regulated FLD1 promoter. *N Biotechnol*, 25(6):396–403, 2009.

S. Rogers and A. Cambrosio. Making a new technology work: the standardization and regulation of microarrys. *Yale Journal of Biology and Medicine*, 80:165–178, 2007.

M. Sauer, P. Branduardi, B. Gasser, M. Valli, M. Maurer, D. Porro, and D. Mattanovich. Differential gene expression in recombinant Pichia pastoris analysed by heterologous DNA microarray hybridisation. *Microb Cell Fact*, 3(1):17, 2004.

U. Sauer. Metabolic networks in motion: ^{13}c-based flux analysis. *Molecular Systems Biology*, (62), 2006.

R. B. Scharpf, C. A. Iacobuzio-Donahue, J. B. Sneddon, and G. Parmigiani. When should one subtract background fluorescence in 2-color microarrays? *Biostatistics*, 8 (4):695–707, 2007.

T. Schlitt and A. Brazma. Modelling gene networks at different organisational levels. *GEBS Letters*, 579:1859–1866, 2005.

S. C. Schuster. Next-generation sequencing transforms todays biology. *Nature Methods*, 5(1):16–18, 2008.

P. M. Sharp, T. M. F. Tuohy, and K. R. Mosurski. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, 14(13): 5125–5143, 1986.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.*, 3(3), 2004.

G. K. Smyth, Y. H. Yang, and T. Speed. Statistical issues in cDNA microarray data analysis. *Functional Genomics*, 224:111–136, 2003.

S. B. Sohn, A. B. Graf, T. Y. Kim, B. Gasser, M. Maurer, P. Ferrer, D. Mattanovich, and S. Y. Lee. Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production. *Biotechnology Journal*, (5), 2010.

A. Solà, H. Maaheimo, K. Ylonen, P. Ferrer, and T. Szyperski. Amino acid biosynthesis and metabolic flux profiling of Pichia pastoris. *Eur J Biochem*, 271(12):2462–70, 2004.

A. Solà, P. Jouhten, H. Maaheimo, F. Sánchez-Ferrando, T. Szyperski, and P. Ferrer. Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology*, 153:281–290, 2007.

M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5): 637–644, 2008.

L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. The generic genome browser: A building block for a model organism system database. *Genome Research*, 10:1599–1610, 2002.

C. Steinhoff and M. Vingron. Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics*, 7(2):166–177, 2006.

R. Steuer. Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry*, 68:2139–2151, 2007.

A. L. Tarca, R. Romero, and S. Draghici. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195:373–388, 2006.

I. Thiele, N. Jamshidi, R. M. T. Fleming, and B. . Palsson. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization. *PLoS Computational Biology*, 5(3), 2009.

Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

S. Wildt and T.U. Gerngross. The humanization of N-glycosylation pathways in yeast. *Nature Reviews Microbiology*, 3:119–128, 2005.

M. Xu, D. Fujita, and N. Hanagata. Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *small*, 5(23):2638–2649, 2009.

Y. Yamada, M. Matsuda, K. Maeda, and K. Mikata. The phylogenetic relationship of methanol-assimilating yeasts based on the partial sequences of 18s and 26s ribosomal RNAs: the proposal of Komagataella gen. nov. (Saccharomycetaceae). *Bioscience Biotechnology Biochemestry*, 59:439–444, 1995.

M. Zahurak, G. Parmigiani, Y. Wayne, R. B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer, and L. Cope. Pre-processing agilent microarray data. *BMC Bioinformatics*, 8(142), 2007.

E. M. Zdobnov and R. Apweiler. Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–8, 2001.

W. Zhang, H. L. Zhao, C. Xue, X. H. Xiong, X. Q. Yao, X. Y. Li, H. P. Chen, and Z. M. Liu. Enhanced secretion of heterologous proteins in Pichia pastoris following overexpression of Saccharomyces cerevisiae chaperone proteins. *Biotechnol Prog.*, 22 (4):1090–5, 2006.

5 Appendix

5.1 Abbreviation

2D-DIGE	2-D DIfference in Gel Electrophoresis
AUC	Area Under the Curve (from ROC-curve)
BLAST	Basic Local Alignment Tool
ChIP	Chromatine Immuno Precipitation
DAG	Directed Acyclic Graph
DTT	Dithiothreitol
EC	Enzyme Commission
FBA	Flux Balance Analysis
FDR	False Discovery Rate
FE	Feature Extraction software by Agilent
GO	Gene Ontology
GPR	Gene Protein Reaction
GSA	Gene Set Analysis
HSA	Human Serum Albumin
hSOD	human Superoxide Dismutase
KEGG	Kyoto Encyclopedia of Genomes and Genes
LC-ESI-MS	Liquid Chromatography-Electrospray Ionization-Mass Spectrometry
MoMA	Minimization of Metabolic Adjustment
MS	Mass Spectrometry
NGS	Next Generation Sequencing
NMR	Nuclear Magnetic Resonance
ORF(s)	Open Reading Frame(s)
PCR	Polymerase Chain Reaction
ROC	Receiver Operating Characteristic
ROOM	Regulatory On Off Minimization
RSCU	Relative Synonymous Codon Usage
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
TRAC	TRanscript analysis by aid of Affinity Capture
UPR	Unfolded Protein Response

5.2 GO Terms

GO number	Biological process
GO:0000746	conjugation
GO:0000910	cytokinesis
GO:0006091	generation of precursor metabolites and energy
GO:0006259	DNA metabolic process
GO:0006350	transcription
GO:0006399	tRNA metabolic process
GO:0006412	translation
GO:0006457	protein folding
GO:0006464	protein modification process
GO:0006468	protein amino acid phosphorylation
GO:0006486	protein amino acid glycosylation
GO:0006519	cellular amino acid and derivative metabolic process
GO:0006725	cellular aromatic compound metabolic process
GO:0006766	vitamin metabolic process
GO:0006810	transport
GO:0006811	ion transport
GO:0006839	mitochondrial transport
GO:0006897	endocytosis
GO:0006950	response to stress
GO:0006970	response to osmotic stress
GO:0006974	response to DNA damage stimulus
GO:0006979	response to oxidative stress
GO:0006986	response to unfolded protein
GO:0006997	nuclear organization and biogenesis
GO:0007005	mitchondrion organization and biogenesis
GO:0007010	cytoskeleton organization and biogenesis
GO:0007029	ER organization and biogenesis
GO:0007031	peroxisome organization and biogenesis
GO:0007033	vacuole organization and biogenesis
GO:0007034	vacuolar transport
GO:0007049	cell cycle
GO:0007059	chromosome segregation
GO:0007114	cell budding
GO:0007124	pseudohyphal growth
GO:0007126	meiosis
GO:0007165	signal transduction
GO:0009408	response to heat
GO:0009409	response to cold
GO:0015031	protein transport
	continued on next page...

GO number	**Biological process**
GO:0016044	cellular membrane organization
GO:0016050	vesicle organization and biogenesis
GO:0016070	RNA metabolic process
GO:0016071	mRNA metabolic process
GO:0016072	rRNA metabolic process
GO:0016192	vesicle-mediated transport
GO:0016567	protein ubiquitination
GO:0016568	chromatin modification
GO:0016570	histone modification
GO:0019725	cellular homeostasis
GO:0030435	sporulation resulting in formation of a cellular spore
GO:0031505	fungal-type cell wall organization
GO:0032196	transposition
GO:0032989	cellular component morphogenesis
GO:0042221	response to chemical stimulus
GO:0042254	ribosome biogenesis
GO:0042594	response to starvation
GO:0043543	protein amino acid acylation
GO:0044255	cellular lipid metabolic process
GO:0044257	cellular protein catabolic process
GO:0044262	cellular carbohydrate metabolic process
GO:0045333	cellular respiration
GO:0046483	heterocycle metabolic process
GO:0048193	golgi vesicle transport
GO:0051169	nuclear transport
GO:0051186	cofactor metabolic process
GO:0051276	chromosome organization and biogenesis
GO:0070271	protein complex biogenesis

6 Publications

6.1 Reviews

Graf A, Dragosits M, Gasser B, Mattanovich D. *Yeast systems biotechnology for the production of heterologous proteins.* FEMS Yeast Res. 2009 May; 9(3):335-48

Contribution: I researched and wrote the part on: 'New (post) genomic approaches to systems biotechnology'.

MINIREVIEW

Yeast systems biotechnology for the production of heterologous proteins

Alexandra Graf[1,2], Martin Dragosits[1], Brigitte Gasser[1] & Diethard Mattanovich[1,2]

[1]Department of Biotechnology, Institute of Applied Microbiology, University of Natural Resources and Applied Life Sciences, Vienna, Austria; and [2]School of Bioengineering, University of Applied Sciences FH-Campus, Vienna, Austria

Correspondence: Diethard Mattanovich, Department of Biotechnology, Institute of Applied Microbiology, University of Natural Resources and Applied Life Sciences, Muthgasse 18, A-1190 Vienna, Austria.
Tel.: +43 1 360 06 6569;
fax: +43 1 369 7615;
e-mail: diethard.mattanovich@boku.ac.at

Received 9 December 2008; revised 23 February 2009; accepted 26 February 2009.
First published online 1 April 2009.

DOI:10.1111/j.1567-1364.2009.00507.x

Editor: Teun Boekhout

Keywords
systems biotechnology; heterologous protein; yeast; *Pichia pastoris*; systems biology; metabolic engineering.

Abstract

Systems biotechnology has been established as a highly potent tool for bioprocess development in recent years. The applicability to complex metabolic processes such as protein synthesis and secretion, however, is still in its infancy. While yeasts are frequently applied for heterologous protein production, more progress in this field has been achieved for bacterial and mammalian cell culture systems than for yeasts. A critical comparison between different protein production systems, as provided in this review, can aid in assessing the potentials and pitfalls of applying systems biotechnology concepts to heterologous protein producing yeasts. Apart from modelling, the methodological basis of systems biology strongly relies on postgenomic methods. However, this methodology is rapidly moving so that more global data with much higher sensitivity will be achieved in near future. The development of next generation sequencing technology enables an unexpected revival of genomic approaches, providing new potential for evolutionary engineering and inverse metabolic engineering.

Introduction

The unexpectedly fast progress in genome sequencing over the last decade has provided an invaluable source of information on the physiology of microorganisms, including a comprehensive overview on cellular endowment with metabolic enzymes. Simultaneously, metabolic modelling has been developed and applied to the mathematical description of the central metabolic processes of bacteria (Edwards & Palsson, 2000) and yeast (Förster *et al*., 2003). Together with extensive work on genomic data to address ideally all metabolic processes of a cell, these metabolic models led to the concept of systems biology (Westerhoff & Palsson, 2004). Several systems biology models of *Saccharomyces cerevisiae* have previously been described and recently unified to one comprehensive model (Herrgård *et al*., 2008, and references therein). To acquire data for dynamic modelling, postgenomic analyses at the transcriptomics, proteomics and metabolomics level are implemented.

These models offer the opportunity to predict cellular processes and are therefore regarded as highly valuable resources for strain optimization. Since 1990, the concepts of metabolic engineering have been developed and applied as the knowledge-based improvement of cell factories using genetic engineering (Bailey, 1991; Nielsen, 2001). Extending the concepts of metabolic engineering to a broad system basis has led to the conception of systems biotechnology (Lee *et al*., 2005; Nielsen & Jewett, 2008). While systems biology aims ideally at the global understanding and modelling of the entire cellular network of reactions, systems biotechnology will rather accept gaps in the description of cellular processes, as long as the processes related to product formation can be mapped. The systems biotechnology approach can be seen as an iterative, cyclic process, integrating high throughput data generation with metabolic modelling and production strain optimization (Fig. 1).

The concepts of metabolic engineering were initially mostly applied to the production of metabolites. In 2000, heterologous proteins were introduced as a new class of products to be addressed by metabolic engineering (Ostergaard *et al*., 2000). It is obvious that the complexity of the protein production and secretion process (Fig. 2) renders it

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

much more challenging to be addressed by tools of rational and quantitative analysis, as summarized in Table 1.

In cases where the genetic traits controlling complex cellular responses are not known, researchers have applied random mutagenesis and selection schemes to engineer metabolic pathways by modifying enzymes, transporters or regulatory proteins. This method has been termed evolutionary engineering (Sonderegger & Sauer, 2003), but its application to protein production is not straight forward as protein overexpression is usually not advantageous for the cell. Single cell sorting of large cell populations may be applied to overcome this limitation (Mattanovich & Borth, 2006). However, as the changes are not directed, it is often difficult to determine the genetic modification that is responsible for the improvement (Nevoigt, 2008). Understanding the biological system as a whole greatly facilitates the rational understanding of such mutants. Genome-wide analysis methods also made another biotechnological approach, namely inverse metabolic engineering (Bailey *et al.*, 1996), much more feasible, where phenotypic differences serve as the basis for elucidating genetic modifications needed to optimize production strains (Bro & Nielsen, 2004).

Yeasts are attractive hosts for production of heterologous proteins (Porro *et al.*, 2005). However, a number of bottlenecks and stress factors limit the full potential of this class of organisms (Mattanovich *et al.*, 2004), and systems biotechnology will offer new opportunities for modelling, analysis and optimization of protein production systems. In the

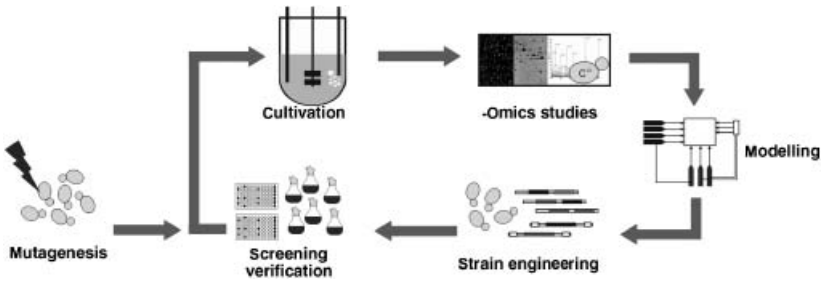


Fig. 1. The systems biotechnology circle. A primary production strain is cultivated under relevant conditions. Omics methods feed models, which aid the design of strain engineering and screening of new, improved strains. Random mutagenesis serves to increase variability, which can also be achieved by evolutionary engineering.

Table 1. Challenges for systems biotechnology research for heterologous protein production

Key elements for systems biotechnology	Challenges faced in protein expression and secretion
Metabolites and enzymes	Molecular players only partly defined
Metabolic pathways	Pathways less clear than those for metabolic processes
Stoichiometry of metabolic reactions	Stoichiometry difficult to define
Metabolic fluxes	Fluxes and concentrations of participating 'metabolites' difficult or not yet measurable

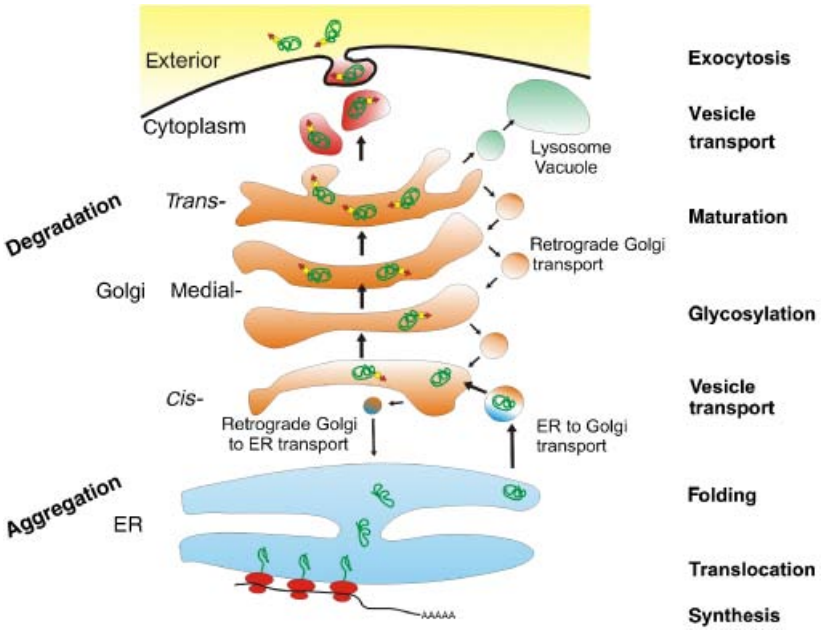


Fig. 2. Protein synthesis and secretion in eukaryotic cells. Pathways and key 'metabolic' processes of secreted proteins are indicated, including unproductive steps such as degradation or aggregation. These pathways should serve as a basis for the development of a model of the protein production and secretion process.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

following we will describe the applications of systems biotechnology to yeasts with a strong emphasis on heterologous protein production, highlighting work on other classes of host organisms also, and provide an outlook as to where the development and integration of new methodology can lead this field.

Impact of systems biology on yeast biotechnological processes

Systems biology is not a purely academic research area as the quantitative description of microbial production cell lines is also already of interest for biotechnological industry. As pointed out before, yeasts producing heterologous proteins have rarely been investigated on a systems level so far, while the importance of systems biotechnology for industrial applications others than protein production has been well documented in the recent years (Pizarro *et al.*, 2007; Takors *et al.*, 2007; Mukhopadhyay *et al.*, 2008). These applications span mainly the production of primary metabolites such as amino acids, alcohols or organic acids, often employing bacteria as production hosts. Global analysis of the host cell metabolism can help to improve the production of metabolites, which may consequently require further cell engineering to resist high concentrations of possibly toxic chemical compounds. Alper *et al.* (2006) recently showed how engineering of the global transcription machinery can help to improve ethanol resistance in yeast cells.

Additionally, systems biotechnology already impacts on the production of yeast-based alcoholic beverages as a system-wide understanding of the molecular basis of the production process can lead to improved sensory qualities for consumer demands. Proteomic and transcriptomic methods have been applied to investigate wine and beer fermentations (Kobi *et al.*, 2004; Beltran *et al.*, 2006; Zuzuarregui *et al.*, 2006). It becomes clear from the system-wide analysis of wine fermentations that microbial cells encounter many different kinds of stresses during the fermentation process. However, it is obvious that a deeper understanding of the cellular reactions to environmental stresses is also crucial for other biotechnologically relevant batch and fed-batch processes such as amino acid, biofuel and, of course, protein production. The yeast stress response has been a topic of detailed investigations in the recent years. Both, transcriptomic and proteomic approaches have been used to investigate the effect of temperature (Gasch *et al.*, 2000; Tai *et al.*, 2007), high osmolarity (Blomberg, 1995; Chen *et al.*, 2003; Gori *et al.*, 2007; Kim *et al.*, 2007), hydrostatic pressure (Fernandes *et al.*, 2004) and nutrient limitations (Kolkman *et al.*, 2006) in several yeast and fungal species in recent years. It has been outlined earlier that cellular reactions to environmental stresses are mainly a transient response, on which most studies have focused. However, during industrial processes, rather constant suboptimal growth conditions that are far from the natural environment of the cells are imposed (Mattanovich *et al.*, 2004). For example, several reports indicate a positive effect of reduced growth temperature on the production rate of heterologous proteins in the yeast *Pichia pastoris* (Li *et al.*, 2001; Jahic *et al.*, 2003; Shi *et al.*, 2003). However, these data are not fully conclusive as the authors suggest that lower activity of proteases in the culture supernatant or decreased cell death rate is responsible for increases in productivity. On the other hand, Hohenblum *et al.* (2003) showed that, at least for *P. pastoris*, significant cell death occurs at low pH while temperature does not influence viability. None of these studies applied system-wide analysis of the host organism, so that the underlying biology remained unexplored. Newer studies may shed light on such contradictory data. Recently, Tai *et al.* (2007) performed steady-state cultivations of *S. cerevisiae* at different temperatures and analysed them with microarrays. Although not using heterologous protein-producing strains, these experiments allow conclusions on the long-term adaptation of production cells to suboptimal conditions. The authors observed an upregulation of ribosome biogenesis genes and a downregulation of environmental stress response genes at a low temperature, which differed largely from previous results on rapid changes of temperature. Similarly, Gasser *et al.* (2007a) found stress response genes downregulated at lower temperature in steady-state cultures of *P. pastoris* expressing an antibody Fab fragment, while the specific productivity of this protein was increased.

However, transcriptome, proteome and metabolome data from small-scale laboratory experiments might differ significantly from cellular regulatory patterns as they occur during large-scale industrial processes. Furthermore there might be crucial genotypic differences between laboratory and industrial strains. As the majority of the mentioned studies were performed in laboratory strains, the direct applicability of these results on industrial strains is questionable. Production strain optimization can be achieved if the cellular metabolism and the regulatory networks of an industrial cell line are considered and investigated (Takors *et al.*, 2007). Up to now, such results are rarely obtained in academia, as the genomic information of important industrial strains is still missing. Omics approaches as well as fast modern sequencing techniques bear the potential to change the methodical approaches here, as will be discussed below.

The external stresses mentioned above, and intrinsic stress mediated by protein overproduction, play a major role for the physiological constraints of a protein production system (Mattanovich *et al.*, 2004). As these constraints share similar patterns among different classes of host organisms, we will also highlight research with non-yeast hosts for protein production in the following chapter.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Application of systems biotechnology to heterologous protein production

Engineering of recombinant yeasts based on genome-wide analysis

Applications of genome-wide technologies in yeasts are scarce in the field of recombinant protein production. Some of the rare examples analysing cellular responses due to protein overproduction are reported for the nonconventional yeasts *P. pastoris* (Sauer *et al.*, 2004; Gasser *et al.*, 2007a; Dragosits *et al.*, 2009) and *Kluyveromyces lactis* (van Ooyen *et al.*, 2006). The analysis of the cellular proteome during a fermentation of a chymosin expressing *K. lactis* strain indicated stress during protein production (upregulation of Hsp26p and Sod2p; van Ooyen *et al.*, 2006), whereas the *P. pastoris* work outlines how environmental factors such as temperature and pH affect protein expression and secretion on a transcriptomic (Sauer *et al.*, 2004; Gasser *et al.*, 2007a) and proteomic level (Dragosits *et al.*, 2009). Alternatively, metabolic flux analyses of protein-producing yeasts were performed. These focused, on the one hand, on the synthesis of high levels of intracellular human superoxide dismutase in *S. cerevisiae* (Gonzalez *et al.*, 2003), and, on the other, on core metabolic processes of *P. pastoris* during growth on glycerol and methanol (Solà *et al.*, 2004, 2007). However, apart from one exception, no strain improvement strategies resulted out of all these studies so far (Table 2).

The comparison of the differential transcriptome of a *P. pastoris* strain overexpressing human trypsinogen vs. a nonexpressing strain did reveal a network of genes being influenced due to the exploitation of the cellular expression machinery. This knowledge was further exploited to elucidate novel secretion helper factors that allowed the removal of bottlenecks in protein expression. Thirteen out of the 524 significantly regulated genes were selected and overexpressed in a *P. pastoris* strain producing a human antibody Fab fragment. Five previously characterized secretion helpers (Pdi1, Ero1, Sso2, Kar2/BiP and Hac1), as well as six novel, hitherto unidentified, factors, more precisely Bfr2 and Bmh2 involved in protein transport, the chaperones Ssa4 and Sse1, the vacuolar ATPase subunit Cup5 and Kin2, a protein kinase connected to exocytosis, increasing both specific production rates as well as volumetric productivity of an antibody fragment up to 2.5-fold in fed batch fermentations (Gasser *et al.*, 2007b). Very recently, a similar approach was leading to improved membrane protein production in *S. cerevisiae*, based on engineered expression of *BMS1*, involved in ribosomal subunit assembly (Bonander *et al.*, 2009). A convincing example of evolutionary engineering was based on random mutagenesis and screening for overproduction of human serum albumin in *S. cerevisiae*, followed by the identification of four genes related to Kar2 ATPase activity, which were upregulated in the selected mutant strain. Overexpression of these genes in other *S. cerevisiae* strains led to increased production of three different heterologous proteins (Payne *et al.*, 2008).

Protein folding and secretion appear to be major limitations for yeast expression systems, while the main concerns for other systems are growth, viability and metabolic burden (see Mammalian cells and Bacteria). For yeast production hosts these problems are not as crucial for the production of recombinant proteins, which can be regarded as one reason why systems biotechnology-based strain engineering has hardly been applied so far in this area.

Another important aspect that explains the lack of omics-based cell engineering in yeasts is the degree of availability of omics tools for yeast and other fungal species. Out of the 82 presently sequenced ascomycetes genomes, only 15% are biotechnologically used organisms, whereas the majority (54%) were pathogens, and the remaining were sequenced for comparative genomic studies (*Saccharomyces sensu stricto* group). The lack of published genome sequences is reflected in a lack of commercially available microarrays for most yeasts species. Some exceptions are arrays available for *S. cerevisiae* and *Schizosaccharomyces pombe* (Affymetrix and Agilent), or *Candida albicans* (Washington University, St. Louis). Proteomic studies are also hampered as they rely on annotated genome sequences for efficient performance. Alternatively, research groups performed transcriptional profiling by either heterologous hybridization to commercial *S. cerevisiae* arrays (e.g. for *P. pastoris*, Sauer *et al.*, 2004; and for *K. lactis*, Rosende *et al.*, 2008), or by designing custom microarrays. As an example, *P. pastoris* microarrays have been developed by our group, and are available for research applications (Graf *et al.*, 2008). While the first approach can only capture genes that are in common with *S. cerevisiae* [therefore excluding species-specific genes such as the assimilation pathways for methanol (*P. pastoris* and *Hansenula polymorpha*), hydrocarbons (*Yarrowia lipolytica*), or xylose (*Pichia stipitis*)], the latter often made the custom-made arrays unavailable for other groups, thereby limiting research activities in the field. Consistently, proteomics were mainly performed for pathogenic species (reviewed by Josic & Kovac, 2008).

Another drawback in the fungal kingdom is the high genetic diversity between the individual species, even within the phylum *Ascomycota*. The average sequence identity of orthologous proteins among the hemiascomycota is 50–60%, which is less than the *c*. 70% identity between man and fish, not to speak of 94% identity between man and mouse (Dujon, 2006). Accordingly, the DNA sequence identity among rodents is much higher, making heterologous omics between the hamster-derived Chinese hamster ovarian cells (CHO) or baby hamster kidney (BHK) cells and mouse or rat more feasible than among yeasts.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Table 2. Overview of systems biotechnological approaches for improved recombinant protein production in different hosts

Host	Analyses	Engineering	Results	References
E. coli	Transcriptome, proteome	Ribosomal genes, amino acid biosynthesis genes, IbpAB	Up to fourfold higher productivity	Reviewed in Park *et al*. (2005)
E. coli	Proteome	Controlled coexpression of PspA		Aldor *et al*. (2005)
E. coli	Proteome, transcriptome	Use of rare codons	Eight times more recombinant protein	Lee & Lee (2005)
E. coli	Proteome	New promoter (aldA)	30-fold higher product levels	Han *et al*. (2008)
E. coli	Secreted proteome	OsmY as fusion partner	High-level secretion	Qian *et al*. (2008)
B. megaterium	Metabolic fluxes	Pyruvate as carbon source	17-fold more secreted product, less protease activity	Fürch *et al*. (2007)
S. cerevisiae	Random mutagenesis	Co-chaperone genes related to Kar2 activity	1.5-fold increase of expression	Payne *et al*. (2008)
P. pastoris	Transcriptome	overexpression of secretion factors	2.5-fold increase of secretion	Gasser *et al*. (2007b)
A. niger	'Genomic methods' genome sequence and transcriptome	Disruption of protease genes	1.4-fold increased secretion	Wang *et al*. (2008)
A. niger	Transcriptome and proteome 'integrative genomics'	Knock-out of ERAD factor doaA and overexpression of oligosaccharyltransferase sstC	Improved intracellular production	Jacobs *et al*. (2009)
A. oryzae	Transcriptome	Knock-down of protease genes	1.2-fold increased secretion	Kimura *et al*. (2008)
CHO	Transcriptome	Overexpression of antiapoptotic and knock-down of proapoptotic genes	Higher viability leading to 2.5-fold higher titers	Wong *et al*. (2006)
CHO	Transcriptome	Stress markers for early clone screening	Time for clone establishment? Better and earlier clone selection	Trummer *et al*. (2008)
NS0	Transcriptome, proteome	Genes related to cholesterol dependence	Cholesterol-independent cell lines	Seth *et al*. (2006)

Filamentous fungi

A recent publication summarized 'The first 50 microarray studies in filamentous fungi', starting with an incomplete microarray for *Trichoderma reesei* in 2002 (Breakspear & Momany, 2007) and stated that for filamentous fungi, so far no engineering based on omics existed. However, in 2008 the first reports of proteome and transcriptome studies resulting in clear engineering strategies were published (Table 2): 132 protease genes were monitored during the production of human lysozyme in *Aspergillus oryzae* on a microarray and compared with degradation conditions, and upregulated protease cluster were identified. Out of these disruption targets three genes were already known to improve heterologous protein production, but the knock-down of one novel protease improved secreted yields of human lysozyme by 22% (Kimura *et al*., 2008). Wang *et al*. (2008) identified four protease genes of *Aspergillus niger* with genomic methods, which, upon disruption, led to increased protein secretion up to 40%. Another study determined the effect of enzyme overproduction in three different strains of *A. niger*, and extracted two engineering targets out of the upregulated genes and proteins involved in protein folding and the endoplasmic reticulum (ER)-associated protein degradation (ERAD) pathway. Combined engineering by knock-out of the ERAD factor doaA and overexpression of the oligosaccharyltransferase sttC led to improved production of a heterologous protein (Jacobs *et al*., 2009).

However, it should be noted that a number of genome-wide studies have revealed important aspects concerning protein production and its limitation in fungal expression systems. While not being applied directly, they have contributed significantly to the understanding of the protein production process, and led to strain engineering later on. The transcriptomic changes upon protein overexpression have been described for *Aspergillus nidulans* (Sims *et al*., 2005), *T. reesei* (Arvas *et al*., 2006) and *A. niger* (Guillemette *et al*., 2007), highlighting the impact of unfolded protein response (UPR) on protein folding, glycosylation, vesicle transport and ERAD, and identifying significant differences of UPR regulation between *S. cerevisiae* and filamentous fungi.

If we look beyond fungi, examples for systems biotechnological approaches in the field of recombinant protein production become more prevalent. Since the late 1990s, transcriptomics and proteomics were applied to bacterial and mammalian cultures used for heterologous protein production in order to elucidate cell physiology. Although numerous studies exist that describe the physiological

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

behaviour of cells to certain stresses – data also available for *S. cerevisiae* – activity beyond pure description is concentrated to a limited number of research groups. Recent reviews by these groups highlight that it is crucial to use state-of-the-art omics tools for physiological understanding and gaining insights into the host, as only detailed understanding of host cell physiology makes subsequent metabolic or cell engineering possible.

Mammalian cells

In mammalian cell culture most proteomic and transcriptomic analyses were performed to address problems or phenomena that have been observed previously on a 'macroscopic' level (e.g. metabolic shift, fed batch cultivation, apoptosis and stress conditions brought up by elevated osmolarity, sodium butyrate and low temperature). Since Korke *et al.* (2002) predicted the implementation of genomics and proteomics in cell culture engineering – for example for the selection of production cell lines (identification of gene regulation leading to adaptation to serum-free growth, or to adaptation to suspension growth) – and for bioprocess engineering both methodology and answers have evolved. Several excellent reviews summarizing these recent advances have been published during the last years (Griffin *et al.*, 2007; Gupta & Lee, 2007; Kuystermans *et al.*, 2007; Jaluria *et al.*, 2008). A list of papers dealing with genome-wide analysis of different mammalian cell lines (mainly murine myeloma cells NS0 and CHO) can be found in Kuystermans *et al.* (2007). In agreement with Gupta & Lee (2007), who pinpoint that a large number of omics approaches only generate lists of genes without direct application, we confirm that only four out of the 21 cited studies have resulted in an actual strategy for cell line engineering (Table 2).

Dinnis & James (2005) asked if one should learn 'lessons from nature' for engineering of antibody secreting mammalian cells. At least two lessons have been learned: induction of the UPR in order to reflect B-cell development (van Anken *et al.*, 2003) leads to increased secretion of several recombinant proteins (reviewed by Dinnis & James, 2005; Khan & Schroder, 2008). Another lesson that has been learned was the overexpression of anti-apoptotic genes, and knock-down of pro-apoptotic genes identified by microarray analysis of CHO cells, leading to prolonged cell viability and consequently up to 2.5-fold higher titres of interferon γ (Wong *et al.*, 2006).

Engineering mammalian cell culture based on genome-scale technologies was mainly applied to improve cell metabolism and growth (Griffin *et al.*, 2007), upstream cell culture conditions (e.g. temperature, hyperosmotic pressure and impact of small chemical compounds), downstream product quality (mainly assessed by proteomics), and cell culture media requirements (Gupta & Lee, 2007). As an example, the transcriptional analysis and proteomics did not stop at the identification of the responsible genes for the cholesterol dependence of NS0 cells. Subsequent engineering of the identified genes allowed cholesterol-independent cell growth (Seth *et al.*, 2006). Cross-species microarrays of high producer clones of EPO-Fc producing CHO lead to the identification of three ER stress marker genes correlated to insufficient resistance to shear stress in the early stage of clone selection before the respective phenotype could be observed (Trummer *et al.*, 2008).

Proteomic and transcriptional profiling of high- and low-productivity cell lines, or cells cultivated under conditions that lead to high specific productivity (q_P) (e.g. treatment with sodium butyrate and low-temperature cultivation) were carried out to discover the target genes leading to the super-secreting cells. Common features correlated to high productivity were the upregulation of secretory pathway proteins (especially chaperones and foldases) and cytoskeletal proteins in high-producing cell lines, as well as higher abundance of proteins belonging to the functional groups redox balance and vesicular transport. Additionally, decreased growth rate-related genes/proteins and decreased levels of stress genes were reported to occur in concordance with higher q_P (Kuystermans *et al.*, 2007; Seth *et al.*, 2007). While there were speculations that high-producing cell lines are likely to have a higher vesicle traffic and membrane recycling activity (Yee *et al.*, 2008), other attempts to identify correlations between single genes and improved secretory capacity failed. When trying to integrate all available genome-scale information of high-producing mammalian cell lines to find the genetic events leading to the super-secreting cells, it had to be concluded that there is no direct relation between a distinct set of genes and a trait, that there are no 'hyperproductivity master genes' (Seth *et al.*, 2007). On the contrary, multiple contributing pathways, even alternative pathways may lead to improved q_P Therefore the authors highlight the importance of data analysis approaches going beyond the identification of differentially expressed genes such as pattern discovery, pathway and network analysis in order to grasp the complexity of the gene–trait relationship (Seth *et al.*, 2007).

Bacteria

A general overview about whole systems level metabolic engineering in bacteria based on omics, including potential applications, was given by Park *et al.* (2005) and Gupta & Lee (2007). One common feature in bacteria, for example *Escherichia coli*, is that protein overproduction usually leads to a (severe) decrease in specific growth rate due to a shortage of energy and precursors.

Very early proteomic work in *E. coli* and *Bacillus subtilis* overproducing a heterologous protein accumulating in the cytoplasm as inclusion bodies showed that both species react to the recombinant protein with increased levels of heat

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

shock proteins and chaperones, whereas no clear picture regarding the regulation of ribosomal proteins emerged, as higher abundance in *B. subtilis*, and decreased levels in *E. coli* were observed (Jürgen *et al*., 2000, 2001). These studies can be seen as initiating a comprehensive understanding of the cellular responses to protein overproduction in bacteria, and although no direct engineering benefits were achieved, they gave rise to improvement of production strains (reviewed by Chou, 2007).

Since then various studies investigated the response of bacteria to several different proteins, but hardly any new hypotheses or applications evolved out of these studies. Some rare exceptions to this include the overexpression of ribosomal genes downregulated during insulin-like growth factor 1 expression in high cell density cultivation of *E. coli* leading to enhanced productivity and the engineering of small heat shock proteins IbpAB identified in inclusion bodies during overexpression of recombinant proteins in *E. coli* (all summarized in Park *et al*., 2005). By proteome profiling Aldor *et al*. (2005) identified the phage shock protein PspA to be coregulated with heterologous protein expression, and improved the yield by controlled coexpression of PspA. However, most studies are conducted on a case-by-case basis, and are not leading to improved production platforms. Alternatively, genome-wide analysis of cellular reactions to protein production may allow for the identification of marker genes that signal cellular stress as a response to protein overexpression. Their monitoring during protein production processes should allow to react on the bioprocess level before the stressful conditions, and consequently reduced cell growth and reduced viability will occur (Dürrschmid *et al*., 2008; Nemecek *et al*., 2008).

Interestingly, genome-wide analyses were also performed to analyse the behaviour of mutant strains with superior production characteristics. As an example, the proteome of a *Bacillus megaterium* chemical mutant exhibiting higher production levels of recombinant intracellular dextranesucrase and better cultivation behaviour, showed higher abundance of proteins related to protein synthesis and protein translocation (Wang *et al*., 2006). The observation of reduced levels of tRNA synthetases both on the proteomic and the mRNA level of an *E. coli* mutant secreting four times more α-haemolysin (HylA) in comparison with its parental strain led to an alternative metabolic engineering strategy, namely to use rare codons to slow down translation, which improved HylA secretion eight times in the parental strain (Lee & Lee, 2005).

In a different application of omics technologies, proteome analysis of *E. coli* in response to oleic acid was used to select oleic acid-inducible promoters for recombinant protein production. The use of the aldehyde dehydrogenase *aldA* promoter increased green fluorescent protein fluorescence intensity 30-fold compared with the IPTG-inducible tac promoter while applying the cheaper inducer, oleic acid (Han *et al*., 2008). On the other hand, a screen of the extracellular proteome of *E. coli* identified naturally secreted proteins as fusion partner for recombinant proteins in order to stimulate secretion. Out of 12 tested low-molecular-weight fusion partners, OsmY proved to be the best secretion partner resulting in high-level excretion of three model proteins into the culture supernatant of *E. coli* (Qian *et al*., 2008).

Alternatively, metabolic flux calculations can be carried out with the aim to identify bottlenecks in the fermentation that may need to be eliminated by genetic engineering. A recent study investigated the influence of two different carbon sources (glucose and pyruvate) on metabolic fluxes and productivity in *B. megaterium*, and concluded that pyruvate improves recombinant protein production 17-fold as less protease secretion and enhanced energy and reduction equivalent metabolism occurred. Additionally, the authors state that the overproduction of the recombinant protein increases the flux through the TCA and glycolysis, and reduces the flux through gluconeogenesis and the pentose phosphate pathway (Fürch *et al*., 2007).

New (post) genomic approaches to systems biotechnology

Systems biology as well as application-oriented systems biotechnology depend essentially on omics methods. A critical overview on current developments in this area and the potentials and pitfalls of current and upcoming methods is provided in the following, and summarized in Table 3.

Genomes

As already discussed in the previous section systems biotechnology has focused on certain organisms simply because their genome was sequenced and at least partly annotated. This is understandable because many omics methods can only be utilized to their full extent if the genome sequence is accessible and information about the positions of functional elements in the DNA is available. Furthermore, though the genome of *S. cerevisiae* has been very well studied, it is not a typical example for many yeast species that are used for protein production (Blank *et al*., 2005). Therefore, it is vital for systems biotechnology to create reference genomes with high-quality annotation of yeast species used in protein production. Sequencing technologies have made tremendous progress in the last few years, rendering the technology significantly cheaper, faster and more flexible than the traditional Sanger method, making it feasible for small scale studies with limited resources. With a reference genome at hand resequencing becomes an integral part of the workflow in systems biotechnology as shown in Fig. 3, thus expanding the systems biotechnology cycle. First studies applying the technology to selected mutants to understand the genetic

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Table 3. Critical summary of omics methods for systems biotechnology

Field	Methods	Advantages and disadvantages
Genomics	NGS	Fast and cheap method for whole-genome (re)sequencing without cloning bias
		Advantageous for mutation and subsequent strain analysis for inverse metabolic engineering purposes
Transcriptomics	Expression microarrays	Whole genome transcriptomics, cheap solution for in-house pipeline
		Susceptible to noise and bias
	Ref-Seq	Better correlation to qPCR results, large dynamic range limited only by sequencing depth
		Little background noise
		Can be used to detect splicing variants and 5′- and 3′-UTR boundaries
		Quantification is feasible even for mRNAs expressed at low levels
		Loss of strand-specific information
Proteomics	DIGE (gel-based systems)	Large number of different proteins over a large mass range can be detected
		Information about physicochemical properties
		Expensive and biased towards high-abundant proteins
		Membrane-bound and hydrophobic as well as small proteins cause problems
	MS	Mass and structure information of proteins
		Amino acid composition
		Detection of post-translational modifications
	Quantitative MS	Labelling (*in vivo* – SILAC, *in vitro* – ICAT, iTRAQ): increases the dynamic range of the analysis, but more expensive and detected proteins depend on the labelling method; *in vivo* labelling is not suitable for industrial processes
		Label-free quantification: quantification of a large number of proteins to characterize cells in different states; but less accurate and problematic to identify low-abundant proteins
	Protein Microarrays	Can only detect selected proteins due to lack of highly specific capture reagents and a lack in sensitivity
		Difficulties in retaining protein functionality
Interactomics	ChIP-chip	Regulatory DNA–protein binding interactions
		Chromatin packaging
	ChIP-Seq	Better resolution, less input material needed than ChIP-chip
		Usable for organisms without available genomic sequence
		Quantification is possible
Metabolomics	Metabolic modelling	Understanding regulatory pathways
		Identifying key players
		Simulation of system- wide reactions (either through logical networks or flux analysis) before biotechnological engineering is possible
		Creation of metabolic/signaling networks is complex and time consuming

background of their improved phenotype are already available for ethanol-producing *P. stipitis* (Smith *et al*., 2008).

Next generation sequencing (NGS)

Though NGS has a high potential of revolutionizing genetics, it comes with a set of pitfalls. All NGS methods create much shorter sequence reads (35–400 bp) than the Sanger method (*c*. 750–900 bp). This is especially a problem for *de novo* sequencing because even short repeats will make an assembly impossible, resulting in a high number of contigs. A comprehensive summary of the NGS technologies that are currently available was published by Mardis (2008). Third-generation sequencing (TGS) or also called next-next generation sequencing methods aim to further reduce the cost and run time of sequencing while improving the ease of handling the method. Additionally, most of these new technologies promise to have much longer reads than the NGS methods and thereby eliminating the problems related to short read length. Variants of currently pursued TGS technologies could be commercially available within the next 5 years (Gupta, 2009).

After the first excitement of the 1990s about the possibilities of sequencing and the completion of the human genome project in 2003, it was believed that the postgenomic era had begun. Now, a few years later, the picture looks somewhat different. It seems that with the emergence of NGS, the chapter of genomics has to be reopened again. At the moment, there are 873 completely sequenced genomes, of which about 83% are bacterial species having a rather small genome, and 4135 ongoing genome projects with a much lower proportion (50%) of bacterial species (http://www.genomesonline.org/gold.cgi). The rapidly increasing amount of available genomic data poses a challenging problem for data storage, management and interpretation, shifting the bottleneck towards bioinformatics, annotation and analysis tools.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

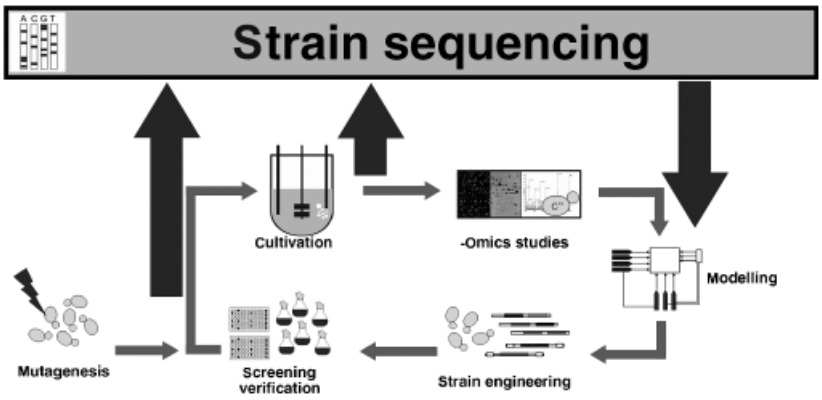


Fig. 3. A second genomics based systems biotechnology circle. NGS methods enable the resequencing of selected mutants to superimpose genotype on phenotype. Thus, evolutionary engineering and inverse metabolic engineering will gain enormous new potential.

Transcriptomics

Apart from *de novo* sequencing, NGS and TGS methods can be applied to many other questions that are relevant for the optimization of protein production, for example copy number variation, transcription factor binding, noncoding RNAs as well as expression profiles. Up to now DNA microarrays were the technologies of choice in the field of transcriptomics. For organisms for which microarrays are not available but that have an accessible genome sequence, bioinformatics tools (gene prediction and oligo design) make it possible to generate comprehensive expression data at a relatively low cost, as recently shown for *P. pastoris* (Graf *et al.*, 2008). Microarray experiments generally suffer from the existence of nonbiological variation, high background noise and only limited comparability due to a multitude of possible data-processing methods (Kawasaki, 2006). Sources for nonbiological variability and possible computational solutions have been extensively discussed in the literature (e.g. Draghici *et al.*, 2006; Johnson *et al.*, 2008). NGS solves many of the technical problems that microarrays suffer from but is still more expensive and many research groups and companies have an established microarray-processing pipeline in place. Also, some companies that offer NGS-sequencing machines offer microarrays as well, which indicates that a complementary use of both technologies is the most probable future.

Proteomics

Changes in the behaviour of a production system are often affected by post-translational modifications of proteins and therefore not visible on a transcriptome level. Besides there is no quantitative correlation between transcript expression levels and the amount of protein in the cell (Hartmann *et al.*, 2008). Analysing the genome and transcriptome alone is therefore not sufficient to understand or predict regulatory mechanism of cells or organisms. Unfortunately, though proteomics provides valuable insights for systems biotechnology, studies of changes during protein production are largely still missing (Josic & Kovac, 2008). The classical technique of proteome research, two-dimensional (2D) gel electrophoresis, suffers from a bias towards specific protein classes and towards highly abundant proteins (Bro & Nielsen, 2004) on the experimental level. On the data-processing level analysis of 2D gel images comprises many of the problems of microarray analysis, especially the difficulty to compare results due to differences between platforms and data-processing methods (Elrick *et al.*, 2006). In recent years, proteomics moved from a local (analysing a limited selection of proteins) to a global (analysing the whole proteome) technology. Quantitative mass spectroscopy has become the method of choice, and coupled with more sensitive labelling methods it facilitates high throughput proteome analysis (Elrick *et al.*, 2006). Stable isotope labelling by amino acids in cell culture (SILAC) is an *in vivo* labelling technique that requires the cells to be cultivated in media-containing labelled amino acids, thereby rendering it not feasible for large-scale production analyses. *In vitro* labelling techniques are more promising, although they still have shortcomings. In isotope-coded affinity tags (ICAT) cysteine residues are tagged. Because these residues are rare, it simplifies the peptide mixtures but proteins that do not contain cystein cannot be measured, and, furthermore, the small number of peptides for each protein compromises measurement reliability. Another *in vitro* labelling method is isobaric tags for relative and absolute quantification (iTRAQ). Here the N-terminus and sidechain amines are tagged with at least four different masses. Because these amines are more frequent, the cysteine-based restriction of ICAT is removed (Bachi & Bonaldi, 2008). Because all approaches involving isotopes are cost-intensive, label-free quantification methods using spectrum counts, integrated ion intensities or spectral feature analysis are sometimes preferable. The drawbacks of label-free methods include increased computational complexity due to lower accuracy and reproducibility of the data and the inability to quantify low-abundance proteins (Nesvizhskii *et al.*, 2007). With the move towards high throughput, it is essential to develop and

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Table 4. Standards for the publication of omics data

Short name	Designation	References or web addresses
MIAME	Minimum information about microarray experiments	http://www.mged.org/Workgroups/MIAME/miame.html
MIAPE	The minimum information about a proteomics experiment	http://www.psidev.info/index.php?q=node/91
MIRIAM	Minimum information requested in the annotation of biochemical models	http://www.ebi.ac.uk/compneur-srv/miriam/, proposed by Le Novère *et al.* (2005)
MIAMET	Minimum information on a metabolomics experiment	Proposed by Bino *et al.* (2004)

use the proper bioinformatics tools to efficiently process and statistically validate the generated data.

Metabolomics and metabolic modelling

Metabolomics and fluxomics go a step further and use information gained from the other omics to build a model of certain processes of the cell or ideally of the whole cell. Whereas metabolomics focuses on the metabolites involved, fluxomics predicts flux distributions within the cell based on measured rates of metabolites and their mass balance (Kim *et al.*, 2008). Several types of metabolic networks exist. Stoichiometric models and dynamic or kinetic models are the more traditional approach for which comprehensive knowledge about the players and their relationship is necessary. These models consist of a relatively small number of reactions or elements and their quality depends to a large extent on the quality of experimental data (Steuer, 2007). The advantage is of course that, if a high-quality model is achieved, reasonable predictions about phenotypic behaviour can be made. Early metabolic networks were limited to a few pathways of the core metabolism, but the availability of genome sequences and extensive omics data made it possible to fully describe the core metabolism and move to new areas such as signalling (Arga *et al.*, 2007) or lipid metabolism (Nookaew *et al.*, 2008). Protein expression and secretion networks are still lacking. Some of the reasons why this topic was avoided by scientists for so long are listed in Table 1. Despite these pitfalls, it can be anticipated that model development of the protein production pathways will significantly contribute to a better quantitative understanding of the contributing reactions and their relevance.

Topological networks on the other hand consist of many nodes that represent genes or proteins and edges representing the connection between those nodes. Such models can be much larger than stoichiometric or kinetic models and can make use of high throughput data but are only static descriptions and contain no information about the type, time or place of interaction between two nodes. At least the type of interaction can be modelled using control logic, described by Schlitt & Brazma (2005). With a sequenced and annotated genome at hand, a topological or control logic network can be computed for all known cellular functions. The challenge now is to combine the different approaches into a predictive dynamic model of the whole cell. The first such approach was taken for *A. niger* (Andersen *et al.*, 2008), while Herrgård *et al.* (2008) took the first step towards such a model for *S. cerevisiae* by combining two existing metabolic models (at the control logic state) into a consensus yeast metabolic network and implementing it in systems biology markup language (SBML), which is a widely used data format for metabolic networks.

With the pace at which omics methods develop and the amount of data they produce, it is important to keep in mind that each of the omics fields only shows us an isolated part of the picture. To improve our understanding of the function of organisms, it is essential to be able to give meaning to the data representing concentrations of genes, proteins and metabolites. Therefore, it will become even more important in the future to merge information from different omics sources into a coherent picture. More effort has to be put in developing methods that can integrate and validate these data as well as help managing the fast-increasing flood of information. This in mind, several standards have been developed to avoid a confusing mess of data sources, treatment and analysis variants (Table 4). While these standards define a minimal frame on experimental and computational quality and data deposition, they cannot cope with systematic differences between different omics platforms. These systematic differences call for cautiousness in data comparison between omics platforms and data-processing methods. It should be emphasized that the aim of standardization is not to define data quality, but rather the information on data assessment, and to guarantee that raw data are made available to the scientific community.

Conclusions

Current status of systems biotechnology for protein production

Systems biotechnology has proven its value for strain design and optimization. System-wide approaches to complex cellular processes such as protein production are still in their infancy. Interestingly, more progress in this field has been made for bacteria and mammalian cells than for yeasts. This may be due to a lack of genomic information and postgenomic tools for industrially relevant yeast species. The

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

revival of genomics through NGS methods is about to close this gap, and a critical review of the state of research with bacterial and cell culture host systems provides guidance as to where to direct research with yeasts in this field.

Lessons to learn from non-yeast expression systems

At present, most limitations in protein production in yeasts are attributed to bottlenecks during folding and secretion. Thus, engineering of yeast protein factories mainly means knowledge-based engineering of chaperones and ER resident folding catalysts. While these approaches were verified by transcriptomic profiling of yeasts and other fungal production hosts recently, they may not convey the full picture. As can be seen in non-yeast expression systems, the application of integrated genomic approaches allows looking beyond the borders of the secretory compartments during the production of recombinant proteins. While engineering strategies leading to higher viability and higher stress resistance in mammalian cells may not be directly applied to yeast systems, they reveal the potential that apparently unconnected cellular processes can be manipulated in order to increase protein yields/productivity. Bacterial studies revealed a shortage of ribosomes, energy and precursors during recombinant protein production, a possibility yet underestimated for fungal hosts. The availability of improved metabolic models – as they exist for bacteria – for *A. niger* and *S. cerevisiae*, and their applicability for related fungal species, may allow a considerable progress regarding the behaviour of the core metabolism and energy supply during protein overexpression. Alternatively, production processes can be monitored by methods of metabolic flux analysis and controlled to improve protein yields.

The importance of systems level screening during clone selection in mammalian cells may also be converted to fungal production hosts, as soon as the respective genetic/molecular traits leading to high secretory capacity are identified. Once again, the importance of thorough and comparative data analysis should be stressed in this respect, as the pathways identified upon protein overproduction in yeasts and filamentous fungi are very similar to those regulated in mammalian cells.

With the advent of cheap and fast sequencing technologies, and consequently higher coverage of biotechnologically relevant fungal species, resequencing of improved mutant strains and subsequent inverse metabolic engineering also become feasible. Additionally, this information may contribute to the big search for the 'holy grail' of protein production – the 'hyperproductivity master genes' – or at least lead to a better understanding of the cellular pathways influencing productivity.

Finally, the impact of systems biotechnology on the improvement of bioprocess performance, for example by the identification of novel stress markers such as those shown in *E. coli*, better strain performance (mammalian cells) or prevention of proteolytic degradation by disruption of cellular proteases as reported for filamentous fungi and bacterial systems should be highlighted. All these approaches can be readily transferred to fungal production processes by applying the respective systems biotechnological tools.

In contrast to mammalian and bacterial protein production, application of systems level approaches for the targeted engineering of yeasts and other fungal production hosts is still at an early stage. However, expectations are high that the recent advances emerging in the fungal field will just be the beginning of the 'systems biotechnological' age for improved protein production strains.

Outlook

A major focus of future work should be the quantitative understanding of molecular principles behind protein synthesis, modification and secretion, derived from basic production strains as well as mutants and rationally engineered strains.

NGS methods provide the tool to rationalize inverse metabolic engineering approaches so that they can be implemented in future into rational system-wide modelling and optimization strategies.

Acknowledgements

Research on heterologous protein production in yeasts in our laboratory is supported by the Austrian Science Fund (project no. I37-B03), the European Science Foundation (programme EuroSCOPE), the Austrian Research Promotion Agency (programme FHplus), Polymun Scientific GmbH and Boehringer Ingelheim Austria GmbH.

References

Aldor I, Krawitz D, Forrest W, Chen C, Nishihara J, Joly J & Champion K (2005) Proteomic profiling of recombinant *Escherichia coli* in high-cell-density fermentations for improved production of an antibody fragment biopharmaceutical. *Appl Environ Microb* **71**: 1717–1728.

Alper H, Moxley J, Nevoigt E, Fink G & Stephanopoulos G (2006) Engineering yeast transcription machinery for improved ethanol tolerance and production. *Science* **314**: 1565–1568.

Andersen M, Nielsen M & Nielsen J (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol Syst Biol* **4**: 178.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Arga K, Onsan Z, Kirdar B, Ulgen K & Nielsen J (2007) Understanding signaling in yeast: insights from network analysis. *Biotechnol Bioeng* **97**: 1246–1258.

Arvas M, Pakula T, Lanthaler K *et al.* (2006) Common features and interesting differences in transcriptional responses to secretion stress in the fungi *Trichoderma reesei* and *Saccharomyces cerevisiae*. *BMC Genomics* **7**: 32.

Bachi A & Bonaldi T (2008) Quantitative proteomics as a new piece of the systems biology puzzle. *J Proteomics* **71**: 357–367.

Bailey J (1991) Toward a science of metabolic engineering. *Science* **252**: 1668–1675.

Bailey J, Sburlati A, Hatzimanikatis V, Lee K, Renner W & Tsai P (1996) Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes. *Biotechnol Bioeng* **52**: 109–121.

Beltran G, Novo M, Leberre V *et al.* (2006) Integration of transcriptomic and metabolic analyses for understanding the global responses of low-temperature winemaking fermentations. *FEMS Yeast Res* **6**: 1167–1183.

Bino R, Hall R, Fiehn O *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* **9**: 418–425.

Blank L, Lehmbeck F & Sauer U (2005) Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS Yeast Res* **5**: 545–558.

Blomberg A (1995) Global changes in protein synthesis during adaptation of the yeast *Saccharomyces cerevisiae* to 0.7 M NaCl. *J Bacteriol* **177**: 3563–3572.

Bonander N, Darby RAJ, Grgic L *et al.* (2009) Altering the ribosomal subunit ratio in yeast maximizes recombinant protein yield. *Microb Cell Fact* **8**: 10.

Breakspear A & Momany M (2007) The first fifty microarray studies in filamentous fungi. *Microbiology* **153**: 7–15.

Bro C & Nielsen J (2004) Impact of 'ome' analyses on inverse metabolic engineering. *Metab Eng* **6**: 204–211.

Chen D, Toone W, Mata J *et al.* (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell* **14**: 214–229.

Chou C (2007) Engineering cell physiology to enhance recombinant protein production in *Escherichia coli*. *Appl Microbiol Biot* **76**: 521–532.

Dinnis D & James D (2005) Engineering mammalian cell factories for improved recombinant monoclonal antibody production: lessons from nature? *Biotechnol Bioeng* **91**: 180–189.

Draghici S, Khatri P, Eklund A & Szallasi Z (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* **22**: 101–109.

Dragosits M, Stadlmann J, Albiol J *et al.* (2009) The effect of temperature on the proteome of recombinant *Pichia pastoris*. *J Proteome Res* **8**: 1380–1392.

Dujon B (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22**: 375–387.

Dürrschmid K, Reischer H, Schmidt-Heck W, Hrebicek T, Guthke R, Rizzi A & Bayer K (2008) Monitoring of transcriptome and proteome profiles to investigate the cellular response of E. coli towards recombinant protein expression under defined chemostat conditions. *J Biotechnol* **135**: 34–44.

Edwards J & Palsson B (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *P Natl Acad Sci USA* **97**: 5528–5533.

Elrick M, Walgren J, Mitchell M & Thompson D (2006) Proteomics: recent applications and new technologies. *Basic Clin Pharmacol* **98**: 432–441.

Fernandes P, Domitrovic T, Kao C & Kurtenbach E (2004) Genomic expression pattern in *Saccharomyces cerevisiae* cells in response to high hydrostatic pressure. *FEBS Lett* **556**: 153–160.

Förster J, Famili I, Fu P, Palsson B & Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* **13**: 244–253.

Fürch T, Wittmann C, Wang W, Franco-Lara E, Jahn D & Deckwer W (2007) Effect of different carbon sources on central metabolic fluxes and the recombinant production of a hydrolase from *Thermobifida fusca* in *Bacillus megaterium*. *J Biotechnol* **132**: 385–394.

Gasch A, Spellman P, Kao C *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**: 4241–4257.

Gasser B, Maurer M, Rautio J *et al.* (2007a) Monitoring of transcriptional regulation in *Pichia pastoris* under protein production conditions. *BMC Genomics* **8**: 179.

Gasser B, Sauer M, Maurer M, Stadlmayr G & Mattanovich D (2007b) Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts. *Appl Environ Microb* **73**: 6499–6507.

Gonzalez R, Andrews B, Molitor J & Asenjo J (2003) Metabolic analysis of the synthesis of high levels of intracellular human SOD in *Saccharomyces cerevisiae* rhSOD 2060 411 SGA122. *Biotechnol Bioeng* **82**: 152–169.

Gori K, Hébraud M, Chambon C, Mortensen H, Arneborg N & Jespersen L (2007) Proteomic changes in *Debaryomyces hansenii* upon exposure to NaCl stress. *FEMS Yeast Res* **7**: 293–303.

Graf A, Gasser B, Dragosits M *et al.* (2008) Novel insights into the unfolded protein response using *Pichia pastoris* specific DNA microarrays. *BMC Genomics* **9**: 390.

Griffin T, Seth G, Xie H, Bandhakavi S & Hu W (2007) Advancing mammalian cell culture engineering using genome-scale technologies. *Trends Biotechnol* **25**: 401–408.

Guillemette T, van Peij N, Goosen T *et al.* (2007) Genomic analysis of the secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. *BMC Genomics* **8**: 158.

Gupta P (2009) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26**: 602–611.

Gupta P & Lee K (2007) Genomics and proteomics in process development: opportunities and challenges. *Trends Biotechnol* **25**: 324–330.

Han M, Lee J, Lee S & Yoo J (2008) Proteome-level responses of *Escherichia coli* to long-chain fatty acids and use of fatty acid

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

inducible promoter in protein production. *J Biomed Biotechnol* **2008**: 735101.

Hartmann M, Roeraade J, Stoll D, Templin M & Joos T (2008) Protein microarrays for diagnostic assays. *Anal Bioanal Chem* **393**: 1407–1416.

Herrgård M, Swainston N, Dobson P *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* **26**: 1155–1160.

Hohenblum H, Borth N & Mattanovich D (2003) Assessing viability and cell-associated product of recombinant protein producing *Pichia pastoris* with flow cytometry. *J Biotechnol* **102**: 281–290.

Jacobs DI, Olsthoorn MM, Maillet I *et al.* (2009) Effective lead selection for improved protein production in *Aspergillus niger* based on integrated genomics. *Fungal Genet Biol* **46**: 141–152.

Jahic M, Wallberg F, Bollok M, Garcia P & Enfors S (2003) Temperature limited fed-batch technique for control of proteolysis in *Pichia pastoris* bioreactor cultures. *Microb Cell Fact* **2**: 6.

Jaluria P, Chu C, Betenbaugh M & Shiloach J (2008) Cells by design: a mini-review of targeting cell engineering using DNA microarrays. *Mol Biotechnol* **39**: 105–111.

Johnson D, Li W, Gordon D *et al.* (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* **18**: 393–403.

Josic D & Kovac S (2008) Application of proteomics in biotechnology – microbial proteomics. *Biotechnol J* **3**: 496–509.

Jürgen B, Lin H, Riemschneider S *et al.* (2000) Monitoring of genes that respond to overproduction of an insoluble recombinant protein in *Escherichia coli* glucose-limited fed-batch fermentations. *Biotechnol Bioeng* **70**: 217–224.

Jürgen B, Hanschke R, Sarvas M, Hecker M & Schweder T (2001) Proteome and transcriptome based analysis of *Bacillus subtilis* cells overproducing an insoluble heterologous protein. *Appl Microbiol Biot* **55**: 326–332.

Kawasaki E (2006) The end of the microarray Tower of Babel: will universal standards lead the way? *J Biomol Tech* **17**: 200–206.

Khan SU & Schroder M (2008) Engineering of chaperone systems and of the unfolded protein response. *Cytotechnology* **57**: 207–231.

Kim T, Sohn S, Kim H & Lee S (2008) Strategies for systems-level metabolic engineering. *Biotechnol J* **3**: 612–623.

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

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

Kobi D, Zugmeyer S, Potier S & Jaquet-Gutfreund L (2004) Two-dimensional protein map of an "ale"-brewing yeast strain: proteome dynamics during fermentation. *FEMS Yeast Res* **5**: 213–230.

Kolkman A, Daran-Lapujade P, Fullaondo A, Olsthoorn M, Pronk J, Slijper M & Heck A (2006) Proteome analysis of yeast response to various nutrient limitations. *Mol Syst Biol* **2**: 2006. 0026.

Korke R, Rink A, Seow T, Chung M, Beattie C & Hu W (2002) Genomic and proteomic perspectives in cell culture engineering. *J Biotechnol* **94**: 73–92.

Kuystermans D, Krampe B, Swiderek H & Al-Rubeai M (2007) Using cell engineering and omic tools for the improvement of cell culture processes. *Cytotechnology* **53**: 3–22.

Lee P & Lee K (2005) Engineering HlyA hypersecretion in *Escherichia coli* based on proteomic and microarray analyses. *Biotechnol Bioeng* **89**: 195–205.

Lee S, Lee D & Kim T (2005) Systems biotechnology for strain improvement. *Trends Biotechnol* **23**: 349–358.

Le Novère N, Finney A, Hucka M *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**: 1509–1515.

Li Z, Xiong F, Lin Q, d'Anjou M, Daugulis A, Yang D & Hew C (2001) Low-temperature increases the yield of biologically active herring antifreeze protein in *Pichia pastoris*. *Protein Expres Purif* **21**: 438–445.

Mardis E (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**: 133–141.

Mattanovich D & Borth N (2006) Applications of cell sorting in biotechnology. *Microb Cell Fact* **5**: 12.

Mattanovich D, Gasser B, Hohenblum H & Sauer M (2004) Stress in recombinant protein producing yeasts. *J Biotechnol* **113**: 121–135.

Mukhopadhyay A, Redding A, Rutherford B & Keasling J (2008) Importance of systems biology in engineering microbes for biofuel production. *Curr Opin Biotech* **19**: 228–234.

Nemecek S, Marisch K, Juric R & Bayer K (2008) Design of transcriptional fusions of stress sensitive promoters and GFP to monitor the overburden of *Escherichia coli* hosts during recombinant protein production. *Bioproc Biosyst Eng* **31**: 47–53.

Nesvizhskii A, Vitek O & Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Meth* **4**: 787–797.

Nevoigt E (2008) Progress in metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol Mol Biol R* **72**: 379–412.

Nielsen J (2001) Metabolic engineering. *Appl Microbiol Biot* **55**: 263–283.

Nielsen J & Jewett M (2008) Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*. *FEMS Yeast Res* **8**: 122–131.

Nookaew I, Jewett M, Meechai A *et al.* (2008) The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC Syst Biol* **2**: 71.

Ostergaard S, Olsson L & Nielsen J (2000) Metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol Mol Biol R* **64**: 34–50.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

Park S, Lee S, Cho J, Kim T, Lee J, Park J & Han M (2005) Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Appl Microbiol Biot* **68**: 567–579.

Payne T, Finnis C, Evans LR, Mead DJ, Avery SV, Archer DB & Sleep D (2008) Modulation of chaperone gene expression in mutagenized *Saccharomyces cerevisiae* strains developed for recombinant human albumin production results in increased production of multiple heterologous proteins. *Appl Environ Microb* **74**: 7759–7766.

Pizarro F, Vargas F & Agosin E (2007) A systems biology perspective of wine fermentations. *Yeast* **24**: 977–991.

Porro D, Sauer M, Branduardi P & Mattanovich D (2005) Recombinant protein production in yeasts. *Mol Biotechnol* **31**: 245–259.

Qian Z, Xia X, Choi J & Lee S (2008) Proteome-based identification of fusion partner for high-level extracellular production of recombinant proteins in *Escherichia coli*. *Biotechnol Bioeng* **101**: 587–601.

Rosende SS, Becerra M, Salgado MT, Lamas-Maceiras M, González M & Picos MAF (2008) Growth phase-dependent expression of *Kluyveromyces lactis* genes and involvement of 3′-UTR elements. *Process Biochem* **43**: 1153–1157.

Sauer M, Branduardi P, Gasser B, Valli M, Maurer M, Porro D & Mattanovich D (2004) Differential gene expression in recombinant *Pichia pastoris* analysed by heterologous DNA microarray hybridisation. *Microb Cell Fact* **3**: 17.

Schlitt T & Brazma A (2005) Modelling gene networks at different organisational levels. *FEBS Lett* **579**: 1859–1866.

Seth G, Ozturk M & Hu W (2006) Reverting cholesterol auxotrophy of NS0 cells by altering epigenetic gene silencing. *Biotechnol Bioeng* **93**: 820–827.

Seth G, Charaniya S, Wlaschin K & Hu W (2007) In pursuit of a super producer-alternative paths to high producing recombinant mammalian cells. *Curr Opin Biotechnol* **18**: 557–564.

Shi X, Karkut T, Chamankhah M, Alting-Mees M, Hemmingsen S & Hegedus D (2003) Optimal conditions for the expression of a single-chain antibody (scFv) gene in *Pichia pastoris*. *Protein Expres Purif* **28**: 321–330.

Sims AH, Gent ME, Lanthaler K, Dunn-Coleman NS, Oliver SG & Robson GD (2005) Transcriptome analysis of recombinant protein secretion by *Aspergillus nidulans* and the unfolded-protein response *in vivo*. *Appl Environ Microb* **71**: 2737–2747.

Smith D, Quinlan A, Peckham H *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* **18**: 1638–1642.

Solà A, Maaheimo H, Ylönen K, Ferrer P & Szyperski T (2004) Amino acid biosynthesis and metabolic flux profiling of *Pichia pastoris*. *Eur J Biochem* **271**: 2462–2470.

Solà A, Jouhten P, Maaheimo H, Sánchez-Ferrando F, Szyperski T & Ferrer P (2007) Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology* **153**: 281–290.

Sonderegger M & Sauer U (2003) Evolutionary engineering of *Saccharomyces cerevisiae* for anaerobic growth on xylose. *Appl Environ Microb* **69**: 1990–1998.

Steuer R (2007) Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry* **68**: 2139–2151.

Tai S, Daran-Lapujade P, Walsh M, Pronk J & Daran J (2007) Acclimation of *Saccharomyces cerevisiae* to low temperature: a chemostat-based transcriptome analysis. *Mol Biol Cell* **18**: 5100–5112.

Takors R, Bathe B, Rieping M, Hans S, Kelle R & Huthmacher K (2007) Systems biology for industrial strains and fermentation processes – example: amino acids. *J Biotechnol* **129**: 181–190.

Trummer E, Ernst W, Hesse F *et al.* (2008) Transcriptional profiling of phenotypically different Epo-Fc expressing CHO clones by cross-species microarray analysis. *Biotechnol J* **3**: 924–937.

van Anken E, Romijn E, Maggioni C, Mezghrani A, Sitia R, Braakman I & Heck A (2003) Sequential waves of functionally related proteins are expressed when B cells prepare for antibody secretion. *Immunity* **18**: 243–253.

van Ooyen A, Dekker P, Huang M, Olsthoorn M, Jacobs D, Colussi P & Taron C (2006) Heterologous protein production in the yeast *Kluyveromyces lactis*. *FEMS Yeast Res* **6**: 381–392.

Wang W, Hollmann R & Deckwer W (2006) Comparative proteomic analysis of high cell density cultivations with two recombinant *Bacillus megaterium* strains for the production of a heterologous dextransucrase. *Proteome Sci* **4**: 19.

Wang Y, Xue W, Sims AH *et al.* (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.

Westerhoff H & Palsson B (2004) The evolution of molecular biology into systems biology. *Nat Biotechnol* **22**: 1249–1252.

Wong D, Wong K, Nissom P, Heng C & Yap M (2006) Targeting early apoptotic genes in batch and fed-batch CHO cell cultures. *Biotechnol Bioeng* **95**: 350–361.

Yee J, de Leon Gatti M, Philp R, Yap M & Hu W (2008) Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol Bioeng* **99**: 1186–1204.

Zuzuarregui A, Monteoliva L, Gil C & del Olmo M (2006) Transcriptomic and proteomic approach for understanding the molecular basis of adaptation of *Saccharomyces cerevisiae* to wine fermentation. *Appl Environ Microb* **72**: 836–847.

© 2009 Federation of European Microbiological Societies
Published by Blackwell Publishing Ltd. All rights reserved

6.2 Research Publications

6.2.1 First or equally contributing author papers

Graf A, Gasser B, Dragosits M, Sauer M, Leparc GG, Tüchler T, Kreil DP, Mattanovich D. *Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.* BMC Genomics 2008, August **9**:390

Contribution: I was responsible for the gene prediction and functional annotation in the commercial draft sequence of *P. pastoris*, I conducted the oligo design and was, to a minor part, involved in the sample preparation, and microarray design. Furthermore, I was responsible for the statistical data analysis, and evaluation of the results.

Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, Mattanovich D, Lee SY. Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. Biotechnology Journal 2010, Jul;5(7):705-15.

Contribution: I worked on the annotation of the genome (as described in [Mattanovich et al., 2009a]), and was responsible for the gene to protein to reaction relationship on which the model is based. This included the comparison of the two sequenced *P. pastoris* strains. I coordinated the project on the side of the University of Natural Ressources and Life Sciences Vienna, and was responsible for data quality.

BMC Genomics

BioMed Central

Research article

Open Access

Novel insights into the unfolded protein response using *Pichia pastoris* specific DNA microarrays

Alexandra Graf[†1], Brigitte Gasser[†1], Martin Dragosits[1], Michael Sauer[2], Germán G Leparc[3], Thomas Tüchler[3], David P Kreil[3] and Diethard Mattanovich[*1,2]

Address: [1]Institute of Applied Microbiology, Department of Biotechnology, University of Natural Resources and Applied Life Sciences Vienna, Muthgasse 18, 1190 Vienna, Austria, [2]School of Bioengineering, University of Applied Sciences FH Campus Vienna, Muthgasse 18, 1190 Vienna, Austria and [3]Vienna Science Chair of Bioinformatics, Department of Biotechnology, University of Natural Resources and Applied Life Sciences Vienna, Muthgasse 18, 1190 Vienna, Austria

Email: Alexandra Graf - alexandra.graf@boku.ac.at; Brigitte Gasser - brigitte.gasser@boku.ac.at; Martin Dragosits - martin.dragosits@boku.ac.at; Michael Sauer - michael.sauer@fh-campuswien.ac.at; Germán G Leparc - german.leparc@boku.ac.at; Thomas Tüchler - thomas.tuechler@boku.ac.at; David P Kreil - pichia08@kreil.org; Diethard Mattanovich* - diethard.mattanovich@boku.ac.at

* Corresponding author †Equal contributors

Published: 19 August 2008

BMC Genomics 2008, **9**:390 doi:10.1186/1471-2164-9-390

Received: 8 February 2008
Accepted: 19 August 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/390

© 2008 Graf et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: DNA Microarrays are regarded as a valuable tool for basic and applied research in microbiology. However, for many industrially important microorganisms the lack of commercially available microarrays still hampers physiological research. Exemplarily, our understanding of protein folding and secretion in the yeast *Pichia pastoris* is presently widely dependent on conclusions drawn from analogies to *Saccharomyces cerevisiae*. To close this gap for a yeast species employed for its high capacity to produce heterologous proteins, we developed full genome DNA microarrays for *P. pastoris* and analyzed the unfolded protein response (UPR) in this yeast species, as compared to *S. cerevisiae*.

Results: By combining the partially annotated gene list of *P. pastoris* with *de novo* gene finding a list of putative open reading frames was generated for which an oligonucleotide probe set was designed using the probe design tool TherMODO (a thermodynamic model-based oligoset design optimizer). To evaluate the performance of the novel array design, microarrays carrying the oligo set were hybridized with samples from treatments with dithiothreitol (DTT) or a strain overexpressing the UPR transcription factor *HAC1*, both compared with a wild type strain in normal medium as untreated control. DTT treatment was compared with literature data for *S. cerevisiae*, and revealed similarities, but also important differences between the two yeast species. Overexpression of *HAC1*, the most direct control for UPR genes, resulted in significant new understanding of this important regulatory pathway in *P. pastoris*, and generally in yeasts.

Conclusion: The differences observed between *P. pastoris* and *S. cerevisiae* underline the importance of DNA microarrays for industrial production strains. *P. pastoris* reacts to DTT treatment mainly by the regulation of genes related to chemical stimulus, electron transport and respiration, while the overexpression of *HAC1* induced many genes involved in translation, ribosome biogenesis, and organelle biosynthesis, indicating that the regulatory events triggered by DTT treatment only partially overlap with the reactions to overexpression of *HAC1*. The high reproducibility of the results achieved with two different oligo sets is a good indication for their robustness, and underlines the importance of less stringent selection of regulated features, in order to avoid a large number of false negative results.

Background

Transcriptomics, the parallel quantification of many, or all transcripts of an organism in given conditions, has become a favorite tool for basic research [1]. Messenger-RNA regulation patterns of model organisms under many different conditions have become available during the last years. However, these methods are still not applicable for many industrially important organisms, mainly due to the lack of DNA microarrays targeting these organisms. A typical example is the yeast *Pichia pastoris*, which is widely applied for the production of recombinant proteins. Several approaches have been taken to derive transcriptomic data without specific microarrays. Sauer et al. [2] have applied heterologous hybridization of *P. pastoris* samples to *Saccharomyces cerevisiae* microarrays. Alternative methodological concepts like Transcript Analysis with the Aid of Affinity Capture (TRAC) [3] may be applied preferentially to subsets of the transcriptome [4], provided that genome sequence data are available. If this is not the case, total cDNA may be utilized as a source of probes, either by applying expressed sequence tags to microarrays [5] or employing RNA fingerprinting like cDNA-amplified fragment length polymorphism (cDNA-AFLP) [6], which has recently been applied to *Trichoderma reesei* [7]. These unannotated methods bear of course the disadvantage that specific hits may only be identified after sequencing their respective probes.

Therefore oligonucleotide microarrays have become the method of choice for many applications, although their design depends on the availability of a genomic sequence with good gene identification and annotation. The genome sequence of *P. pastoris* is not published yet. The data available from Integrated Genomics (IG, Chicago, IL, USA; [8]) contain a partial gene identification and annotation, so that additional effort in this direction was a first step necessary towards development of comprehensive DNA microarrays for this yeast species. There is a wide choice of computational gene finders available at the moment which can be classified into intrinsic and extrinsic prediction programs. Intrinsic or *de novo* gene finder only use information from the sequences to be studied, building statistical models to distinguish between coding and non-coding regions of the genome on the basis of biological sequence patterns [9-11]. Extrinsic gene finder utilize homology search to determine where protein coding regions are in the genome. Their applicability is therefore limited to organisms that have homologs in current databases that are correctly annotated. Because of this limitation it is common to integrate homology search with *de novo* prediction [12]. Most state of the art gene finders use a form of Hidden Markov Model (HMM) differing in the implementation and complexity of the model as well as the ease in which users can adapt the application to their needs [13].

It is well known that cross-hybridization can confound microarray results rendering good probe design an essential requirement for accurate microarray analyses. The specificity of oligonucleotides is determined by the Gibbs free energy (ΔG) of the hybridization reaction between potential binding partners. Highly specific probes will bind their target transcript much more strongly than any other transcript. Considering that microarray experiments are non-equilibrium measurements, it is desirable that microarray probes exhibit uniform thermodynamic properties, which many probe design tools aim to achieve by demanding a narrow distribution of the probe-target melting temperature T_m. Ideally, probes should have a uniform binding free energy at the hybridization temperature T_{hyb} [14].

Previous studies have demonstrated that industrial production strains may behave quite differently to laboratory strains and model organisms [15], which emphasizes the importance of analytical tools for industrially relevant strains and species. As an example, the unfolded protein response (UPR), a regulation circuit of high relevance for heterologous protein production in eukaryotic cells [16], has been shown to be differentially regulated in *P. pastoris* [4] compared to *S. cerevisiae* [17], which is the typical model species for hemiascomycete yeasts. The development of specific microarrays for *P. pastoris* was intended to allow a detailed analysis of UPR regulation in *P. pastoris*. As in previous transcriptomics work with *S. cerevisiae* the induction of UPR was either accomplished by addition of dithiothreitol (DTT) or tunicamycin, this work aimed at a comparison of DTT induced gene regulation in *P. pastoris* to that in *S. cerevisiae* published by Travers et al. [17]. Finally we aimed at the comparison of DTT induced regulation to the regulatory response to overexpression of *HAC1*, the transcription factor controlling the UPR. Transcriptional regulation of *HAC1* overexpression has not been studied for yeasts so far, so that we expected valuable data to better define the core UPR regulated transcriptome.

Results and Discussion

Gene prediction and Oligo Design

To evaluate available gene finders for their performance on yeast genomes, three *de novo* gene finders (GeneMark, Glimmer3, GlimmerHMM) were tested on the genome sequence of *S. cerevisiae*. GeneMark and Glimmer3 work with a prokaryotic Hidden Markov Model (HMM) whereas GlimmerHMM employs a eukaryotic gene model. GeneMark was trained with coding and non-coding sequences of *S. cerevisiae*, building an HMM transition probability matrix of the 7th order. Glimmer3 and GlimmerHMM could be trained directly on the genome in question without specifying coding and non-coding regions. In Lomsadze et al. [18] and Besemer and Boro-

dovsky [9] the difficulty of eukaryotic gene finders in the prediction of genes for organisms with few introns is discussed and linked to a lack of data for representative exon – intron models. Our results confirmed that a gene finder written for eukaryotes (GlimmerHMM) could not be trained well on yeast genomes, introducing far too many introns into the predicted genes. Both prokaryotic versions performed much better, with GeneMark predicting less false negatives but more false positives than Glimmer3 (Table 1). Even though the positive prediction value was somewhat lower with GeneMark it was more important not to miss true positives than to achieve a lower rate of false positives. A further improvement could be achieved by a GeneMark model for lower eukaryotes, in which the prokaryotic algorithm is modified to use Kozak start sites instead of prokaryotic ribosomal binding sites. *P. pastoris* genes were predicted using this version of GeneMark with the lowest possible threshold (probability score t = 0.05) so that filter conditions could be better controlled at a later state. The prediction yielded a total of 26,471 putative genes for the genome of *P. pastoris*.

In a WU-BLASTN search against *S. cerevisiae*, 6,374 sequences that were predicted by GeneMark, and 3,964 of the IG predictions produced hits with *S. cerevisiae* using an *E* value (Expectation value, [19]) of $< 10^{-4}$, a hit length > 100 nucleotides and an identity of >50%. To reduce the redundancy within the data set the predicted genes were clustered into groups sharing more than 90% similarity using cd-hit [20]. From a total of 31,896 candidate sequences (GeneMark and IG predictions), 22,020 cd-hit groups were obtained. From the cluster file it was clear that some of the clusters had to be analyzed further before selecting target sequences for the oligo design. After the removal of all sequences that had a short length and a low prediction value, complex clusters were defined as clusters for which the minimum relative length of all sequences was smaller than 0.9. A total of 2,612 clusters fell into this category and were excluded at a first design stage.

Finally 19,508 predicted target sequences remained to be tested in the first microarray experiments. OligoArray 2.1 [21] was able to design oligonucleotide probes for 17,161 sequences ranging in length from 57 to 60 nucleotides.

Validation arrays for the first list of predicted transcript sequences (Same-Same experiment)

With these probes 4 × 44 K slides were produced on the Agilent microarray platform and employed for an initial validation of the predicted transcript sequences by hybridization with the Pool samples of *P. pastoris* (for preparation of Pool samples see Material and Methods). One slide had to be discarded because of quality issues. For the remaining 12 arrays the number of probes showing a signal varied between 10,708 and 15,598. Of these, 7,980 had a signal on all 12 arrays, and only 951 probes showed no hybridization on all 12 arrays.

Second, curated list of predicted target sequences and second oligo design

The results of the initial validation arrays were utilized to adapt the list of predicted genes, keeping all predictions for which a hybridization signal could be observed for all arrays plus all predictions with significant sequence similarity to annotated genes as well as all sequences with an average gene prediction score > 0.5. This approach allows for the fact that not all genes will have been actively expressed in the target samples. Additionally, predicted transcripts resulting from a subsequent analysis of the complex clusters were included at this stage. Of the 2,612 complex cluster that were not included in the design for the first batch of arrays, only 223 contained more than 2 sequences and for a further 14 no subsequence match of at least 60 nucleotides could be found within the last 1000 bases at the 3'-end. These 237 clusters were manually curated while the rest could be automatically reduced to one sequence. To make full use of the 15,208 features available on the Agilent microarray platform, it was decided to also include predicted sequences with somewhat lower gene prediction score that showed a hybridization signal in at least 8 of the 12 arrays. Finally, a selected set of 15,253 predicted transcript sequences was used as targets for probe design of a comprehensive *P. pastoris* microarray. While it is obvious that this list is larger than the expected number of open reading frames (6,000–7,000), as judged in comparison to other yeast species [22], we intentionally included more putative transcript sequences, as false positives with a distinct sequence will not negatively affect microarray design or

Table 1: Comparison of gene finder performance on yeast genomic sequence data

Gene finder	True positives	Partly	False positives	False negatives	Sensitivity (%)	Positive prediction value (%)
Glimmer3	75	13	21	31	73.9	68.8
GlimmerHMM	1	3	68(234)	115	3.2	1.4
GeneMark	81	16	32	22	81.5	62.7

Three different gene finders were tested on the genome sequence of *S. cerevisiae* chromosome 1 to evaluate the quality of gene prediction. Sensitivity = TP/(TP + FN), positive prediction value = TP/(TP + FP); For Glimmer HMM the column False Positives contains the number of genes and in brackets the number of exons.

experiments, in contrast to the damage of falsely excluding a potential transcript target.

Oligonucleotide probes were designed using a probe design tool developed in-house, a thermodynamic model-based oligoset optimizer ('TherMODO', [23]). TherMODO designed probes for 15,035 sequences, of which only 665 were predicted as having cross-hybridization potential. The TherMODO design was compared to probe design with eArray [24]. The distributions of ΔG and T_m of both designs are shown in additional file 1. Clearly the TherMODO designed probes are more uniform in respect to the Gibbs free energy ΔG, indicating a superior hybridization performance [14].

The final probe design was manufactured on 8 × 15 K slides by Agilent, and evaluated for reproducibility and biological meaningfulness. Pool samples were applied to 2 arrays on 2 slides each, including dye swap. The scatterplots show uniformly high correlations > 97% both within and between arrays, both on same and different slides, indicating high reproducibility of hybridization signals between identical samples. Exemplarily, a scatterplot of signal intensities derived from the same samples (wild type strain untreated) is shown in Figure 1. For the final gene list the annotation was improved in addition to the annotation provided by IG. This resulted in 3954 annotated ORFs, of which 2989 had an IG annotation. 965 newly annotated ORFs were found, and the annotation of 288 hypothetical proteins was confirmed. All annotated genes are listed in Additional file 2.

Biological evaluation of the new microarrays

The performance of the new arrays was examined by a hybridization experiment using samples, for which transcript regulation data have been obtained before [4]. The biological question evaluated was the regulatory response of *P. pastoris* to constitutive overexpression of the active form of *S. cerevisiae HAC1*, the transcription factor controlling UPR target genes. By this approach, the regulation of 52 genes which have been studied before using TRAC [3] could be verified, with 80% of these genes showing the same regulation pattern for both methods (genes highlighted in bold in Additional file 2). This correlation is statistically significant based on calculating the regression ($p = 8.8 \cdot 10^{-6}$).

The similarities and differences of UPR induction and reaction to DTT stress have been discussed before [4,25,26]. To achieve further insight into this technologically relevant issue, we compared the gene regulation patterns of a *HAC1* overexpressing strain *vs* wildtype control with the regulation pattern of the wildtype treated with DTT for 60 min *vs* the untreated control. Genes were qualified as significantly regulated with a *p*-value < 0.05 (adjusted for multiple testing). 11,262 of all features on the microarrays appeared as differentially regulated either upon DTT treatment or *HAC1* overexpression, or both. 8,480 reacted to *HAC1*, and 6,870 to DTT, with an overlap of 4,088. Considering only the 3,954 annotated genes, a similar pattern is observed with roughly half of the regulated genes overlapping between DTT and *HAC1*, and another half being typical only for either of the treatments

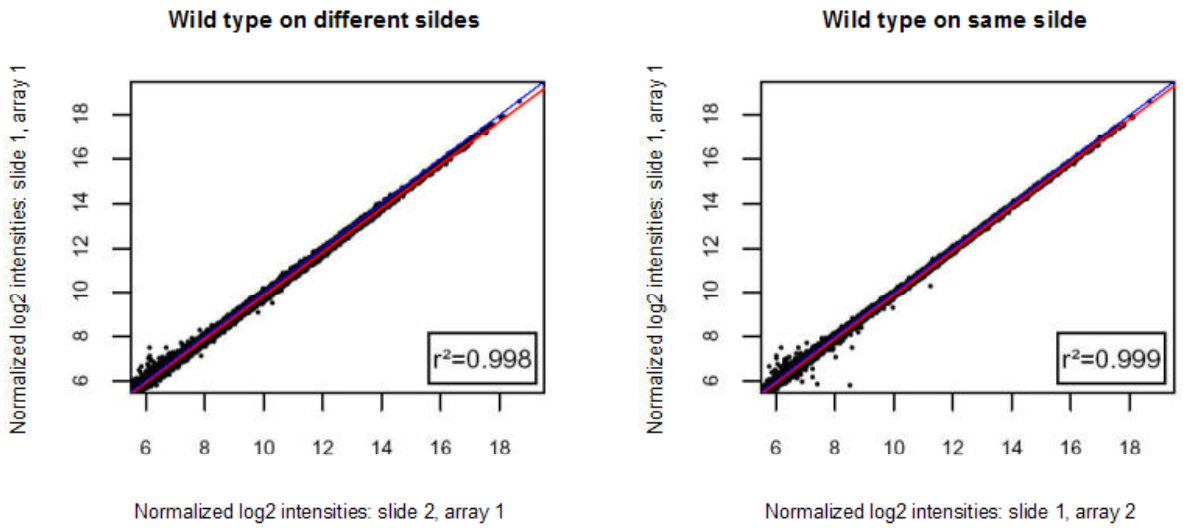


Figure 1
Correlation of signal intensities. Scatterplots of untreated wild type strain samples on (A) different arrays of the same slide; (B) different arrays on different slides. Red line: linear regression of the data; blue line: theoretical perfect correlation.

(Figure 2). Accordingly, the correlation of log fold changes of the two treatments is apparent but rather weak (Figure 3). While DTT treatment is widely accepted as a standard inducer of UPR, these observations indicate that the gene regulation pattern triggered by the UPR transcription factor Hac1 differs to a significant extent from that exerted by DTT.

As previous research on transcriptome regulation upon UPR induction usually employs a fold change (FC) cut-off to highlight the strongly regulated genes, we decided to introduce FC > 1.5 as a second criterion to identify more strongly regulated genes for further detailed analysis (Volcano plots visualizing the two criteria are provided in Additional file 3). Although the introduction of a FC cut-off alters the absolute number of regulated genes, it does not alter the relative distribution of regulated genes categorized into functional groups (GO slim biological process), as can be seen in Figure 4 and Additional file 4.

Comparison of UPR induction by DTT in P. pastoris *and* S. cerevisiae

In order to compare the effects of DTT treatment in *S. cerevisiae* with those in *P. pastoris*, the data published by Travers et al. [17] for 60 min treatment of *S. cerevisiae* with DTT were evaluated alongside with our results for *P. pastoris*. All genes of *S. cerevisiae* which were listed in [17] and for which homologs in *P. pastoris* were identified were

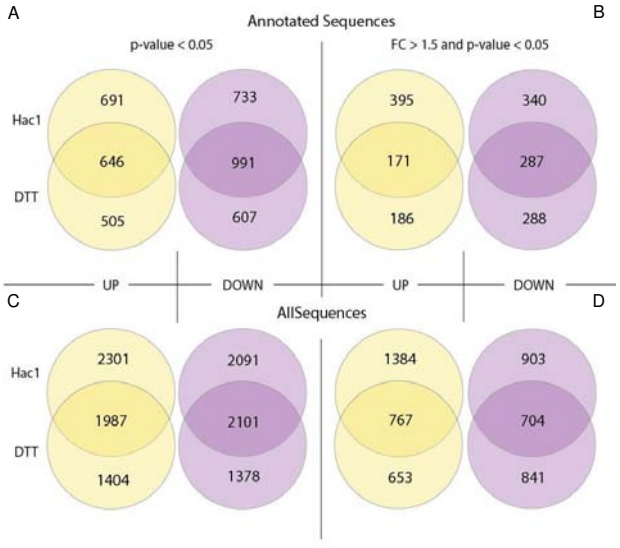


Figure 2
Venn diagrams of differentially expressed genes upon DTT treatment or *HAC1* overexpression. (A, B) Regulated hits with annotation; (C, D) all regulated features; (A, C) cut-off adjusted *p*-value < 0.05; (B, D) cut-off adjusted *p*-value < 0.05 and FC > 1.5.

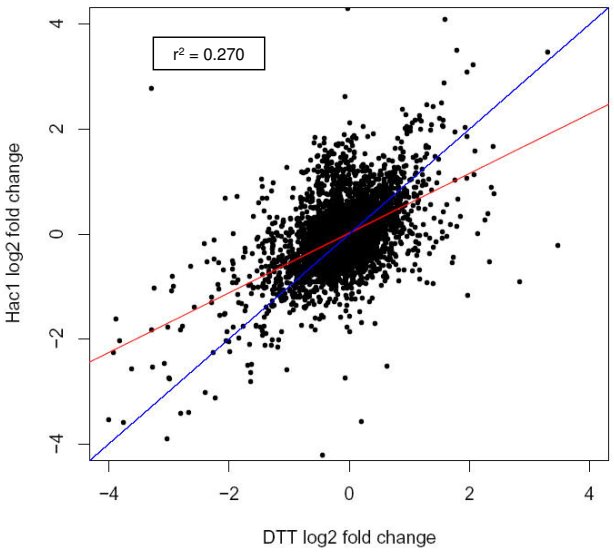


Figure 3
Comparison of expression changes induced by DTT treatment and *HAC1* overexpression, respectively. Log_2 values of expression changes (log_2 FC) caused by DTT (DTT treated wildtype *vs* untreated wildtype) and by Hac1 (*HAC1* overexpression *vs* wildtype) are compared. The correlation coefficient r^2 is indicated. Red line: linear regression of the data; blue line: theoretical perfect correlation.

classified as upregulated, downregulated or unregulated. In order to compare the two data sets, a cutoff of 1.5 fold differential expression was set in both to define regulated genes. A significance threshold on *p*-values could not be employed, as these data were not provided for *S. cerevisiae*. 48% of these genes defined as regulated or unregulated reacted in *P. pastoris* just as in *S. cerevisiae*.

A closer evaluation revealed that certain GO groups were regulated very similarly in both yeast species, while others showed only a low degree of similarity (Table 2). Fisher's exact test was performed to evaluate the significance of groups with low similarity. Especially the GO groups 'translocation', 'protein folding', 'protein degradation', and to some extent 'glycosylation' and 'transport' showed high degrees of similarity. In some GO groups, only some subgroups reacted similarly while others behaved differently in the two yeasts. Of the 'glycosylation' group, core oligosaccharide synthesis and glycosyltransferase genes behaved very similarly, while glycoprotein processing, GPI anchoring and O-glycosylation related genes were regulated significantly different ($p < 0.05$). In the 'protein degradation' group, more similarity was observed for ERAD genes than for ubiquitin/proteasome related genes. Among the 'transport' gene group, budding, fusion and retrieval of ER to Golgi showed a high degree of similar regulation, contrary to the subgroup distal secretion. Low

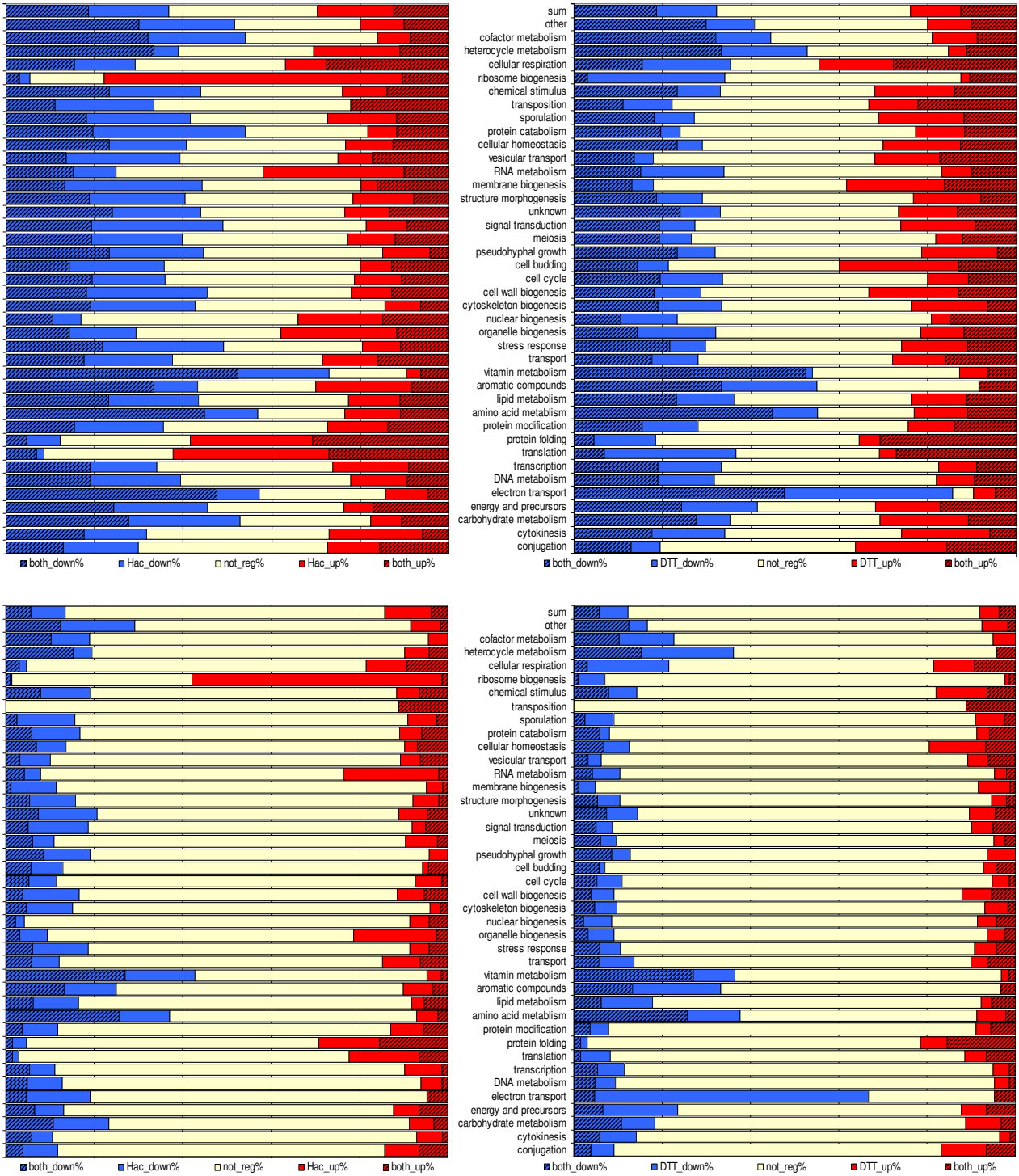


Figure 4

Fractions of up- and downregulated genes in functional groups. Relative numbers of upregulated (red), downregulated (blue) and unregulated (yellow) genes categorized in GO biological process terms upon *HAC1* overexpression (left panel) and DTT treatment (right panel). Shaded in black: regulated in both treatments. Upper panels: cut-off *p*-value < 0.05, lower panel cut-off *p*-value < 0.05 and FC > 1.5. The results of significance testing (Fisher's exact test) are given in additional file 4.

Table 2: Similarity of gene regulation between *P. pastoris* and *S. cerevisiae* upon DTT treatment

Function	Subfunction	No. of similarly regulated/total	% similar regulation
Translocation	total	4/6	67
Glycosylation	total	11/22	50
	Core oligosaccharide synthesis	3/4	75
	Oligosaccharyltransferase	4/4	100
	Glycoprotein processing	1/5	20
	GPI anchoring	1/4	25
	Golgi/O-linked	2/5	40
Protein Folding	total	5/8	63
	Chaperones	3/5	60
	Disulfide bond formation	2/3	67
Protein Degradation	total	4/5	80
	ERAD	3/3	100
	Ubiquitin/Proteasome	1/2	50
Transport	total	11/20	55
	Budding (ER-Golgi)	4/7	57
	Fusion (ER-Golgi)	1/1	100
	Retrieval (ER-Golgi)	4/5	80
	Distal secretion	2/7	29
Lipid Metabolism	total	5/18	28
	Fatty acid metabolism	0/4	0
	Heme biosynthesis	2/5	40
	Phospholipid biosynthesis	2/6	33
	Sphingolipid biosynthesis	0/1	0
	Sterol metabolism	1/2	50
Vacuolar Protein Sorting	total	1/4	25
Cell Wall Biogenesis	total	4/10	40

All genes that were indicated in [17] as core UPR genes in *S. cerevisiae* and having an annotation in *P. pastoris* were grouped by their GO process functions. Similar regulation of a gene means upregulated, downregulated or below cut-off, respectively, in both yeasts.

similarities were observed for 'lipid metabolism', 'vacuolar protein sorting' and 'cell wall biogenesis' genes. It becomes obvious that core UPR genes related to protein translocation, folding and ER transport, as well as core N-glycosylation react similarly to DTT treatment in *P. pastoris* as compared to *S. cerevisiae*, while genes involved in processes which are more distal from ER protein folding behave more differently, indicating that those processes (like functions in the Golgi, [27]) differ significantly between the two yeasts.

Overexpression of Hac1 triggers a different regulation pattern compared to DTT treatment

In most previous studies of the UPR in lower eukaryotic cells, treatment with DTT or tunicamycin, or heterologous protein expression has been employed to trigger the UPR. This study clearly indicates that the set of regulatory events triggered by DTT analysis only partially overlaps with the reactions to constitutive expression of the activated form of the UPR transcription factor Hac1 (see Figures 2 and 3). Interestingly, both treatments resulted in the same amount of genes being down-regulated as being up-regulated, a fact that has been neglected to some extent in the existing literature.

Those genes appearing beyond the threshold (p-value < 0.05 and FC >1.5) were subjected to a more detailed comparison between the effects of DTT treatment and Hac1 induced regulation. The relative numbers of up- and down-regulated genes in each GO biological process term based on the SGD GO slim tool [28] are depicted in Figure 4.

A pattern common to both treatments is the down-regulation of major metabolic processes like carbohydrate, amino acid and lipid metabolism, as well as that of vitamins, cofactors and aromatic and heterocyclic compounds. This makes it obvious that the UPR has a major impact on decreasing both catabolic and anabolic processes. On the other side, both treatments lead to up-regulation of protein folding and vesicular transport. These effects are in line with the published literature, indicating the cellular reaction towards alleviation of the UPR [4,25,26,17].

As expected, the genes coding for classical UPR targets are induced both in Hac1 overproducing and in DTT stressed cells, and genes underlined in the following paragraphs have been identified as UPR targets in previous studies. Especially the ER folding catalysts *PDI1* and *ERO1*, the

DnaJ homologs *IEM1* and *SCJ1*, the ER resident chaperones *CNE1* (calnexin), *KAR2*/BiP and *LHS1* and the mitochondrial chaperones *HSP60* and *SSC1* are significantly up-regulated in both conditions. Among the functional group of 'protein modification' the majority of up-regulated genes belong to the core oligosaccharide synthesis (*DPM1*, *DIE2*), oligosaccharyltransferase complex (*OST1*, *OST2*, *OST3*, *SWP1*, *STT3*, *WBP1*), glycoprotein processing (*ALG2*, *ALG7*, *SEC53*), GPI anchor biosynthesis (*GPI2*, *GPI14*, *PSA1*) and Golgi/O-linked glycosylation (*PMT1*, *PMT2*, *PMT4*, *PMT6*). Besides these, several genes coding for the translocon pore complex (*SEC61*, *SEC62*, *SEC63*, *SEC72*, *SSS1*), which aid the translocation of nascent polypeptides into the ER, are induced. Higashio and Kohno [29] describe the stimulation of ER-to-Golgi transport through the UPR by inducing COPII vesicle formation. In this context, we see *SEC23*, *SEC24*, *SFB2*, *YIP3*, and *ERV2* upregulated. However, also proteins building the COPI coatomer, which are required for retrograde Golgi-to-ER transport, show increased transcription levels upon ER stress in our experiments (*COP1*, *RET2*, *SEC21*, *SEC27*).

While we cannot give any information on ERAD regulation, as *HRD1* is the only annotated gene of this protein degradation process (up-regulated in the Hac1 strain), we observed the down-regulation of some components involved in the assembly of the 20 S core of the 26 S proteasome (*ADD66*, *PRE1*, *PRE4*, *SCL1*) and ubiquitin *UBI4* upon constitutive UPR activation. In this context, Shaffer et al. [30] describe reduced degradation of newly synthesized proteins in XBP1-overexpressing human Raji cells.

Induction of genes encoding cytosolic chaperones (Cns1, Jjj3, Hsp82, Ssa1, Ssa2, Sse1, Ydj1, Zuo1) can only be seen in the Hac1-overproducing strain. Additionally, the ER-resident Pdi homolog Mpd1 and two members of the PPIases (*FPR4* and *CPR6*) are only up-regulated in the engineered strain, but not upon DTT addition.

One of the most striking patterns is the significant up-regulation of a large number of genes with functions in ribosomal biogenesis (233 genes assigned to the GO-categories 'ribosome biogenesis and assembly' and 'RNA metabolic process'). Most of these genes are contributing to rRNA processing (RRP family) and ribosome subunit nuclear export and assembly, while the ribosomal proteins (RPS and RPL families) themselves are not among the regulated genes for *P. pastoris* (see Additional file 2). No genes with a function in mRNA decay show increased transcription levels. The induction of the above functional categories came as a surprise, as translational down-regulation of proteins involved in ribosomal biogenesis was recently reported when *S. cerevisiae* cell were treated with DTT [31]. In contrast, the transcription levels of 9 out of the 16 mRNAs listed by these authors are enhanced in our study. Transcriptional down-regulation of ribosomal proteins during ER stress conditions was also revealed when reanalysing the raw data provided by Travers et al. [17]. However, Shaffer et al. [30] describe an increase in total protein synthesis as well as in the number of assembled ribosomes upon the overexpression of the mammalian Hac1 homolog XBP1 in Raji cells, but did not observe upregulation of genes related to ribosome biogenesis. A similar effect was observed after XBP1 overexpression in CHO-K1 cells [32]. These results may be an indication that the positive effect of overexpression of the UPR transcription factor on heterologous protein production [33,34,16,35] results not just from stimulation of folding and secretion of proteins but also their synthesis. The induction of protein folding related genes upon Hac1 overexpression is in line with the literature on UPR effects, while an impact on organelle biosynthesis other than ER and Golgi has so far only been described for mammalian cells.

The stimulatory effects of XBP1 induction on ribosomes and organelle synthesis in mammalian cells like lymphocytes have been attributed to their function as dedicated protein factories. On the other hand the UPR in lower eukaryotes should rather serve to alleviate the load of unfolded, aggregation prone protein. It will be of interest in the future to investigate whether Hac1 stimulates ribosome biogenesis in other yeasts and fungi as well, and whether this leads to increased translation.

In this context, it is worthwhile to mention the induction of two pathways leading to the unusual post-translationally modified amino acid derivatives diphthamide and hypusine which are exclusively found in eukaryotic translation elongation factors 2 (eEF2) and 5 (eEF5), respectively [36,37]. As these biosynthetic pathways are rather complex, and outstanding in the otherwise downregulated group of 'amino acid biosynthesis', this induction underlines the increased demand for protein synthesis.

Furthermore, we observe that ER stress leads to increased transcription of genes coding for the large and small subunits of the mitochondrial ribosomes (*MRPS*, *RSM* and *MRPL* families), mitochondrial translation initiation and elongation factors (*IFM1*, *MEF1*, *MEF2*) and mitochondrial DNA polymerase (*MIP1*). Several essential constituents of the mitochondrial inner membrane presequence translocase (*TIM* family) are also up-regulated, indicating increased necessity for protein import into the mitochondria. Similarly, XBP1 was shown to increase mitochondrial mass and function in two types of mammalian cells [30].

While previous studies analysing UPR regulation mainly focus on up-regulated genes [17], more than half of the genes identified in our study to be regulated are strongly down-regulated (at least 1.5 fold). As can be seen in Figure 4, anabolic processes such as vitamin production, amino

acid and aromatic compound biosynthesis, heterocycle metabolic processes, carbohydrate, lipid and cofactor metabolism are among the most prominent repressed classes in both DTT-treated as well as Hac1-overproducing cells. The down-regulation of energy consuming biosynthetic pathways emerges as a general picture during ER stress conditions. However, it becomes obvious that the response to the folding perturbation agent DTT strongly differs from constitutive UPR induction by Hac1-overproduction. Especially the prominent down-regulation of genes belonging to 'electron transport' and 'cellular respiration' can easily be explained by the strong reducing capacities of DTT. Prominent members of the mitochondrial inner membrane electron transport chain such as subunits of the cytochrome c oxidase (*COX4*, *COX4*, *COX5A*, *COX13*) and the ubiquinol cytochrome-c reductase complex (*COR1*, *QRC6*, *QRC7*, *QRC9*, *RIP1*) are significantly repressed upon DTT treatment. Additionally, cytochrome c (*CYC1*), cytochrome c1 (*CYT1*) and cytochrome c heme lyase (*CYC3*) are only under DTT-dependent repression (GO: 'generation of precursor metabolites and energy'). The reducing features of DTT are most probably also the reason for the up-regulation of genes involved in the upkeeping of 'cellular homeostasis' and clearly, addition of DTT is provoking a 'response to a chemical stimulus'.

Down-regulated genes appearing in both Hac1 and DTT in the 'protein modification' group focus on protein kinases (*CDC5*, *CDH1*, *DBF2*) and components of the ubiquitinylation complex (*BUL1*, *CUL3*) involved in cell cycle regulation driving the cells towards mitotic exit (*CDC5*, *CDH1*, *MOB1*). These effects are even more pronounced in the Hac1-strain, where several more histone modifying enzymes as well as cycline-dependent protein kinases and components of the protein kinase C signalling pathway show reduced transcription levels compared to the wild type. Unlike reported for the filamentous fungi *T. reesei* [7] and *A. nidulans* [26], genes encoding the histones H2A, H2B, H3 and H4 appear to be down-regulated upon secretion stress in *P. pastoris*.

No clear picture emerges regarding the regulation of 'lipid metabolism': While sterol and ergosterol biosynthesis tend to be inhibited, the production of sphingolipid precursor substances is enhanced. On the other hand, a down-regulation of the major cell wall constituents (β-1,3 glucanases *BGL2* and *EXG1*, cell wall mannoproteins *CCW12*, *CWP2* and *TPI1*, GPI-glycoproteins *GAS1* and *SED1*, *PST1*) and genes coding for proteins required for the transport of cell wall components to the cell surface (*SBE22*) is manifest. Taken together, these results indicate a significant remodelling process regarding the *P. pastoris* cell envelope during ER stress conditions.

Interestingly, the major groups of metabolic genes were down-regulated upon Hac1 overexpression, indicating a decrease of the supply of metabolites. However, it should be noted that no reduction of the specific growth rate was observed as compared to the wild type strain ($\mu = 0.37$ and 0.39 h^{-1}, respectively). A reduction of metabolic processes, and amino acid synthesis in particular, is contradictory to translation stimulation. Further research will be needed to elucidate the overall regulatory pattern of UPR in respect to protein synthesis.

Conclusion

Additional gene finding and annotation added to the available data for *P. pastoris* lead to a list of approximately 4,000 genes with a putative identification of their function, and 11,000 more potential open reading frames. An oligonucleotide probe set was designed, the hybridization results were evaluated for reproducibility, and results from a biologically relevant analysis were tested for meaningfulness. In a direct comparison to *S. cerevisiae* employing DTT treatment for UPR induction, 45 out of 93 genes reacted similarly. The differences thus observed between *P. pastoris* and *S. cerevisiae* underline the importance of DNA microarrays for industrial production strains. *HAC1* overexpression in *P. pastoris* obviously leads to induction of many genes involved in translation: most genes of ribosome biogenesis, as well as many related to RNA metabolism and translation were up-regulated, an effect that has never been observed in yeasts and filamentous fungi so far.

The upregulation of ribosomal biogenesis, RNA metabolism, translation, and organelle biosynthesis is specific for *HAC1* overexpression and not observed with DTT treatment, while the latter leads specifically to the upregulation of genes related to chemical stimulus, and the downregulation in the groups electron transport and respiration, so that these reactions have to be regarded as specific for the treatment with a reducing agent rather than UPR regulated.

Methods

Gene Prediction and Sequence Selection

Gene prediction and the selection of sequences for oligonucleotide probes were based on sequenced contigs of the *P. pastoris* genome including predictions of protein coding genes, available through Integrated Genomics [8]. The number of predicted genes was 5,425 of which 3,680 had an assigned function. The ORFs were made up of experimentally identified genes, as well as ORFs predicted by a proprietary gene finder [38].

To validate and possibly improve these predictions, *de novo* gene finding was conducted. First three *de novo* gene

finder (GeneMark, Glimmer3, GlimmerHMM) were tested on the genome sequence of *S. cerevisiae* (data from BioMart, [39]) to evaluate their performance on yeast genomes. As described in Results and Discussion, GeneMark [40] was selected for further gene prediction on the *P. pastoris* genome sequence. To run the gene prediction it was necessary to train GeneMark on *S. cerevisiae* by building a matrix with transition probabilities for coding and non-coding regions used by the Hidden Markov Model (HMM) of the program. With the amount of data available we were able to generate a matrix of the 7th order. The genes of *P. pastoris* were predicted using the *S. cerevisiae* matrix and the lowest possible probability score cut-off (t = 0.05). In the initial stage of the microarray design the aim was to predict as many putative ORFs as possible. In this context a higher false positives rate was accepted in order to keep the false negatives rate as low as possible.

The predicted sequences were merged with data from IG and clustered by running cd-hit [20] with a similarity cut-off of 90%. For all of the resulting sequences a BLASTX search was done against *S. cerevisiae* using WU-BLAST [19]. Blast data was further filtered for length (cutoff 55 bp) and low prediction score. Clusters comprised of more than one gene were represented by the longest sequence, or curated manually, if appropriate.

From this first gene list (PpaV1) microarrays were analyzed as described below. Spots with a positive signal were determined using the mean plus one standard deviation of the negative control probes as a cut-off. Sequences were selected if they were positive in at least 8 out of 12 arrays. This criterion was chosen to fill the array capacity. Additionally all sequences with a probability score higher than 0.5 or having an annotation were kept for the second set of sequences (PpaV2).

Annotation

For the PpaV2 sequence set the program cd-hit-est [20] was used to find all ORFs that had a global identity of > 80% with *S. cerevisiae*. WU-BLASTX and WU-TBLASTN searches were conducted against *S. cerevisiae*, using a low complexity filter and E < 10^{-7}. For all the sequences that did not have a match with *S. cerevisiae* under these conditions the two BLAST searches were repeated against the SwissProt/TrEMBL [41] database. A perl script was developed to summarize and compare the BLAST results.

Oligo Design and Array platform

Oligos for the PpaV1 sequences were designed with the Program OligoArray 2.1 [21] to match the melting temperature distribution of Agilent's *S. cerevisiae* oligos on the Yeast Oligo Microarray (V2), design number 013384.

The oligo-set for the PpaV2 sequence set was designed using the thermodynamic model-based oligoset optimizer 'TherMODO'. This tool incorporates advanced quantitative models for probe-target binding region accessibility and position-dependent target labelling efficiency, and replaces the common greedy search algorithm by a global set optimization step, achieving high discrimination power for particularly uniform probe sets [23]. Probes for Agilent arrays are limited to a maximum length of 60 nucleotides by the manufacturing process. For increased flexibility in the probe design, the oligoset design optimization considered probes ranging in length from 57 to 60 nucleotides.

These arrays were produced on Agilent 60 mer oligonucleotide high density arrays 4 × 44 K (with 42,034 available features) for PpaV1 and 8 × 15 K (with 15,208 available features) for PpaV2.

Experimental Design

For the first batch of arrays a same-same design was used, employing six replicates each of Pool 1 and of Pool 2. The aim of this experiment was to determine which of the probes hybridize to *P. pastoris* targets. For the second batch of arrays a two-state comparison set up was chosen with 6 replicates for each experiment of which 3 were dye swapped.

Strains und Cultures

For the first batch of arrays the aim was to determine which of the predicted probes hybridize with targets from *P. pastoris*. To make sure that many genes were active it was important to pool samples from various conditions of the cells. Samples were taken from two different *P. pastoris* strains, X-33 and CBS2612, grown on different media and taken at both exponential and stationary growth phase. The media were YP Medium (1% yeast extract, 2% peptone and either 2% glucose, 2% glycerol or 0.5% methanol as carbon source), Buffered Minimal Medium (1.34% yeast nitrogen base, 4 × 10^{-5}% biotin, 100 mM potassium phosphate pH 6.0 and either 2% glucose, 2% glycerol or 0.5% methanol as carbon source), and Buffered Minimal Medium described above supplemented with amino acids (0.005% of L-glutamic acid, L-methionine, L-lysine, L-leucine and L-isoleucine). The samples were combined into two pools with Pool 1 containing 18 samples from the exponential growth phase and Pool 2 containing 18 samples from the stationary phase. Both pools additionally contained seven chemostat samples of the strain X-33 3H6Fab, grown as in [42].

For the UPR experiments, strains GS115 HAC1, constitutively overproducing the activated form of *S. cerevisiae* Hac1, as described in Gasser et al. [33,4], as well as GS115

transformed with the empty vector pGAPHIS (a histidine prototrophic isogenic strain of GS115) were cultivated in YPD (YP as above with glucose) at 28°C. After growing the cultures to an OD_{600} = 5.7, dithiothreitol (2.5 mM) was added where appropriate. After 1 more hour of cultivation, 1 ml culture was added to 0.5 ml precooled phenol solution (5% in absolute ethanol) and centrifuged immediately for 30 sec at 13.000 rpm. After discarding the supernatants the pellets were frozen at -80°C.

RNA Isolation

All samples were resuspended with 1 mL TRI Reagent (Sigma). Cells were disrupted after addition of 500 μL glass-beads with a Thermo Savant Fastprep FP120 Ribolyzer by treatments of 2 × 20 sec at 6.5 ms^{-1}. RNA was extracted with chloroform, precipitated with isopropanol, washed with 75% ethanol and dissolved with diethylpyrocarbonate treated water. The extracted RNAs were quantified via absorption at 260 and 280 nm. The quality of the RNA samples was verified with the Agilent Bioanalyzer 2100 and RNA 6000 Nano Assay kit (Agilent Technologies, California).

Labeling and Hybridization

Hybridization targets for *P. pastoris* microarrays were prepared according to Agilent's Two-Color Microarray-Based Gene Expression Analysis protocol (Version 5.5, February 2007). Purification of the labelled and amplified RNA was conducted using RNeasy mini spin columns (Qiagen). The quality of labelled cRNA was evaluated on the Agilent Bioanalyzer 2100 and quantified using a ND-1000 NanoDrop Spectrophotometer. Fragmented cRNA samples were applied to the individual arrays. The slides were placed into Agilent hybridization oven and hybridized for 17 h, at 65°C and 10 rpm.

Microarray Analysis

Slides were scanned with an Agilent MicroArray Scanner and intensities were extracted using Agilent's Feature Extraction software (version 9.1). The resulting data was imported into R where data pre-processing and normalization was performed. In the pre-processing step all outliers and saturated spots were given the weight zero. After plotting the data we decided to refrain from background correction since it has the tendency to add more noise to the data [43]. The data were normalized using locally weighted MA-scatterplot smoothing (LOESS) followed by a between array scale normalization. Both functions are available within the limma package of R [44]. For the selection of differentially expressed genes linear models were fitted to the log-ratios of the expression data separately for each gene. An empirical Bayes approach was used to shrink the probe-wise sample variances towards a common value yielding a moderated *t*-statistic per gene [45]. *P*-values were corrected for multiple testing using Holm's method [46]. Features were defined as differentially expressed if they had a *p*-value < 0.05. For the identification of stronger regulatory effects an additional cut-off for the fold change (FC) of 1.5 > FC > 1/1.5 was applied. Description of the platform, array, raw data as well as processed data were deposited at ArrayExpress [47] under the accession numbers A-MEXP-1157.

All annotated *P. pastoris* genes were categorized into GO biological process terms using the SGD GO slim tool [28], whereby *P. pastoris* specific genes were included into the term 'other'. The significance of a deviation of the number of up- or downregulated genes in each group from the average was verified with a Fisher test (Additional file 4).

Authors' contributions

AG performed gene finding and annotation, statistical data analysis, supported data evaluation, and drafted part of the manuscript. BG performed data evaluation, supported annotation, study design, array design and drafted part of the manuscript. MD performed the cultivations and hybridizations. MS contributed to study design, annotation and array design. GGL developed the employed probe design tool and supported the array design. TT developed the quantitative model for the position-dependent target labelling efficiency and adapted it for the relevant end-primed labelling protocol. DPK supervised gene identification and annotation, supervised and contributed to the development of the employed probe design tool, and contributed to the manuscript. DM conceived of the study, and participated in data evaluation and manuscript drafting. All authors read and approved the final manuscript.

Additional material

Additional File 1

Thermodynamic properties of the TherMODO probe design compared to probes designed through Agilent's eArray. *Distribution of Gibbs free energy ΔG (A) and the probe-target melting temperature T_m (B) of the oligo sets. The upper row (1) shows the oligos designed through eArray and the lower row (2) the oligos designed with TherMODO. PpaV2 is the name of the second set of sequences as described in the Materials and Methods section.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-390-S1.pdf]

Additional File 2

Differential expression values of all annotated genes upon DTT treatment and* HAC1 *overexpression. *Differential expression values and adjusted* p*-values of all annotated genes of* P. pastoris*, denominated with the gene name of their respective* S. cerevisiae *homolog. Genes that were tested with TRAC as a different method for transcript quantification are highlighted in bold letters. Legend of headers: id – internal unique identifier of sequence; sequ_id – ERGO identifier (RPPA.) or gene finder identifier (orf.) respectively; DTT_logFC – log_2 fold change of DTT treatment compared to control;* HAC1 *logFC – log_2 fold change of* HAC1 *overexpression compared to control; Gene name – Standard gene name or if missing systematic ORF name according to* S. cerevisiae *nomenclature; GO – Gene Ontology term (for descriptions see additional file 4). If a gene is present in a certain GO group it has a 1 in the respective column, if not it has a 0.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-390-S2.xls]

Additional File 3

Volcano plots of fold change vs. adjusted* p*-values. *(A) DTT treatment; (B)* HAC1 *overexpression. Blue line:* p*-value cut-off* p > *0.05; red lines: optional fold change cut-off FC > 1.5.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-390-S3.pdf]

Additional File 4

Fisher's exact test of the up/down regulated gene groups upon DTT treatment and* HAC1 *overexpression. *Fisher's exact test was applied to test significance of the up- and downregulated gene groups displayed in figure 4.* p_{adj} *values are given for each GO group. Legend of headers: group – Gene Ontology term; Description – Gene Ontology description; odds.ratio – measure of independence between variables; adj.p – Holm adjusted* p*-value;* HAC1 *up/down – up/down regulated in* HAC1 *overexpression experiment; DTT up/down – up/down regulated in DTT experiment. The first work sheet represents results using only a* p*-value cut-off* p > *0.05, the second work sheet represents results using a* p*-value cut-off* p > *0.05 and a fold change cut-off FC > 1.5.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-390-S4.xls]

Acknowledgements

This work was supported by the Austrian Science Fund (project No. I37-B03), the European Science Foundation (programme EuroSCOPE), and the Austrian Research Promotion Agency (programme FHplus).

The Vienna Science Chair of Bioinformatics gratefully acknowledges support by the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres (ARC) Seibersdorf, and Austrian Centre of Biopharmaceutical Technology (ACBT). TT acknowledges partial funding by the GenAu BIN-II PhD programme.

References

1. Ye RW, Wang T, Bedzyk L, Croker KM: **Applications of DNA microarrays in microbial systems.** *J Microbiol Methods* 2001, **47(3):**257-272.
2. Sauer M, Branduardi P, Gasser B, Valli M, Maurer M, Porro D, Mattanovich D: **Differential gene expression in recombinant Pichia pastoris analysed by heterologous DNA microarray hybridisation.** *Microb Cell Fact* 2004, **3(1):**17.
3. Rautio JJ, Kataja K, Satokari R, Penttila M, Soderlund H, Saloheimo M: **Rapid and multiplexed transcript analysis of microbial cultures using capillary electophoresis-detectable oligonucleotide probe pools.** *J Microbiol Methods* 2006, **65(3):**404-416.
4. Gasser B, Maurer M, Rautio J, Sauer M, Bhattacharyya A, Saloheimo M, Penttilä M, Mattanovich D: **Monitoring of transcriptional regulation in *Pichia pastoris* under protein production conditions.** *BMC Genomics* 2007, **8:**179.
5. Sims AH, Robson GD, Hoyle DC, Oliver SG, Turner G, Prade RA, Russell HH, Dunn-Coleman NS, Gent ME: **Use of expressed sequence tag analysis and cDNA microarrays of the filamentous fungus *Aspergillus nidulans*.** *Fungal Genet Biol* 2004, **41(2):**199-212.
6. Bachem CW, van der Hoeven RS, de Bruijn SM, Vreugdenhil D, Zabeau M, Visser RG: **Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development.** *Plant J* 1996, **9(5):**745-753.
7. Arvas M, Pakula T, Lanthaler K, Saloheimo M, Valkonen M, Suortti T, Robson G, Penttila M: **Common features and interesting differences in transcriptional responses to secretion stress in the fungi *Trichoderma reesei* and *Saccharomyces cerevisiae*.** *BMC Genomics* 2006, **7:**32.
8. Genomics I: **ERGO bioinformatics suite.** [http://ergo.integratedgenomics.com/ERGO/].
9. Besemer J, Borodovsky M: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W451-4.
10. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20(16):**2878-2879.
11. Mathé C, Sagot MF, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30(19):**4103-4117.
12. Majoros WH, Pertea M, Salzberg SL: **Efficient implementation of a generalized pair hidden Markov model for comparative gene finding.** *Bioinformatics* 2005, **21(9):**1782-1788.
13. Pedersen JS, Hein J: **Gene finding with a hidden Markov model of genome structure and evolution.** *Bioinformatics* 2003, **19(2):**219-227.
14. Kreil DP, Russell RR, Russell S: **Microarray oligonucleotide probes.** *Methods Enzymol* 2006, **410:**73-98.
15. Erasmus DJ, van der Merwe GK, van Vuuren HJ: **Genome-wide expression analyses: Metabolic adaptation of *Saccharomyces cerevisiae* to high sugar stress.** *FEMS Yeast Res* 2003, **3(4):**375-399.
16. Valkonen M, Penttilä M, Saloheimo M: **Effects of inactivation and constitutive expression of the unfolded- protein response pathway on protein production in the yeast Saccharomyces cerevisiae.** *Appl Environ Microbiol* 2003, **69(4):**2065-2072.
17. Travers KJ, Patil CK, Wodicka L, Lockhart DJ, Weissman JS, Walter P: **Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation.** *Cell* 2000, **101(3):**249-258.
18. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucleic Acids Res* 2005, **33(20):**6494-6506.
19. University W: **WU-BLAST.** [http://blast.wustl.edu/].
20. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22(13):**1658-1659.
21. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.** *Nucleic Acids Res* 2003, **31(12):**3057-3062.
22. Arvas M, Kivioja T, Mitchell A, Saloheimo M, Ussery D, Penttila M, Oliver S: **Comparison of protein coding gene contents of the fungal phyla Pezizomycotina and Saccharomycotina.** *BMC Genomics* 2007, **8:**325.
23. Leparc GG, Tuechler T, Striedner G, Bayer K, Sykacek P, Hofacker I, Kreil DP: **Model based probe set optimization for high-performance microarrays.** *Nucleic Acids Research* 2008, **submitted:**.
24. Agilent: **eArray.** [https://earray.chem.agilent.com/earray/].
25. Guillemette T, van Peij NN, Goosen T, Lanthaler K, Robson GD, van den Hondel CA, Stam H, Archer DB: **Genomic analysis of the**

secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. *BMC Genomics* 2007, **8:**158.
26. Sims AH, Gent ME, Lanthaler K, Dunn-Coleman NS, Oliver SG, Robson GD: **Transcriptome analysis of recombinant protein secretion by *Aspergillus nidulans* and the unfolded-protein response *in vivo*.** *Appl Environ Microbiol* 2005, **71(5):**2737-2747.
27. Rossanese OW, Soderholm J, Bevis BJ, Sears IB, O'Connor J, Williamson EK, Glick BS: **Golgi structure correlates with transitional endoplasmic reticulum organization in *Pichia pastoris* and *Saccharomyces cerevisiae*.** *J Cell Biol* 1999, **145(1):**69-81.
28. SGD: **SGD Gene Ontology Slim Mapper.** [http://db.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl].
29. Higashio H, Kohno K: **A genetic link between the unfolded protein response and vesicle formation from the endoplasmic reticulum.** *Biochem Biophys Res Commun* 2002/08/15 edition. 2002, **296(3):**568-574.
30. Shaffer AL, Shapiro-Shelef M, Iwakoshi NN, Lee AH, Qian SB, Zhao H, Yu X, Yang L, Tan BK, Rosenwald A, Hurt EM, Petroulakis E, Sonenberg N, Yewdell JW, Calame K, Glimcher LH, Staudt LM: **XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation.** *Immunity* 2004, **21(1):**81-93.
31. Payne T, Hanfrey C, Bishop AL, Michael AJ, Avery SV, Archer DB: **Transcript-specific translational regulation in the unfolded protein response of *Saccharomyces cerevisiae*.** *FEBS Lett* 2008.
32. Tigges M, Fussenegger M: **Xbp1-based engineering of secretory capacity enhances the productivity of Chinese hamster ovary cells.** *Metab Eng* 2006.
33. Gasser B, Maurer M, Gach J, Kunert R, Mattanovich D: **Engineering of Pichia pastoris for improved production of antibody fragments.** *Biotechnol Bioeng* 2006, **94(2):**353-361.
34. Gasser B, Sauer M, Maurer M, Stadlmayr G, Mattanovich D: **Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts.** *Appl Environ Microbiol* 2007/09/04 edition. 2007, **73(20):**6499-6507.
35. Valkonen M, Ward M, Wang H, Penttilä M, Saloheimo M: **Improvement of foreign-protein production in *Aspergillus niger* var. awamori by constitutive induction of the unfolded-protein response.** *Appl Environ Microbiol* 2003, **69(12):**6979-6986.
36. Mattheakis LC, Sor F, Collier RJ: **Diphthamide synthesis in *Saccharomyces cerevisiae*: structure of the DPH2 gene.** *Gene* 1993, **132:**149-154.
37. Wolff EC, Kang KR, Kim YS, Park MH: **Posttranslational synthesis of hypusine: Evolutionary progression and specificity of the hypusine modification.** *Amino Acids* 2007, **33(2):**341-350.
38. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E Jr., Liolios K, Joukov V, Kaznadzey D, Anderson I, Bhattacharyya A, Burd H, Gardner W, Hanke P, Kapatral V, Mikhailova N, Vasieva O, Osterman A, Vonstein V, Fonstein M, Ivanova N, Kyrpides N: **The ERGO genome analysis and discovery system.** *Nucleic Acids Res* 2003, **31(1):**164-171.
39. (EBI) EBI, (CSHL) CSHL: **BioMart.** [http://www.biomart.org/biomart/martview/].
40. Technology GI: **GeneMark.** [http://exon.gatech.edu/GeneMark/].
41. (SIB) SIB: **Swiss-Prot/TrEMBL.** [http://www.expasy.ch/sprot/].
42. Baumann K, Maurer M, Dragosits M, Cos O, Ferrer P, Mattanovich D: **Hypoxic fed batch cultivation of *Pichia pastoris* increases specific and volumetric productivity of recombinant proteins.** *Biotechnol Bioeng* 2008, **100(1):**177-83.
43. Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, Shabbeer S, Cope L: **Pre-processing Agilent microarray data.** *BMC Bioinformatics* 2007, **8:**142.
44. Smyth GK, Speed T: **Normalization of cDNA microarray data.** *Methods* 2003, **31(4):**265-273.
45. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article3.
46. Holm S: **A Simple Sequentially Rejective Bonferroni Test.** *Scandinavian Journal of Statistics* 1979:65 -670.
47. EMBL-EBI: **ArrayExpress.** [http://www.ebi.ac.uk/microarray-as/aer/].

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Research Article

Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production

Seung Bum Sohn[1,2*], *Alexandra B. Graf*[3,4*], *Tae Yong Kim*[1,2], *Brigitte Gasser*[3], *Michael Maurer*[4], *Pau Ferrer*[5], *Diethard Mattanovich*[3**] *and Sang Yup Lee* [1,2,6]

[1] Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon, Republic of Korea
[2] Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea
[3] Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Austria
[4] School of Bioengineering, University of Applied Sciences FH-Campus Wien, Vienna, Austria
[5] Department d'Enginyeria Química, Escola d'Enginyeria, Universitat Autònoma de Barcelona, Bellaterra, Spain
[6] Department of Bio and Brain Engineering and Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea

The methylotrophic yeast *Pichia pastoris* has gained much attention during the last decade as a platform for producing heterologous recombinant proteins of pharmaceutical importance, due to its ability to reproduce post-translational modification similar to higher eukaryotes. With the recent release of the full genome sequence for *P. pastoris,* in-depth study of its functions has become feasible. Here we present the first reconstruction of the genome-scale metabolic model of the eukaryote *P. pastoris* type strain DSMZ 70382, PpaMBEL1254, consisting of 1254 metabolic reactions and 1147 metabolites compartmentalized into eight different regions to represent organelles. Additionally, equations describing the production of two heterologous proteins, human serum albumin and human superoxide dismutase, were incorporated. The protein-producing model versions of PpaMBEL1254 were then analyzed to examine the impact on oxygen limitation on protein production.

Received 3 May 2010
Revised 13 May 2010
Accepted 18 May 2010

Supporting information available online

Keywords: Genome-scale metabolic model · Heterologous protein production · HSA · hSOD · *Pichia pastoris*

1 Introduction

Genome-scale metabolic models have been utilized with great success in the last decade in a number of different applications from microbial metabolic engineering for producing various bioproducts to the discovery of new information from the biological system. Use of the genome-scale metabolic models has enabled designing strategies for increasing the production of target compounds [1], such as nutritional supplements [2] and biofuels [3]. Metabolic models have also been utilized for identifying targets for new drugs against pathogenic microorganisms [4], and have been useful in dis-

Correspondence: Professor Sang Yup Lee, Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea
E-mail: leesy@kaist.ac.kr
Fax: +82423503910

Abbreviations: HSA, human serum albumin; **hSOD**, human superoxide dismutase

* These authors contributed equally to this work
** Additional corresponding author: Professor Diethard Mattanovich
E-mail: diethard.mattanovich@boku.ac.at

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

covering new information on the biological systems that they represent. Currently, most of the genome-scale metabolic models describe bacterial systems, due to their widespread use in the field of biotechnology and their lower degree of cellular complexity compared to eukaryotic systems. A few eukaryotic metabolic models developed to date have been limited to those of model organisms, such as *Saccharomyces cerevisiae*, *Aspergillus nidulans*, *Mus musculus*, *Arabidopsis thaliana* and *Homo sapiens* [5–10]. Here we present the reconstruction of the genome-scale metabolic model of an industrially important methylotrophic yeast *Pichia pastoris* type strain DSMZ 70382.

The methylotrophic yeast *P. pastoris* is well known for its ability to efficiently produce heterologous proteins [11], and has been recently described for metabolic engineering applications [12]. Its high capability for producing heterologous proteins is due to the presence of its highly inducible expression systems. Proteins destined for therapeutic applications require the proper post-translational modification to be effective and safe for human usage. *P. pastoris* is capable of modifying the recombinant proteins post-translationally so that they are structurally and functionally similar, if not identical, to their human counterparts [13, 14]. For example, the *N*-glycosylation pathway in *P. pastoris* has been engineered successfully to yield heterologous proteins with human-like *N*-glycan structures [15–17]. This makes *P. pastoris* invaluable for the production of heterologous proteins intended for pharmaceutical applications [18–20]. These characteristics make *P. pastoris* an ideal platform for the production of recombinant therapeutic proteins [19]. With the complete genome sequence of *P. pastoris* being available, a greater understanding of *P. pastoris* can be attained and consequently used for metabolic engineering of this organism [21, 22].

Here, we report the reconstruction of the genome-scale metabolic model of *P. pastoris*, PpaMBEL1254, which consists of 1254 metabolic reactions and 1147 metabolites divided among 8 different compartments to represent eight different organelles or regions in the cell. The metabolic model was reconstructed from the annotated genome, supplemented with knowledge available in databases [23–25]. The metabolic model was validated using experimental data obtained from this study and previous reports. Additionally, production of two heterologous proteins was implemented and flux variability analysis was performed to study the physiology and protein production capability of *P. pastoris* under oxygen-limited conditions.

2 Materials and methods

2.1 Model reconstruction

The initial reconstruction of the metabolic model was performed using the set of biochemical reactions annotated from the genome based on the gene to protein to reaction (GPR) relationship [26]. Biochemical functions were assigned to functionally ambiguous proteins based on sequence homology to proteins with known functions found in other species. Databases were employed to supplement the genome annotation during the reconstruction process, such as TransportDB for transport reactions [24]. Literature data were also employed to fill in the gaps in the annotation (Fig. 1).

Water and hydroxyl ions were not balanced as it was assumed that there are other non-enzymatic functions in the cell that use these molecules, so that they do not need to be balanced in the set composed of enzymatic reactions. The complete list of metabolic reactions and metabolites in PpaMBEL1254 can be found in the Supporting Information S1. The biomass reaction was assembled using biomass components, such as carbohydrates, amino acids and fatty acids, and their respective contributions to the formation of biomass are indicated. Data on the composition of the biomass components were taken from experimental measurements and literature data (see Supporting Information S2) [27].

2.2 Strains

Two strains of *P. pastoris* were used in this work, the type strain (DSMZ 70382, also known as CBS704) and strain GS115. The latter is the most frequently employed strain for protein production. The genome sequence of the *P. pastoris* type strain DSMZ 70382 was used and its annotation is the basis of this genome-scale metabolic model reconstruction. Because of the high degree of identity of the coding sequences between the *P. pastoris* type strain DSMZ 70382 and the *P. pastoris* GS115 (93.7% average identity on amino acid level) as well as the functional annotation, the model is suitable for representing both strains. Experimental work was performed with strain GS115 and its derivatives.

2.3 Culture condition

P. pastoris was cultured as follows. An overnight shake flask culture was used to inoculate a defined 2 L batch medium with 40 g glucose/L as a sole carbon source, to a starting optical density (OD_{600}) of

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

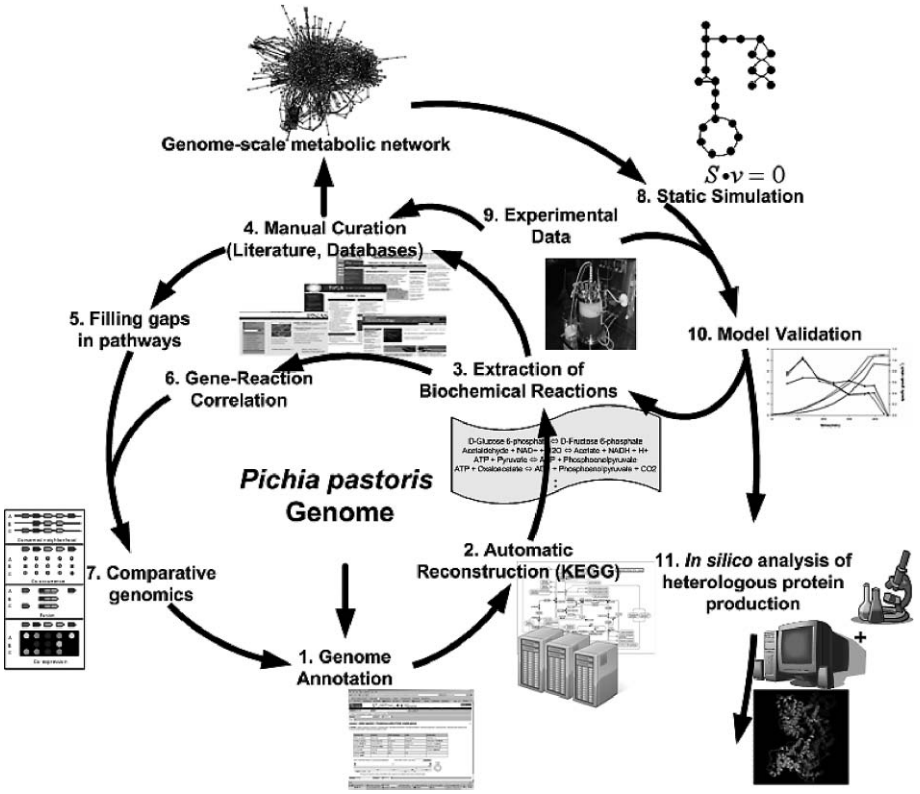


Figure 1. Metabolic reconstruction process starting with the annotated *P. pastoris* genome in the center of the circle. With the full genome sequence, the reconstruction of the metabolic model proceeds in an iterative manner, represented by the loop, with updates to the metabolic model being incorporated as new data and information are uncovered.

1.0 [28]. The cultivation was carried out in a 5.0-L bioreactor (Infors, Bottmingen-Basel, Switzerland). The fermentation temperature was controlled at 25°C, and the pH was kept at 5.0 with addition of 25% ammonium hydroxide. The dissolved oxygen concentration (DOC) was maintained above 20% of saturation by controlling the stirrer speed between 250 and 1200 rpm and the air flow between 2.0 and 5.0 L/min. Oxygen and carbon dioxide in the exiting gas were monitored during the whole process (BlueSense, Herten, Germany). Chemostat cultures were performed by adjusting the flow rate of chemostat medium [29] and the harvest rate to maintain a dilution rate D of 0.1/h after the end of the batch phase.

2.4 Sampling and analysis

Samples were taken frequently over the whole cultivation process and analyzed as described below. Three 10-mL aliquots of culture broth were centrifuged and the supernatant was saved for HPLC analysis. The pellets were washed in distilled water and re-centrifuged, transferred into weighed beakers and dried at 105°C until a constant weight was attained. The supernatant was analyzed for extracellular metabolites (glucose, ethanol, acetic acid, propionic acid, formic acid, pyruvic acid and acetaldehyde) by HPLC (Prominence, Shimadzu, Japan) using an ion exchange column Aminex HPX-87H (Bio-Rad). The mobile phase was an isocratic flow of 0.6 mL/min of 4 mM sulfuric acid. The peaks were detected by a refraction index detector (RID-10A Shimadzu, Japan) and quantified by the peak height.

2.5 Energetic parameter calculation

ATP requirement for cellular survival was determined by energetic parameters that were calculated from experimental data. There are two forms of energetic parameters: growth-associated maintenance energy (GAME) and non-growth-associated maintenance energy (NGAME). GAME (g/g DCW) is represented in the biomass equation as ATP and NGAME (mmol/g DCW/h) is represented as an independent ATP consumption reaction. The flux for the ATP consumption reaction representing NGAME is fixed to a specific value, which is calculated from experimental data. The data used to calculate the energetic parameters for *P. pastoris* were taken from glucose-limited chemostat fermentation (see Supporting information S3). The energetic parameters were calculated according to the equation

$$r_{ATP} = Y_{xATP} \cdot \mu + m_{ATP} \quad (1)$$

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

where Y_{xATP} corresponds to the GAME, μ is the specific growth rate, m_{ATP} is the NGAME and r_{ATP} is the rate of ATP being utilized by the *P. pastoris*. The values for Y_{xATP} and m_{ATP} were adjusted until prediction values matched the results obtained from the chemostat data. Detailed procedure in calculating the energetic parameters can be found elsewhere [30, 31].

2.6 Constraints-based flux analysis

For the analysis of the genome-scale metabolic model, constraints-based flux analysis was used where the internal metabolites are first balanced under the assumption of pseudo-steady state [32]. This results in a stoichiometric model $S_{ij} \cdot v_j = 0$, in which S_{ij} is a matrix of the stoichiometric coefficients of all the metabolic reactions, while i represents the metabolite index of the j^{th} reaction. An element in the vector v_j represents the flux of the j^{th} reaction given in mmol/g DCW/h. Linear programming (LP), subject to the constraints pertaining to mass conservation, reaction thermodynamics and the strain's physiological characteristics, was carried out to determine the fluxes [31]. Reaction fluxes were calculated towards the maximization or minimization of an objective function, usually the maximization of the biomass formation rate, under the constraints of upper and lower fluxes bounds for each reaction j ($v_{j,min} \leq v_j \leq v_{j,max}$). The upper and lower bounds for each metabolic reaction were set to ∞ and −∞, respectively, for reversible metabolic reactions and ∞ and zero for irreversible metabolic reactions. Constraints for glucose and oxygen uptake rates were set to values determined experimentally including the constraint for the metabolic reaction representing the maintenance energy requirement. The flux of any irreversible reaction is considered to be positive; for reversible reactions, a negative flux signifies the reverse direction of the reaction.

Growth medium in the *in silico* environment was determined by applying constraints to represent the presence or absence of specific compounds. The minimal media contained ammonia, phosphate, sulfate, oxygen and a carbon source (usually glucose). In carbon utilization studies, glucose was replaced with the carbon substrate of interest. The uptake rates for ammonia, phosphate, and sulfate were not constrained, allowing the *in silico* cell to freely utilize them. The glucose uptake rate was set to 2.88 mmol/g DCW/h, which was determined from batch cultures (Fig. 2); it was determined from the change in glucose concentrations during the exponential growth phase (time 18 and 24 h). The upper limit for oxygen uptake rate was set to 4.14 mmol/g DCW/h, which was also determined from the fermentation results.

The production of heterologous proteins was implemented by the formulation of additional reactions that describe biosynthesis of the proteins of interest. The metabolite components incorporated

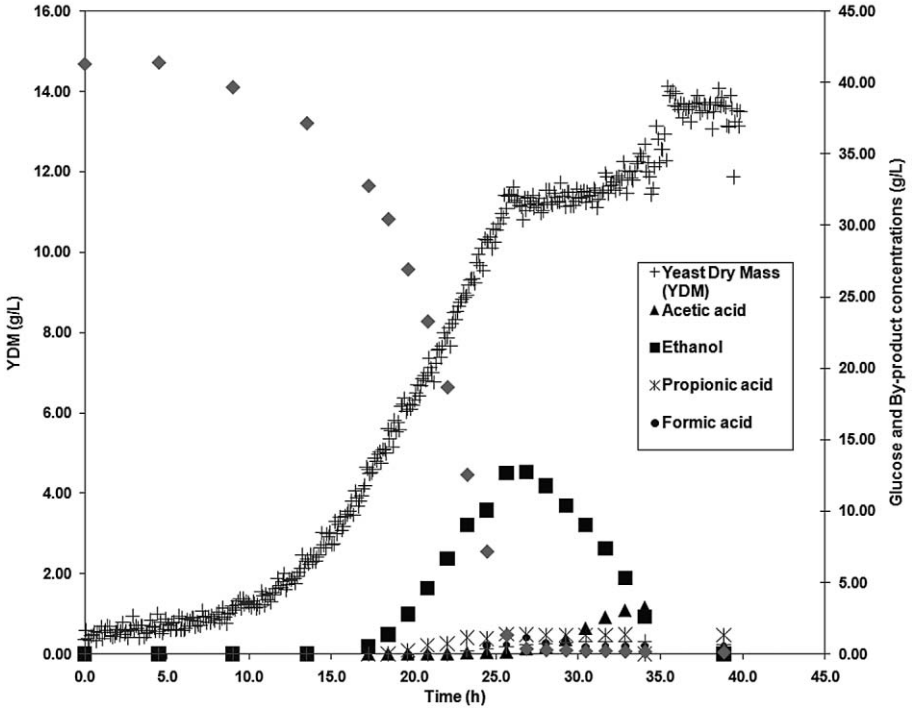


Figure 2. Batch fermentation profile of *P. pastoris* grown in a minimal medium with glucose from which the parameters that were used as constraints in PpaMBEL1254 were calculated. The region of interest for genome-scale *in silico* metabolic simulation is the exponential growth phase where the maximum growth rate occurs.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

into the reaction include nucleotides, to represent the recombinant DNA and mRNA sequence, and amino acids to represent the polypeptide sequence of the recombinant protein. Additionally, energetic requirements for the polymerization of the polypeptide were included into the reaction. The stoichiometric relation of DNA, mRNA and protein was calculated based on literature and experimental data. As the model was based on the constitutive expression system using the GAP promoter and glucose as a substrate, 100 transcript copies were assumed per gene copy, as described for the Gapdh gene of *S. cerevisiae*. For both human serum albumin (HSA) and human superoxide dismutase (hSOD) approximately 10^5 protein copies per gene copy were calculated for steady state based on the experimental data described by Marx *et al*. [33]. The protein production reaction is summarized in equation (2)

$$a\sum\mathrm{dNTP(DNA)} + b\sum\mathrm{NTP(mRNA)} + c\sum\mathrm{(Amino\ acid)} + x\mathrm{GTP(Energy\ for\ protein\ synthesis)} \rightarrow c\mathrm{Protein} + x\mathrm{GDP} + x\mathrm{Pi} \quad (2)$$

where *a, b,* and *c* are coefficients which represent the ratios at which DNA, mRNA and protein are found in the cell, respectively, and *x* is the amount of cellular energy required to drive this process. This was determined using the conversion of 2 mol GTP/1 mol amino acid in the protein sequence. Cellular energy for the polymerization of the nucleotides was considered to be negligible due to the relatively low molar ratios of DNA and mRNA to proteins. As outlined above, the molar ratio of DNA to mRNA to proteins was determined to be approximately 1:100:100 000 and the number of gene copies encoding for the proteins are 5 and 15 for HSA and hSOD, respectively. Detailed information on the nucleotide and amino acid sequence of HSA and hSOD can be found in Supporting Information S4.

3 Results and discussion

3.1 Characteristics of the genome-scale metabolic network, PpaMBEL1254

The initial draft model of PpaMBEL1254 was reconstructed by assembling a list of metabolic reactions based on the annotated genome sequence of *P. pastoris* DSMZ 70382 [21, 22]. The draft model was further supplemented with information found in biochemical and genetic databases, such as KEGG [23] and the *Pichia* genome database [25], as well as biochemical information obtained from experimental studies of *P. pastoris* or yeast-specific literature. Compartments were used to represent the different organelles present in the eukaryotic cell. Assignment for the metabolic reactions to each compartment was based on information available in literature complemented by computational localization using a combination of motif finding and prediction tools described in Mattanovich *et al*. [21]. Localization information from *S. cerevisiae* was used if no organelle localization information was available for the metabolic reaction. The metabolic reactions localized in the cytoplasm, mitochondria, nucleus, and peroxisome are represented by more than 97% of the metabolic reactions and the remaining 3% are represented by reactions in the extracellular environment, vacuole, endoplasmic reticulum, and Golgi apparatus. These percentages do not include the exchange reactions between the compartments.

The genome-scale model for the yeast *P. pastoris*, PpaMBEL1254, contains 1254 metabolic reactions and 1147 metabolites. The model accounts for 540 genes from *P. pastoris*, or approximately 9.9% of the total number of predicted ORFs (Table 1). A total of 144 metabolic reactions were also included in the model despite the absence of their respective annotations based on physiological evidence in the literature and from experiments. The metabolic reactions were divided into 50 different pathways based on their functional roles in the metabolism (see Supporting information S1).

An important metabolic reaction included in PpaMBEL1254 is the reaction representing the biosynthesis of biomass specific to *P. pastoris*. All known biomass components from different parts of the metabolic network as well as their corresponding contributions to the synthesis of biomass were combined together to construct the biomass reaction. Data taken from literature as well as experimental data measuring the components from cultures of *P. pastoris* were utilized (see Supporting information S2) [34]

3.2 Comparison of PpaMBEL1254 with the *S. cerevisiae* metabolic models

The properties of PpaMBEL1254 were compared to two previously published metabolic models of *S. cerevisiae* (Table 1) [7, 35]. In comparing the metabolic models of the two species, it was found that *S. cerevisiae* has a higher ORF representation (16% and 20% for *i*FF708 and *i*MM904, respectively) than *P. pastoris*, which has an ORF representation of 9.9% of the total number of ORF predicted. This high representation in *S. cerevisiae* can be explained by the large number of redundant genes

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Metabolic network characteristics of *P. pastoris* and comparison with *S. cerevisiae*

	S. cerevisae		*P. pastoris*
	iFF708 [39]	**iMM904 [35]**	**PpaMBEL1254**
Metabolic reactions	1175	1413	1254
Transport	349	395	293
Cytoplasmic	702	709	573
Mitochondrial	124	175	155
Other compartments	–	134	292
Metabolites	733	1230	1147
Gene coverage	708 (16%)	904 (20%)	540 (9.9%)

present in *S. cerevisiae* for several metabolic reactions [7, 35]. Additionally, the larger volume of information pertaining to *S. cerevisiae* also contributes to the larger representation due to more detailed characterization of its metabolism and genetic characteristics.

S. cerevisiae has been widely employed as a host for heterologous protein production due to its status as a model organism for eukaryotes. As *P. pastoris* has been gaining momentum as an attractive host for heterologous protein production, we compared the maximum capabilities of *P. pastoris* to produce all 20 amino acids with those of *S. cerevisiae*. Results show that the maximum production capacities for producing amino acids were the same between the *P. pastoris* and *S. cerevisiae,* with the exception of cysteine and asparagine (Fig. 3). This suggests that both yeasts possess similar capacities in synthesizing amino acids, and subsequently proteins, and thus other characteristics such as post-translational modification and high expression systems will have to be considered in choosing a suitable host for heterologous protein production.

Simulations were also performed to determine gene essentiality in minimal medium with glucose. Gene knockout simulations using constraints-based flux analysis have identified 123 essential reactions in *P. pastoris* and 1041 non-essential reactions when simulating growth in *in silico* minimal medium supplemented with glucose. The remaining 83 reactions show retarded growth compared to the 'wild type' when they were knocked out *in silico*. Thus, 9.4% of the reactions are essential and 90.6% are non-essential or partially essential to *P. pastoris* (see Supporting information S5 for detail). These percentages of essential and non-essential reactions are similar to those for *S. cerevisiae*, where 12.9% of the genes in the *S. cerevisiae* model are essential and 87.1% are non-essential [35].

The essential metabolic reactions were also examined based on the functional categories of the metabolic reactions (Fig. 4). Single knockout simulations were performed using PpaMBEL1254 and *i*MM904 [35] under the same *in silico* glucose minimal media conditions. The ratios of essential, partially essential and non-essential reactions are shown for *P. pastoris* (Fig. 4A) and *S. cerevisiae* (Fig. 4B). Comparison of the distribution of the essential reactions in each functional category between *P. pastoris* and *S. cerevisiae* shows an almost identical distribution, with the exception of the pentose phosphate pathway. Analysis of the metabolic reactions in the pentose phosphate pathway shows that the results from the knockout simulations of the *S. cerevisiae* model incorrectly predicted three reactions, glucose-6-phosphate isomerase, glucose-6-phosphate dehydrogenase, and 6-phosphogluconate dehydrogenase, to be com-

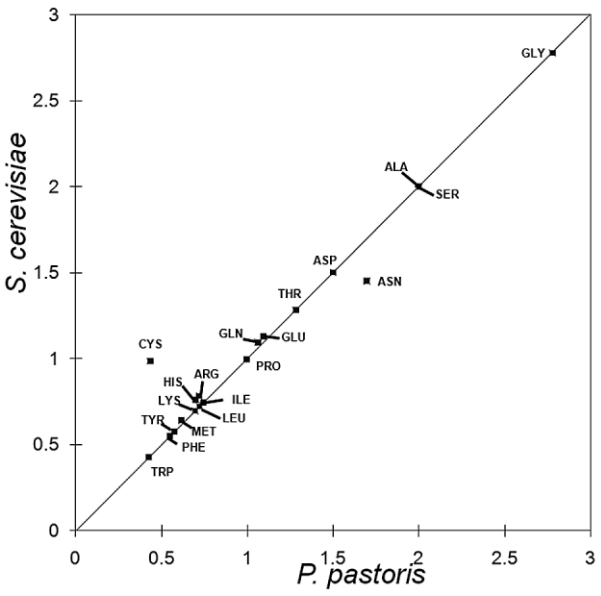


Figure 3. Maximum amino acid production capacity in *P. pastoris* and *S. cerevisiae* (mol/mol glucose). Amino acid abbreviations are as follows: ALA, alanine; ARG, arginine; ASN, asparagine; ASP, aspartate; CYS, cysteine; GLN, glutamine; GLU, glutamate; GLY, glycine; HIS, histidine; ILE, isoleucine; LEU, leucine; LYS, lysine; MET, methionine; PHE, phenylalanine; PRO, proline; SER, serine; THR, threonine; TRP, tryptophan; TYR, tyrosine; VAL, valine.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

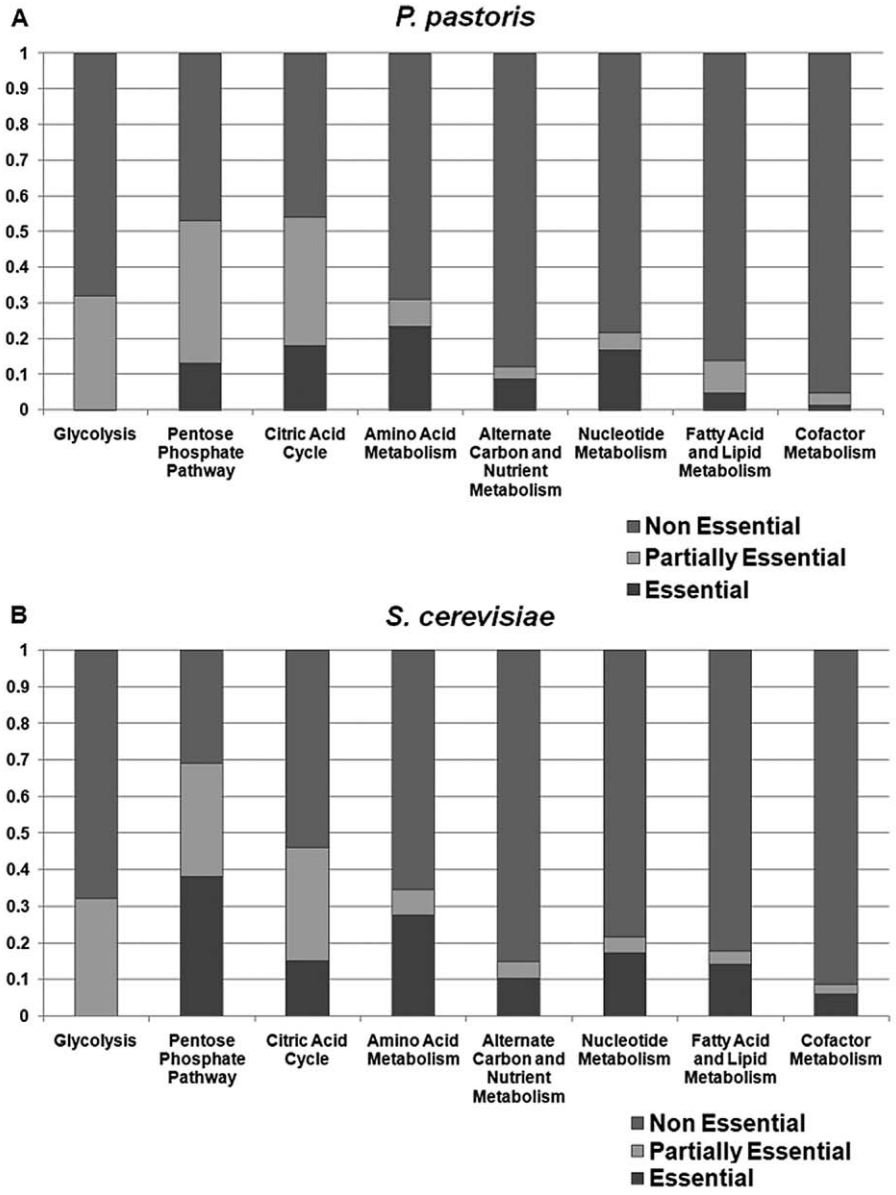


Figure 4. Results from simulations of determining essential metabolic reactions. (A) The ratios of essential reactions to partially essential reactions to non-essential reactions in each functional category in *P. pastoris* using PpaMBEL1254. (B) The ratios of essential reactions to partially essential reactions to non-essential reactions in each functional category in *S. cerevisiae* using *i*MM904.

pletely essential and was confirmed using the *S. cerevisiae* database [36]. While the results from the knockout simulation of the PpaMBEL1254 model were in good agreement with data found with *S. cerevisiae,* there is a possibility that these results are not true. Therefore, further experimental studies detailing the knockout phenotypes of *P. pastoris* are required.

3.3 Carbon source utilization

The *in silico* growth capabilities of *P. pastoris* represented by PpaMBEL1254 were examined and compared with information found in literature [37]. Growth characteristics with 25 different carbon sources were examined by substituting each compound into the *in silico* minimal medium (Table 2). Among the 25 different carbon sources examined, 3 of them, 1,2 propanediol, L-rhamnose, and mannitol, failed to support growth of *in silico P. pastoris*, which contradicts the experimental evidence [37]. Investigation of the metabolic pathways for these substrates elucidated several essential metabolic reactions in the metabolism of these compounds that were unannotated and absent from the metabolic model. Addition of the respective non-gene-associated metabolic reactions allowed *in silico P. pastoris* to grow on these three compounds (Table 2). The remaining 22 substrates were in agreement with data regarding the substrate uti-

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. List of substrates which can or cannot be metabolized by *P. pastoris*

	Experimental	*In silico*
D-Glucose	+	+
DL-Lactate	+	+
Ethanol	+	+
Glycerol	+	+
Sorbitol	+	+
Succinate	+	+
Trehalose	+	+
D-1,2 Propanediol	+	+
L-Rhamnose	+	+
Mannitol	+	+
Starch	+	+
D-Arabinose	–	–
D-Ribose	–	–
D-Xylose	–	–
Galactose	–	–
Lactose	–	–
L-Arabinose	–	–
L-Sorbose	–	–
L-Tryptophan	–	–
Maltose	–	–
Sucrose	–	–
Thiamine	–	–
Xylitol	–	–

+, Cell growth is observed experimentally or *in silico;* –, cell growth is not observed experimentally or *in silico.*

lization capability of *P. pastoris*. Of the 22 substrates, 7 of the substrates displayed biomass formation and the remaining 15 substrates were unable to generate biomass.

Using glucose as a carbon source, the specific growth rate (*e.g.*, biomass formation rate) was estimated *in silico* using the PpaMBEL1254. At the experimentally measured glucose uptake rate of 2.88 mmol/g DCW/h, which was determined from the batch fermentation run in Fig. 2 during the exponential growth phase (18 and 24 h), and the oxygen uptake rate of 4.14 mmol/g DCW/h, the growth rate calculated using the objective function of maximizing the biomass formation was 0.244/h, which is close to the experimentally measured maximum specific growth rate of 0.2/h; thus, the maximum experimentally observed specific growth rate is 83% of the value predicted *in silico*.

3.4 Simulation of heterologous protein production

P. pastoris has been receiving much attention as a platform for the production of heterologous proteins for biotechnological applications. Here, we used PpaMBEL1254 as a tool to design strategies for engineering *P. pastoris* towards enhanced heterologous protein production. Production of heterologous proteins was implemented in the metabolic model by introducing several metabolic reactions as described in the Materials and methods section. Key components represented in the biosynthesis of the recombinant protein include the nucleotides necessary to represent the genes encoding the target protein, the mRNA composition, and amino acids that make up the protein. The molar ratios for the components were measured in the cell (see Materials and methods for details). Additionally, the energy requirement for the polymerization of the protein was calculated and included in the metabolic network. Because each protein sequence is unique, different versions of the metabolic model were constructed to reflect those differences (Table 3). Here, the production of the proteins HSA and hSOD is represented in the metabolic models PpaMBEL1254_HSA and PpaMBEL1254_hSOD, respectively [27, 38]. Before examining the effect of oxygen limitation on the production of heterologous proteins, the *in silico* maximum production rates were calculated. Using the objective function of maximizing the intracellular production rate of HSA or hSOD, the simulation results showed that the maximum production rates of HSA and hSOD were 0.28 and 0.32 g/g DCW/h, respectively. The proteins of interest to be produced by *P. pastoris* are usually targeted for secretion. Thus, an interesting future study will be to correlate the protein production with secretion.

3.5 *In silico* analysis of heterologous protein production in *P. pastoris* under oxygen limitation

Production of heterologous proteins is impacted by a number of factors. In *P. pastoris*, it has been reported that hypoxic culture condition leads to the increase in productivity of the heterologous proteins by two- to threefold over fully aerobic condition [38]. Thus, flux variability analysis was carried out to investigate the impact that oxygen uptake rate has on protein production within a range of possible oxygen uptake rates; the oxygen uptake rate was used as a constraint to simulate the hypoxic environment for PpaMBEL1254_HSA and PpaMBEL1254_hSOD. The growth rate was also examined for each state to determine an optimal physiological state that balances growth and heterologous protein production. Also, it was noted from fed-batch cultures during the heterologous protein production that ethanol is also produced [38]. In that experiment, the ethanol concentration increased from 2 to 10 g/L when the cell concentration increased from 65 to 71 g DCW/L. This gives

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3. Metabolic reactions representing the production of the heterologous proteins HSA and hSOD

Protein name	Metabolic reaction
Human serum albumin	8.05e-05 pDNA + 8.41e-03 pRNA + 0.99 pAA → HSA
DNA	0.923 damp + 0.695 dcmp + 0.923 dtmp + 0.695 dgmp → pDNA
RNA	0.889 damp + 0.669 dgmp + 0.669 dcmp + 0.889 dump → pRNA
Amino acid	0.8928 alaL + 0.4032 argL + 0.2448 asnL + 0.5184 aspL + 0.504 cysL + 0.288 gluL + 0.8928 glnL + 0.1872 glyL + 0.2304 hisL + 0.1296 ileL + 0.9216 leuL + 0.864 lysL + 0.1008 metL + 0.504 pheL + 0.3456 proL + 0.4032 serL + 0.4176 thrL + 0.0288 trpL + 0.2736 tyrL + 0.6192 valL + 1.22 gtp → pAA + 1.22 gdp +1.22 pi
Human superoxide dismutase	8.97e-05 pDNA + 9.46e-03 pRNA + 0.99 pAA → hSOD
DNA	0.823 damp + 0.795 dcmp + 0.823 dtmp + 0.795 dgmp → pDNA
mRNA	0.791 damp + 0.765 dgmp + 0.765 dcmp + 0.791 dump → pRNA
Amino acid	0.6284 alaL + 0.3625 argL + 0.4399 asnL + 0.6912 aspL + 0.2514 cysL + 0.6284 gluL + 0.1885 glnL + 1.571 glyL + 0.5027 hisL + 0.5656 ileL + 0.5656 leuL + 0.6912 lysL + 0.0628 metL + 0.2514 pheL + 0.3142 proL + 0.0628 serL + 0.5027 thrL + 0.0628 trpL + 0 tyrL + 0.8798 valL + 0.31 gtp → pAA + 0.31 gdp + 0.31 pi

an ethanol secretion rate of 0.511 mmol/g DCW/h. Thus, an additional constraint of ethanol production (0.5 mmol/g DCW/h) was introduced during the simulation.

Results from the simulations are represented in contour plots with the oxygen uptake rate and protein production rate on the horizontal and vertical axes, respectively, and the specific growth rate represented by the contour lines (Fig. 5). The plot shows that the specific growth rate decreases as oxygen uptake rate decreases as expected. Likewise, the specific growth rate also decreases as the protein production rate increases for high levels of oxygen uptake rate (more than 5 mmol/g DCW/h). However, when the oxygen uptake rate is less than 5 mmol/g DCW/h, the specific growth rate increases or remains the same as the protein production rate increases for a given oxygen uptake rate, up to an optimal point. This optimal point displays the maximum rate of protein production that can be achieved without retarding the growth rate at a specified oxygen uptake rate. Beyond the optimal point for a given oxygen uptake rate, the specific growth rate begins to decrease, as observed for higher oxygen uptake rate cases. The optimal points can be seen to follow an inverse linear path where the optimal protein production rate increases as the oxygen uptake rate decreases as indicated by the solid line in Fig. 5. This relationship between the optimal protein production rate and the oxygen uptake rate was observed in hypoxic fed-batch cultures where the productivities of three different proteins produced by *P. pastoris* increased when oxygen limitation was applied [38].

While both HSA and hSOD showed increasing optimal protein production rates as the oxygen uptake rate decreased, the production rates of the two proteins responded differently to decreasing levels

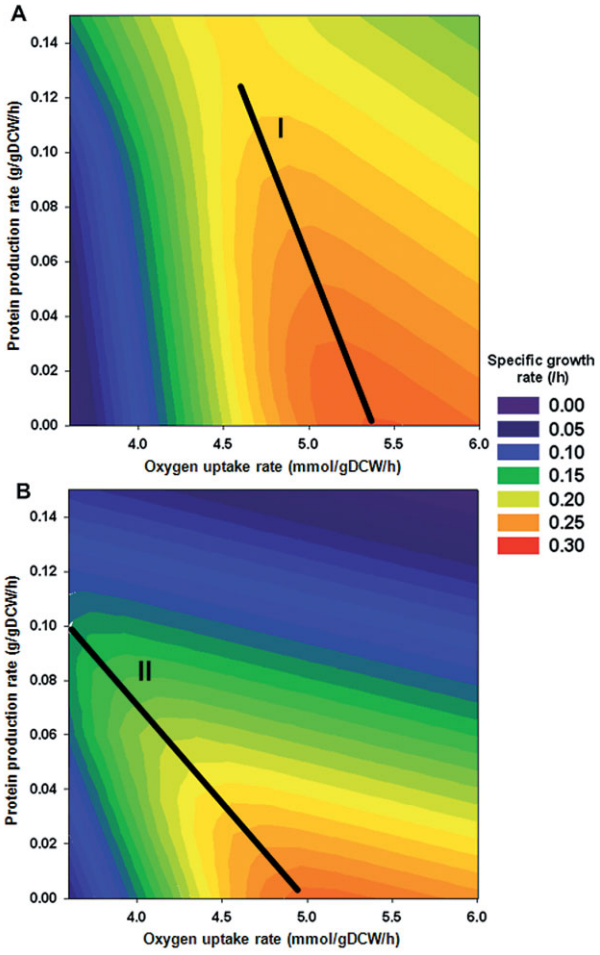


Figure 5. Simulation results showing the impact of oxygen limitation on protein production and growth rate. The solid black line represents the points along which the optimal protein production rate occurs for a given oxygen uptake rate. (A) Flux variability analysis of HSA production and specific growth rate to different oxygen uptake rates, and (B) flux variability analysis of hSOD production and specific growth rate to different oxygen uptake rates.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

of oxygen uptake rates. The line representing the optimal protein production for HSA (Fig. 5A I) has a steeper slope compared with the line for hSOD production (Fig. 5B II). This indicates that the optimal production level for HSA is more sensitive to changing oxygen uptake rate compared with hSOD. On the other hand, the HSA production rate is less sensitive to the specific growth rate compared to the hSOD production rate; the hSOD production rate increase is accompanied with the rapid change in specific growth rate. These differences can be attributed to the different properties in expressing these proteins in *P. pastoris* and is incorporated in the protein production reactions in the metabolic model. In *P. pastoris,* there are 15 copies for the gene that encodes for hSOD, compared to 5 copies for the gene encoding for HSA. Due to the higher number of gene copies, hSOD production would require more cellular resources including DNA, RNA and amino acids towards the production of the protein and thus negatively affect cell growth (Table 3). HSA production, on the other hand, requires less precursors, and thus, it would not redirect too much cellular resources away from cell growth.

4 Concluding remarks

In this paper, we have presented the first reconstruction of the genome-scale metabolic model of the methylotrophic yeast *P. pastoris*, PpaMBEL1254, based on the annotation of the recently sequenced genome of *P. pastoris* and available literature information. As this is the first reconstruction of the metabolic model for this species, we take another step towards attaining a more comprehensive understanding of its metabolic capabilities. Here, we incorporated the production of two heterologous proteins, HSA and hSOD, and investigated how the oxygen supply affects its physiology and capabilities in producing these two proteins. Analysis of the physiology under oxygen limiting conditions determined an optimum state that maximizes protein production without decreasing the growth rate for a given oxygen uptake rate. As a result, fermentation strategies can be designed to optimize protein production and growth rate by limiting the oxygen supply. With this metabolic model as a basis, further studies in understanding and engineering *P. pastoris* to increase protein production can be explored, such as gene knockout and gene overexpression simulations.

This work was supported by the Korean Systems Biology Research Project (20100002164) of the Ministry of Education, Science and Technology (MEST) through the National Research Foundation of Korea, by the Austrian Science Fund (FWF), project I37-B03, and the Austrian Research Promotion Agency (program FHplus).

The authors have declared no conflict of interest.

5 References

[1] Kim, T. Y., Sohn, S. B., Kim, H. U., Lee, S. Y., Strategies for systems-level metabolic engineering. *Biotechnol. J.* 2008, *3*, 612–623.

[2] Park, J. H., Lee, K. H., Kim, T. Y., Lee, S. Y., Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 7797–7802.

[3] Lee, J. Y., Jang, Y. S., Lee, J., Papoutsakis, E. T., Lee, S. Y., Metabolic engineering of *Clostridium acetobutylicum* M5 for highly selective butanol production. *Biotechnol. J.* 2009, *4*, 1432–1440.

[4] Kim, H. U., Kim, T. Y., Lee, S. Y., Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen *Acinetobacter baumannii* AYE. *Mol. Biosyst.* 2010, *6*, 339–348.

[5] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I. *et al.*, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 1777–1782.

[6] Sheikh, K., Forster, J., Nielsen, L. K., Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol. Prog.* 2005, *21*, 112–121.

[7] Duarte, N. C., Herrgard, M. J., Palsson, B. O., Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 2004, *14*, 1298–1309.

[8] David, H., Ozcelik, I. S., Hofmann, G., Nielsen, J., Analysis of *Aspergillus nidulans* metabolism at the genome-scale. *BMC Genomics* 2008, *9*, 163.

[9] Ma, H., Sorokin, A., Mazein, A., Selkov, A. *et al.*, The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* 2007, *3*, 135.

[10] de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., Nielsen, L. K., AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 2010, *152*, 579–589.

[11] Graf, A., Dragosits, M., Gasser, B., Mattanovich, D., Yeast systems biotechnology for the production of heterologous proteins. *FEMS Yeast Res.* 2009, *9*, 335–348.

[12] Marx, H., Mattanovich, D., Sauer, M., Overexpression of the riboflavin biosynthetic pathway in *Pichia pastoris*. *Microb. Cell Fact.* 2008, *7*, 23.

[13] Cregg, J. M., Cereghino, J. L., Shi, J., Higgins, D. R., Recombinant protein expression in *Pichia pastoris*. *Mol. Biotechnol.* 2000, *16*, 23–52.

[14] Macauley-Patrick, S., Fazenda, M. L., McNeil, B., Harvey, L. M., Heterologous protein production using the *Pichia pastoris* expression system. *Yeast* 2005, *22*, 249–270.

[15] Hamilton, S. R., Davidson, R. C., Sethuraman, N., Nett, J. H. *et al.*, Humanization of yeast to produce complex terminally sialylated glycoproteins. *Science* 2006, *313*, 1441–1443.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[16] Hamilton, S. R., Gerngross, T. U., Glycosylation engineering in yeast: the advent of fully humanized yeast. *Curr. Opin. Biotechnol.* 2007, *18*, 387–392.
[17] Jacobs, P. P., Geysens, S., Vervecken, W., Contreras, R., Callewaert, N., Engineering complex-type *N*-glycosylation in *Pichia pastoris* using GlycoSwitch technology. *Nat. Protoc.* 2009, *4*, 58–70.
[18] Gasser, B., Mattanovich, D., Antibody production with yeasts and filamentous fungi: On the road to large scale? *Biotechnol. Lett.* 2007, *29*, 201–212.
[19] Gerngross, T. U., Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nat. Biotechnol.* 2004, *22*, 1409–1414.
[20] Potgieter, T. I., Cukan, M., Drummond, J. E., Houston-Cummings, N. R. *et al.*, Production of monoclonal antibodies by glycoengineered *Pichia pastoris*. *J. Biotechnol.* 2009, *139*, 318–325.
[21] Mattanovich, D., Graf, A., Stadlmann, J., Dragosits, M. *et al.*, Genome, secretome and glucose transport highlight unique features of the protein production host *Pichia pastoris*. *Microb. Cell Fact.* 2009, *8*, 29.
[22] De Schutter, K., Lin, Y. C., Tiels, P., Van Hecke, A. *et al.*, Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat. Biotechnol.* 2009, *27*, 561–566.
[23] Aoki, K. F., Kanehisa, M., Using the KEGG database resource. *Curr. Protoc. Bioinformatics* 2005, *Chapter 1*, Unit 1 12.
[24] Ren, Q., Chen, K., Paulsen, I. T., TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* 2007, *35*, D274–279.
[25] Mattanovich, D., Callewaert, N., Rouze, P., Lin, Y. C. *et al.*, Open access to sequence: browsing the *Pichia pastoris* genome. *Microb. Cell Fact.* 2009, *8*, 53.
[26] Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* 2003, *4*, R54.
[27] Carnicer, M., Baumann, K., Toplitz, I., Sanchez-Ferrando, F. *et al.*, Macromolecular and elemental composition analysis and extracellular metabolite balances of *Pichia pastoris* growing at different oxygen levels. *Microb. Cell Fact.* 2009, *8*, 65.
[28] Maurer, M., Kuhleitner, M., Gasser, B., Mattanovich, D., Versatile modeling and optimization of fed batch processes for the production of secreted heterologous proteins with *Pichia pastoris*. *Microb. Cell Fact.* 2006, *5*, 37.
[29] Dragosits, M., Stadlmann, J., Albiol, J., Baumann, K. *et al.*, The effect of temperature on the proteome of recombinant *Pichia pastoris*. *J. Proteome Res.* 2009, *8*, 1380–1392.
[30] Borodina, I., Krabben, P., Nielsen, J., Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* 2005, *15*, 820–829.
[31] Varma, A., Palsson, B. O., Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 1994, *60*, 3724–3731.
[32] Gombert, A. K., Nielsen, J., Mathematical modelling of metabolism. *Curr. Opin. Biotechnol.* 2000, *11*, 180–186.
[33] Marx, H., Mecklenbrauker, A., Gasser, B., Sauer, M., Mattanovich, D., Directed gene copy number amplification in *Pichia pastoris* by vector integration into the ribosomal DNA locus. *FEMS Yeast Res.* 2009, *9*, 1260–1270.
[34] Dragosits, M., Stadlmann, J., Graf, A., Gasser, B. *et al.*, The response to unfolded protein is involved in osmotolerance of *Pichia pastoris*. *BMC Genomics* 2010, *11*, 207.
[35] Mo, M. L., Palsson, B. O., Herrgard, M. J., Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 2009, *3*, 37.
[36] Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N. *et al.*, CYGD: The Comprehensive Yeast Genome Database. *Nucleic Acids Res.* 2005, *33*, D364–368.
[37] Barnett, J. A., Payne, R. W., Yarrow, D., *Yeasts: Characteristics and Identification,* Cambridge University Press, Cambridge 2000.
[38] Baumann, K., Maurer, M., Dragosits, M., Cos, O. *et al.*, Hypoxic fed-batch cultivation of *Pichia pastoris* increases specific and volumetric productivity of recombinant proteins. *Biotechnol. Bioeng.* 2008, *100*, 177–183.
[39] Forster, J., Famili, I., Palsson, B. O., Nielsen, J., Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *OMICS* 2003, *7*, 193–202.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

6.2.2 Other research papers

Tüchler T, Velez G, Graf A, Kreil DP. *BibGlimpse: The case for a light-weight reprint manager in distributed literature research.* BMC Bioinformatics 2008, October **9**:406

Contribution: I was responsible for implementing a Windows solution for BibGlimpse, using the Cygwin environment.

Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B. *Genome, secretome and glucose transport highlight unique features of the protein production host. Pichia pastoris.* Microbial Cell Factories 2009, June **8**:29

Mattanovich D, Callewaert N, Rouzé P, Lin YC, Graf A, Redl A, Tiels P, Gasser B, De Schutter K. *Open access to sequence: Browsing the Pichia pastoris genome.* Microbial Cell Factories 2009, October **8**:53

Contribution: I evaluated the quality of the assembly and did the functional annotation of the genome, consisting of gene prediction, promoter prediction, tRNA prediction and functional annotation of the genes using homology based tools. I took part in the manual curation of the putative ORFs. I conceived, implemented and evaluated the prediction pipeline for the secretome and was involved in the comparison of the *in-silico* predicted and experimentally measured secretome. I supervised the implementation of the Genome Browser and provided the input data.

Dragosits M, Stadlmann J, Graf A, Gasser B, Maurer M, Sauer M, Kreil DP, Altmann F, Mattanovich D. *The response to unfolded protein is involved in osmotolerance of Pichia pastoris.* BMC Genomics 2010, March **11**:207

Baumann K, Carnicer M, Dragosits M, Graf AB,Stadlmann J, Jouhten P, Maaheimo H, Gasser B, Albiol J, Mattanovich D, Ferrer P. *A multi-level study of recombinant Pichia pastoris in different oxygen conditions as knowledge base for strain improvement.* (submitted)

Contribution: The publications of Dragosits et al. [2010] and Baumann et al. [2010] are part of the GENOPHYS project. In both publications I was involved in the micro-array design and responsible for the statistical analysis of the arrays. For all of the transcriptomics data in this project the analyis pipeline, described in the results section, was used.

BMC Bioinformatics

Software

Open Access

BibGlimpse: The case for a light-weight reprint manager in distributed literature research

Thomas Tüchler*[1], Golda Velez[2], Alexandra Graf[1] and David P Kreil[1]

Address: [1]Chair of Bioinformatics, Boku University, AT-1190 Muthgasse 18, Vienna, Austria and [2]Internet WorkShop, 2921 South Cottonwood Lane, Tucson, AZ 85713, USA

Email: Thomas Tüchler* - thomas.tuechler@boku.ac.at; Golda Velez - gv@btuscon.com; Alexandra Graf - alexandra.graf@boku.ac.at; David P Kreil - bibglimpse08@boku.ac.at

* Corresponding author

Published: 1 October 2008

BMC Bioinformatics 2008, **9**:406 doi:10.1186/1471-2105-9-406

Received: 1 April 2008
Accepted: 1 October 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/406

© 2008 Tüchler et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: While text-mining and distributed annotation systems both aim at capturing knowledge and presenting it in a standardized form, there have been few attempts to investigate potential synergies between these two fields. For instance, distributed annotation would be very well suited for providing topic focussed, expert knowledge enriched text corpora. A key limitation for this approach is the availability of literature annotation systems that can be routinely used by groups of collaborating researchers on a day to day basis, not distracting from the main focus of their work.

Results: For this purpose, we have designed BibGlimpse. Features like drop-to-file, SVM based automated retrieval of PubMed bibliography for PDF reprints, and annotation support make BibGlimpse an efficient, light-weight reprint manager that facilitates distributed literature research for work groups. Building on an established open search engine, full-text search and structured queries are supported, while at the same time making shared collections of annotated reprints accessible to literature classification and text-mining tools.

Conclusion: BibGlimpse offers scientists a tool that enhances their own literature management. Moreover, it may be used to create content enriched, annotated text corpora for research in text-mining.

Background

The published biomedical literature is growing at a tremendous pace [1,2]. Although access has increased considerably with the availability of most published research in electronic form (typically in the Portable Document Format, PDF), researchers now face a considerable challenge in organizing and managing comprehensive and up to date manuscript collections.

Existing literature management tools

Currently, a wide range of software is offered for searching and organizing published manuscripts. With approaches ranging from open-source bibliography managers for the desktop to professional online abstracting services, supported feature sets differ substantially (see Table 1).

While Google Scholar and other web search engines provide a full-text index of public documents online [3], there is no mechanism supporting personal collections of

Table 1: Feature comparison

	Management of personal collections	Full-text search	Bibliography-PDF match automated	Personal annotations	Shared collections	Approximate search patterns	Search with synonyms	Index browsing	Ease of I/O with external tools[10]	Free use or open-source	Requirements
BibGlimpse	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes/Yes	Yes[11]	Bash, Perl, Apache
Abstracting services											
PubMed	Prt[1]	No	Prt[3]	No	No	No	Prt	Yes	No/Yes	Yes	Web
ISI Web of	Prt[1]	No	No	No	No	No	No	Yes	No/Yes	No	Web
Science Search engine											
Google Scholar	No	Yes[2]	Prt[3]	No	No	No	No	No	No/No	Yes	Web
Reference managers											
EndNote	Yes	No	No[4]	Yes	No	No	No	Prt	No/Yes	No	Win/Mac
RefBase	Yes	No	No	Yes	Yes	No	No	Yes[9]	No/Yes	Yes	XAMPP[12]
iPapers	Yes	Yes	No[5]	Yes	No	No	No	No	No/Yes	Yes	Mac
Social bookmarking											
CiteULike	Yes	No	No[6]	Prt[7]	Yes	No	No	Tags	No/Prt	Yes	Web
Connotea	Yes	No	No[6]	Yes	Yes	No	No	Tags	No/Prt	Yes	Web
Digital library											
Greenstone	Yes	Yes	No	Yes[8]	Yes[8]	Yes	No	Yes	Yes/Yes	Yes	Perl, Apache

Feature comparison between BibGlimpse and other bibliographic software tools. 'Prt' indicates that a feature is *partly* supported. [1] Searches and selected references can be stored in collections. [2] Covers the first 120 kB of open access papers and a non-disclosed list of publishers; all recent articles by Elsevier publications are, *e.g.*, excluded. PubMed is indexed with a lagtime of up to a year. [3] Linking bibliography to full text; not the other way around. [4] For a given reference an automated online search for the corresponding full-text can be performed. [5] Requires file named PMID.pdf, where PMID is the PubMed ID, to download bibliography from PubMed. [6] Needs link to website, not link to PDF. Retrieval is not generic, but publisher site tailored. [7] Notes are not searchable. [8] Greenstone is a tool to *build* digital libraries, so library needs to be designed first. [9] MySQL database can be queried directly by passing MySQL search strings. [10] *Input* means that results of external tools can easily be input into the system for subsequent integrated analysis and searches. *Output* means that data in the system can be output to external tools. [11] Code free for non-profit and academic use. [12] Package providing PHP, MySQL and Apache for different platforms.

online documents to be annotated or searched, and many articles are not even publicly available. Desktop search tools, in turn, do not facilitate the sharing of documents and their annotation between collaborators. Dedicated bibliography management software like RefBase and WikIndx (see [4]) or the popular commercial tools EndNote [5] and RefWorks [6] have a different focus: there is no support for full-text search and, depending on the tool, complex queries are not supported or data cannot be shared online. On the other hand, full-featured software for digital libraries (like Greenstone [7]) is not only difficult to set up but also impractical for casual use. Long forms to fill in, which make filing a PDF reprint considerably more tedious than just saving it to disk, reduce acceptance in day-to-day work.

With these challenges and unmet needs in mind, we here introduce the concept of a light-weight reprint manager for the joint creation and exploitation of content enriched collections of expert annotated full-text reprints. The BibGlimpse implementation provides a simple framework for distributed literature research especially designed for this purpose. In particular, besides allowing full-text searches on manuscript collections, the support for searchable personal annotations and the ability of sharing these with colleagues are key features of the system. The automated creation of bibliographic records for a simple PDF reprint, moreover, substantially facilitates the uptake of the system. Such automatic retrieval of bibliographic records from full text PDF files is a feature that has, to the best of our knowledge, not yet been implemented elsewhere.

Implementation

A defining design requirement for BibGlimpse was that users can file a new PDF reprint by *simply saving or copying it to disk*, or by uploading it online without being prompted to fill in any forms. The lightweight filesystem-based approach also means that users can easily import an entire collection of reprints using just a single command or 'drag and drop' operation for copying the directories of their PDF files. Users can of course also manually edit or add bibliographic records, personal annotation, and supplement files. Medline, BibTeX, and RIS formats are supported. When a Medline or RIS format record is available but not a BibTeX record, the system automatically creates one. Extending standard Webglimpse functionality (*cf.* Figure 1), the index is updated in the background to cover the bibliographic record, any user annotation, and the extracted full-text of an article.

Automated Medline retrieval

When a new reprint is detected in the indexed file system, BibGlimpse extracts the plain text from the PDF file and constructs queries to automatically obtain a matching bibliographical record from PubMed. Queries are compiled with a generic pattern recognition approach, avoiding a need to prepare numerous journal specific templates for the extraction of bibliographic information from the PDF [8]. To this end, BibGlimpse first discards unspecific or irrelevant text sections, *i.e.* lines that are presumably not suited for constructing a meaningful PubMed query. This heuristic filter comprises simple rules, like a line must at least contain five characters, two words and at least one word with more than four letters. It also excludes lines that are likely to contain figure captions ('Fig.'), contact or company addresses ('Inc.'). Moreover, since citations can easily confound query construction (and are often found on the first page of articles where articles do not start on a fresh page), lines matching regular expressions that target such citations are equally removed.

By means of further heuristics, several features are then extracted from the prefiltered text in order to come up with query strings for the putative title, authors and abstract of the reprint in question: While identification of the putative manuscript title mainly focuses on the position in the text (*e.g.*, the title is assumed to be located within the top lines of the document, it is supposed not to exceed a certain length and to be separated from the remainder by blank lines), features relevant for finding the manuscript's authors strongly rely on punctuation. To illustrate that, consider the following example line, obtained after converting a reprint PDF to text:

Xiaolei Yu,1 Milorad Susa,2 Cornelius Knabbe,2 Rolf D. Schmid,1 and Till T. Bachmann1*

Without knowing that 'Xiaolei' or 'Cornelius' are names, we can characterize this line by the following features: it contains 6 delimiters (including punctuation like the comma and ampersand symbols as well as the word 'and'), 5 author affiliation symbols including 1 corresponding author asterisk, 2 middle name initials (*e.g.* one in Rolf *D*. Schmid), and 10 out of 13 words start with capital letters (not counting the initials). Based on such observations, we extract the following 8 characteristic features: per-word ratios for delimiters (6/13 in this example), footnote symbols (5/13), middle name initials (2/13), and words starting with capital characters (10/13); moreover, indicators for the existence of an asterisk (1), colon characters (0), and whether the whole line is written in capital letters (0), as well as the distance to the next putative headline (counted in lines).

To construct a PubMed query string with the putative authors of a manuscript, these features are first computed for each line in the PDF and then exploited to identify 'author lines', *i.e.* lines containing the authors of a manuscript. For classification, we trained a radial basis function

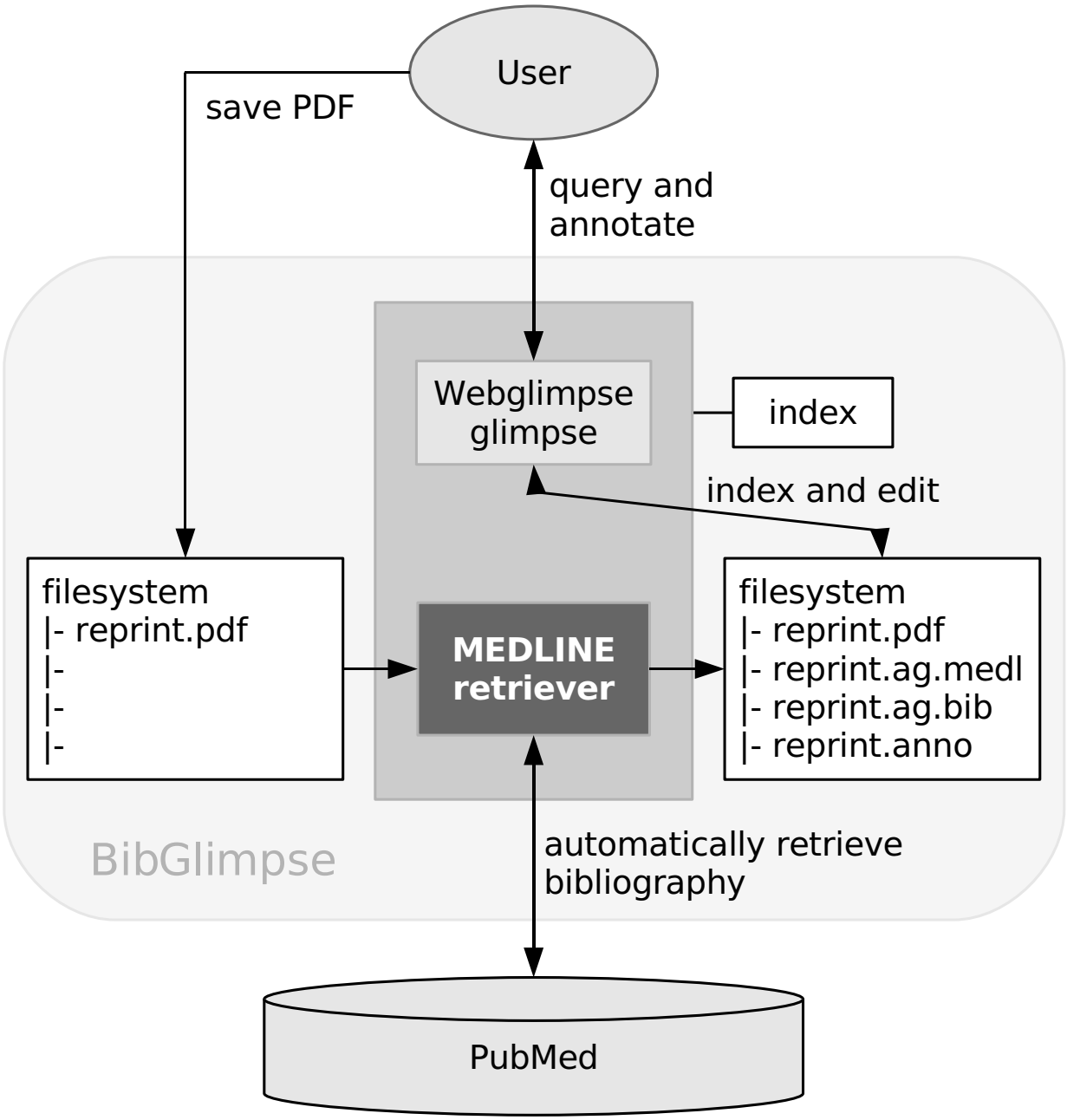


Figure 1
BibGlimpse scheme. The figure schematically illustrates how BibGlimpse incorporates automated Medline retrieval into the Webglimpse search environment. Saved PDFs are automatically matched with a Medline record and indexed. For integration with external tools, all data are directly available in flatfile format.

kernel support vector machine (SVM) on different representative training sets of 20 author and 20 other text lines. Using only the above 8 features, the finally selected support vectors achieved a respectable recall of 95% true positives (TP) with only 8% false positives (FP) as assessed in an independent test set of over 2000 candidate lines. A robust typical recall of more than 85% TP with about 10% FP in an investigation of alternative random training data

indicated a well chosen feature set. Details regarding the classifier and the test corpus employed can be found in the online Supplement.

The good performance achieved in author-line identification allows targeted PubMed queries for authors with only a reasonable number of FP non-author text queries submitted. At this point, we wish to emphasize that PubMed hits are not only gathered by querying for the putative authors, but also from searching for the presumed title, abstract, and the digital object identifier, if available. The aim of the described filters is therefore not to extract the true bibliography directly from the manuscript, but rather to construct a set of query strings, that allow a retrieval of this information from PubMed with as few requests as possible.

Eventually, even obtaining a unique PubMed hit for a query string is not sufficient to assure that manuscript and bibliography match. Each retrieved Medline record is thus additionally cross-checked against the extracted full-text by reverse queries of title, authors and abstract.

Technically, the retrieval of Medline records from PubMed was implemented in Perl, such that it can be run from a single stand-alone script. This makes the feature easily accessible for other environments.

Results and discussion

Performance

The real-world performance of the complete system for the automated retrieval of bibliographic records was assessed on a test set of over 1000 PubMed listed manuscripts covering about 200 different journals. BibGlimpse was able to retrieve the correct PubMed records for 95% of these manuscripts with only 0.5% spurious hits and the remaining 4.5% being tagged as not-found. This shows that the combination of multiple heuristic queries and the cross-checking process yields an overall robust performance. There are some cases, however, where retrieval might not succeed. Consider the following text line returned from a 'pdftotext' conversion:

Gene Expression Profiles in Formalin-Fixed, ParaffinEmbedded Tissues Obtained with a Novel Assay for Microarray Analysis, Marina Bibikova,1 Joanne M. Yeakley,1 Eugene Chudin,1 Jing Chen,1 Eliza Wickham,1 Jessica WangRodriguez,2 and Jian-Bin g Fan1* (1 Illumina, Inc., San Diego, CA)

In this example, no newline character separates the title of the manuscript from the authors and their addresses. Moreover, the corresponding Medline entry (PMID 15563488) lists 'paraffin-embedded tissues' in the title, instead of the poorly converted 'ParaffinEmbedded' in the text version of the PDF. The cross-checking step may in such cases find no match. Also, single PDF files containing multiple short comments or letters to the editor may be assigned to the bibliography for the first comment or letter. Finally, the system may not be able to distinguish preprints from actually published manuscripts if they have the same title, authors, and abstract. In summary, however, these limitations only apply to a very small fraction of reprints files.

Application

Freeing researchers from a need to look up or enter bibliographical records and giving them an opportunity of annotating their reprints together with full-text query capabilities not only raises acceptance of the system in day-to-day usage but has profound practical implications for knowledge discovery, sharing, and retrieval. We illustrate this with a few examples (Figure 2). Important information, such as the cell line types employed, is often not contained in the abstract but can be queried by full-text search (*e.g.*, a query for cell line HCT116, Figure 2). Other information may even only be *implicit* in the full manuscript text but can be captured explicitly by user annotation (*e.g.*, user annotation as 'p53 wildtype').

The challenge of searching natural language text, of course, remains. Depending on the application domain, it may be valuable to consider extending the free-form annotation of articles by keywords from ontologies with controlled vocabularies or controlled subsets of natural language [9]. External tools can easily be integrated to either automatically generate these, or at least assist the user in a manual curation process [10]. This is much facilitated by having all internally stored information available as plain text files on disk. Through structured query support such additional fields can be queried directly. Having straightforward means for the integration of text-mining tools thus allows future developments to further assist users in extracting searchable knowledge from their annotated reprint collections.

The availability of installation support and minimal software prerequisites make BibGlimpse accessible for small groups of collaborating researchers. Building on Webglimpse [11], our system inherits a search engine that has been actively maintained and supported for over ten years, is easy to install and maintain using an administrative web interface, and only needs Perl, a running web server and a utility to extract plain text from PDF files, *e.g.*, from 'xpdf' [12].

Implications for text-mining

Application of BibGlimpse can make expert annotated literature collections available to text-mining. On the one hand, richly annotated text corpora are valued for training

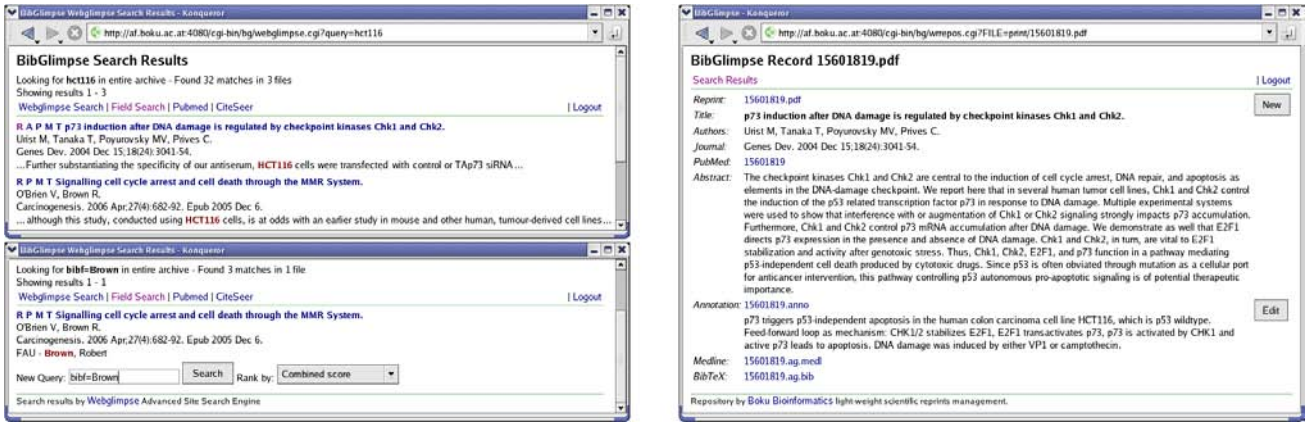


Figure 2
BibGlimpse impressions. The upper left panel shows results of a full-text query for 'HCT116'. A corresponding repository record is depicted on the right, where a domain expert captured relevant information in free-form annotation. Note that the short URLs can easily be sent to collaborating researchers. The lower-left panel demonstrates a structured query, searching only the bibliographic records for 'Brown', which avoids picking up this frequent term in the full-text, *e.g.*, from the citations section.

and testing algorithms [13-15]. On the other hand, text-mining applications benefit even from coarse auxiliary information [16-18]. Yet, the ability to devote scarce resources to annotating literature is often the limiting factor constraining especially machine learning approaches [19,20].

Also it is recognized that full manuscript texts provide more information than abstracts [21-24] and that information retrieval is more successful in domain specific collections [18]. Obtaining access to a comprehensive range of full-text articles, however, can be troublesome due to copyright issues, and the identification of relevant journals can be quite difficult in an interdisciplinary field [25]. Specialist researchers are actually best placed for compiling representative domain specific collections of content enriched full-text articles.

Facilitating the creation process of such annotated collections could hence significantly advance biomedical text-mining. So far, most researchers collect, freely [26] or by subscription, manuscripts of interest to their research area from multiple journals, typically by storing the PDF reprint on their computers. They maintain their personal collections of reprints, notes, and bibliographic records using a variety of tools, ranging from simple text editors or spreadsheets to commercial bibliography management software. But while extensively covering a particular area of interest, representing valuable resources of domain knowledge, such personal repositories are currently hard to search or exploit.

BibGlimpse offers researchers a tool enhancing their own routine literature management and supports the creation of shared domain-specific collections of annotated full-text manuscripts. This benefits both biological researchers as well as the text-mining community.

Conclusion

Considering the benefits of the system's automation and query capabilities in supporting shared literature research, together with its straightforward interfacability with literature classification and text-mining tools, we have reason to expect that BibGlimpse will be widely adopted. We are confident that the concept of a light-weight reprint manager demonstrated in BibGlimpse will transform how research groups collect, manage, and share knowledge from literature research.

Availability and requirements

Project name: BibGlimpse

Project home page: http://bioinf.boku.ac.at/bibglimpse

Operating system(s): UNIX

Programming language: Perl, Bash

Other requirements: Apache 2.2.6 or higher, Perl 5.8.6 or higher, pdftotext (*e.g.*, from xpdf 3.01 or higher)

License: http://webglimpse.net/sublicensing/licensing.html

Any restrictions to use by non-academics: licence needed

Authors' contributions

TT implemented the automated PubMed entry retrieval and drafted the manuscript, GV helped with integrating this feature into the Webglimpse search engine, AG contributed the BibGlimpse on Cygwin package, while DPK devised the concept of BibGlimpse, participated in its implementation and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The Boku Bioinformatics group acknowledges support by the Vienna Science and Technology Fund (WWTF), the Austrian Centre of Biopharmaceutical Technology (ACBT), Austrian Research Centres Seibersdorf (ARCS), and Baxter AG. TT acknowledges support by the GEN-AU project Bioinformatics Integration Network of the Austrian Federal Ministry of Science and Research program.

References

1. Ananiadou S, Kell DB, Tsujii JI: **Text mining and its potential applications in systems biology.** *Trends Biotechnol* 2006, **24(12):**571-579.
2. Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery.** *Nat Rev Genet* 2006, **7(2):**119-29.
3. Giustini D, Barsky E: **A look at Google Scholar, PubMed and Scirus: comparisons and recommendations.** *JCHLA/JABSC* 2005:85-89.
4. **sourceforge** [http://sourceforge.net]
5. **EndNote** [http://www.endnote.com]
6. **RefWorks** [http://www.refworks.com]
7. **Greenstone** [http://greenstone.org]
8. Müller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PLoS Biol* 2004, **2(11):**e309.
9. Kuhn T, Royer L, Fuchs N, Schroeder M: **Improving text mining with controlled natural language: A case study for protein interactions.** *DILS LNBI* 2006.
10. Rebholz-Schuhmann D, Kirsch H, Couto F: **Facts from text-is text mining ready to deliver?** *PLoS Biol* 2005, **3(2):**e65.
11. Velez G: **The Searchable Site.** *Linux Gazette* 2006, **147:**.
12. **xpdf** [http://www.foolabs.com/xpdf/]
13. Kim JD, Ohta T, Tsujii J: **Corpus annotation for mining biological events from literature.** *BMC Bioinformatics* 2008, **9:**10.
14. Kim JD, Tsujii J: **Corpora and their Annotation.** In *Text Mining for Biology and Biomedicine* Edited by: Ananiadou S, McNaught J. Artech House; 2006:179-211.
15. Cohen K: **Corpus design for biomedical natural language processing.** *Proceedings of the ACL workshop on Linking Biological Literature, Ontologies and Databases: mining biological semantics, Association for Computational Linguistics* 2005:38-45.
16. Bockhorst J, Craven M: **Exploiting Relations Among Concepts to Acquire Weakly Labeled Training Data.** In *Proceedings of the 19th International Conference on Machine Learning* Morgan Kaufman; 2002:43-50.
17. Suomela BP, Andrade MA: **Ranking the whole MEDLINE database according to a large training set using text indexing.** *BMC Bioinformatics* 2005, **6:**75.
18. Lee M, Wang W, Yu H: **Exploring supervised and unsupervised methods to detect topics in biomedical text.** *BMC Bioinformatics* 2006, **7:**140.
19. Hunter L, Cohen KB: **Biomedical language processing: what's beyond PubMed?** *Mol Cell* 2006, **21(5):**589-94.
20. Wilbur WJ, Rzhetsky A, Shatkay H: **New directions in biomedical text annotation: definitions, guidelines and corpus construction.** *BMC Bioinformatics* 2006, **7:**356.
21. Corney DP, Buxton BF, Langdon WB, Jones DT: **BioRAT: extracting biological information from full-length papers.** *Bioinformatics* 2004, **20(17):**3206-13.
22. Ray S, Craven M: **Learning statistical models for annotating proteins with function information using biomedical text.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S18.
23. Saric J, Jensen L, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2006, **22(6):**645-50.
24. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG: **Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.** *BMC Bioinformatics* 2006, **7:**373.
25. Postma E: **Inflated impact factors? The true impact of evolutionary papers in non-evolutionary journals.** *PLoS ONE* 2007, **2(10):**e999.
26. Wren JD: **Open access and openly accessible: a study of scientific publications shared via the internet.** *BMJ* 2005, **330(7500):**1128.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Microbial Cell Factories

Research

Open Access

Genome, secretome and glucose transport highlight unique features of the protein production host *Pichia pastoris*

Diethard Mattanovich*[1,2], Alexandra Graf[1,2], Johannes Stadlmann[3], Martin Dragosits[1], Andreas Redl[1,2], Michael Maurer[1,2], Martin Kleinheinz[1], Michael Sauer[1,2], Friedrich Altmann[3] and Brigitte Gasser[1]

Address: [1]Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Austria, [2]School of Bioengineering, University of Applied Sciences FH-Campus Wien, Vienna, Austria and [3]Department of Chemistry, University of Natural Resources and Applied Life Sciences, Vienna, Austria

Email: Diethard Mattanovich* - diethard.mattanovich@boku.ac.at; Alexandra Graf - alexandra.graf@boku.ac.at; Johannes Stadlmann - johannes.stadlmann@boku.ac.at; Martin Dragosits - martin.dragosits@boku.ac.at; Andreas Redl - andreas.redl@boku.ac.at; Michael Maurer - michael.maurer@boku.ac.at; Martin Kleinheinz - martin.kleinheinz@boku.ac.at; Michael Sauer - michael.sauer@fh-campuswien.ac.at; Friedrich Altmann - friedrich.altmann@boku.ac.at; Brigitte Gasser - brigitte.gasser@boku.ac.at

* Corresponding author

Published: 2 June 2009

Microbial Cell Factories 2009, **8**:29 doi:10.1186/1475-2859-8-29

Received: 13 May 2009
Accepted: 2 June 2009

This article is available from: http://www.microbialcellfactories.com/content/8/1/29

© 2009 Mattanovich et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: *Pichia pastoris* is widely used as a production platform for heterologous proteins and model organism for organelle proliferation. Without a published genome sequence available, strain and process development relied mainly on analogies to other, well studied yeasts like *Saccharomyces cerevisiae*.

Results: To investigate specific features of growth and protein secretion, we have sequenced the 9.4 Mb genome of the type strain DSMZ 70382 and analyzed the secretome and the sugar transporters. The computationally predicted secretome consists of 88 ORFs. When grown on glucose, only 20 proteins were actually secreted at detectable levels. These data highlight one major feature of *P. pastoris*, namely the low contamination of heterologous proteins with host cell protein, when applying glucose based expression systems. Putative sugar transporters were identified and compared to those of related yeast species. The genome comprises 2 homologs to *S. cerevisiae* low affinity transporters and 2 to high affinity transporters of other Crabtree negative yeasts. Contrary to other yeasts, *P. pastoris* possesses 4 H^+/glycerol transporters.

Conclusion: This work highlights significant advantages of using the *P. pastoris* system with glucose based expression and fermentation strategies. As only few proteins and no proteases are actually secreted on glucose, it becomes evident that cell lysis is the relevant cause of proteolytic degradation of secreted proteins. The endowment with hexose transporters, dominantly of the high affinity type, limits glucose uptake rates and thus overflow metabolism as observed in *S. cerevisiae*. The presence of 4 genes for glycerol transporters explains the high specific growth rates on this substrate and underlines the suitability of a glycerol/glucose based fermentation strategy. Furthermore, we present an open access web based genome browser http://www.pichiagenome.org.

Background

Yeasts have attracted renewed interest in the last few decades as production hosts for biopharmaceutical proteins as well as for bulk chemicals. The methylotrophic yeast *Pichia pastoris* (Guillermond) Phaff (1956) is well reputed for efficient secretion of heterologous proteins [1], and has come into focus for metabolic engineering applications recently. Especially reengineering of the N-glycosylation pathway has enabled the production of heterologous proteins with human-like N-glycan structures [2-4]. While protein production is the major application of *P. pastoris*, production of metabolites has come into research focus recently too [5,6]. Apart from these biotechnological applications, it is widely used as a model for peroxisome [7] and secretory organelle research [8]. *P. pastoris* has recently been reclassified into a new genus, *Komagataella* [9], and split into three species, *K. pastoris*, *K. phaffii*, and *K. pseudopastoris* [10]. Strains used for biotechnological applications belong to two proposed species, *K. pastoris* and *K. phaffii*. The strains GS115 and X-33 are *K. phaffii*, while the SMD series of protease deficient strains (most popular SMD1168) is classified into the type species, *K. pastoris*. Apart from these strains which have been made available by Invitrogen, research labs and industry use different other strains belonging to either of these two species, and no trend towards a superior expression level of one of the two species has been observed. In order to provide a common information basis across the different strains, we have performed this work with the type strain (DSMZ 70382) of the type species *K. pastoris*, which is the reference strain for all the available *P. pastoris* strains. In coherence with the published literature, we name all strains *P. pastoris*, which thus stands for the entire genus *Komagataella*. As other strains, DSMZ 70382 was isolated from tree exudate, in this case from the chestnut tree.

The majority of *P. pastoris* processes described so far utilize methanol as substrate and inducer for heterologous protein production. While tight gene regulation and high product titers can be achieved with this strategy, the disadvantages as large scale use of a flammable substrate, high heat production and oxygen consumption, and significant cell lysis have been reported. Apart from technological challenges in large scale fermentation, this leads to significant contamination of culture supernatants with intracellular proteins including proteases [11]. *P. pastoris* has been described to secrete some heterologous proteins like human serum albumin [12] or as recently reported glycoengineered antibodies [13] in the g L^{-1} range, while naturally secreted proteins account only for low amounts [14], which supports the easy production of highly pure proteins. However, several secreted *P. pastoris* proteins are observed as contaminants in culture supernatants, requiring elaborate product purification and analytical effort. A detailed characterization of the secretome would significantly improve production and quality control of biopharmaceuticals produced with this expression system. The secretomes of few yeasts and filamentous fungi have been analyzed experimentally. Computational analyses of yeast genomes predicted approximately 200 potentially secreted proteins [15,16]. Secretomes of filamentous fungi contain numerous enzymes for degradation of starch, cellulose, lignin and similar plant polymers [17-19]. However, these predictions suffer from some limitations. As certain targeting sequences are not recognized, the predictions may contain proteins which are retained in cellular organelles. Most cell wall associated proteins can be predicted, but due to the fluctuating nature of the cell wall during growth and budding a fraction of these may be released from the cell wall structure and add to the secretome. Finally the actual composition of the secretome will depend on growth conditions and the actual expression of the genes encoding potentially secreted proteins. Therefore the extracellular proteome of *P. pastoris* was analyzed here and compared to the predicted secretome.

Substrate uptake kinetics determine growth kinetics and the characteristics of biotechnological processes. *P. pastoris* is described as a Crabtree-negative yeast, featuring respiratory metabolism under glucose surplus [20]. A major reason for the easy growth to high biomass concentrations is assumed in the endowment with hexose transporters and their features. We report here the determination and analysis of the *P. pastoris* draft genome sequence and its application in correlating *in silico* and mass spectrometric analysis of the extracellular proteome. Furthermore, a comparative analysis of hexose transporters allows drawing conclusions towards glucose uptake kinetics, a major determinant of growth and bioprocess characteristics in relation to substrate supply. Additionally, a web based database with search functions and annotation data for analysis of the genome sequence is reported.

Results

Sequencing

The genome of *P. pastoris* was sequenced using two next generation sequencing methods. First a Roche GS-FLX run was used to take advantage of the longer reads (400 nts) of this method, which was then complemented by a paired end run with the short read method of Illumina Genome Analyzer (36 nts) to improve the quality of the sequence. The combined result of both methods was a draft genome of 326 assembled contigs of which 93 were larger than 10 kb and 60 between 1 and 10 kb. The longest contig comprised 419,475 nts and the shortest 128 nts. 125 of the 326 contigs could be aggregated into 38 supercontigs. Overall 9,405,451 bases were sequenced with a coverage of 22× with Roche GS-FLX and 60× with Illu-

mina GA. Key statistical data of the draft genome are presented in table 1.

Gene prediction

We initially predicted 7,935 open reading frames using two different gene finders. Manual curation reduced this number to 5,450 ORFs. The eukaryotic gene finder Augustus has been pre-trained on a number of datasets including various yeast species. Of these, *Candida guilliermondii*, *Debariomyces hansenii* and *Pichia stipitis* were selected for their relatively close relation to *P. pastoris* (based on sequence similarity), and *Saccharomyces cerevisiae* as a reference yeast species with the best sequence annotation. In addition the prokaryotic gene finder Glimmer3 was applied since many eukaryotic gene finders overpredict intron containing genes. As yeast genomes are generally compact a large amount of intron containing genes was not expected. All putative ORFs < 100 nts or comprising a starting codon other than ATG were excluded from the set except for genes on contig borders. 194 of the predicted genes are truncated because they crossed contig borders. Ribosomal RNAs were annotated by homology to *S. cerevisiae* rRNAs. Contrary to *S. cerevisiae*, the 5S rRNA is not part of the cluster containing 18S, 26S and 5.8S rRNA but spread across the genome. 149 transfer RNAs were identified using tRNA Scan, which is lower than the average number of tRNAs identified in other yeasts (216 on average).

Table 1: Genome statistics overview

Sequencing Data:	
Total DNA bases after Roche GS FLX	9,408,251
Average coverage Roche GS FLX	22
Total DNA bases after Illumina GA	9,405,451
Average coverage Illumina GA	60
Number of reads Roche FLX	562,515
Number of reads Illumina GA	15,761,520
Number of contigs	326
Contigs > 1 kbp	153
Largest contig	419,475
Smallest contig	128
Average contig size	28,906
GC content	41.34%
Gene Prediction Data:	
Predicted ORFs	7,935
Manually curated number of ORFs	5,450
Thereof ORFs with introns	741
Truncated ORFs	194
ORFs with annotation	4,257
GC content coding regions	41.90%
RNA Prediction Data:	
tRNA genes	149
5S rRNA	14
rRNA cluster (18S, 26S, 5.8S rRNA)	1

Functional Annotation

Functional annotation was performed computationally with a reciprocal best hit (RBH) strategy, using BLAST [21] searches against a selected dataset of the subphylum *Saccharomycotina* to which *P. pastoris* also belongs, and the Uniprot database. All *P. pastoris* genes and proteins that were publicly available at the NCBI (National Center for Biotechnology Information) were manually compared against our predictions. The native genes and proteins were present in our set. The average identity between these genes deposited in NCBI and their homologs in the present genome sequence was 95%. For all proteins that were predicted to be secreted and all others that are discussed here the functional annotation was manually curated. The distribution in GO functional terms of all functionally annotated ORFs was compared to *S. cerevisiae* (figure 1). The distribution is rather similar with differences observed mainly in the groups organelle organization, protein modification, lipid, amino acid and cofactor metabolism.

Secretome

To validate the secretome prediction pipeline (see Materials and Methods) used for *P. pastoris*, it was applied to the *S. cerevisiae* proteome beforehand. The majority of proteins which were described to be extracellular in the Saccharomyces genome database SGD [22] were found in the secreted dataset, for the rest a GPI-anchor signal was predicted. Due to the good performance of the prediction pipeline with *S. cerevisiae* and the successful application of similar methods for *K. lactis* [15] and *C. albicans* [16] respectively, a high accuracy for the secretome predictions was expected for *P. pastoris* as well. The predicted secretome of *P. pastoris* comprises 88 putative proteins of which 55 could be functionally annotated. Additionally, 172 ORFs were predicted to encode proteins entering the general secretion pathway but being localized in different cellular compartments (for the complete list see additional file 1). Obviously the secretome prediction cannot easily discriminate between ER/Golgi localized and secreted proteins, as the chaperone Kar2 and protein disulfide isomerise (Pdi1) appear among the predictions. Therefore the experimental determination of the extracellular proteins is essential for an assessment.

To identify the extracellular secretome of *P. pastoris*, the strain DSMZ 70382 was grown in chemostat culture on glucose as limiting carbon source, reaching 26.4 ± 0.1 g L^{-1} dry biomass (YDM). The supernatants contained 407 mg L^{-1} total protein. Analysis by SDS-PAGE indicated that approximately 15 distinct protein bands, ranging from 12 kDa to 170 kDa, were present in the culture supernatant (figure 2a). On 2D gels, 28 protein spots were visible at higher abundance, at least 7 thereof being obviously isoforms of other protein spots with identical MW but differ-

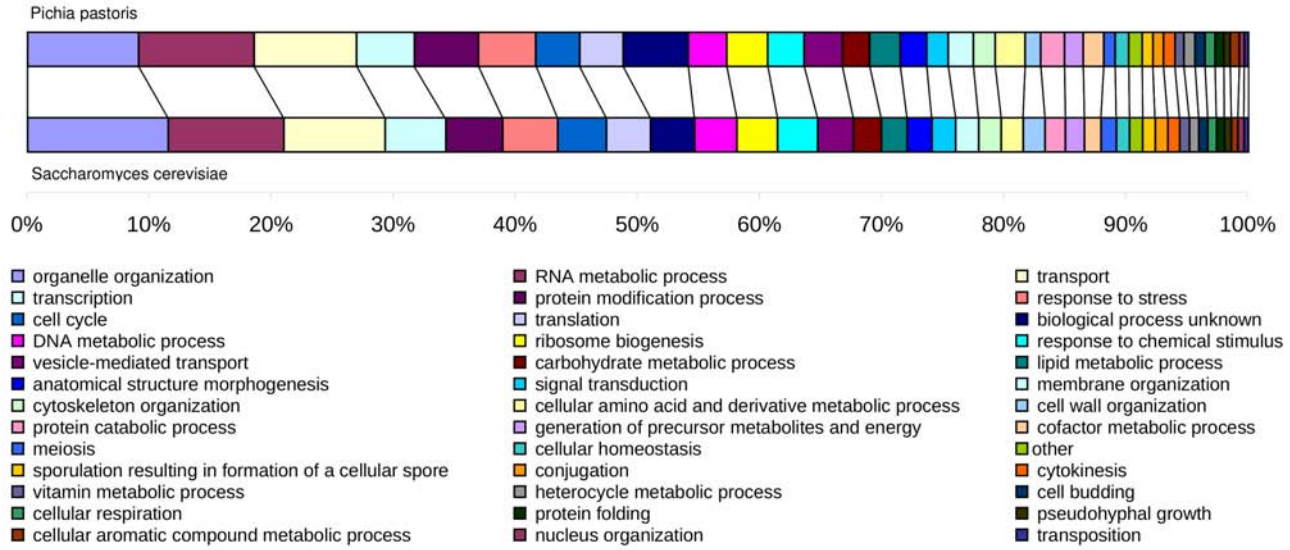


Figure 1
Categorization of the *P. pastoris* annotated genome compared to *S. cerevisiae*. The GO functional groups are displayed based on their relative representation with annotated ORFs.

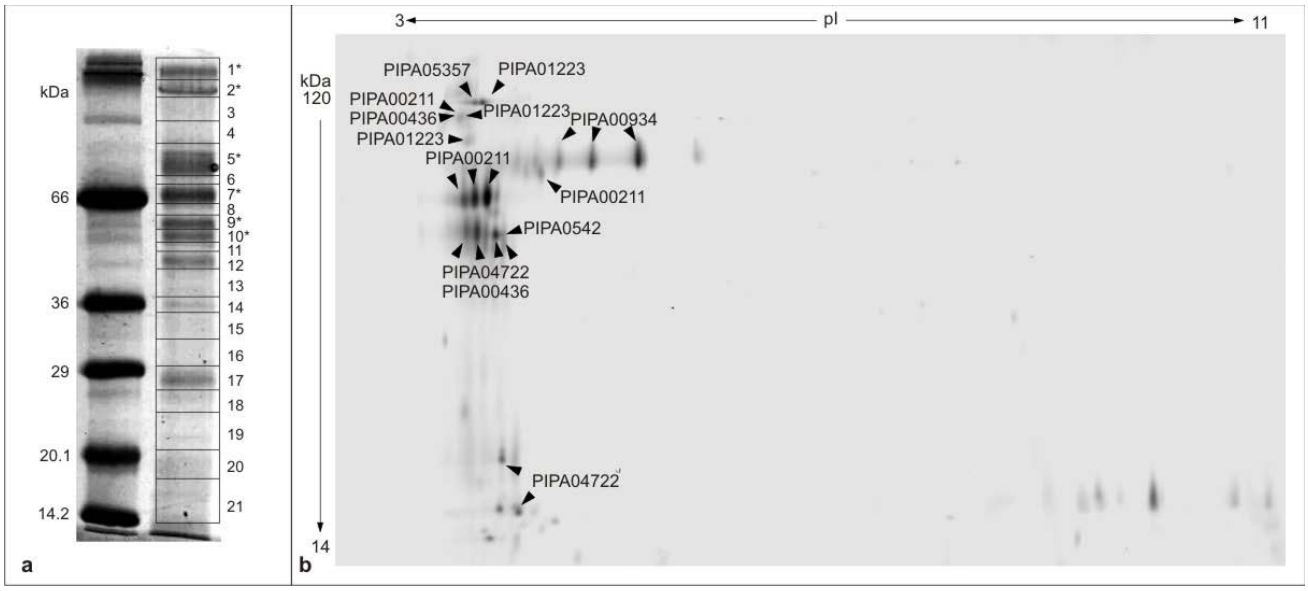


Figure 2
Secretome of *P. pastoris*. (a) SDS-polyacrylamide gel. Left lane: molecular weight marker, right lane: supernatant of *P. pastoris* chemostat culture. Boxes indicate the gel slices used for LC-MS protein identification. Bands corresponding to glycoproteins are marked with an asterisk. (b) 2D electrophoresis gel of *P. pastoris* culture supernatants. Proteins identified by LC-MS are indicated.

Table 2: Secreted proteins of *P. pastoris*

PIPA ID	Predicted function	theoretical pI/MW [kDa]	Predicted N-glycosylation sites	Predicted localization
PIPA00211	Covalently-bound cell wall protein of unknown function	5.01/45.73	1	secreted
PIPA00246	hypothetical fungal hexokinase	5.98/24.92	1	no SP
PIPA00436	Cell wall protein related to glucanases	4.83/36.07	0	secreted
PIPA00545	Cell wall protein related to glucanases	4.33/45.02	2	secreted
PIPA00748	O-glycosylated protein required for cell wall stability	4.22/31.86	1	secreted
PIPA00934	SCP-domain family protein, unknown function, extracellular	5.55/31.72	0	secreted
PIPA00956	60S ribosomal protein L18A	9.92/21.82	1	no SP
PIPA01008	GAS1; Beta-1,3-glucanosyltransferase	3.98/57.20	4	secreted
PIPA01010	GAS1; Beta-1,3-glucanosyltransferase	3.99/58.37	5	secreted
PIPA01223	potential cell wall glucanase	4.34/49.39	0	secreted
PIPA01958	Endo-beta-1,3-glucanase	4.03/33.76	1	secreted
PIPA02332	no similarity found	6.01/23.64	2	no SP
PIPA02510	Glyceraldehyde-3-phosphate dehydrogenase	6.24/35.74	1	no clear SP
PIPA02524	glucan 1,3-beta-glucosidase similar to *S. cerevisiae* EXG1 (YLR300W)	4.51/46.22	1	secreted
PIPA02544	aldehyde dehydrogenase, Adh2p [*S. cerevisiae*]	6.00/36.86	0	no SP
PIPA03955	endo-1,3-beta-glucanase [*P. stipitis* CBS 6054], Dse4p [*S. cerevisiae*]	4.70/109.45	5	secreted
PIPA04722	Cell wall protein with similarity to glucanases	5.18/32.95	0	secreted
PIPA05357	no similarity found	4.25/66.46	1	no SP, 2 TM
PIPA05673	YLR286Cp-like protein [*S. cerevisiae*], endochitinase	4.05/71.87	1	no clear SP
PIPA05771	Chitin deacetylase, Cda2p [*S. cerevisiae*]	5.25/34.66	2	secreted, lower probability

List of identified secreted proteins, with theoretical pI and theoretical MW, and information on the predicted localization (SP = signal peptide, TM = transmembrane domain).

ent pI (figure 2b). Almost all highly abundant proteins ran at low pI values between 3 and 5.5. As the cellular viability was 99% throughout the cultivation, and total DNA content of the supernatants was 1.12 ± 0.03 μg mL^{-1}, a maximum of 1% lysed cells was estimated, accounting for maximally 10% of total protein in the supernatant. Therefore, the potential contamination by intracellular protein was assumed to be minor. A 1D SDS PAGE gel was cut into 21 slices and analyzed by LC-ESI-MS/MS. Detailed data on protein identification are found in additional file 2. Twenty different proteins were identified (table 2), 12 of which appeared in more than one gel slice (additional file 2). Nine proteins ran at higher molecular weight than predicted from the sequence. Eight out of these proteins contained potential N-glycosylation sites (table 2 and additional file 2) and corresponded to detected glycoproteins (figure 2a). Apparently 6 of these proteins were subject to proteolysis. However, the proteolytic activity in the supernatants was very low (equivalent to 11 ± 0.9 ng mL^{-1} trypsin), and in contrast to other yeast secretomes, no protein with putative proteolytic activity was identified. Fourteen of the proteins identified by homology are obviously secreted or cell wall bound, 6 of them with homology to glucanases. The other proteins with extracellular localization comprise 7 cell wall modifying enzymes and 1 secreted protein of unknown function. Four proteins are homologous to intracellular proteins (including glyceraldehyde phosphate dehydrogenase which has been described to be also located at the cell wall in *S. cerevisiae* [23], and for 2 no similarity was found. The putative intracellular proteins mainly comprise glycolytic enzymes and ribosomal proteins which are highly abundant on glucose [24]. A comparison of predicted to identified secretome reveals a good correlation of prediction, putative function, and experimentally determined localization (table 2). All proteins homologous to intracellular proteins were predicted to be intracellular, and only for 2 of the 14 putatively secreted proteins the prediction was unclear or slightly below threshold.

Hexose transporters

Fourteen putative sugar transporters all belonging to the major facilitator superfamily (MFS) were identified by sequence similarity. All *P. pastoris* sugar transporters feature the classical 12 transmembrane domains, and contain the PESP motif and at least one of the two sugar transporter signature sequences. Contrary to *S. cerevisiae*, which comprises 20 isogenes for low and high affinity hexose transport, only two putative transporters with sequence similarity to *S. cerevisiae* transporters are present in the *P. pastoris* genome. While PIPA00236 possesses more than 60% identity to *S. cerevisiae* HXT-family proteins, and the low-affinity transporters of *Kluyveromyces lactis* Rag1 [25] and *Hansenula polymorpha* Hxt1 [26] on the amino acid level, PIPA08653 shows only low similarity (max. 37% identity/58% positives) to these proteins as well as to other *P. pastoris* sugar transporters. Although all 5 conserved amino acids that have been postulated to be required for high affinity transporters in *S. cerevisiae* Hxt2 [27] are present also in the respective translated protein sequence of *P. pastoris* gene PIPA00236, disruption of the gene led to impaired growth on high concentrations of glucose (2%). Disruption of PIPA08653 did not show a distinct growth phenotype. This indicates that PIPA00236 encodes the major low affinity glucose transporter in *P. pastoris*.

For high affinity transport, two *P. pastoris* proteins (PIPA02561 and PIPA00372) with high sequence similarity (>65% identity) to *K. lactis* high affinity glucose transporter Hgt1 were identified (see figure 3). The potential transporter-like hexose sensor is encoded by PIPA01691, and lacks the C-terminal "glucose sensor domain" as do the respective orthologous sensors in *H. polymorpha* (Hxt1) and *Candida albicans* [26]. Additionally a gene with similarity to quinate permease of *P. stipitis* and filamentous fungi was identified, which has putative orthologs in many other yeast species, but is missing in *S. cerevisiae*. According to Barnett et al. [28] *P. pastoris* cannot utilize quinate as a carbon source, although some of the genes required for the utilization of quinate are part of the shikimate pathway leading to the production of aromatic amino acids, and are present as part of the pentafunctional AROM protein. However, regulatory proteins of the quinate pathway are missing in the genome of *P. pastoris*. Interestingly, *P. pastoris* possesses four transporters that are highly similar to putative glycerol transporters from *K. lactis* (KLLA0A03223g) and *Yarrowia lipolytica* (YALI0F06776g), and weakly similar to the *S. cerevisiae* glycerol transporter Slt1. Sequence similarities of the proteins discussed above to their respective orthologs in *S. cerevisiae*, *P. stipitis*, *H. polymorpha*, *K. lactis*, and *Emericella nidulans* are illustrated in figure 3.

Database, genome browser

To make the genomic data accessible it was loaded into a relational database. For visualization a genome browser was installed on a web server and connected to the database.

The genome browser of *P. pastoris* is publicly available at http://www.pichiagenome.org [29].

The draft genome sequence data are deposited at EMBL-EBI, accession number CABH01000001 – CABH01000326.

Discussion

The predicted size of the haploid genome of *P. pastoris* [30] was confirmed here to comprise 9.4 Mb, which is

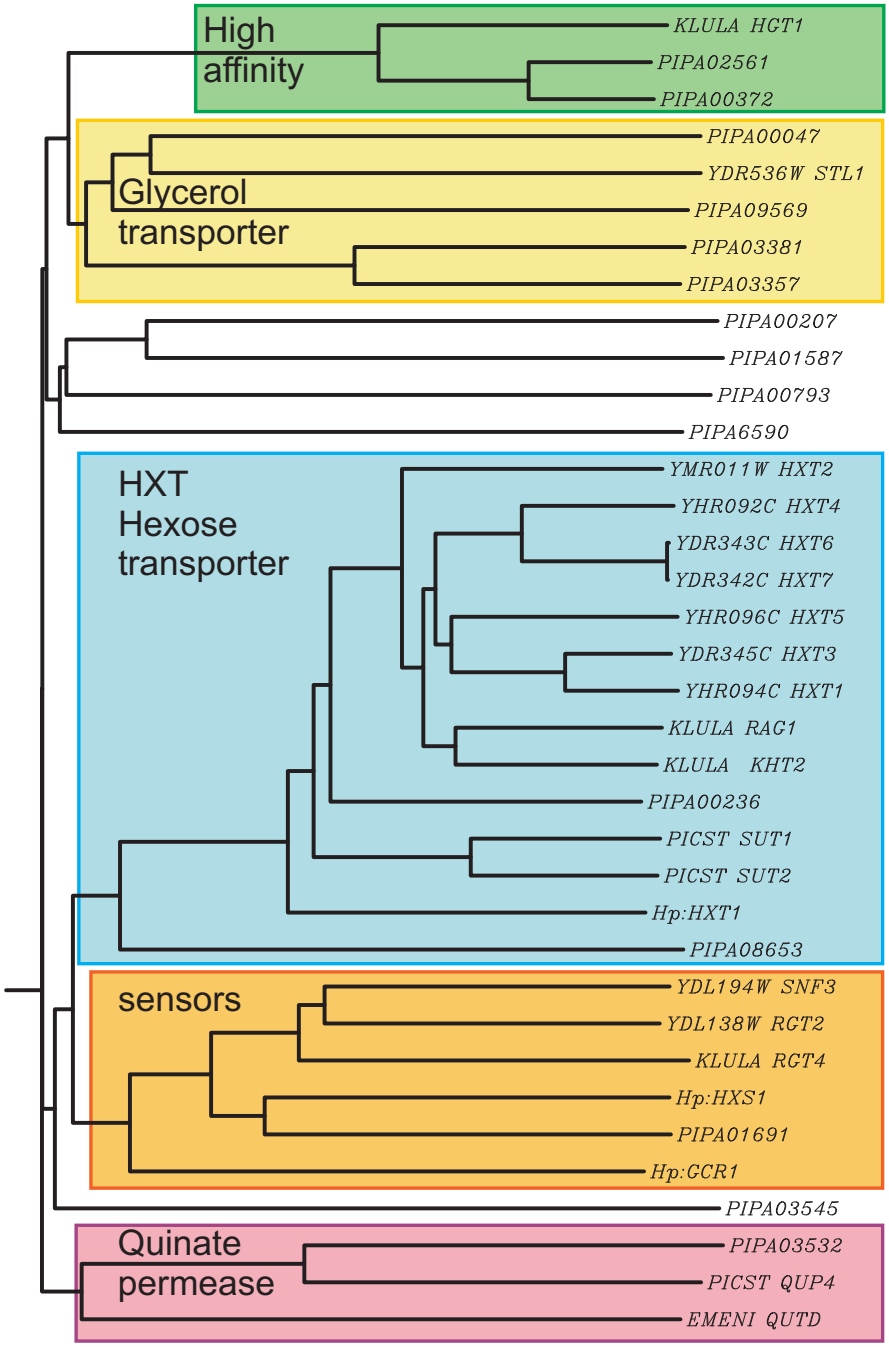


Figure 3
Branch length dendrogram of sugar transporters and related proteins of different yeasts. Putative hexose transporters and sensors and related proteins were aligned with ClustalW, and clusters of functional categories are highlighted. High affinity = high affinity glucose transporters; glycerol transporters = H^+/glycerol symporter; HXT = low affinity *S. cerevisiae* hexose transporter family; sensors = transporter-like glucose sensors; quinate permease = homologs to fungal quinate permeases. ORF IDs relate to: PIPA = *P. pastoris*; Ynnnnnn = *S. cerevisiae*; KLULA = *K. lactis*; PICST = *P. stipitis*; Hp = *H. polymorpha*; EMENI = *Emericella nidulans*. ORFs not highlighted are homologous to other substrate transporters with sequence similarity to hexose transporters.

smaller than the genomes of other yeasts, spanning from 10–20 Mb [31]. Nevertheless the number of functionally annotated genes is comparable to other yeasts, which can be attributed to the fact that *P. pastoris* contains fewer genome redundancies compared e.g. to S. *cerevisiae* and *D. hansenii*, which have undergone genome duplications followed by partial genome losses during evolution [32]. While *P. pastoris* contains specific subclasses of genes for methanol metabolism and peroxisome synthesis, structure and degradation which are present only in methylotrophic yeasts, most metabolic enzymes are present only in single copies, and the number of secreted proteins is low. To verify the quality of gene prediction, all 173 *P. pastoris* genes and 245 proteins currently deposited in NCBI were BLAST searched among the predicted gene list. All of the *P. pastoris* specific genes were present, indicating a high quality of gene prediction.

The secretomes of *K. lactis* and *C. albicans* have been predicted computationally [15,16], yielding 178 ORFs of *K. lactis* and 283 of *C. albicans*. The *C. albicans* secretome apparently is more complex and contains numerous lipases, proteases and agglutinin-like proteins, while both for *K. lactis* and *P. pastoris* only few enzymes apart from glucanases and chitin modifying enzymes appear. As *P. pastoris* utilizes only few carbon sources [28] it appears obvious that neither proteolytic, lipolytic or saccharolytic activities are secreted for substrate utilization. Yeast glucanases and chitinases are required for cell wall plasticity during cell growth and division [33]. While these enzymes are commonly regarded to be cell wall associated, it is plausible that they reach the culture supernatant during cell wall remodelling, indicating that a distinct border cannot be drawn between cell wall and the exterior space.

Fourteen of the 20 proteins identified in the culture supernatant of *P. pastoris* were homologous to proteins implicated in cell wall or extracellular functions. No other secretory enzyme homologs were identified, further indicating that cell wall associated proteins are the essential constitutents of the secretome of glucose grown *P. pastoris*. The computationally predicted secretome contains all secreted proteins plus mainly soluble cellular proteins containing a signal peptide but no transmembrane domains. Thus these predictions obviously overestimate secretory proteomes (figure 4). The culture supernatant of *K. lactis* contained significantly more (82) proteins [15] of which 34 were predicted to be secreted or cell wall bound, and the rest were assumed to be localized either to the ER or the cytosol. The latter group of proteins indicates a significant release of intracellular proteins in this study, probably by cell lysis due to the culture conditions.

The low concentration, together with the small number of actually secreted proteins from *P. pastoris* highlights a major advantage of this protein production system, as secreted products are much less contaminated with host cell protein. Jahic et al. [34] have shown that host cell protein released from *P. pastoris* grown on methanol mainly derives from cell lysis, which occurs to a much lower extent upon growth on glucose. Combined with the fact that strong promoters for use on glucose are available [34,35], these data provide convincing arguments for a reconsideration of methanol based protein production with *P. pastoris*. The toxicity of methanol and several of its metabolites is the main reason for cell lysis and consequently also protease leakage to the culture supernatant. Additionally other host cell proteins are released, leading to significant contamination of protein products. A common approach to reduce product proteolysis is the knock out of protease genes. However, multiple protease knockout strains tend to be growth retarded, so that it appears reasonable to employ a production strategy based on glucose media which avoids the detrimental effects of methanol at all. Detailed knowledge of the secreted host cell proteins, as presented here, can have a strong positive effect on product purification and quality control, as specific assays can be developed. Additionally a knock out of major secreted proteins can reduce the host cell protein load significantly [36].

Substrate uptake kinetics determines growth kinetics and the characteristics of biotechnological processes. The fermentative (Crabtree-positive) yeast *S. cerevisiae* consumes glucose at high rates when supplied with high concentrations. This exceptionally high glucose uptake rate is attributed to high abundance of hexose transporters, encoded by more than 10 isogenes [37]. Respiratory (Crabtree-negative) yeasts limit glucose uptake, as they contain few hexose transporter genes, encoding energy dependent symporters with high affinity to glucose [38]. The endowment of *P. pastoris* with hexose transporters is in good accordance to other respiratory yeasts such as *K. lactis*, *H. polymorpha* and *P. stipitis*, all having a reduced number of hexose transporters in comparison to *S. cerevisiae*. Moreover, Crabtree-negative yeasts usually exhibit K_m values in the micromolar range for glucose [37], due to their very high-affinity transporters such as *K. lactis* Hgt1, which is an ortholog of *P. pastoris* PIPA02561 and PIPA00372. While K_m values for *P. pastoris* specific transporters remain to be determined in future, conclusions to glucose uptake behavior can be drawn. Accordingly, specific glucose uptake rate is limited to q_{Smax} = 0.35 g g^{-1} YDM h^{-1} (at growth rates near μ_{max} = 0.193 h^{-1}) in *P. pastoris* chemostat cultivations [39], in comparison to q_{Smax} = 2.88 g g^{-1} YDM h^{-1} in fully aerobic *S. cerevisiae* [40]. The limited glucose uptake prevents Crabtree-negative yeasts such as *P. pastoris* from extensive overflow metabolism, which leads to the aerobic formation of ethanol and a reduced biomass yield at high external glucose concentrations in *S. cerevisiae*.

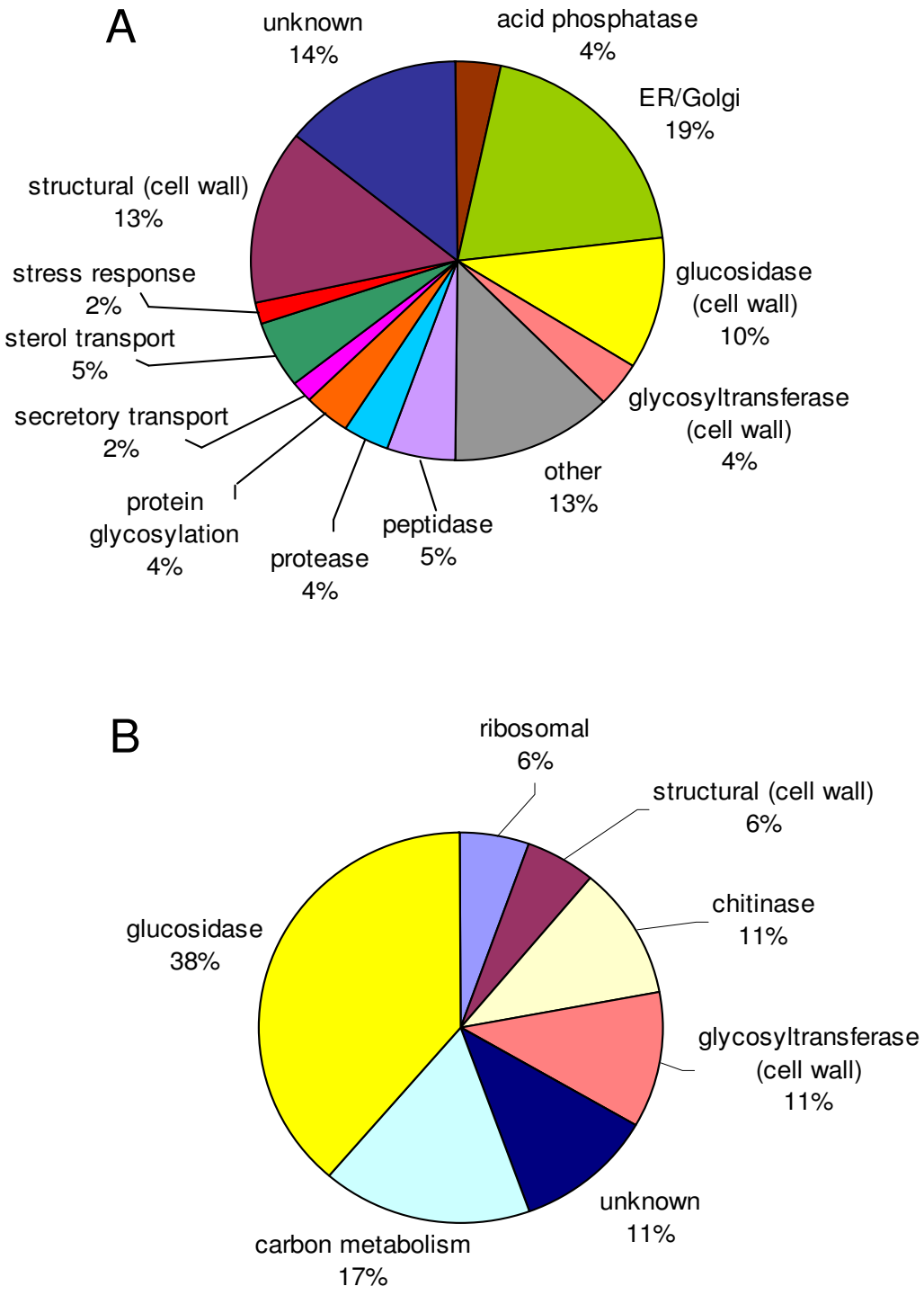


Figure 4
Categorization of *P. pastoris* secretome. (a) predicted and (b) detected secretome based on GO terms. Proteins without *S. cerevisiae* homologs are classified as "unknown".

This difference is also reflected in the very high biomass concentrations (more than 100 g l^{-1}) that can be achieved in *P. pastoris* cultivations. For heterologous protein production, aerobic ethanol formation is a substantial problem, because it lowers the yield of the desired product due to a lower biomass concentration.

Interestingly, *P. pastoris* contains four genes encoding putative H^+/glycerol symporters, contrary to all other sequenced yeasts up-to-date. Consequently, the maximum glycerol uptake rate of *P. pastoris* is $q_{Glycerol\_max}$ = 0.37 g g^{-1} YDM h^{-1}. This is substantially higher than the uptake rates reported for *S. cerevisiae* ($q_{Glycerol\_max}$ = 0.046 g g^{-1} YDM h^{-1}) and many other yeast species [41]. The ability to grow on glycerol as a single carbon and energy source – a mode of cultivation widely applied for generation of biomass with *P. pastoris* prior to methanol induction or glucose fed batch – is dependent on the activity of a constitutive salt-independent active glycerol transport by the H^+/glycerol symport and has also been reported for *Pichia sorbitophila* and *Pichia jadinii* [41]. Specific growth rates of these yeasts on glycerol are similar to the specific growth rates that can be obtained on glucose (e.g. for *P. pastoris* on mineral media $\mu_{Glycerol\_max}$ = 0.26 h^{-1}, $\mu_{Glucose\_max}$ = 0.19 h^{-1}), whereas yeasts lacking the activity of such a type of carrier have significantly reduced growth rates on glycerol. The high specific glycerol uptake rate, enabled by the exceptional endowment with specific transporters emphasizes the suitability of glycerol as a substrate for biomass growth.

Conclusion

The availability of genome data has become an essential tool for cell and metabolic engineering of biotechnological production organisms. This work highlights major advantages of *P. pastoris* as a protein production platform and the benefits of glycerol/glucose based production technology. Apart from lower heat production and oxygen demand compared to methanol based processes, glucose grown cultures display higher viability and essentially no protease release to the culture supernatant. Furthermore detailed insights into the sugar transport will enable rational modulation of substrate fluxes, especially for efficient metabolite production.

Material and methods

Strain

The *P. pastoris* type strain (DSMZ 70382 = CBS704) was selected as the source of genomic DNA, and used for all experimental work. Genomic DNA was prepared as described in Hohenblum et al. using the Qiagen Genomic G-20 kit [42].

Sequencing

Genomic DNA was sequenced by GATC Biotech AG, Konstanz (Germany) with a Roche GS FLX-Titanium Series complemented by an Illumina Genome Analyzer paired end run. The reads were assembled with SeqMan NGen by DNASTAR. To verify the sequencing quality all *P. pastoris* gene and protein sequences available at NCBI were downloaded and the sequences were compared using BLAST searches.

Gene prediction and annotation

Gene prediction was performed with the eukaryotic gene finder Augustus [43] using the option for overlapping genes as well as the prokaryotic gene finder Glimmer3 [44]. Predicted open reading frames were kept if they were longer than 100 nucleotides and started with ATG, except for genes predicted on contig boarders. The ORF sets were merged and made non redundant using the clustering program cd-hit-est [45] with a similarity cut-off of 95%.

Annotation was done by a reciprocal protein BLAST against a dataset consisting of the publicly available *Saccharomycotina* species and the UNIPROT protein database with an E-value threshold of 10^{-10}. All *P. pastoris* proteins and genes available at NCBI, all proteins that were predicted to be secreted and all sugar transporters were manually curated. Gene Ontology annotation was done for all proteins with a homolog in *S. cerevisiae*.

Ribosomal RNA annotation was done through homology with *S. cerevisiae* using nucleotide BLAST against the *P. pastoris* contigs, and the results were manually analyzed. tRNAs were localized using the program tRNAscan-SE [46]. Gene predictions were manually curated using BLASTx.

In silico *secretome prediction*

A similar method was used as described to predict the secretomes of *K. lactis* [15] and *C. albicans* [16], respectively. The prediction pipeline included SignalP 3.0 [47,48] to identify the N-terminal signal peptide, Phobius [49] to predict the transmembrane topology, GPI-SOM [50] and the fungal version of big-PI [51] for GPI anchor prediction, TargetP [52] to exclude all proteins with predicted mitochondrial localization. Additionally WoLF PSORT [53] was used for general localization prediction.

Proteins were considered to be secreted when an N-terminal signal peptide existed but neither a transmembrane domain (except one within the first 40 residues), nor a GPI-anchor, nor any localization signal to other organelles were identified.

The prediction pipeline was tested on an *S. cerevisiae* dataset of 5,884 proteins which was downloaded from the Saccharomyces Genome Database SGD [22].

Experimental secretome analysis

P. pastoris DSMZ 70382 was grown in fully aerobic chemostat cultures on minimal medium with glucose as carbon source until steady state (biomass yield and RQ constant for at least 2 residence times). Detailed data on media compositions, fermentation data and the analysis of culture supernatant can be found in additional file 3. Culture supernatants were concentrated by acetone precipitation and subjected to 1D SDS-PAGE on a 12% PAA gel and 2D-DIGE, respectively. For 2D-DIGE supernatant protein was Cy5 labelled and separated on a IPGDryStrip (3-11NL) in the first dimension, followed by SDS-PAGE on a 12% PAA gel as described in Dragosits et al. [24]. 1D gel lanes were cut into 21 slices, and protein spots from CBB stained 2D gels were picked. After tryptic digest, samples were analyzed by reversed-phase chromatography (UltiMate 3000 Capillary LC-system, Dionex) coupled with ESI MS/MS analysis (Q-TOF Ultima Global, Waters). The obtained mass spectra were subsequently analysed using X!Tandem 2008.12.01 [54]. The identified proteins had to meet the following criteria: protein score e-value $\leq 10^{-5}$ with at least 2 peptides per protein. Glycoproteins were detected by SDS-PAGE and blotting of proteins onto a nitrocellulose membrane followed by detection via Concanavalin A and Horseradish peroxidase. Putative N-glycosylation sites were identified with NetNGlyc 1.0 server [55].

Analysis of hexose transporters

P. pastoris ORFs encoding putative sugar transporters were identified by sequence similarity using BLAST. Multiple sequence alignment of the respective protein sequences to previously identified hexose transporters and sensors from other yeasts was performed by ClustalW [56] using BLOSUM weight matrix, and a dendrogram with branch length was generated. Additionally an integrated search in PROSITE [57], Pfam, PRINTS and other family and domain databases was performed with InterProScan [58] for all these protein sequences.

Disruption cassettes for PIPA00236 and PIPA08653 were generated by PCR (primers: PIPA08653FW: ATGGCAGGTATTAAAGTTGGATC; PIPA08653BW: TACTGCCATCTGCTTCTTTC; PIPA00236FW: GCAGGAGAATAGTCCAGTTTAC; PIPA00236BW: TTCATAGCCTCGTCGACTCTG). 200–300 bp each up- and downstream of the start codon were exchanged for the Zeocin resistance cassette. These cassettes were introduced into the genome of *P. pastoris* DSMZ 70382 by electroporation, and clones were selected on YP plates containing 1% yeast extract, 2% peptone, 2% agar-agar, 2% glycerol and 25 μg mL^{-1} Zeocin. Positively growing clones were then analyzed for their growth behavior on YP plates containing either 2% glycerol, 2% glucose or 0.01% glucose for 48 h at 28°C.

Genome Database

The gene predictions were parsed into GFF file format and loaded into a Chado [59] database which is designed especially to hold a wide variety of biological data.

Gbrowse [60], the Generic Genome Browser, was installed on a web server in the latest stable version (1.69) and configured to display the genomic data from the Chado database.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DM initiated and coordinated the *P. pastoris* genome project. AG and AR were responsible for genome annotation and analysis. AG predicted the secreted proteins. MD performed the chemostat cultivations and 2D-gel electrophoresis. AR developed the genome database. JS performed the MS identification of the secreted proteins. FA coordinated and supervised proteomics. MM, MK and MS contributed to annotation. BG carried out the analysis of the hexose transporters and contributed to gene annotation. DM, AG, MD, MM and BG wrote the final text of the manuscript.

Additional material

Additional file 1

***Predicted secretome of* P. pastoris**. *Predicted localization of all genes containing a predicted signal peptide. The output of the prediction pipeline is given, as well as ORF and gene names and descriptions of* S. cerevisiae *homologs, if available.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2859-8-29-S1.xls]

Additional file 2

Summary of identified proteins. *List of mass spectrometry identified proteins on both 1D and 2D gels, including protein scores and all individual peptides with corresponding peptide scores.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2859-8-29-S2.xls]

Additional file 3

Chemostat cultivation data. *Detailed chemostat cultivation data including culture medium composition and evaluation of DNA, RNA and protein content of the supernatant.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2859-8-29-S3.xls]

Acknowledgements

This work has been supported by the European Science Foundation (ESF, program EuroSCOPE), the Austrian Science Fund (FWF), project no. I37, and the Austrian Resarch Promotion Agency (Program FHplus). Special thanks to Harald Pichler (TU Graz) for critically reviewing the manuscript.

Addendum

During revision of this manuscript, De Schutter et al. have published the genome sequence of *K. phaffii* (*P. pastoris*) strain GS115 (Nat. Biotechnol. doi:10.1038/nbt.1544).

References

1. Cereghino G, Cereghino J, Ilgen C, Cregg J: **Production of recombinant proteins in fermenter cultures of the yeast Pichia pastoris.** *Curr Opin Biotechnol* 2002, **13(4):**329-332.
2. Hamilton S, Davidson R, Sethuraman N, Nett J, Jiang Y, Rios S, Bobrowicz P, Stadheim T, Li H, Choi B, *et al.*: **Humanization of yeast to produce complex terminally sialylated glycoproteins.** *Science* 2006, **313(5792):**1441-1443.
3. Hamilton S, Gerngross T: **Glycosylation engineering in yeast: the advent of fully humanized yeast.** *Curr Opin Biotechnol* 2007, **18(5):**387-392.
4. Jacobs P, Geysens S, Vervecken W, Contreras R, Callewaert N: **Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology.** *Nat Protoc* 2009, **4(1):**58-70.
5. Marx H, Mattanovich D, Sauer M: **Overexpression of the riboflavin biosynthetic pathway in Pichia pastoris.** *Microb Cell Fact* 2008, **7:**23.
6. He J, Deng J, Zheng Y, Gu J: **A synergistic effect on the production of S-adenosyl-L-methionine in Pichia pastoris by knocking in of S-adenosyl-L-methionine synthase and knocking out of cystathionine-beta synthase.** *J Biotechnol* 2006, **126(4):**519-527.
7. Dunn WJ, Cregg J, Kiel J, Klei I van der, Oku M, Sakai Y, Sibirny A, Stasyk O, Veenhuis M: **Pexophagy: the selective autophagy of peroxisomes.** *Autophagy* 2005, **1(2):**75-83.
8. Payne W, Kaiser C, Bevis B, Soderholm J, Fu D, Sears I, Glick B: **Isolation of Pichia pastoris genes involved in ER-to-Golgi transport.** *Yeast* 2000, **16(11):**979-993.
9. Yamada Y, Matsuda M, Maeda K, Mikata K: **The phylogenetic relationships of methanol-assimilating yeasts based on the partial sequences of 18S and 26S ribosomal RNAs: the proposal of Komagataella gen. nov. (Saccharomycetaceae).** *Biosci Biotechnol Biochem* 1995, **59(3):**439-444.
10. Kurtzman C: **Description of Komagataella phaffii sp. nov. and the transfer of Pichia pseudopastoris to the methylotrophic yeast genus Komagataella.** *Int J Syst Evol Microbiol* 2005, **55(Pt 2):**973-976.
11. Curvers S, Brixius P, Klauser T, Thömmes J, Weuster-Botz D, Takors R, Wandrey C: **Human chymotrypsinogen B production with Pichia pastoris by integrated development of fermentation and downstream processing. Part 1. Fermentation.** *Biotechnol Prog* **17(3):**495-502.
12. Kobayashi K, Kuwae S, Ohya T, Ohda T, Ohyama M, Ohi H, Tomomitsu K, Ohmura T: **High-level expression of recombinant human serum albumin from the methylotrophic yeast Pichia pastoris with minimal protease production and activation.** *J Biosci Bioeng* 2000, **89(1):**55-61.
13. Potgieter T, Cukan M, Drummond J, Houston-Cummings N, Jiang Y, Li F, Lynaugh H, Mallem M, McKelvey T, Mitchell T, *et al.*: **Production of monoclonal antibodies by glycoengineered Pichia pastoris.** *J Biotechnol* 2009, **139(4):**318-325.
14. Macauley-Patrick S, Fazenda ML, McNeil B, Harvey LM: **Heterologous protein production using the Pichia pastoris expression system.** *Yeast* 2005, **22(4):**249-270.
15. Swaim C, Anton B, Sharma S, Taron C, Benner J: **Physical and computational analysis of the yeast Kluyveromyces lactis secreted proteome.** *Proteomics* 2008, **8(13):**2714-2723.
16. Lee S, Wormsley S, Kamoun S, Lee A, Joiner K, Wong B: **An analysis of the Candida albicans genome database for soluble secreted proteins using computer-based prediction algorithms.** *Yeast* 2003, **20(7):**595-610.
17. Martinez D, Berka R, Henrissat B, Saloheimo M, Arvas M, Baker S, Chapman J, Chertkov O, Coutinho P, Cullen D, *et al.*: **Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina).** *Nat Biotechnol* 2008, **26(5):**553-560.
18. Ravalason H, Jan G, Mollé D, Pasco M, Coutinho P, Lapierre C, Pollet B, Bertaud F, Petit-Conil M, Grisel S, *et al.*: **Secretome analysis of Phanerochaete chrysosporium strain CIRM-BRFM41 grown on softwood.** *Appl Microbiol Biotechnol* 2008, **80(4):**719-733.
19. Oda K, Kakizono D, Yamada O, Iefuji H, Akita O, Iwashita K: **Proteomic analysis of extracellular proteins from Aspergillus oryzae grown under submerged and solid-state culture conditions.** *Appl Environ Microbiol* 2006, **72(5):**3448-3457.
20. Porro D, Sauer M, Branduardi P, Mattanovich D: **Recombinant protein production in yeasts.** *Mol Biotechnol* 2005, **31(3):**245-259.
21. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3):**403-410.
22. **Saccharomyces Genome Database** [http://www.yeastgenome.org]
23. Delgado M, O'Connor J, Azorín I, Renau-Piqueras J, Gil M, Gozalbo D: **The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins.** *Microbiology* 2001, **147(Pt 2):**411-417.
24. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P, Mattanovich D: **The Effect of Temperature on the Proteome of Recombinant Pichia pastoris.** *J Proteome Res* 2009 in press.
25. Wesolowski-Louvel M, Goffrini P, Ferrero I, Fukuhara H: **Glucose transport in the yeast Kluyveromyces lactis. I. Properties of an inducible low-affinity glucose transporter gene.** *Mol Gen Genet* 1992, **233:**1-2.
26. Stasyk OG, Maidan MM, Stasyk OV, Van Dijck P, Thevelein JM, Sibirny AA: **Identification of hexose transporter-like sensor HXS1 and functional hexose transporter HXT1 in the methylotrophic yeast Hansenula polymorpha.** *Eukaryot Cell* 2008, **7(4):**735-746.
27. Kasahara T, Maeda M, Ishiguro M, Kasahara M: **Identification by comprehensive chimeric analysis of a key residue responsible for high affinity glucose transport by yeast HXT2.** *J Biol Chem* 2007, **282(18):**13146-13150.
28. Barnett JA, Payne RW, Yarrow D: **Yeasts: Characteristics and Identification.** *Volume 3*. 3rd edition. Cambridge, UK: Cambridge University Press; 2000.
29. **Pichia pastoris genome browser** [http://www.pichiagenome.org]
30. Ohi H, Okazaki N, Uno S, Miura M, Hiramatsu R: **Chromosomal DNA patterns and gene stability of Pichia pastoris.** *Yeast* 1998, **14(10):**895-903.
31. Jeffries T, Grigoriev I, Grimwood J, Laplaza J, Aerts A, Salamov A, Schmutz J, Lindquist E, Dehal P, Shapiro H, *et al.*: **Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis.** *Nat Biotechnol* 2007, **25(3):**319-326.
32. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, *et al.*: **Genome evolution in yeasts.** *Nature* 2004, **430(6995):**35-44.
33. Adams DJ: **Fungal cell wall chitinases and glucanases.** *Microbiology* 2004, **150(Pt 7):**2029-2035.
34. Jahic M, Wallberg F, Bollok M, Garcia P, Enfors S: **Temperature limited fed-batch technique for control of proteolysis in Pichia pastoris bioreactor cultures.** *Microb Cell Fact* 2003, **2(1):**6.
35. Waterham HR, Digan ME, Koutz PJ, Lair SV, Cregg JM: **Isolation of the Pichia pastoris glyceraldehyde-3-phosphate dehydrogenase gene and regulation and use of its promoter.** *Gene* 1997, **186(1):**37-44.
36. Nombela C, Gil C, Chaffin WL: **Non-conventional protein secretion in yeast.** *Trends Microbiol* 2006, **14(1):**15-21.
37. Boles E, Hollenberg CP: **The molecular genetics of hexose transport in yeasts.** *FEMS Microbiol Rev* 1997, **21(1):**85-111.
38. van Urk H, Postma E, Scheffers W, van Dijken J: **Glucose transport in crabtree-positive and crabtree-negative yeasts.** *J Gen Microbiol* 1989, **135(9):**2399-2406.
39. Maurer M, Kuehleitner M, Gasser B, Mattanovich D: **Versatile modeling and optimization of fed batch processes for the produc-**

tion of secreted heterologous proteins with Pichia pastoris. *Microb Cell Fact* 2006, **5:**37.
40. Otterstedt K, Larsson C, Bill RM, Stahlberg A, Boles E, Hohmann S, Gustafsson L: **Switching the mode of metabolism in the yeast Saccharomyces cerevisiae.** *EMBO Rep* 2004, **5(5):**532-537.
41. Lages F, Silva-Graca M, Lucas C: **Active glycerol uptake is a mechanism underlying halotolerance in yeasts: a study of 42 species.** *Microbiology* 1999, **145(Pt 9):**2577-2585.
42. Hohenblum H, Gasser B, Maurer M, Borth N, Mattanovich D: **Effects of gene dosage, promoters, and substrates on unfolded protein stress of recombinant Pichia pastoris.** *Biotechnol Bioeng* 2004, **85(4):**367-375.
43. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19(Suppl 2):**ii215-225.
44. Delcher A, Bratke K, Powers E, Salzberg S: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007, **23(6):**673-679.
45. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide.** *Bioinformatics* 2006, **22(13):**1658-1659.
46. Lowe T, Eddy S: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25(5):**955-964.
47. Bendtsen J, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340(4):**783-795.
48. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10(1):**1-6.
49. Käll L, Krogh A, Sonnhammer E: **Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server.** *Nucleic Acids Res* 2007:W429-432.
50. Fankhauser N, Mäser P: **Identification of GPI anchor attachment signals by a Kohonen self-organizing map.** *Bioinformatics* 2005, **21(9):**1846-1852.
51. Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F: **A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and Schizosaccharomyces pombe.** *J Mol Biol* 2004, **337(2):**243-253.
52. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300(4):**1005-1016.
53. Horton P, Park K, Obayashi T, Fujita N, Harada H, Adams-Collier C, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic Acids Res* 2007:W585-587.
54. **X!Tandem** [http://www.thegpm.org/tandem/]
55. **NetNGlyc 1.0 server** [http://www.cbs.dtu.dk/services/NetNGlyc/]
56. **ClustalW** [http://align.genome.jp]
57. Bairoch A: **PROSITE: a dictionary of sites and patterns in proteins.** *Nucleic Acids Res* 1992, **20(Suppl):**2013-2018.
58. **InterProScan** [http://www.ebi.ac.uk/Tools/InterProScan/]
59. Mungall C, Emmert D: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information.** *Bioinformatics* 2007, **23(13):**i337-346.
60. Stein L, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich J, Harris T, Arva A, *et al.*: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10):**1599-1610.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Microbial Cell Factories

BioMed Central

Commentary

Open Access

Open access to sequence: Browsing the *Pichia pastoris* genome

Diethard Mattanovich*[1,2], Nico Callewaert[3,4], Pierre Rouzé[5,6], Yao-Cheng Lin[5,6], Alexandra Graf[1,2], Andreas Redl[1,2], Petra Tiels[3,4], Brigitte Gasser[1] and Kristof De Schutter[3,7]

Address: [1]Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Austria, [2]School of Bioengineering, University of Applied Sciences FH-Campus Wien, Vienna, Austria, [3]Unit for Molecular Glycobiology, Department for Molecular Biomedical Research, VIB, Ghent-Zwijnaarde, Belgium, [4]Unit for Molecular Glycobiology, L-ProBE, Department of Biochemistry and Microbiology, Ghent University, Ghent-Zwijnaarde, Belgium, [5]Department of Plant Systems Biology, VIB, Ghent-Zwijnaarde, Belgium, [6]Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium and [7]Department for Biomedical Molecular Biology, Ghent University, Ghent-Zwijnaarde, Belgium

Email: Diethard Mattanovich* - diethard.mattanovich@boku.ac.at; Nico Callewaert - Nico.Callewaert@dmbr.vib-ugent.be; Pierre Rouzé - pierre.rouze@psb.vib-ugent.be; Yao-Cheng Lin - yao-cheng.lin@psb.vib-ugent.be; Alexandra Graf - alexandra.graf@boku.ac.at; Andreas Redl - andreas.redl@boku.ac.at; Petra Tiels - petra.tiels@dmbr.vib-ugent.be; Brigitte Gasser - brigitte.gasser@boku.ac.at; Kristof De Schutter - kristof.deschutter@dmbr.vib-ugent.be

* Corresponding author

Published: 16 October 2009

Microbial Cell Factories 2009, **8**:53 doi:10.1186/1475-2859-8-53

Received: 25 December 2008
Accepted: 16 October 2009

This article is available from: http://www.microbialcellfactories.com/content/8/1/53

© 2009 Mattanovich et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The first genome sequences of the important yeast protein production host *Pichia pastoris* have been released into the public domain this spring. In order to provide the scientific community easy and versatile access to the sequence, two web-sites have been installed as a resource for genomic sequence, gene and protein information for *P. pastoris*: A GBrowse based genome browser was set up at http://www.pichiagenome.org and a genome portal with gene annotation and browsing functionality at http://bioinformatics.psb.ugent.be/webtools/bogas. Both websites are offering information on gene annotation and function, regulation and structure.

In addition, a WiKi based platform allows all users to create additional information on genes, proteins, physiology and other items of *P. pastoris* research, so that the *Pichia* community can benefit from exchange of knowledge, data and materials.

Commentary

Modern biological research requires genome sequence information of the organisms of interest for numerous applications: the development of transcriptomic methods like DNA microarrays relies on genome data, proteomics needs a genome sequence for efficient identification of proteins, metabolic modelling and flux analysis is based on the knowledge of ideally all enzymatic reactions encoded in the genome of an organism. Systems biology, as the synthesis of the above mentioned techniques [1], relies on comprehensive genome sequence data. Systems biology is most advanced for a few model organisms, for which genome sequencing has been an international challenge funded with public support. Systems biotechnology, the application of these approaches to biotechnological strain and process development, faces the same needs [2]. However, genome sequencing of biotechnologically relevant organisms has mainly been pur-

sued with corporate support, and the results were kept confidential over years for commercial exploitation. A major disadvantage of this strategy is the delay of basic research related to these organisms, negatively affecting the knowledge of organisms with the highest relevance for industry.

One such example is the yeast *Pichia pastoris*, widely used for heterologous protein production (reviewed in [3,4]), but also for the production of metabolites [5,6]. The major research areas towards implementing *P. pastoris* as a production host for heterologous proteins are engineering of glycosylation [7-9] and protein folding and secretion (reviewed in [10]). A draft genome sequence has been available commercially since appr. 5 years and omics methods have been developed based on this sequence (transcriptomics [11,12]; proteomics [13]; metabolic flux analysis ()[14,15]), but the strict obligation to keep sequence information confidential has hampered publication of relevant data and collaborations, so that the community could not benefit from exchange of knowledge, data and materials.

To bridge this gap we have published the genome sequences of two *P. pastoris* strains, DSMZ 70382 [16] and GS115 [17], obtained with next generation sequencing technologies. Versatile access to genome sequences is a prerequisite for efficient utilisation of the information. Therefore a genome browser was set up at http://www.pichiagenome.org[18] with a main focus on *P. pastoris* DSMZ 70382 and a genome portal with the gene annotation and browsing functionality for *P. pastoris* GS115 at http://bioinformatics.psb.ugent.be/webtools/bogas[19].

Both of these *Pichia* sites serve as a resource for genomic sequence data and gene and protein information for *P. pastoris*. The genome browser (GBrowse for DSMZ 70382 and AnnoJ [20] for GS115) allows users to view and navigate genomic sequences including non-translated regions of the genome. BLAST searches for comparing any query sequence against the *P. pastoris* dataset, full text searches and gene/sequence resources (Get Sequence) serve to retrieve, display and analyze a gene or sequence in many ways, such as protein translation. In the near future, a comparison of the genome of different strains will be added to both genome browsers.

The genome browser of *P. pastoris* DSMZ 70382 is based on the Generic Genome Browser (GBrowse) which consists of a web interface and a database backend. The system was developed by the Generic Model Organism Database project [21,22] for the purpose of exploring genomic sequences together with annotated data. GBrowse has already been used successfully in various genome database projects like SGD, FlyBase or WormBase and its functionality will therefore be familiar to many researchers. The browser simultaneously provides a bird's eye view and detailed views of the genome and facilitates easy navigation through the genome using its zoom capacity. A flexible display of a variety of features, including genes, proteins, RNAs, GC content and restriction sites, on separated customizable tracks permits the user to adapt the browser to his or her needs. The visualization of Microarray probe locations allow for the direct access to specific probe sequence and location of published microarray designs [12]. The *Pichia* Genome Browser further allows locating DNA or protein sequence patterns, to design sequencing and PCR primers and to display restriction maps for a sequence. Several search functions are implemented, including a full text search of the gene annotation. Each gene has a details page where further information about the gene such as its annotation or assigned Gene Ontology (GO) terms [23] is displayed. Apart from the DNA, the coding and the translated sequence of a gene, an up- or downstream region can be specified to be displayed on this page. At the bottom of each details page, links allow users to directly send the specific sequence to other analysis tools such as BLAST. Furthermore, the results of a precalculated InterProScan pattern search [24] are displayed for each annotated protein and can be accessed through the respective link. A comments section enables researchers to add information to their genes of choice. Data downloads are available either in the format of decorated FASTA files or gff files which include gene annotation. Future work on the genome browser of *P. pastoris* DSMZ 70382 will include a genome snapshot which will summarize the status of annotation and the distribution of gene products among functional groups. Batch download processes and an extension of the tools section are planned as well as a platform for the community to share experiences and knowledge in order to promote collaboration. Tutorials for GBrowse are available at [25] or [26].

Except the basic genome browsing and search function, the genome portal of GS115 strain also provides a comprehensive protein-coding gene annotation by the BOGAS (Bioinformatics Gent Online Genome Annotation System). The BOGAS is a gene centric concept, which means the information is provided based on the information related to the gene. Each gene has it's own annotation page which provides an overview of the gene information including the annotator, gene function, gene ontology, protein domain, protein homologs, gene structure, CDS and protein. The annotator information tells who and when annotated this gene and the history log to go back to previous version. Gene function field is filled by anno-

tators with the full gene function and a dictionary to provide a standardized gene nomenclature (short name). The BOGAS system automatically updates the protein information to provide the gene ontology and protein domain by InterProScan, the protein homologs and the multiple alignment by BLASTP and MUSCLE [27] when the user updates the gene structure.

The most important feature of BOGAS system is that it allows the registered users to update the information. Users can correct existing gene structure or create new genes by the annotation software (Artemis [28] or GenomeView [29]) and contribute their expert biological domain in the gene function field. Since the BOGAS provides the history log function, other experts can update the information and people in the community can trace these changes in few clicks. The full text search function in BOGAS can search across locus id, protein domain, genomic location and annotator information. The BLAST function also provides bidirectional link between the query sequence and the possible gene or genomic region. After running the sequence similarity search to fish out the candidate gene or genomic sequence, the user will be linked between the BLAST search result and the corresponding gene region.

As it has been adopted already to a large extent, we suggest that *P. pastoris* gene names should follow the format established for *S. cerevisiae* gene names. A detailed guide to *S. cerevisiae* nomenclature has been published in Trends in Genetics [30]. The gene name should consist of three letters followed by an Arabic number (e.g. *TPI1*). Where *P. pastoris* and *S. cerevisiae* genes appear to be orthologous, they should share the same gene name. The use of prefixes adds clarity to papers discussing genes from different species that share a name (e.g., Pp*URA3* vs. Sc*URA3*), but the gene names themselves do not include the prefix.

These two *Pichia pastoris* genome sites have been developed as a service for the scientific community. The remote annotations can be added either by informing the authors or through the BOGAS system. The WiKi based platform will allow to create additional information on genes, proteins, physiology and other items of *P. pastoris* research. We invite the *P. pastoris* community to join our efforts by providing new information on gene annotation, function, regulation and structure.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to this manuscript.

References

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology.** *Annu Rev Genomics Hum Genet* 2001, **2:**343-372.
2. Lee S, Lee D, Kim T: **Systems biotechnology for strain improvement.** *Trends Biotechnol* 2005, **23(7):**349-358.
3. Cereghino JL, Cregg JM: **Heterologous protein expression in the methylotrophic yeast Pichia pastoris.** *FEMS Microbiol Rev* 2000, **24(1):**45-66.
4. Macauley-Patrick S, Fazenda ML, McNeil B, Harvey LM: **Heterologous protein production using the Pichia pastoris expression system.** *Yeast* 2005, **22(4):**249-270.
5. Marx H, Mattanovich D, Sauer M: **Overexpression of the riboflavin biosynthetic pathway in Pichia pastoris.** *Microb Cell Fact* 2008, **7:**23.
6. Hu H, Qian J, Chu J, Wang Y, Zhuang Y, Zhang S: **DNA shuffling of methionine adenosyltransferase gene leads to improved S-adenosyl-L-methionine production in Pichia pastoris.** *J Biotechnol* 2009, **141(3-4):**97-103.
7. Hamilton S, Davidson R, Sethuraman N, Nett J, Jiang Y, Rios S, Bobrowicz P, Stadheim T, Li H, Choi B, *et al.*: **Humanization of yeast to produce complex terminally sialylated glycoproteins.** *Science* 2006, **313(5792):**1441-1443.
8. Hamilton S, Gerngross T: **Glycosylation engineering in yeast: the advent of fully humanized yeast.** *Curr Opin Biotechnol* 2007, **18(5):**387-392.
9. Jacobs P, Geysens S, Vervecken W, Contreras R, Callewaert N: **Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology.** *Nat Protoc* 2009, **4(1):**58-70.
10. Gasser B, Saloheimo M, Rinas U, Dragosits M, Rodríguez-Carmona E, Baumann K, Giuliani M, Parrilli E, Branduardi P, Lang C, *et al.*: **Protein folding and conformational stress in microbial cells producing recombinant proteins: a host comparative overview.** *Microb Cell Fact* 2008, **7:**11.
11. Gasser B, Maurer M, Rautio J, Sauer M, Bhattacharyya A, Saloheimo M, Penttilä M, Mattanovich D: **Monitoring of transcriptional regulation in Pichia pastoris under protein production conditions.** *BMC Genomics* 2007, **8:**179.
12. Graf A, Gasser B, Dragosits M, Sauer M, Leparc G, Tuechler T, Kreil D, Mattanovich D: **Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.** *BMC Genomics* 2008, **9(1):**390.
13. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P, Mattanovich D: **The effect of temperature on the proteome of recombinant Pichia pastoris.** *J Proteome Res* 2009:1380-92.
14. Solà A, Maaheimo H, Ylönen K, Ferrer P, Szyperski T: **Amino acid biosynthesis and metabolic flux profiling of Pichia pastoris.** *Eur J Biochem* 2004, **271(12):**2462-2470.
15. Solà A, Jouhten P, Maaheimo H, Sánchez-Ferrando F, Szyperski T, Ferrer P: **Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates.** *Microbiology* 2007, **153(Pt 1):**281-290.
16. Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B: **Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris.** *Microb Cell Fact* 2009, **8:**29.
17. De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouze P, Peer Y Van de, Callewaert N: **Genome sequence of the recombinant protein production host Pichia pastoris.** *Nat Biotechnol* 2009, **27(6):**561-566.
18. **Pichia Genome browser** [http://www.pichiagenome.org]
19. **BOGAS** [http://bioinformatics.psb.ugent.be/webtools/bogas]
20. **Anno-J** [http://www.annoj.org/]
21. **GMOD** [http://www.gmod.org/]
22. Stein L, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich J, Harris T, Arva A, *et al.*: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10):**1599-1610.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.

24. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17(9):**847-848.
25. **GBrowse Tutorial** [http://www.openhelix.com/gbrowse]
26. **GBrowse Tutorial** [http://gmod.org/wiki/Gbrowse]
27. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
28. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16(10):**944-945.
29. **GenomeView** [http://genomeview.sourceforge.net]
30. Cherry JM: **Genetic nomenclature guide. Saccharomyces cerevisiae.** *Trends Genet* 1995:11-12.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Dragosits *et al. BMC Genomics* 2010, **11**:207
http://www.biomedcentral.com/1471-2164/11/207

BMC Genomics

RESEARCH ARTICLE Open Access

The response to unfolded protein is involved in osmotolerance of *Pichia pastoris*

Martin Dragosits[1], Johannes Stadlmann[2], Alexandra Graf[1,3], Brigitte Gasser[1], Michael Maurer[3], Michael Sauer[3], David P Kreil[4], Friedrich Altmann[2] and Diethard Mattanovich*[1,3]

Abstract

Background: The effect of osmolarity on cellular physiology has been subject of investigation in many different species. High osmolarity is of importance for biotechnological production processes, where high cell densities and product titers are aspired. Several studies indicated that increased osmolarity of the growth medium can have a beneficial effect on recombinant protein production in different host organisms. Thus, the effect of osmolarity on the cellular physiology of *Pichia pastoris*, a prominent host for recombinant protein production, was studied in carbon limited chemostat cultures at different osmolarities. Transcriptome and proteome analyses were applied to assess differences upon growth at different osmolarities in both, a wild type strain and an antibody fragment expressing strain. While our main intention was to analyze the effect of different osmolarities on *P. pastoris* in general, this was complemented by studying it in context with recombinant protein production.

Results: In contrast to the model yeast *Saccharomyces cerevisiae*, the main osmolyte in *P. pastoris* was arabitol rather than glycerol, demonstrating differences in osmotic stress response as well as energy metabolism. 2D Fluorescence Difference Gel electrophoresis and microarray analysis were applied and demonstrated that processes such as protein folding, ribosome biogenesis and cell wall organization were affected by increased osmolarity. These data indicated that upon increased osmolarity less adaptations on both the transcript and protein level occurred in a *P. pastoris* strain, secreting the Fab fragment, compared with the wild type strain. No transcriptional activation of the high osmolarity glycerol (HOG) pathway was observed at steady state conditions. Furthermore, no change of the specific productivity of recombinant Fab was observed at increased osmolarity.

Conclusion: These data point out that the physiological response to increased osmolarity is different to S. *cerevisiae*. Increased osmolarity resulted in an unfolded protein response (UPR) like response in *P. pastoris* and lead to pre-conditioning of the recombinant Fab producing strain of *P. pastoris* to growth at high osmolarity. The current data demonstrate a strong similarity of environmental stress response mechanisms and recombinant protein related stresses. Therefore, these results might be used in future strain and bioprocess engineering of this biotechnologically relevant yeast.

Background

The response of cells to high osmotic pressure and increased salinity has been a subject of close investigation in many different organisms [1-4]. Depending on the intensity of the osmotic shock the immediate response to high osmolarity usually includes the activation of the environmental stress response (ESR) and of the high osmolarity glycerol (HOG) pathway to induce changes that are necessary to cope with this stressful environmental condition in *Saccharomyces cerevisiae* [5,6]. In batch culture, osmotic shock usually implies a temporary growth arrest to adapt the cellular metabolism [7]. Major adjustments of gene transcription in *Saccharomyces cerevisiae* and other yeasts include the induction of glycerol-3-phosphate dehydrogenase *GPD1* transcription [8], transcriptional repression of the plasma membrane glycerol efflux channel *FPS1* [2], but also the adjustment of ribosome biogenesis and the translation and protein folding machinery [9]. Glycerol production, but also the production of other small organic molecules, is induced in

* Correspondence: diethard.mattanovich@boku.ac.at
[1] Department of Biotechnology, BOKU-University of Natural Resources and Applied Life Sciences, Vienna, Austria
Full list of author information is available at the end of the article

BioMed Central © 2010 Dragosits et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

different yeast species to compensate variations of osmotic conditions [10]. Polyols, such as glycerol, pertain to a class of small molecules known as compatible solutes, which, in contrast to inorganic ions, can be safely accumulated and degraded in the cell without impairing cellular function or having detrimental effects on protein and nucleic acid stability [11]. Furthermore, biomass yield is reduced upon exposure to high osmolarity because of higher maintenance energy in both, batch and chemostat cultures [7,12]. However, it is known that after the immediate shock response, transcript levels of many stress responsive genes return to near basal levels after cells have adapted to the new environmental conditions [6].

The effect of osmolarity on cellular physiology is not only of particular interest for the basic research community. As biotechnological production processes aim at high cell and product concentrations, cultivation media usually employ high concentrations of nutrient salts and carbon source resulting in high osmolarities. Additionally, there is some evidence that exposure to osmotic stress can have a beneficial effect on recombinant protein production in bacterial, yeast and mammalian host organisms [13-16]. Unfortunately, the positive effect of increased osmolarity on heterologous protein production is, at least in mammalian cells, often cell line specific [17] and e.g. in case of the yeast *Pichia pastoris* it remains anecdotal. The genome sequence of *P. pastoris* has been recently published [18,19] and with a publicly available sequence at hand thorough physiological investigations and characterization of this biotechnologically relevant organism becomes feasible.

In this context the effect of osmolarity on the physiology of *P. pastoris* was analyzed in both a non-expressing wild type (wt) strain and a recombinant protein secreting strain. The protein secretion strain expressed the antibody Fab fragment 3H6 [20,21] under the control of the constitutive glyceraldehyde-3-phosphate dehydrogenase (GAP) promoter. The effect of osmolarity was monitored in steady state by applying chemostat cultivation in both strains. Although chemostat cultivation differs from batch and fed batch systems, which are usually applied for large scale production of recombinant proteins, long term suboptimal growth conditions as they occur during batch and fed batch cultivation can also be applied in steady state chemostat conditions [22]. Furthermore, chemostat cultivation offers the advantage that growth rate related effects, which otherwise would interfere with high throughput protein and mRNA analytics, can be avoided [23].

To analyze the effect of increased osmolarity on host cell physiology, 2D Fluorescence Difference Gel Electrophoresis (2D-DIGE) and DNA microarray analyses were applied. These techniques have already been successfully applied to monitor the effect of environmental factors, such as temperature and osmolarity in yeasts [6,24-26]. Furthermore, HPLC analysis was applied to analyze to intracellular polyol and trehalose contents.

The obtained data indicated an unfolded protein response (UPR) like response upon growth at increased osmolarity in the non-expressing wt strain of *P. pastoris*. In the recombinant protein secreting strain, the UPR was obviously already induced due to protein overexpression. The observed overlap of the response to increased osmolarity and the response to recombinant protein production, lead to less adaptations/changes upon high osmolarity on both, the transcriptome and proteome scale, in the Fab secreting strain than in the non-expressing wt strain.

Results

General characteristics of cultures at different osmotic conditions

Chemostat cultivations of *P. pastoris* were performed at three different osmotic conditions, which were achieved by different concentrations of KCl in the growth medium. This resulted in supernatant osmolarities of approximately 140, 850 and 1350 mOsmol kg^{-1}, which will be named low, medium and high osmolarity thereafter. Samples were taken at steady state, which means fully adapted cells were analyzed. The characteristics of the chemostat cultures did not dramatically change upon cultivation at different osmotic conditions (Table 1). Biomass yield decreased with increasing osmolarity in the wt strain and Fab 3H6 producing strain. However, the decrease in biomass yield was only statistically significant ($p \leq 0.05$) between cultivations at low and high osmolarity in both strains. The amount of total protein secreted into the culture supernatant did not change upon higher osmotic pressure, but was generally higher in the Fab 3H6 production strain. Osmolarity of the growth medium did not significantly influence specific productivity (q_P) of the Fab 3H6. Generally, increased osmotic pressure poses a severe stress condition to cells [2]. Although the osmolarity was increased approximately 6-fold and ten-fold in the current study, no decrease of cell viability was observed. More than 97% of the cells in chemostat culture represented viable cells throughout all cultivations (Table 1). According to flow cytometry data, increased osmolarity resulted in a decrease of the mean cell size as indicated by a decrease of the mean forward scatter of the cells (Table 1).

Production of compatible solutes and trehalose in P. pastoris upon growth at different osmolarities

In yeasts, glycerol is a very common solute but other polyols such as arabitol, mannitol and erythritol are also produced in some yeast species [10]. To analyze whether

Table 1: Characteristics of *P. pastoris* X-33 grown in carbon limited chemostat cultures at different osmolarities.

Clone	Osmolarity [mOs kg-1]	YDM [g L-1]	Total protein supernatant [mg L-1]	viability [%]	forward scatter	qP FabYDM-1h-1
wt	149 +/- 11,4	27.7 +/- 0.3	0.38 +/- 0.01	97.5 +/- 0.3	443 +/- 95	-
wt	865.7 +/- 3,2	27.3 +/- 0.2	0.36 +/- 0.02	98.8 +/- 0.2	258 +/- 15	-
wt	1351.3 +/- 2,0	26.0 +/- 0.3	0.38 +/- 0.00	98.1 +/- 0.2	297 +/- 4	-
Fab	135.2 +/- 3.3	27.8 +/- 0.2	0.47 +/- 0.04	97.2 +/- 0.4	355 +/- 60	0.039 +/- 0.004
Fab	857.3 +/- 8.5	26.8 +/- 0.4	0.44 +/- 0.05	97.9 +/- 0.6	225 +/- 30	0.042 +/- 0.002
Fab	1352 +/- 10,1	24.9 +/- 0.3	0.45 +/- 0.05	97.8 +/- 0.5	206 +/- 12	0.047 +/- 0.006

wt represents the non-expressing wild type strain. Fab represents the recombinant protein producing strain. Osmolarity values represent actual measured values of the culture supernatant. +/- represents the standard error of the mean. - not applicable.

P. pastoris produces any of these substances, cell extracts were analyzed by HPLC.

It turned out that very low levels of mannitol and nearly no detectable amounts of erythritol were present in *P. pastoris* cells (Additional file 1). Intracellular glycerol levels were higher than mannitol and erythritol levels and a slight significant increase in the wt strain from low to high and medium to high osmolarity occurred ($p \leq 0.05$), whereas no significant changes of glycerol content were observed in the Fab 3H6 expressing strain (Figure 1A). Surprisingly, arabitol was the most abundant compound of the analyzed polyols in *P. pastoris* cells (five fold higher basal level then glycerol) and showed statistically significant increased levels ($p \leq 0.05$) when shifting growth conditions towards high osmolarity (Figure 1B). A 3-fold increase of intracellular arabitol levels was observed when comparing low and medium osmolarity conditions and a 4-fold increase when comparing cells grown at low and high osmolarity.

Furthermore, intracellular levels of trehalose were analyzed as trehalose is thought to be involved in relieving or impeding protein folding stress [27], which may also occur during salt stress [9]. Intracellular trehalose levels were in the same range as glycerol levels but showed a significant trend ($p \leq 0.05$) towards decreased concentrations at medium and high osmolarity growth conditions in the wt strain but slightly missed the threshold p-value in the 3H6 Fab expressing strain (Figure 1C and Additional file 1)

The effect of osmolarity on the P. pastoris intracellular proteome

As 2D-DIGE has already been successfully applied to monitor changes in the *P. pastoris* proteome upon growth at different temperatures [26], this method was also applied to track changes upon growth at elevated osmolarity. In the wt strain of *P. pastoris* approximately 300 proteins passed the criteria (see experimental procedures section for details), whereas about 150 proteins passed these criteria in the recombinant protein producing strain. Most of the protein spots represented low abundant proteins with too small quantities of proteins on the 2D gels to be confidently identified, resulting in 37 successfully identified proteins (Table 2). A list of all identified protein spots with corresponding peptides, obtained from MS/MS analysis, is available in Additional file 2. As already reported in previous studies [26,28], additionally to most likely full length proteins, protein fragments were identified according to the spot position on the gel (e.g Spot 4-Aco1, Spot 20 and 21-Ino1 and Spot 30-Ssb1; Table 2). Furthermore, protein levels showed the largest changes when low and medium and low and high osmolarity cultivations were compared, whereas only minor changes occurred when comparing medium and high osmolarity setpoints (see Additional file 3 for a complete table of relative protein levels).

As can be seen in Figure 2, the major impact of osmolarity on the *P. pastoris* proteome was on proteins involved in energy metabolism and protein folding. Whereas protein levels of a major spot of aconitate hydratase (Aco1p) were increased at medium and high osmolarity in the wt strain, they were not significantly affected by osmolarity in the Fab 3H6 producing strain. In contrast, in the Fab 3H6 expressing strain three minor isoforms or degradation products were significantly down-regulated at higher osmolarity, but showed no altered abundance in the wt strain. Furthermore, citrate synthase (Cit1p) protein levels were decreased at high salt concentrations in the wt strain but showed no significant change in the Fab producing strain. Formate dehydrogenase (Fdh1p), glycerol kinase (Gut1p), and isocitrate lyase (Icl1p) showed similar trends towards lower protein levels during medium and high osmolarity cultivations in both strains, whereas pyruvate kinase (Cdc19p) and phosphoglycerate kinase (Pgk1p) were generally up-regulated at higher osmolarity (it should be noted that Pgk1p

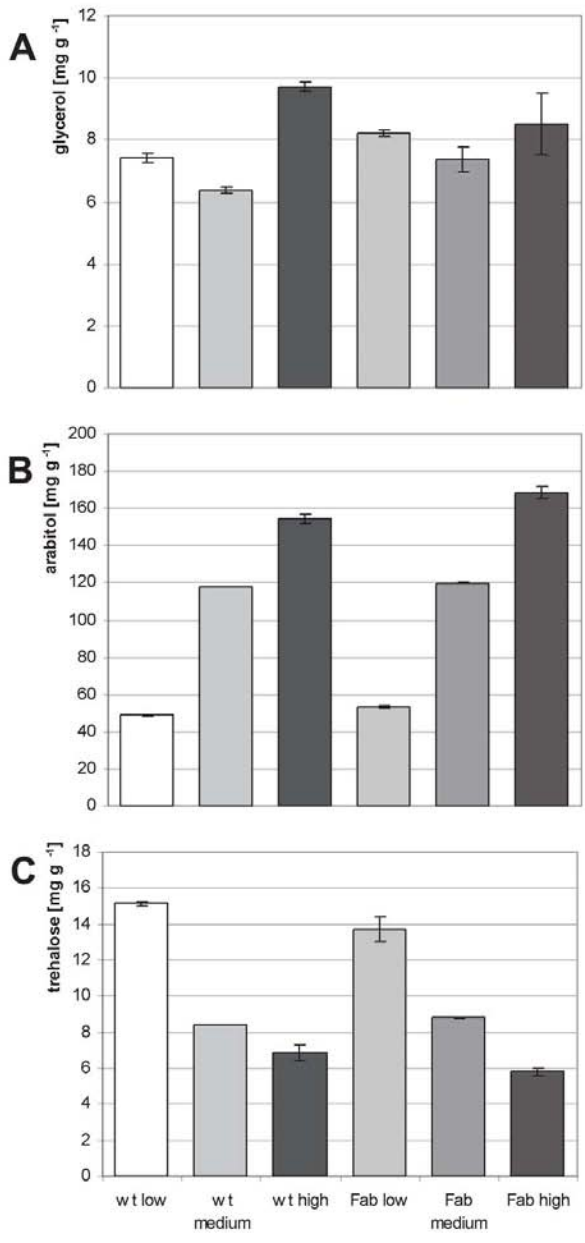


Figure 1 Intracellular glycerol (A), arabitol (B) and trehalose (C) levels during carbon limited steady state cultivation of *P. pastoris* at different osmolarities (low, medium and high). wt - wild type strain. Fab 3H6 - Fab 3H6 strain. Error bars represent the standard error of the mean.

levels returned to levels similar to low osmolarity cultivation in the production strain). Phosphoglucose isomerase (Pgi1p) showed lower levels at medium osmolarity in the production strain and a protein identified as Atp3p (a subunit of the mitochondrial F0F1 ATPase) was massively down-regulated at medium and high salt concentrations in the wt strain. Furthermore, alcohol oxidase (Aox1p), a key enzyme in methanol utilization, was down-regulated at medium and high osmolarity in the wt strain but did not show a significant change in the Fab 3H6 expressing strain. It should be pointed out that *AOX1* transcription is thought to be repressed during growth on glucose and that the current study was performed with glucose as carbon source to constitutively express the heterologous protein under the control of the GAP-promoter. However, it was shown previously that basal levels of Aox1p were actually present during glucose limited growth of *P. pastoris* [26].

Similar discrepancies between the wt and the recombinant protein expressing strain were observed for proteins involved in protein folding and secretion and folding stress response. Whereas the major ER chaperone and unfolded protein response (UPR) sensor Kar2p/BiP and the protein disulfide isomerase Pdi1p were up-regulated at medium and high osmolarity in the wt strain, no changes of these two proteins were observed in the Fab producing strain. More prominently increased levels of cytosolic and mitochondrial chaperones Ssc1p, Sse1p, Ssz1p and Hsp60p were observed at medium and high osmolarity in the wt strain than in the Fab 3H6 producing strain. The stress induced chaperone Ssa4p showed increased levels at medium salt concentrations in both strains but returned to below basal levels at high salt conditions in the Fab producing strain. Ino1p, a protein involved in synthesis of inositol phosphates and inositol-containing phospholipids and which is linked to the UPR [29], was also up-regulated at high salt concentrations in both strains analyzed.

Other protein spots that changed their abundance upon cultivation at increased salt concentrations were Agx1p and Gdh1p (both involved in amino acid synthesis). Both of them showed higher protein levels during growth at high osmolarity.

Figure 2 summarizes the osmolarity-induced effects observed on the proteome level of *P. pastoris*.

The effect of osmolarity on the P. pastoris transcriptome

To analyze the effect of different osmolarities on the *P. pastoris* transcriptome, *P. pastoris* specific microarrays were applied (Agilent platform). To support microarray analysis, real-time PCR was performed. Real-time PCR data proved to be consistent with microarray results (Additional file 4). More significant changes on the transcriptome level were observed in the wt strain than in the recombinant protein expressing strain (Table 3). Low *p*-values can result from a high technical variation within the replicates or reflect the biological truth within the samples. To determine if the lower amount of significantly regulated genes in the Fab expressing strain is a technical artefact, the correlation, standard deviation and the coefficient of variation for the replicates of the wt and expressing strain were compared. Correlation of intensity

Table 2: Proteins that were affected by growth at different osmolarities in carbon limited chemostat cultures of *P. pastoris* X-33.

				wt low/high		Fab low/high	
Spot no	**Protein**	**description**	**MW/pI**	**Av ratio**	**1-ANOVA**	**Av ratio**	**1-ANOVA**
1	Aco1	aconitase	84.5/5.93	-2.25	**3.70E-04**	1.01	**7.60E-04**
2	Aco1	aconitase	84.5/5.93	-1.13	5.80E-02	1.54	**4.40E-05**
3	Aco1	aconitase	84.5/5.93	-1.23	8.40E-02	1.49	**5.10E-04**
4	Aco1	aconitase	84.5/5.93	1.10	2.10E-01	1.51	**2.30E-04**
5	Agx1	alanine:glyoxylate aminotransferase	31.0/6.36	-1.79	**1.00E-04**	-1.43	**8.80E-03**
6	Aox1	alcohol oxidase	73.8/6.41	2.34	**1.90E-02**	-1.36	1.80E-01
7	Atp3	F1F0 ATPase subunit	31.6/7.74	3.28	**3.20E-06**	1.26	4.20E-01
8	Cdc19	pyruvate kinase	49.6/6.24	-1.81	**3.40E-03**	-1.26	**9.10E-03**
9	Cit1	citrate synthase	51.9/8.32	3.23	**2.90E-04**	-1.04	1.30E-01
10	Eft2	Elongation Factor 2	93.6/6.29	-1.93	**3.40E-04**	1.14	6.30E-02
11	Erg10	acetyl CoA acetyltransferase	41.7/6.10	-1.83	**7.00E-05**	-1.20	5.30E-02
12	Faa2	long chain fatty acyl-CoA synth.	25.4/6.73	-1.65	**3.40E-03**	1.16	5.10E-01
13	Fdh1	formate dehydrogenase	40.3/6.61	1.79	**1.30E-03**	1.35	**3.20E-03**
14	Gdh1	glutamate dehydrogenase	49.3/5.67	-2.38	**3.40E-02**	-1.45	**2.20E-04**
15	Gut1	glycerol kinase	68.2/5.33	1.31	**5.90E-03**	1.18	**1.20E-03**
16	Hbn1	nitroreductase (similar to bacterial)	21.8/6.30	1.12	2.10E-01	1.36	**3.00E-08**
17	Hsp60	heat shock protein 60	60.2/5.08	-1.83	**1.20E-05**	-1.13	**1.40E-04**
18	Icl1	isocitrate lyase	61.5/6.15	1.41	**2.10E-03**	1.56	**1.90E-05**
19	Ino1	inositol-1-P synthase	58.4/5.26	-2.61	**1.30E-05**	-1.88	**1.30E-04**
20	Ino1	inositol-1-P synthase	58.4/5.26	-1.78	**1.80E-05**	-1.11	**1.60E-04**
21	Ino1	inositol-1-P synthase	58.4/5.26	-1.06	3.60E-01	1.41	**9.00E-04**
22	Kar2	BiP	74.2/4.79	-2.95	**4.70E-05**	-1.05	5.90E-01
23	Pab1	poly A binding protein	68.6/5.07	-1.77	**5.00E-04**	-1.03	**3.70E-05**
24	Pdi1	protein disulfide isomerase	57.8/4.63	-1.66	**2.10E-05**	1.08	**1.00E-02**
25	Pgi1	phosphoglucose isomerase	61.9/5.83	1.12	1.00E-01	1.07	**7.00E-04**
26	Pgk1	phosphoglycerate kinase	44.1/7.77	-1.97	**3.30E-05**	-1.03	**2.40E-02**
27	Pil1	Primary component of eisosomes	35.3/5.03	1.35	**2.60E-02**	1.12	**6.20E-03**
28	Rib3	DHBP synthase/ riboflavin	22.9/5.09	-1.84	**7.80E-06**	-1.16	7.60E-02

Table 2: Proteins that were affected by growth at different osmolarities in carbon limited chemostat cultures of *P. pastoris* X-33. (Continued)

29	Sor2	similar to sorbitol dehydrogenase	38.6/5.76	1.34	**3.50E-02**	1.24	**2.50E-03**
30	Ssa4	heat shock protein	70.3/5.12	-1.20	**4.90E-02**	1.15	**1.10E-05**
31	Ssb1	heat shock protein	66.5/5.12	-1.02	8.70E-01	1.17	**1.70E-03**
32	Ssb1	heat shock protein	66.5/5.12	1.58	**1.20E-03**	1.03	**1.50E-05**
33	Ssc1	mitochondrial matrix ATPase	69.7/5.71	-4.7	**3.80E-03**	-1.20	**5.40E-05**
34	Sse1	hsp70 family ATPase	78.7/5.11	-2.79	**5.80E-05**	-1.14	**3.70E-04**
35	Ssz1	hsp70 family ATPase	57.9/4.83	-1.81	**8.00E-06**	-1.21	**4.30E-05**
36	Tfs1	carboxypeptidase Y inhibitor	24.2/4.92	-2.00	**3.90E-04**	-1.39	**2.00E-02**
37	Ymr090W	unknown function	25.1/6.91	-1.41	**3.40E-05**	-1.23	**1.30E-04**

Analysis was performed by 2D-DIGE and subsequent LC ESI-MS/MS identification. Protein standard name (according to the SGD, http://www.yeastgenome.org), protein functional description, theoretical molecular weight (MW) and theoretical isoelectric point (p*I*), average expression values and 1-ANOVA values are shown. Significant ANOVA values are indicated in bold letters. The average protein abundance fold change of the comparison of low and high osmolarity setpoints (Av ratio) are shown. Additional protein fold change data are available through Additional file 3.

values was generally high between all microarrays of one group (wt/expressing red channel/green channel (see Additional file 5) with r^2 values between 0.95 and 0.97. The values for standard deviation and coefficient of variation (CV) indicated that the variance in replicates of the expressing strain was slightly but consistently higher than for the wt strain (on average CV 0.18 for the wt and CV 0.28 for the expressing strain, Additional file 5). Based on these results additional microarray experiments were performed to exclude any bias in the data. These additional data did not change the result or number of regulated genes, suggesting a true biological difference. To eliminate the possibility that the samples of the expressing strain vary more than the ones of the wt strain, hierarchical cluster analysis (HCA) and gene set analysis (GSA) were performed on the fold change data and indicated that regulation was indeed different in the two strains analyzed (Additional file 5).

Because most of the genes that were regulated when comparing low to medium osmolarity were also regulated when comparing low to high osmolarity, the following data presentation and discussion will focus on the effects that were observed when low and high osmotic conditions were compared.

To get an overview of the general adaptations during steady-state cultivation, Fisher's exact test was performed to identify cellular processes, which were affected by different osmolarities on the transcript level. A total of 23 GO categories were either affected in both or at least in one of the analyzed strains (Additional file 5). Concordant with the mere number of regulated genes, there appeared more significantly affected cellular processes in the wt strain than in the heterologous protein expressing strain. Only 3 GO categories occurred to be affected in both strains, namely GO:0006811 (ion transport), GO:0007047 (cell wall organization) and GO:0019725 (cellular homeostasis). Additionally, in the wt strain the GO terms GO:0005975 (carbohydrate metabolism), GO:0006350 (transcription), GO:0006412 (translation) and GO:0042254 (ribosome biogenesis and assembly) were affected by increased extracellular osmolarity.

Figure 3 summarizes the important changes at the mRNA level of *P. pastoris*, grown in carbon-limited chemostat cultures when comparing high to low osmolarity, whereas microarray data for the discussed genes can be found in Additional file 6.

Regarding ion transport, uptake and metabolism, high osmolarity resulted in increased expression of the iron transporters *FTR1*, *SIT1* and the vacuolar iron reductase *FRE6* in both strains. Calcium ion homeostasis and calcium dependent signal transduction were obviously affected by high osmolarity in the wt strain as the Ca^{2+} transporter *PMC1* and Calcineurin A (*CNA1*) were down-regulated at high osmolarity.

A major effect was apparent for genes involved in cell wall organization and its biogenesis. Whereas 21% of the genes belonging to this GO group were down-regulated at high osmolarity in the wt strain, a similar effect, albeit with fewer significant genes, was observed in the 3H6 Fab secreting strain (Figure 4A and 4B). Additionally, a putative extracellular or cell wall associated protein with homology to the *S. cerevisiae PRY1* gene was up-regu-

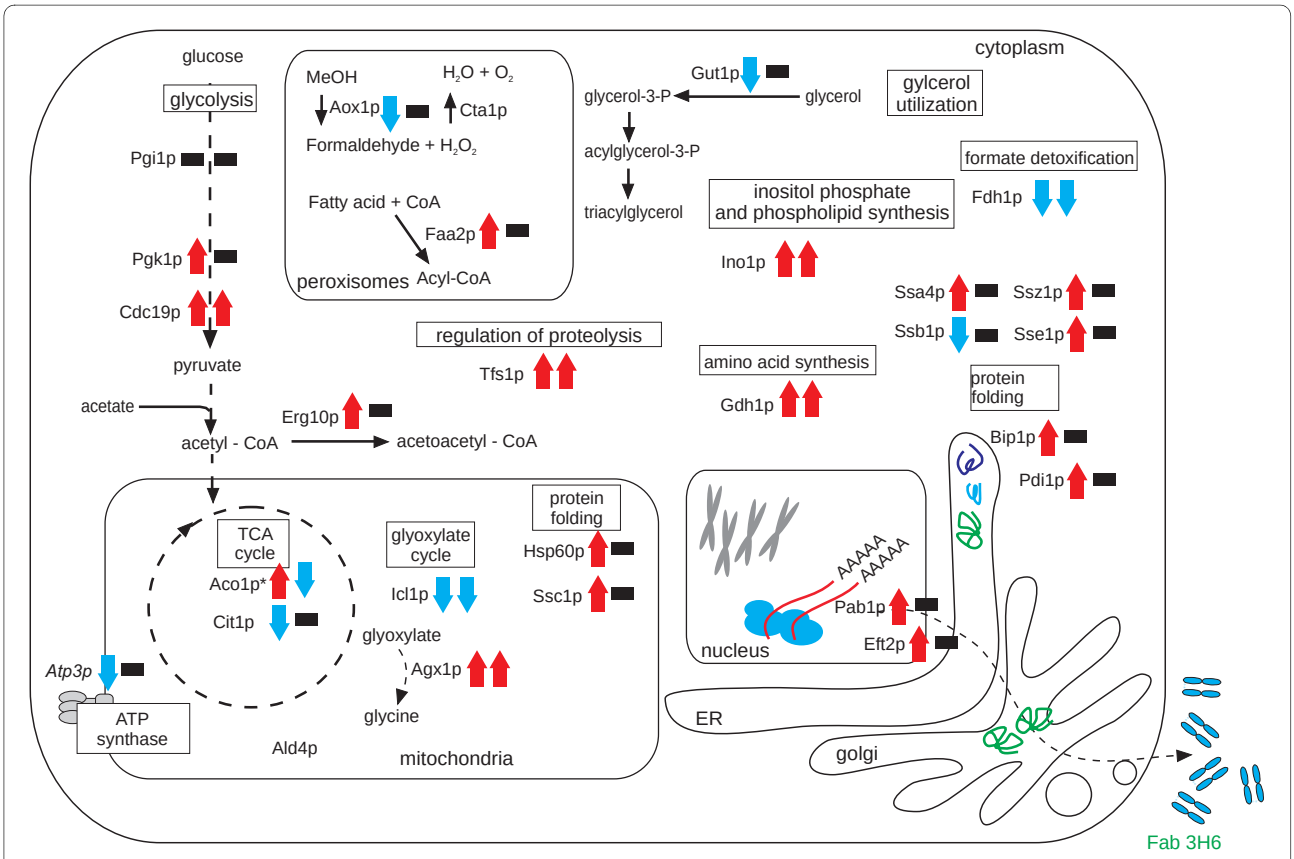


Figure 2 Schematic representation of significant differences on the proteome level between high and low osmolarity cultivation of *P. pastoris* in carbon limited chemostats. Left arrow represents changes in wt strain of *P. pastoris*, and the right arrow represents changes in the Fab 3H6 secreting strain. Red upward arrow - higher abundance at high osmolarity; Blue downward arrow - lower abundance at high osmolarity; black bar - no change in abundance.

lated in both strains at high osmolarity. Increased levels of a gene with homology to *S. cerevisiae PRY1* upon increased salinity have also been reported for the halotolerant yeast *Hortaea werneckii* previously [30]. However, no changes in the protein pattern, indicating higher protein levels, were observed by SDS-PAGE of the culture supernatant (data not shown).

A signaling cascade for sensing and adaptation to osmotic stress in *S. cerevisiae* has already been established based on the available data [2,31] and genes with homology to the corresponding genes in *S. cerevisiae* were also identified in *P. pastoris*. None of the genes involved in osmotic stress sensing upstream of the mitogen activated protein kinase (MAPK) Hog1 showed significant regulation in the Fab 3H6 expressing strain, whereas *SHO1*, *SSK1* and *PTP3* were up-regulated and *STE50* was down-regulated at increased osmolarity in the wt strain (Figure 4A).

Table 3: Number of regulated annotated genes (up- and down-regulated) in the wt strain and the Fab expressing strain at different osmolarities during carbon limited chemostat cultivation.

strain	low/high up	low/high down	low/medium up	low/medium down	medium/high up	medium/high down
Wt	226	165	153	50	1	1
Fab	27	13	13	10	8	6
common genes	22	7	11	4	0	0

Wt - non-expressing wild type strain. Fab - Fab 3H6 expressing strain. The number of genes that are similarly regulated in both strains are listed as common genes.

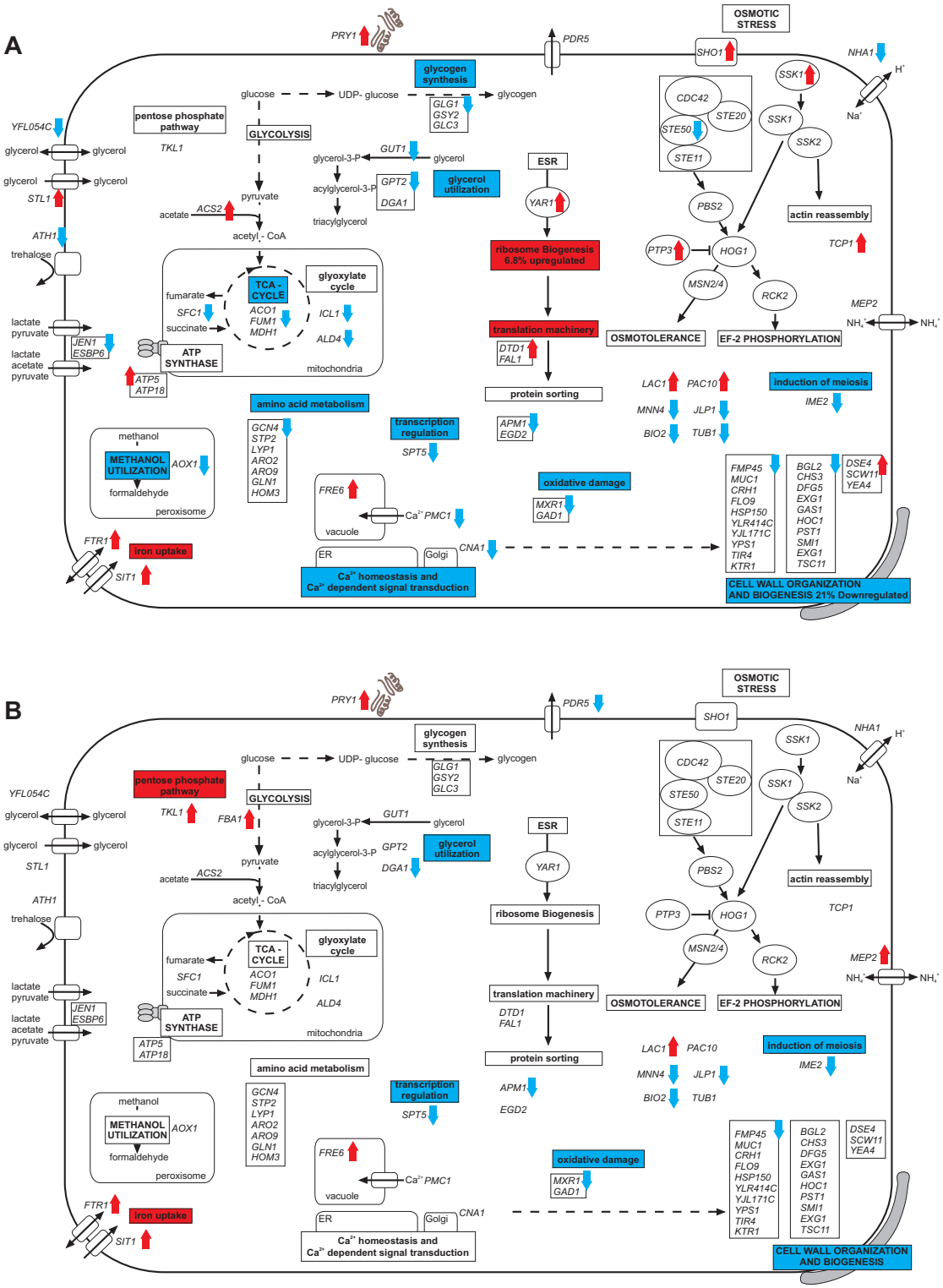


Figure 3 Schematic representation of significant changes on the transcriptome level between high and low osmolarity cultivation of *P. pastoris* in carbon limited chemostats. Wt strain (A). 3H6 Fab expressing strain (B). Only statistically significant genes are represented (cut-off $q < 0.05$). Blue downward arrows indicate down-regulation of genes at high osmolarity. Red upward arrows indicate up-regulated genes at high osmolarity. No arrow indicates no significant regulation of the genes or gene groups.

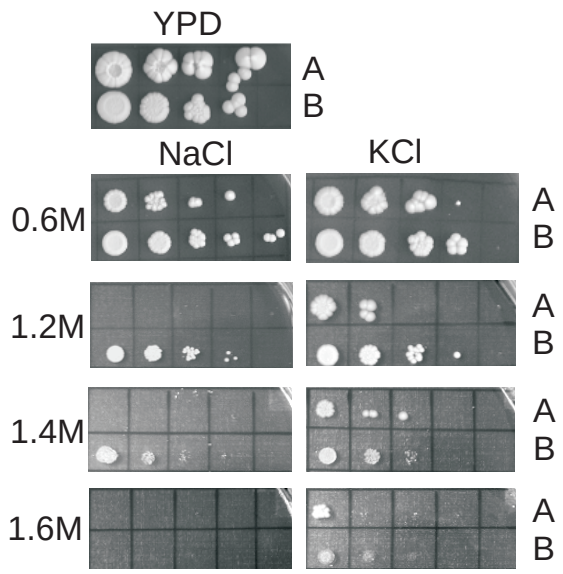


Figure 4 Growth of *S. cerevisiae* HA232 (A) and *P. pastoris* X-33 (B) on YPD agar with different concentrations of NaCl or KCl. YPD agar plates and plates containing 0.6 M NaCl or KCl were incubated at 28°C for 4 days, whereas all other plates were incubated at 28°C for 6 days.

Several genes involved in energy metabolism and storage carbohydrate metabolism were affected by increased osmolarity. *FBA1*, a key enzyme in glycolysis and gluconeogenesis, was up-regulated in the heterologous protein expressing strain at high osmolarity. The acetyl-coA synthetase *ACS2* was up-regulated in the wt strain at high osmolarity. Transcript levels of several genes involved in the tricarboxylic acid (TCA) cycle and the glyoxylate cycle, namely *ACO1*, *FUM1*, *MDH1*, *SFC1* and *ICL1* were reduced and subunits of the ATP synthase (*ATP5* and *ATP18*) were up-regulated during growth at high osmolarity in the *P. pastoris* wt strain. In the Fab producing strain, *TKL1*, involved in the pentose phosphate (PP) pathway was significantly up-regulated. Glycogen synthesis was also affected by high osmolarity as *GLG1*, *GSY2* and *GLC3* showed decreased transcript levels during growth at high osmolarity. Decreased levels of *DGA1*, *GUT1* and *GTP2* indicated changes in glycerol and lipid metabolism. Additionally, a homologue to the *S. cerevisiae* putative passive glycerol channel *YFL054C* was down-regulated and the active glycerol importer *STL1* was up-regulated during growth at high osmotic conditions in the wt strain. Furthermore, significant down-regulation of the *P. pastoris* alcohol oxidase *AOX1* was observed in the wt strain at high osmolarity, whereas no significant regulation was observed in the Fab 3H6 expressing strain. These data are concordant with data on the proteome level (Table 2).

In the wt strain approximately 7% of the genes involved in ribosome biogenesis and assembly were up-regulated during steady state cultivation at high osmolarity.

The effect of recombinant protein production on the transcriptome of P. pastoris at low osmolarity

As the experiment was performed with a non-expressing wt and a recombinant protein expressing strain of *P. pastoris* the effect of recombinant protein expression itself on the cellular transcriptome at low osmolarity could also be analyzed. The most prominent effect of recombinant protein production on host cell physiology is the induction of the unfolded protein response (UPR) [32-35]. Therefore, we analyzed whether this effect could be observed at low osmolarity in glucose limited chemostat cultures of *P. pastoris*. It turned out that the transcriptional response to recombinant protein production during chemostat cultivation at low osmolarity was low. Only 7 mRNAs were significantly regulated when comparing the two strains using a strict cut-off (a Benjamini-Yekutieli corrected p-value of $q \leq 0.05$). Using a less strict cut-off (unadjusted p-value of ≤ 0.001) 79 genes were significantly regulated (Additional file 7). *HAC1* was up-regulated in the Fab expressing strain and indicated the induction of the UPR in the Fab expressing strain. In contrast to *HAC1* other members of the core UPR and targets of the *HAC1* transcription factor showed no significant response by applying a p-value cut-off. However, Gene Set Analysis (GSA) applied on the microarray data revealed increased expression of genes related to the GO term GO:0006986 (Response to unfolded protein) (Additional file 7). The UPR (GO:0030968) is a specific response to unfolded protein in the ER and is a subcategory of GO:0006986.

The up-regulation of the *HAC1* transcript as well as the significant upregulation of the response to unfolded protein indicated the induction of the UPR in the Fab 3H6 expressing strain. The effect was lower than in previous studies, which in fact were performed in non-carbon limited batch cultures. It is known that during carbon-limited chemostat cultivation of *S. cerevisiae* metabolic control can be more important than gene regulation [36]. Furthermore, it is known that part of the regulation of the UPR can be performed on a post-transcriptional or even post-translational level in *S. cerevisiae* and *Aspergillus niger* [37,38]. Less information about control mechanisms are available for *P. pastoris*. We conclude that similar to these previous studies, a substantial part of regulation, including the regulation of the UPR, is achieved on a post-transcriptional level during glucose-limited chemostat cultivation of *P. pastoris*.

Regarding the response to unfolded protein, GSA indicated significantly increased transcription of genes related to this GO category (GO:0006986) in the wt strain

at high osmolarity. In contrast, higher growth medium osmolarity did not result in an induction of the response to unfolded protein in Fab 3H6 expressing strain (Additional file 5).

Salt tolerance of Pichia pastoris

As no data on the salt tolerance of *P. pastoris* compared to *S. cerevisiae* were found in literature, a growth test on YPD agar plates, containing different amounts of NaCl or KCl, was performed. This growth test indicated higher tolerance to growth on NaCl and a less pronounced higher tolerance to growth on KCl of *P. pastoris* compared to *S. cerevisiae* (Figure 4).

Discussion

Production of Compatible Solutes

To counterbalance the osmotic pressure by high or low salt or solute concentrations in the growth medium, microorganisms produce various compatible solutes. In *S. cerevisiae* and many other organisms glycerol is the main osmolyte accumulated during osmotic stress. However, we found that intracellular glycerol levels were low at all osmotic conditions in *P. pastoris*. On the other hand, arabitol was more abundant than glycerol, even at low osmolarity, and it was accumulated in *P. pastoris* during growth at elevated osmotic pressure (Figure 1A and 1B). Glycerol production in *S. cerevisiae* depends on the increased expression of glycerol-3-phosphate dehydrogenase *GPD1* and glycerol-3-phosphatase *GPP2* [7,25]. Nevertheless, we could not find neither of the two genes involved in glycerol metabolism, *GPD1* and *GPP2*, to be up-regulated on the transcript level and did not identify protein spots with altered expression which would match these two genes in *P. pastoris*. Furthermore, it was shown in *Debaryomyces hansenii* that NaCl stress lead to increased levels of proteins involved in the upper part of glycolysis and down-regulation of proteins involved in the TCA-cycle [24]. It was concluded that these changes may favor the accumulation of dihydroxyacetonephosphate and consequently the production of glycerol [39]. No changes related to the upper part of glycolysis were observed in the current study, although it is clear that the need for compatible solutes leads to a redirection of a part of the carbon source to alleviate the stress induced by increased osmolarity. This would make sense as arabitol obviously plays a more important role as compatible solute than glycerol in *P. pastoris*. However, the regulation of genes involved in glycerol uptake and efflux, such as the up-regulation of *STL1* and the down-regulation of *YFL054C* may be beneficial for *P. pastoris* as well. The loss of minor osmolytes may result in detrimental effects on cellular integrity at high KCl concentrations. No significant regulation was observed for other putative glycerol transporters of *P. pastoris* recently described by Mattanovich and co-workers [18]. Arabitol synthesis is linked to the pentose phosphate (PP) pathway. However, no changes possibly linked to the PP pathway and arabitol synthesis were observed on the transcript or the proteome level. The regulation of arabitol synthesis might be mainly achieved on a post-transcriptional level by increased translation or by protein modification and changes of enzyme activity during chemostat cultivation of *P. pastoris*. Nevertheless, concordant with other studies, arabitol obviously is of particular importance for the metabolism of *P. pastoris* as it is also secreted into the supernatant at certain growth conditions, such as low oxygenation [40].

Trehalose has been previously shown to play an important role in heat shock induced refolding of proteins in baker's yeast [41] and *in vitro* [27]. Furthermore, trehalose may also be involved in the response to temperature induced stress in *P. pastoris* as intracellular levels increased at elevated temperature (own unpublished data). However, as trehalose levels were lower during growth at high osmolarity and no changes of the stress induced cytosolic trehalase *NTH1* [42] were observed on the protein or mRNA level, trehalose may not be directly involved in the protection of proteins against osmotic induced protein denaturation or damage. It is more likely that, similar to *S. cerevisiae*, trehalose degradation may play a role during growth at elevated osmolarity [43], or that trehalose levels may be simply lower due to a redirection of carbon source to the production of arabitol rather then to the production of trehalose.

Effect on Energy Metabolism

The differential response of Aco1p and the differences of transcript levels of genes involved the TCA cycle to different osmotic conditions between the wt and the Fab 3H6 expressing strain lead to the conclusion that recombinant protein production influenced the osmo-dependent adaptation of the energy metabolism. Previous data already indicated a metabolic burden and influence of recombinant protein production on energy metabolism in *P. pastoris* [26,44]. Furthermore, the key enzyme of methanol utilization, *AOX1*, was differently regulated in the two strains and indicated significant differences in the regulation of energy metabolism. Protein and transcript levels of the alcohol oxidase (*AOX1*) were significantly negatively affected by growth at high osmolarity in the wt strain but not in the Fab 3H6 secreting strain. *P. pastoris* Aox1 seems to be tightly regulated upon exposure to various stresses and might represent an ideal candidate as a marker gene/protein to monitor diverse external and internal stresses in *P. pastoris*. Apart from these additional data supporting the idea of a metabolic burden during recombinant protein production in *P. pastoris*, no clear interpretation about the changes of energy metabo-

lism upon growth at different osmolarities in chemostat cultures emerged. Further investigations using a different approach to the one used in the current study will be necessary to elucidate the effect of osmolarity on the energy metabolism of *P. pastoris*.

Activation of translation, ribosome biogenesis and the response to unfolded protein at high osmolarity

Another major effect was the massive increase of chaperones and UPR related proteins at high osmolarity. The UPR, including heat shock proteins and cellular chaperones, plays an essential role in the response to various stresses [45]. Apart from its role in the ESR of unicellular organisms, the UPR is also of great importance in human disease as highlighted by its involvement in the development of several human maladies such as diabetes, neurodegenerative disorders and cancer [46,47]. The observation of increased levels of molecular chaperones during growth at high osmolarity is concordant with previous results for *Aspergillus nidulans* [1] and similar to results obtained for *D. hansenii* [24] and *S. cerevisiae* [25] in batch culture. Furthermore, high osmotic pressure resulted in increased levels of Pdi1p and Kar2p, indicating Endoplasmic Reticulum related protein folding stress. Unlike *S. cerevisiae*, UPR induction has been reported to be a main event upon exposure to salt stress in the halotolerant yeast *Rhodotorula mucilaginosa* [9]. Generally, the induction of the UPR may not only be a result of high concentrations of ionic solutes such as salts but the response to unfolded proteins is also triggered by high osmotic pressure induced by other substances such as sugar compounds. It has been reported for mammalian cells that low as well as high hexose concentrations can lead to UPR induction [48,49]. The UPR has been described to be mainly a transcriptional response, but recently post-transcriptional and post-translational regulation has been described for fungal organisms [37,38]. In this context, mRNA levels of the UPR transcription factor *HAC1* are increased in the Fab expressing strain as a reaction to recombinant protein production. On the other hand, GSA of the microarray data also showed induction of "responses to unfolded protein" at high osmolarity in the wt strain. These results are supported by the effects observed at the proteome level. A conventional induction of the UPR by increased *HAC1* levels was not observed at increased osmolarity in neither of the strains. However, increased osmolarity resulted in increased Kar2p and Pdi1p on the proteome level in the wt strain. As this clearly demonstrated ER associated protein folding stress we refer to a UPR-like response of *P. pastoris* wild type cells at high osmolarity. Although a direct comparison between the wt and recombinant protein producing strain was not possible on the proteome level, this comparison was possible on the transcript level and strongly indicated the upregulation of processes involved in response to unfolded protein in the recombinant Fab producing strain. Thus, we hypothesize that the up-regulation of these ER resident proteins as well as other chaperones was obviously not necessary in a Fab 3H6 producing strain at high osmolarity, as these changes had already been triggered by the UPR that was induced by the recombinant protein.

Additionally to this UPR-like response, the induction of ribosome biogenesis and translation were apparent on the transcript level in the wt strain. The up-regulation of genes involved in protein synthesis during osmotic stress has been reported for the salt-tolerant yeasts *H. werneckii* and *D. hansenii* [30,50]. Furthermore, studies on brewing strains of *S. cerevisiae* concluded that the faster adaption to higher salt concentration compared to a laboratory strain was achieved by higher expression levels of genes involved in protein synthesis [51]. Similar to other environmental factors, such as temperature [52], translation might become a rate-limiting factor during growth at high osmolarities because of stress related to decreased intracellular water availability. Boosting the protein synthesis machinery might be necessary for growth of *P. pastoris* at elevated osmolarity. The Fab 3H6 expressing strain of *P. pastoris* did not show this increase of the protein synthesis machinery. We have shown previously that over-expression of the transcription factor *HAC1* in *P. pastoris* batch cultures resulted in increased expression of genes involved in mRNA translation and to a massive increase of genes involved in ribosome biogenesis and assembly [53]. Obviously the up-regulation of ribosome biogenesis, translation and other co-regulated processes at high osmolarity was not necessary in the Fab 3H6 producing strain as these changes had already been induced by recombinant protein production itself.

This UPR-like response at high osmolarity also points to the fact that, similar to halotolerant yeast species like *R. mucilaginosa*, *P. pastoris* might use different mechanisms for gaining osmotic stress resistance than *S. cerevisiae*. This hypothesis was supported by the growth tests for salt tolerance, which were performed with *P. pastoris* and *S. cerevisiae*. *P. pastoris* showed indeed higher resistance to increased salt concentrations in the growth medium than *S. cerevisiae* (Figure 4). Many changes observed in the wt upon a change from low to high osmolarity were not observed in the recombinant protein expressing strain. Although high osmolarity triggered the response to unfolded protein, ribosome biogenesis and translation in the wt strain, the activation of these apparently co-regulated processes was compensated in the Fab 3H6 strain by recombinant protein induced UPR.

Other cellular responses

Increased osmolarity also influenced other cellular mechanisms, such as some parts of the oxidative damage response, which apparently were not co-regulated with the protein synthesis and folding machinery. Therefore, these processes which seem to be an essential part of the response to increased osmolarity were monitored in both strains of *P. pastoris*. The interrelation of salt and oxidative stress is already established in plants [54] and the interrelation and cross-talk of the HOG pathway and other pathways such as protein kinase C (PKC) and calcineurin dependent signaling are also established in yeasts [2,55,56]. Changes in cell wall integrity signaling, which were evident by altered expression levels of cell wall components in both strains, may be directly related to the changes of the *CNA1* and *PMC1* transcripts, as some of these cell wall synthesis related genes are dependent on calcineurin signaling [56]. For example, decreased transcript levels of *CRH1* and *GAS1* at high osmolarity may indicate a change of cell wall rigidity at high osmolarity. High osmolarity results in decreased turgor pressure when compared with low or hypo-osmotic conditions. Thus, increased osmolarity results in cell shrinkage and smaller cells, which in fact was obvious by a decreased mean forward scatter of the cells in the present study (Table 1). In this context the down-regulation of cell wall components, which would result in lower cell wall rigidity, at high osmolarity makes sense. It has been shown in *Aspergillus nidulans* that salt addition to the growth medium resulted in decreased cell wall rigidity [57]. A further indication that high osmolarity is compatible with lower cell wall rigidity is the fact that the swollen cellular phenotype of Gas1 mutant cells, a gene which is also downregulated at high osmolarity in the current study, in *S. cerevisiae* is compensated by growth at high osmolarity [58]. Although this effect was very evident on the transcript level we were not able to monitor it on the proteome level. This may be simply due to the preparation of protein samples and the resulting absence of cell wall and membrane proteins, which are rather difficult to extract by standard protein preparation methods.

Additionally to these events, the induction of iron transporters at high osmolarity also occurred in both strains. A proteomic study of *Bacillus subtilis* recently highlighted that salt stress had an impact on iron homeostasis [59]. As already concluded for *B. subtilis*, also in *P. pastoris* the induction of iron uptake mechanisms might be of importance for growth in natural environments, where iron availability is generally scarce and may become even more limiting during growth implying high osmotic stress.

Conclusion

Although the central ESR pathways are well conserved among fungi, the up- and downstream elements can be significantly different among species to satisfy niche-specific requirements [60]. Most notably, the presented data demonstrate a very high similarity and/or cross-talk of the stress induced by recombinant protein production and the reaction to elevated osmolarity in *P. pastoris*. Growth at high osmolarity resulted in the induction of the response to unfolded proteins. Additionally ribosome biogenesis and translation processes were upregulated, whereas genes involved in cell wall synthesis were down-regulated at high osmolarity. Osmotic stress is a common condition for biotechnological production processes, due to high nutrient concentrations. In this light it is interesting to observe that *P. pastoris* is more osmo-tolerant than *S. cerevisiae*, and employs another main osmolyte, namely arabitol instead of glycerol, to compensate for osmotic stress.

The recombinant Fab 3H6 secreting *P. pastoris* strain was less prone to osmotic induced stress. Distinct differences, especially in the central carbon metabolism and processes linked to the UPR existed between the wt and the 3H6 Fab producing strain, which can be at least partially explained as response to unfolded protein is significantly induced in the Fab producing strain even at low osmolarity. Although in the current study elevated osmolarity did not result in increased productivity of recombinant Fab 3H6, the obtained data might be useful to explain the results of other research groups. It has been reported previously that osmotic stress applied prior to induction of protein secretion resulted in higher levels of scFv antibody in *P. pastoris* in batch culture [13]. Because osmotic stress obviously results in a UPR-like response in *P. pastoris*, it seems plausible that cells may be preconditioned for recombinant protein production as folding competence of the host cells may be increased compared to untreated cells. In this respect, the data obtained in the present study might be exploited not only for improved bioprocesses, but also for novel routes of strain engineering.

According to the data presented in this study, post-translational control mechanisms play an essential role in *P. pastoris*, especially during chemostat cultivation. Other proteomic methods such as the analysis of the phosphoproteome [61] might be very useful to gain detailed insight into these yet non-established mechanisms. However, the current data represent a first step towards a systems wide approach to assess the response to environmental stresses, as well as their overlap with recombinant protein induced stress, in *P. pastoris*.

Methods

Materials

All chemicals for yeast cultivations were molecular biology grade and were purchased from Roth, Germany. All chemical reagents for two-dimensional gel electrophoresis were high purity grade and were purchased from Sigma, unless stated otherwise.

Yeast Strains

Two strains, which have been described recently [26], have been used in this study. For secreting the Fab 3H6, both the light and the heavy chain of the Fab fragment were expressed under the control of the constitutive GAP-promoter using the pGAPZαA vector. Secretion was mediated by the *S. cerevisiae* α-mating factor secretion signal. For the non-expressing strain, *P. pastoris* X-33 was transformed with an empty pGAPZαA vector as described by Gasser and co-workers [62].

Chemostat cultivation

For chemostat cultivations a 3.5 L bench-top bioreactor (MBR, Switzerland) was used at a working volume of 1.5 L. A 1000 mL shake flask containing 150 mL YPG medium (2% (w/v) peptone, 1% (w/v) yeast extract, 1% (w/v) glycerol) was inoculated with 1 mL cryostock of the respective *P. pastoris* clones. The cultures were grown for approximately 24 h at 28°C and shaking at 170 rpm, before they were used to inoculate the bioreactor to an optical density (OD_{600}) of 1.0. After a batch phase of approximately 24 hours the continuous culture was started at a dilution rate of D = 0.1 h^{-1} (growth medium flow rate of 150 g h^{-1}). pH was controlled at 5.0 with 25% ammonium hydroxide (w/w). Gas flow rate was kept constant at 1.5 vvm (volume gas per volume medium and minute) and dissolved oxygen was kept at 20% by controlling the stirrer speed. Three chemostat media, with different osmolarities, were used.

Batch medium contained per liter: 39.9 g glycerol, 1.8 g citric acid, 12.6 g $(NH_4)_2HPO_4$, 0.022 g $CaCl_2 \cdot 2H_2O$, 0.9 g KCl, 0.5 g $MgSO_4 \cdot 7H_2O$, 2 mL Biotin (0.2 g L^{-1}), 4.6 mL trace salts stock solution. The pH was set to 5.0 with 25% (w/w) HCl. Osmolarity of the growth medium was controlled by KCl concentration. Chemostat medium contained per liter: 50 g glucose $\cdot 1H_2O$, 0.9 g citric acid, 4.35 g $(NH_4)_2HPO_4$, 0.01 g $CaCl_2 \cdot 2H_2O$, 1.7 (low) or 29.9 (medium) or 48.5 (high) g KCl, 0.65 g$MgSO_4$ $7H_2O$, 1 mL Biotin (0.2 g L^{-1}), and 1.6 mL trace salts stock solution. The pH was set to 5.0 with 25% (w/w) HCl. Trace salts stock solution contained per liter: 6.0 g $CuSO_4 \cdot 5H_2O$, 0.08 g NaI, 3.0 g $MnSO_4 \cdot H_2O$, 0.2 g $Na_2MoO_4 \cdot 2H_2O$, 0.02 g H_3BO_3, 0.5 g $CoCl_2$, 20.0 g $ZnCl_2$, 5.0 g $FeSO_4 \cdot 7H_2O$, and 5.0 mL H_2SO_4 (95-98% w/w).

Three chemostat cultivations were performed for each strain, whereas the osmolarity regime was different for each cultivation to avoid adaptive evolution effects and sample bias due to long term cultivation [63]. Samples were taken at steady state after 8 residence times after a switch of culture medium. Biomass was determined by drying duplicates of 10 mL chemostat culture to constant weight at 105°C in pre-weight beakers. Samples for 2D-DIGE and DNA microarray analysis were taken from the chemostat and immediately frozen at -80°C until use, whereat the samples for transcript analysis were fixed with 5% (v/v) phenol/ethanol prior to freezing. Viability of cells was determined immediately after samples were taken from the chemostat on a FACSCalibur flow cytometer (BD Biosciences) and a cell viability kit (BD Biosciences) as described previously [64].

Determination of culture supernatant osmolarity

To determine the actual osmolarity of the supernatant, supernatant samples were analyzed on a Semi-Microosmometer K-7400 (Knaur).

Analysis of intracellular polyols and trehalose

To quantify intracellular levels of glycerol, arabitol, mannitol, erythritol and trehalose, heat extraction was performed as described by Philips and co-workers [65]. Cell pellets were resuspended in 0.5 M TrisCl pH 7.5, heated to 95°C for 10 min and centrifuged for 10 min to remove cell debris. Supernatants were kept for analysis via HPLC. Isocratic conditions, using 4 mM H_2SO_4 as solvent and a flow rate of 0.6 mL min^{-1} on a Aminex HPX-87H column (Biorad) at 40°C and a Biologic DuoFlow (Biorad) combined with a Smartline RI Detector 2300 (Knauer) were applied to separate and analyze substances (Additional file 1). Concentrations were determined by external standard solutions. Solute concentrations were correlated with biomass.

2D Fluorescence Difference in Gel Electrophoresis (2D-DIGE) and protein identification

2D-DIGE and protein identification were essentially performed as described previously [26]. After adequate sample preparation, cleaning, quantification and Cy-dye labeling, proteins were separated on IPG DryStrips pH 3-11NL (GE Healthcare) on an IPGphor for a total of 65 kVh. 2nd dimension separation was performed by SDS polyacrylamide gel electrophoresis on 12% polyacrylamide gels. Fluorescence gel images were taken at a resolution of 100 μm on a Typhoon 9400 Fluorescence scanner. The DeCyder Software package v.5 (GE Healthcare) was used to analyze the obtained gel images. Significantly regulated protein spots (fold-change ≥ 1.5, 1-way ANOVA ≤ 0.05 in at least one comparison of cultivation conditions and present on at least 80% of the spot maps) were picked

from Coomassie stained gels and after a tryptic digest subjected to reversed phase capillary chromatography (BioBasic C18, 5 μ, 100 × 0.18 mm, Thermo) and ESI-MS/MS on a quadrupole time-of-flight (Q-TOF) Ultima Global (Waters Micromass) mass spectrometer. Mass spectra were analyzed either by using the Protein Lynx Global Server 2.1 software (Waters) or X!Tandem http://www.thegpm.org/tandem/. Only proteins identified by at least 2 peptides were considered to represent confident hits, except for Pdi1, which was verified by Western blotting [26].

DNA microarray analysis

DNA microarray analysis was performed using *P. pastoris* specific microarrays (Agilent) as described by Graf and co-workers [53]. RNA was extracted from ethanol/phenol fixed cell samples. Reverse transcription and synthesis of Cy3/5 labeled cRNA was done using the Low RNA Input Two-Color Amplification Kit (Agilent). cRNAs were purified via RNeasy Mini spin colums (Qiagen). Quality of total RNA and labeled cRNA was confirmed on an Agilent Bioanalyzer 2100 and the RNA Nano 6000 Assay Kit (Agilent). RNA concentrations were determined on a ND-1000 (Nanodrop). After hybridization at 65°C for 17 h, slides were scanned on an Agilent MicroArray scanner and raw data were extracted using Feature Extraction v.9.1 (Agilent). Normalization steps and statistical analysis of microarray data, including Hierarchical cluster analysis, Fisher's exact test and Gene Set Analysis (GSA), were done using the R software package http://www.r-project.org. For identifying differentially expressed genes, the False Discovery Rate was controlled strongly less than 5% ($q < 0.05$) using a Benjamini-Yekutieli correction for multiple testing. For Fisher's exact test and GSA, 63 gene ontology terms were considered. This list of terms was compiled based on the GOslim annotation of the *Saccharomyces* genome database http://www.yeastgenome.org, where some of the larger categories were resolved at a finer gene ontology level. A threshold of $p \leq 0.05$ was chosen to be appropriate to identify significantly regulated GO categories. Microarray data are available in the ArrayExpress database http://www.ebi.ac.uk/arrayexpress under the accession number E-MEXP-2433.

Real-Time PCR

To support microarray data, Real-time PCR was performed. Total RNA was reverse-transcribed using a Superscript III cDNA synthesis kit (Invitrogen). Quantity of cDNA was determined on a ND-1000 (Nanodrop). Real time PCR was performed using the SensiMix Plus PCR premix (GenXpress) on a Rotorgene 6000 (Corbett Life Sciences). The following target genes were selected for Real-time PCR analysis: *ACT1*, *AOX1*, *DGA1*, *GLG1*, *SIT1*, *PDI1*, *HAC1*, *3H6 Fab HC* and *3H6 Fab LC* (Additional file 4). Data were analyzed via the Rotorgene Software package and Microsoft Excel. *ACT1* was chosen as reference to determine relative mRNA levels of the other genes.

Growth tests on different salt concentrations

P. pastoris X-33 and *S. cerevisiae* HA232 http://www.biotec.boku.ac.at/acbr.html were grown in YPD medium (2% (w/v) peptone, 1% (w/v) yeast extract, 2% (w/v) glucose) at 28°C on a shaker at 170 rpm over night. Cultures were diluted to an OD of 0.1 in sterile PBS and 1:10 serially diluted in sterile PBS. 3 μL were spotted onto YPD agar plates (2% (w/v) peptone, 1% (w/v) yeast extract, 1% (w/v) agar, 2% (w/v) glucose) containing 0, 0.6, 1.2, 1.4 and 1.6 M NaCl or KCl. Plates were incubated at 28°C for 4 to 6 days.

3H6 Fab quantification

To analyze the 3H6 Fab produced during chemostat cultivation, a sandwich ELISA was performed as described in previous studies [62].

Additional material

Additional file 1 Determination of intracellular polyol and trehalose content in *P. pastoris* upon growth at different osmolarities. contains data on methodology of HPLC measurements, retention times of analytes and analyte concentrations with corresponding standard errors of the mean.

Additional file 2 Table of peptides of interesting proteins identified by 2D-DIGE and LC-ESI-MS/MS. contains a list of all peptides assigned to the proteins identified by 2D-DIGE and LC-ESI-MS/MS and described in the manuscript (Table 2). Additionally, scores and scoring schemes are indicated.

Additional file 3 Fold-change and one-way ANOVA data for all contrasts of the 2D-DIGE experiment. contains 2D-DIGE data of all comparisons (low to medium, low to high and medium to high) of both strains. Tables contain protein master numbers, short protein names, protein descriptions, fold-changes and corresponding one-way ANOVA values as described for Table 2.

Additional file 4 Real-time PCR results of *P. pastoris* grown at different osmolarities. contains detailed data on real-time PCR. Primers sequences, PCR conditions as well as result diagrams are included.

Additional file 5 Quality of microarray experiments and statistical test for osmolarity experiments. contains supplemental data on microarray analysis: Signal intensity plots, correlation of intensities, standard deviations and variations of the microarray experiment. Additionally, results of Hierarchical Cluster Analysis (HCA), Gene Set Analysis (GSA) and Fisher's exact test for the different osmolarities are included.

Additional file 6 Microarray results of interesting genes. contains gene expression results of interesting genes, which are described and discussed in the manuscript.

Additional file 7 Differences between the non-expressing wt strain and the Fab 3H6 expressing strain of *P. pastoris* at low osmolarity at the transcript level. contains microarray data on the gene expression differences between the two strains. Spreadsheet 1 contains microarray data for differentially regulated genes (p-value ≤ 0.001). Spreadsheet 2 contains results of gene set analysis (GSA).

Authors' contributions

MD performed chemostat cultivation, microarray analysis, Real-time PCR, 2D-DIGE, HPLC analysis and growth tests. JS performed protein identifications by LC-ESI-MS/MS. AG performed the statistical evaluation of the microarray exper-

iments. MM set up media recipes for cultivation and assisted in chemostat cultivations. BG, MS and DM contributed to the design of the study and data interpretation. DPK contributed to the design and advised on the analysis of the microarray experiments. FA supervised MS/MS analytics. MD, BG and DM drafted the manuscript. DM conceived of the study. All authors read and approved the final manuscript.

Acknowledgements
This work has been supported by the European Science Foundation (ESF, program EuroSCOPE), the Austrian Science Fund (FWF), project no. I37-B03 and the Austrian Resarch Promotion Agency (Program FHplus) and is part of the Genophys research project. DPK acknowledges funding by the the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres Seibersdorf, and the Austrian Centre of Biopharmaceutical Technology. Thanks to Martina Chang (Polymun Scientific) and Burghardt Scheibe (GE Healthcare) for their support and advice in 2D-DIGE, Hans Marx (School of Bioengineering, University of Applied Sciences FH-Campus Wien) for his advice in HPLC analytics and Astrid Mecklenbräuker and Corinna Rebnegger (Department of Biotechnology, University of Natural Resources and Applied Life Sciences) for their help concerning real-time PCR.

Author Details
[1]Department of Biotechnology, BOKU-University of Natural Resources and Applied Life Sciences, Vienna, Austria, [2]Department of Chemistry, BOKU-University of Natural Resources and Applied Life Sciences, Vienna, Austria, [3]School of Bioengineering, University of Applied Sciences FH-Campus Wien, Vienna, Austria and [4]Chair of Bioinformatics, BOKU-University of Natural Resources and Applied Life Sciences, Vienna, Austria

Received: 5 November 2009 Accepted: 26 March 2010
Published: 26 March 2010

References

1. Kim Y, Nandakumar M, Marten M: **Proteome map of Aspergillus nidulans during osmoadaptation.** *Fungal Genet Biol* 2007, **44(9):**886-895.
2. Mager W, Siderius M: **Novel insights into the osmotic stress response of yeast.** *FEMS Yeast Res* 2002, **2(3):**251-257.
3. Shen D, Sharfstein S: **Genome-wide analysis of the transcriptional response of murine hybridomas to osmotic shock.** *Biotechnol Bioeng* 2006, **93(1):**132-145.
4. Zeller G, Henz S, Widmer C, Sachsenberg T, Rätsch G, Weigel D, Laubinger S: **Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays.** *Plant J* 2009, **58(6):**1068-1082.
5. O'Rourke S, Herskowitz I: **Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis.** *Mol Biol Cell* 2004, **15(2):**532-542.
6. Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11(12):**4241-4257.
7. Norbeck J, Blomberg A: **Metabolic and regulatory changes associated with growth of Saccharomyces cerevisiae in 1.4 M NaCl. Evidence for osmotic induction of glycerol dissimilation via the dihydroxyacetone pathway.** *J Biol Chem* 1997, **272(9):**5544-5554.
8. Larsson K, Ansell R, Eriksson P, Adler L: **A gene encoding sn-glycerol 3-phosphate dehydrogenase (NAD+) complements an osmosensitive mutant of Saccharomyces cerevisiae.** *Mol Microbiol* 1993, **10(5):**1101-1111.
9. Lahav R, Nejidat A, Abeliovich A: **Alterations in protein synthesis and levels of heat shock 70 proteins in response to salt stress of the halotolerant yeast Rhodotorula mucilaginosa.** *Antonie Van Leeuwenhoek* 2004, **85(4):**259-269.
10. Kayingo G, Kilian S, Prior B: **Conservation and release of osmolytes by yeasts during hypo-osmotic stress.** *Arch Microbiol* 2001, **177(1):**29-35.
11. Yancey P: **Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses.** *J Exp Biol* 2005, **208(Pt 15):**2819-2830.
12. Olz R, Larsson K, Adler L, Gustafsson L: **Energy flux and osmoregulation of Saccharomyces cerevisiae grown in chemostats under NaCl stress.** *J Bacteriol* 1993, **175(8):**2205-2213.
13. Shi X, Karkut T, Chamankhah M, Alting-Mees M, Hemmingsen S, Hegedus D: **Optimal conditions for the expression of a single-chain antibody (scFv) gene in Pichia pastoris.** *Protein Expr Purif* 2003, **28(2):**321-330.
14. Blackwell J, Horgan R: **A novel strategy for production of a highly expressed recombinant protein in an active form.** *FEBS Lett* 1991, **295(1-3):**10-12.
15. Kim N, Lee G: **Response of recombinant Chinese hamster ovary cells to hyperosmotic pressure: effect of Bcl-2 overexpression.** *J Biotechnol* 2002, **95(3):**237-248.
16. Wu M, Dimopoulos G, Mantalaris A, Varley J: **The effect of hyperosmotic pressure on antibody production and gene expression in the GS-NS0 cell line.** *Biotechnol Appl Biochem* 2004, **40(Pt 1):**41-46.
17. Park S, Lee G: **Enhancement of monoclonal antibody production by immobilized hybridoma cell culture with hyperosmolar medium.** *Biotechnol Bioeng* 1995, **48(6):**699-705.
18. Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B: **Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris.** *Microb Cell Fact* 2009, **8:**29.
19. De Schutter K, Lin Y, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouzé P, Peer Y Van de, Callewaert N: **Genome sequence of the recombinant protein production host Pichia pastoris.** *Nat Biotechnol* 2009, **27(6):**561-566.
20. Gach J, Maurer M, Hahn R, Gasser B, Mattanovich D, Katinger H, Kunert R: **High level expression of a promising anti-idiotypic antibody fragment vaccine against HIV-1 in Pichia pastoris.** *J Biotechnol* 2007, **128(4):**735-746.
21. Gach J, Quendler H, Strobach S, Katinger H, Kunert R: **Structural analysis and in vivo administration of an anti-idiotypic antibody against mAb 2F5.** *Mol Immunol* 2008, **45(4):**1027-1034.
22. Mattanovich D, Gasser B, Hohenblum H, Sauer M: **Stress in recombinant protein producing yeasts.** *J Biotechnol* 2004, **113(1-3):**121-135.
23. Regenberg B, Grotkjaer T, Winther O, Fausbøll A, Akesson M, Bro C, Hansen L, Brunak S, Nielsen J: **Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae.** *Genome Biol* 2006, **7(11):**R107.
24. Gori K, Hébraud M, Chambon C, Mortensen H, Arneborg N, Jespersen L: **Proteomic changes in Debaryomyces hansenii upon exposure to NaCl stress.** *FEMS Yeast Res* 2007, **7(2):**293-303.
25. Blomberg A: **Global changes in protein synthesis during adaptation of the yeast Saccharomyces cerevisiae to 0.7 M NaCl.** *J Bacteriol* 1995, **177(12):**3563-3572.
26. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P, Mattanovich D: **The Effect of Temperature on the Proteome of Recombinant Pichia pastoris.** *J Proteome Res* 2009, **8(3):**1380-1392.
27. Zancan P, Sola-Penna M: **Trehalose and glycerol stabilize and renature yeast inorganic pyrophosphatase inactivated by very high temperatures.** *Arch Biochem Biophys* 2005, **444(1):**52-60.
28. Pascoe D, Arnott D, Papoutsakis E, Miller W, Andersen D: **Proteome analysis of antibody-producing CHO cell lines with different metabolic profiles.** *Biotechnol Bioeng* 2007, **98(2):**391-410.
29. Chang H, Jones E, Henry S: **Role of the unfolded protein response pathway in regulation of INO1 and in the sec14 bypass mechanism in Saccharomyces cerevisiae.** *Genetics* 2002, **162(1):**29-43.
30. Vaupotic T, Plemenitas A: **Differential gene expression and Hog1 interaction with osmoresponsive genes in the extremely halotolerant black yeast Hortaea werneckii.** *BMC Genomics* 2007, **8:**280.
31. Gat-Viks I, Shamir R: **Refinement and expansion of signaling pathways: the osmotic response network in yeast.** *Genome Res* 2007, **17(3):**358-367.
32. Hohenblum H, Gasser B, Maurer M, Borth N, Mattanovich D: **Effects of gene dosage, promoters, and substrates on unfolded protein stress of recombinant Pichia pastoris.** *Biotechnol Bioeng* 2004, **85(4):**367-375.
33. Ma Y, Hendershot L: **The unfolding tale of the unfolded protein response.** *Cell* 2001, **107(7):**827-830.
34. Valkonen M, Penttilä M, Saloheimo M: **Effects of inactivation and constitutive expression of the unfolded-protein response pathway on protein production in the yeast Saccharomyces cerevisiae.** *Appl Environ Microbiol* 2003, **69(4):**2065-2072.
35. Kauffman K, Pridgen E, Doyle Fr, Dhurjati P, Robinson A: **Decreased protein expression and intermittent recoveries in BiP levels result from**

cellular stress during heterologous protein expression in Saccharomyces cerevisiae. *Biotechnol Prog* 2002, **18(5):**942-950.
36. Tai S, Daran-Lapujade P, Luttik M, Walsh M, Diderich J, Krijger G, van Gulik W, Pronk J, Daran J: **Control of the glycolytic flux in Saccharomyces cerevisiae grown at low temperature: a multi-level analysis in anaerobic chemostat cultures.** *J Biol Chem* 2007, **282(14):**10243-10251.
37. Guillemette T, van Peij N, Goosen T, Lanthaler K, Robson G, Hondel C van den, Stam H, Archer D: **Genomic analysis of the secretion stress response in the enzyme-producing cell factory Aspergillus niger.** *BMC Genomics* 2007, **8:**158.
38. Payne T, Hanfrey C, Bishop A, Michael A, Avery S, Archer D: **Transcript-specific translational regulation in the unfolded protein response of Saccharomyces cerevisiae.** *FEBS Lett* 2008, **582(4):**503-509.
39. Neves M, Oliveira R, Lucas C: **Metabolic flux response to salt-induced stress in the halotolerant yeast Debaryomyces hansenii.** *Microbiology* 1997, **143(Pt 4):**1133-1139.
40. Carnicer M, Baumann K, Toplitz I, Sanchez-Ferrando F, Mattanovich D, Ferrer P, Albiol J: **Macromolecular and elemental composition analysis and extracellular metabolite balances of Pichia pastoris growing at different oxygen levels.** *Microb Cell Fact* 2009, **8(1):**65.
41. Simola M, Hänninen A, Stranius S, Makarow M: **Trehalose is required for conformational repair of heat-denatured proteins in the yeast endoplasmic reticulum but not for maintenance of membrane traffic functions after severe heat stress.** *Mol Microbiol* 2000, **37(1):**42-53.
42. Zähringer H, Burgert M, Holzer H, Nwaka S: **Neutral trehalase Nth1p of Saccharomyces cerevisiae encoded by the NTH1 gene is a multiple stress responsive protein.** *FEBS Lett* 1997, **412(3):**615-620.
43. Garre E, Pérez-Torrado R, Gimeno-Alcañiz J, Matallana E: **Acid trehalase is involved in intracellular trehalose mobilization during postdiauxic growth and severe saline stress in Saccharomyces cerevisiae.** *FEMS Yeast Res* 2009, **9(1):**52-62.
44. Ramón R, Ferrer P, Valero F: **Sorbitol co-feeding reduces metabolic burden caused by the overexpression of a Rhizopus oryzae lipase in Pichia pastoris.** *J Biotechnol* 2007, **130(1):**39-46.
45. Hohmann S: **Osmotic stress signaling and osmoadaptation in yeasts.** *Microbiol Mol Biol Rev* 2002, **66(2):**300-372.
46. Marciniak S, Ron D: **Endoplasmic reticulum stress signaling in disease.** *Physiol Rev* 2006, **86(4):**1133-1149.
47. Hetz C: **The UPR as a survival factor of cancer cells: More than folding proteins?** *Leuk Res* 2009, **33(7):**880-882.
48. Fonseca S, Fukuma M, Lipson K, Nguyen L, Allen J, Oka Y, Urano F: **WFS1 is a novel component of the unfolded protein response and maintains homeostasis of the endoplasmic reticulum in pancreatic beta-cells.** *J Biol Chem* 2005, **280(47):**39609-39615.
49. Mulhern M, Madson C, Danford A, Ikesugi K, Kador P, Shinohara T: **The unfolded protein response in lens epithelial cells from galactosemic rat lenses.** *Invest Ophthalmol Vis Sci* 2006, **47(9):**3951-3959.
50. Gonzalez N, Vázquez A, Ortiz Zuazaga H, Sen A, Olvera H, Peña de Ortiz S, Govind N: **Genome-wide expression profiling of the osmoadaptation response of Debaryomyces hansenii.** *Yeast* 2009, **26(2):**111-124.
51. Hirasawa T, Nakakura Y, Yoshikawa K, Ashitani K, Nagahisa K, Furusawa C, Katakura Y, Shimizu H, Shioya S: **Comparative analysis of transcriptional responses to saline stress in the laboratory and brewing strains of Saccharomyces cerevisiae with DNA microarray.** *Appl Microbiol Biotechnol* 2006, **70(3):**346-357.
52. Farewell A, Neidhardt F: **Effect of Temperature on In Vivo Protein Synthetic Capacity in Escherichia coli.** *J Bacteriology* 1998, **180:**4707-4710.
53. Graf A, Gasser B, Dragosits M, Sauer M, Leparc G, Tüchler T, Kreil D, Mattanovich D: **Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.** *BMC Genomics* 2008, **9:**390.
54. Gueta-Dahan Y, Yaniv Z, Zilinskas B, Ben-Hayyim G: **Salt and oxidative stress: similar and specific responses and their relation to salt tolerance in citrus.** *Planta* 1997, **203(4):**460-469.
55. García-Rodríguez L, Valle R, Durán A, Roncero C: **Cell integrity signaling activation in response to hyperosmotic shock in yeast.** *FEBS Lett* 2005, **579(27):**6186-6190.
56. Cyert M: **Calcineurin signaling in Saccharomyces cerevisiae: how yeast go crazy in response to stress.** *Biochem Biophys Res Commun* 2003, **311(4):**1143-1150.
57. Zhao L, Schaefer D, Xu H, Modi S, LaCourse W, Marten M: **Elastic properties of the cell wall of Aspergillus nidulans studied with atomic force microscopy.** *Biotechnol Prog* 2008, **21(1):**292-299.
58. Turchini A, Ferrario L, Popolo L: **Increase of external osmolarity reduces morphogenetic defects and accumulation of chitin in a gas1 mutant of Saccharomyces cerevisiae.** *J Bacteriol* 2000, **182(4):**1167-1171.
59. Hoffmann T, Schütz A, Brosius M, Völker A, Völker U, Bremer E: **High-salinity-induced iron limitation in Bacillus subtilis.** *J Bacteriol* 2002, **184(3):**718-727.
60. Nikolaou E, Agrafioti I, Stumpf M, Quinn J, Stansfield I, Brown A: **Phylogenetic diversity of stress signalling pathways in fungi.** *BMC Evol Biol* 2009, **9:**44.
61. Stasyk T, Morandell S, Bakry R, Feuerstein I, Huck C, Stecher G, Bonn G, Huber L: **Quantitative detection of phosphoproteins by combination of two-dimensional difference gel electrophoresis and phosphospecific fluorescent staining.** *Electrophoresis* 2005, **26(14):**2850-2854.
62. Gasser B, Maurer M, Gach J, Kunert R, Mattanovich D: **Engineering of Pichia pastoris for improved production of antibody fragments.** *Biotechnol Bioeng* 2006, **94(2):**353-361.
63. Ferea T, Botstein D, Brown P, Rosenzweig R: **Systematic changes in gene expression patterns following adaptive evolution in yeast.** *Proc Natl Acad Sci USA* 1999, **96(17):**9721-9726.
64. Hohenblum H, Borth N, Mattanovich D: **Assessing viability and cell-associated product of recombinant protein producing Pichia pastoris with flow cytometry.** *J Biotechnol* 2003, **102(3):**281-290.
65. Philips J, Herskowitz I: **Osmotic balance regulates cell fusion during mating in Saccharomyces cerevisiae.** *J Cell Biol* 1997, **5:**961-974.

doi: 10.1186/1471-2164-11-207
Cite this article as: Dragosits *et al.*, The response to unfolded protein is involved in osmotolerance of Pichia pastoris *BMC Genomics* 2010, **11**:207

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

A multi-level study of recombinant *Pichia pastoris* in different oxygen conditions as knowledge base for strain improvement

Kristin Baumann*[1], Marc Carnicer [*1], Martin Dragosits[2], Alexandra Bettina Graf[2, 3], Johannes Stadlmann[4], Paula Jouhten[5], Hannu Maaheimo[5], Brigitte Gasser[2], Joan Albiol[1], Diethard Mattanovich[2, 3], and Pau Ferrer[1§]

[1]Department of Chemical Engineering, Autonomous University of Barcelona, Spain

[2]Institute of Applied Microbiology, Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Austria

[3]School of Bioengineering, University of Applied Sciences, FH Campus Vienna, Austria

[4]Department of Chemistry, University of Natural Resources and Applied Life Sciences, Vienna, Austria

[5]VTT Technical Research Centre of Finland, Espoo, Finland

* Both authors contributed equally to this work

[§] Corresponding author: pau.ferrer@uab.cat

Email addresses:

KB: kristin.baumann@uab.cat

MC: marc.carnicer@uab.cat

MD: martin.dragosits@boku.ac.at

ABG: alexandra.graf@boku.ac.at

JS: johannes.stadlmann@boku.ac.at

PJ: paula.jouhten@vtt.fi

HM: hannu.maaheimo@vtt.fi

BG: brigitte.gasser@boku.ac.at

JA: joan.albiol@uab.cat

DM: diethard.mattanovich@boku.ac.at

PF: pau.ferrer@uab.cat

Abstract

Background

The yeast *Pichia pastoris* is a powerful protein production system and is becoming increasingly attractive for basic research and white biotechnology as well. While many recombinant proteins have already been successfully produced in this host, understanding the mechanisms leading to efficient production of complex proteins still remains a major challenge for researchers. A number of studies have demonstrated that correct protein folding and secretion are highly interrelated with environmental stresses, in particular with oxygen availability. While adaptive responses to such stresses are extensively studied in yeast, little is known about their impact on heterologous protein production.

We could recently show a significant increase in the specific productivity of an antibody Fab fragment produced in *P. pastoris* chemostat cultivations when shifting from respiratory to hypoxic growth conditions. As a consequence, a systems biology approach was used to comprehensively identify cellular responses that lead to increased product formation at low oxygen concentration. Gene expression profiling was combined with differential proteome data, and the experimental determination of steady-state fluxes in the central carbon metabolism.

Results

There was a positive correlation between all three levels of analysis in the central carbon metabolism, showing a considerable increase of the glycolytic activity, a lower contribution of the pentose phosphate pathway to glucose catabolism, and a reduced TCA cycle activity under hypoxic conditions. Transcriptome data further demonstrated important changes in lipid metabolism and membrane biogenesis, as well as a significantly different expression pattern of genes involved in protein folding and trafficking, stress response and transporter systems. In general, our analyses indicate a strong oxygen-dependent expression pattern rather than a protein-production related pattern.

Conclusions

This systems approach helps to understand the cellular mechanisms that lead to increased product formation at low oxygen and might reveal hitherto unidentified key factors or major pacemakers of efficient protein production. In this context, strain engineering studies based on the generated knowledge are under way.

Background

Over the last two decades significant progress has been made in heterologous protein production, particularly due to the initiation of the genomics era. The acquisition of profound knowledge and entire genome sequences of a number of expression platforms has lead to considerable success for the production of many pharmaceutical proteins or industrial enzymes [1-3]. Nevertheless, understanding the mechanisms governing efficient production of very complex proteins as functional entities still remains a major challenge.

Over the past recent years, it has been demonstrated that correct protein folding and secretion are highly interrelated with environmental stress factors [4]. While adaptive responses to such stresses have been studied in the model yeast *Saccharomyces*

cerevisiae [5, 6] little is known about their influence on heterologous protein production. Recently, Dragosits and co-workers [7] described the impact of temperature on the proteome of recombinant *Pichia pastoris* in a chemostat-based study. Their data indicated that a decreased folding stress at lower cultivation temperatures (20ºC) considerably favoured heterologous Fab antibody production. Some other studies described a similar effect of temperature on the yield of recombinant proteins in other hosts [8, 9], but without much information on the underlying physiological reaction of cells, particularly in context of heterologous protein production. Even less is known about the impact of oxygen availability. Oxygen transfer is often described as an important issue in high cell density fermentations. While affecting growth and protein production, oxygen also influences cellular redox reactions, and these are interlinked with protein folding reactions within the cell. Moreover, protein folding related oxidative stress has been described [10-12]. Paradoxically, we have demonstrated recently that hypoxic conditions in chemostat as well as fed batch cultures significantly increased the specific productivity of recombinant *P. pastoris* [13]. This yeast expression system has become increasingly popular as it represents a valuable and cost-effective tool for protein engineering studies with potential of performing many of the posttranslational modifications typically associated with higher eukaryotes (reviewed in [14]). However, it is not straightforward to predict whether this “hypoxic effect” would be also observed in other expression systems like *S. cerevisiae.* Yeasts do not perform in an identical way, but differ in productivity and with regard to their capacity to secrete, to process and to modify proteins in particular cases. As a consequence, it is important to systematically identify the complex mechanisms ruling efficient protein production, and integrated ‘omics’ studies are a valuable tool for the study of biological processes. An

appropriate cultivation system that allows for strictly controlled environmental conditions while changing one single parameter, like it is the case for chemostat cultivations, provides an excellent basis for such a systems biology approach. Of particular importance is the maintenance of a constant specific growth rate in order to avoid growth rate related effects in the data [15].

In the present study, an integrative multilevel analysis of cellular processes involved in the expression of an antibody Fab fragment in *P. pastoris* has been performed in carbon chemostat cultivations with a fixed growth rate at different oxygenation rates. This study is aimed to understand further the global mechanisms connecting protein production to environmental conditions, in particular those that lead to increased product formation under hypoxia, as reported previously [13]. Despite the availability of the strong inducible alcohol oxidase promoter (AOX) that is mostly used in this methylotrophic yeast, we opted for a constitutive expression with the glycolytic glyceraldehyde-3-phosphate dehydrogenase promoter (GAP). It does not only facilitate easier handling of a continuous cultivation, but also benefits the viability, and the reduction of protease release, heat production and oxygen demand [16]. Although transcriptional profiling, proteomics and metabolic flux analysis of our samples indicated potential bottlenecks for heterologous protein production, more detailed analysis will be necessary to determine whether host cell physiology can be improved based on these data.

Results and Discussion

In order to study the global adaptive response of recombinant *Pichia pastoris* to oxygen availability, we integrated transcriptome, proteome and metabolic flux data of cells grown in steady state cultures under normoxic, oxygen limited and hypoxic conditions. Special emphasis was given to the comparison between fully aerobic

(normoxic) and hypoxic conditions, given that this shift contributed to increased recombinant product secretion [13].

Global transcriptional response of recombinant *Pichia pastoris* to hypoxia

The global transcriptional profile of *P. pastoris* grown in different conditions of oxygen availability was studied with samples from three individual chemostat cultivations each. At a first glance, the microarray statistics (see Table 1) of the approximately 3900 annotated sequences for *P. pastoris* displayed more than 600 genes, independent of the strain, that were differently regulated (adjusted p-value $\leq$ 0.05) when comparing normoxic and hypoxic conditions. A fold-change (FC) threshold was not included in this statistics given that also small changes in the gene expression were considered to be crucial for a global view (*e.g.*, for GO term clustering).

As illustrated in the Venn diagrams (Figure 1), only around fifty percent of these genes were identically regulated in both the Fab expression and the control strain, indicating a different behaviour under hypoxic conditions. A direct comparison of the gene regulation pattern between expressing and control strain at the lowest oxygen concentration (8 %), in contrast, revealed only six genes (see Table 2) to be differently expressed (adjusted p-value ≤ 0.05). For all of these six genes the FC was also greater than 1.5. Four of them showed higher transcript levels in the Fab producing strain, specifically, genes involved in a non-classical protein export pathway (*NCE103*), glycolysis (*PFK3*), the ergosterol pathway (*ERG25*) and in multidrug transport (*AQR1*). On the other hand, a high-affinity cysteine-specific transporter (*YCT1*) and a plasma membrane transporter of the major facilitator super family (*FLR1*) were significantly down regulated in the producing strain. In general, these data point at a major impact of oxygen rather than a consequence of heterologous protein production on the global transcriptional response of *P. pastoris*.

To gain an overview of the functional processes that are significantly correlated with a change in oxygen availability, we assigned all regulated genes that passed the p-value threshold to their respective GO functional group(s) (Gene Ontology GO based on Saccharomyces Genome Database SGD). The percentage distribution of the genes in each category is shown in Figure 2A and 2B. The most prominent biological processes that were exclusively induced under hypoxic conditions are *chemical stimulus*, *cell wall biogenesis, heterocycle metabolism, protein folding* and *cellular aromatic compounds*. The regulated genes in the GO groups *cofactor- and carbohydrate metabolic process*, *lipid metabolic process* and *vitamin metabolic process* showed very similar profiles of up and downregulation at the same time. On the other hand, hypoxia decreased the activity of genes involved in *conjugation, sporulation* and *pseudohyphal growth.*

The global tendencies of regulated biological processes overlap quite concordantly in both strains, with some exceptions: interestingly, the number of upregulated genes in the GO groups *RNA metabolism, protein catabolism*, *ribosome biogenesis*, *transcription* and *DNA metabolism* were much higher in the control strain. In contrast, processes like *cellular respiration, carbohydrate metabolism, cellular homeostasis and amino acid metabolism,* showed a higher abundance of downregulated genes in the control strain. On the other hand, four GO categories appeared to have a greater enrichment of downregulated genes in the producing strain, namely *cell cycle, meiosis*, *cytoskeleton biogenesis* and *nuclear biogenesis*.

It has to be emphasized that these results were obtained excluding the fold change threshold to provide a more informative transcriptional profile. Thereafter, two selected GO groups were analyzed in more detail in such a way that all the genes that

were significantly up- or downregulated under hypoxia by at least 1.5-fold were categorized according to their functions (as discussed below). The selection of these groups was either based on major differences between oxygen supply conditions or on the potential impact on protein production, as some of them appeared to be generally stronger regulated in the Fab producing strain.

The profiles of the Fab mRNA were confirmed by quantitative real-time PCR (see Table 4). Both of the encoding genes for the light chain (Fab Lc) and the heavy chain (Fab Hc) demonstrated higher transcripts under hypoxic conditions, which positively correlates with an increase in the specific Fab titre. The increased expression level of Fab under control of the *TDH3* promoter was coherent with a transcriptional upregulation of glycolytic genes (and, particularly, of *TDH3*), as discussed later in the text.

Proteome data of hypoxically grown *Pichia pastoris*

In parallel to the microarray experiments, we employed 2D-DIGE gels to measure the relative protein abundance changes at different levels of oxygen availability. To assure accuracy of the measurements and unbiased data, we labelled the protein extracts with both Cy3 and Cy5 fluorescent dyes and assigned the samples randomly to the protein gels according to the scheme in Additional file 1. We detected 85 spots with a significantly (1-way ANOVA p-value ≤ 0.05) different abundance pattern comparing the expression under normoxic and hypoxic conditions, and a smaller number of spots when comparing the proximate oxygen set points with each other (normoxic vs. oxygen limited, oxygen limited vs. hypoxic). A total of 45 spots could be excised from Coomassie Brilliant Blue stained protein gels and identified by LC-ESI-QTOF Tandem MS (see Additional file 2 for a representative 2D gel image and Additional file 3 for the list of identified protein spots). Some proteins showed more

than one spot on the 2D gels indicating the existence of isoforms which probably derive from posttranslational modification (PTM) events such as phosphorylation, glycosylation or limited proteolysis (see, for example [17]). In order to obtain a simplified structure of the behaviour of all identified proteins under different conditions of oxygen availability, we subjected the relative protein abundances (see Materials and Methods) to principal component analysis (PCA) and heat map clustering. PCA projection demonstrates that the maximum variability in the dataset occurs between oxygen set points 21 % and 8 % (see Figure 3A) with the first component covering 65.8 % of the data variance. This result is also reflected in the heat map, where two major clusters separating the protein abundance profile at normal oxygen levels from that at limiting and hypoxic levels can be observed (see Figure 3B). All the identified proteins exhibited a similar expression profile when comparing the Fab producing strain and the non producing control strain.

Most of the proteins with a high abundance at low oxygen are involved in *glycolysis*, *amino acid metabolism* and *general stress response*. Identified proteins with a low expression in hypoxic conditions mostly belong to the functional processes *TCA cycle*, *vitamin metabolism* and *oxidative stress response*.

When oxygen is scarce, cells have to re-adjust their metabolic fluxes to meet energy demands. They are no longer able to produce ATP through oxidative phosphorylation or from other reactions like fatty acid oxidation and hence boost glycolysis to maintain the energy balance. The glycolytic enzymes Pgi1p, Fba1p, Tdh3p, Gpm1p, Eno1p and Cdc19p were identified to be strongly induced, while TCA cycle proteins Aco1p, Fum1p and Mdh1p show very low abundance in hypoxic conditions, pointing at a redirection of the central carbon metabolism towards fermentation. These results were expected and are consistent with previously obtained data on yeast physiology

under oxygen deprivation, where glycolytic activity in anaerobically grown *S. cerevisiae* was shown to be higher than in aerobiosis on the proteome level [18, 19]. These studies, however, revealed a weak correlation between transcriptome and proteome data of this key cellular process in *S. cerevisiae*, which is in strong contrast to our results where the transcript levels (see also transcriptome data) indeed correlate with the proteome profile and *in vivo* fluxes of the central carbon metabolism, as discussed below.

Along with an adaptation towards a fermentative metabolism cells have to remove excess redox equivalents that accumulate during biomass synthesis and excretion of oxidized metabolites [20]. Anaerobically grown *S. cerevisiae* produces glycerol in order to reoxidize accumulated NADH during amino acid biosynthesis [21]. In contrast, we have recently described that hypoxically growing *P. pastoris* cells secrete arabitol, a 5-carbon sugar alcohol, but not glycerol [22]. Arabitol (and glycerol) accumulation has been previously observed in *Pichia anomala* cultures during growth in high-salt environments and on highly concentrated sugar substrates [23, 24]. Dragosits and co-workers recently demonstrated intracellular arabitol accumulation in *P. pastoris* chemostat cultivations grown under conditions of high osmolarity [25]. In fact, it has been suggested that arabitol has the same physiological role as glycerol in the protection to osmotic stress [26]. Similarly, arabitol might also be involved in maintaining the redox balance during fermentative growth.

There is no identified protein in our data clearly assigned to arabitol biosynthesis. D-arabitol 2-dehydrogenase, the only protein that is described to form arabitol from ribulose in *P. pastoris,* could not be detected in our protein gels. However, we speculate that Ydl124wp, identified as a putative NADPH-dependent alpha-keto

amide reductase and described to have similarity with Gre3p, the major aldose reductase in *S. cerevisiae* [27, 28], may be involved in the formation of the 5-carbon sugar alcohol. Since D-ribulose and D-xylulose from the pentose pathway are the main precursors for the formation of arabitol in many fungi [29, 30] it may be assumed that Ydl124w is involved in these reductive activities.

Gut2p, a mitochondrial enzyme that is associated with redox balance maintenance via the glycerophosphate shuttle under aerobic conditions, showed decreased abundance under hypoxia, as it is not needed under this condition to transfer reducing equivalents to the respiratory chain. The same result was obtained for Gut1p, which together with Gut2p is responsible for glycerol degradation. Fdh1p, a NAD^+-dependent formate dehydrogenase, is also downregulated when oxygen is scarce. Singh and co-workers described a possible relation between pyruvate break-down and the production of formate [31]. Since the pyruvate pathway is directed towards ethanol formation under hypoxic conditions, the synthesis of formate might be reduced and could explain the lower abundance of formate dehydrogenase.

Although the methanol pathway is tightly repressed under growth on glucose, the basic level of Aox1p was strictly downregulated under hypoxic conditions as has been shown previously in glucose limited conditions [7, 25]. A plausible explanation for the low abundance of Aox1p is the higher glycolytic activity under low oxygen conditions and the repressive effect of glucose on *AOX1* transcription. This effect could also be responsible for the weak expression of the proteins Ald4p, a mitochondrial aldehyde dehydrogenase, and Acs1p, an acetyl-CoA-synthetase.

Rpn10p is a regulatory particle of the proteasome and involved in the degradation of ubiquitinated proteins. It was also shown to have low abundance under hypoxic conditions. The same trend was demonstrated for 3 stress-related proteins: Ccp1p and

Prx1p are involved in oxidative stress responses, while Ypr127wp is an uncharacterized protein that shows similarity to the *Schizosaccharomyces pombe* pyridoxal reductase [32].

Other stress-associated proteins were identified to be induced upon a shift to hypoxic conditions as it was the case for Pil1p, Tsa1p and Ssa4p. While Pil1p is a component of the eisosomes that mark endocytic sites in the plasma membrane, Tsa1p and Ssa4p are molecular chaperones. It is worth mentioning that Tsa1p has dual activities [33]. It acts either as ribosome-associated or as free cytoplasmic antioxidant and only self-associates to form a high-molecular weight chaperone complex under oxidative stress. In *Candida albicans* it was further demonstrated that the cell membrane was highly altered in a *tsa1* deletion strain [34]. Another interesting protein with a high abundance at low oxygen conditions was Sam2p, an S-adenosylmethionine synthetase with a broad range of biological functions. Besides its role in the methylation of proteins and lipids, S-adenosylmethionine (AdoMet) is also involved in the synthesis of polyamines and biotin. It was further attributed a role in the synthesis of phospholipids [35], which is also the case for Ino1p, an Inositol 1-phosphate synthase. Pdx3p (sterol uptake regulation), Thi13p (pyrimidine synthesis), Cys4p (cysteine biosynthesis) and the key enzyme in actively fermenting yeast cells, Pdc1p, were also induced by hypoxia. The abundances of the elongation factor 2, which is encoded by *EFT2* and catalyzing ribosomal translocation during protein synthesis, also increased significantly in low oxygen concentrations.

Impact of oxygen availability on particular cellular processes: Central carbon metabolism

The macroscopic growth parameters for both control and Fab-producing *P. pastoris* strains growing under oxygen excess, oxygen-limiting and hypoxic conditions have been recently reported elsewhere [13, 22]. As discussed above, the impact of reduced oxygen

supply on the core metabolism was readily observed, both in the biomass yields and the profile of secreted by-products such as ethanol and arabitol, reflecting the adaptation from a respiratory to a respiro-fermentative metabolism. Since all cultivations were carbon-limited (residual glucose concentration in the reactor under detecting concentrations), the decrease in the biomass yield decreased under oxygen limiting and hypoxic conditions resulted in an increase of specific glucose uptake rates under such conditions.

Biosynthetically directed fractional (BDF) ^{13}C-labeling of proteinogenic amino acids combined with 2D-NMR enabled the analysis of metabolic flux ratios (METAFoR analysis). The metabolic flux ratios were calculated using the relative abundances (*f*-values) of intact carbon fragments arising from a single source molecule of glucose (Additional file 5). The calculated flux ratios are shown in Table 3. As expected, the *f*-values obtained for this series of cultivations confirm that the proteinogenic amino acids are primarily synthesized in *P. pastoris* according to the pathways documented for *S. cerevisiae,* as previously reported [36].

In ^{13}C-based metabolic flux analyses (^{13}C-MFA), the metabolic flux ratios determined by METAFoR were used as additional constraints for the stoichiometric equation system to be able to solve the metabolic flux distribution without including cofactors (NADH, NADPH and ATP), O_2 and CO_2 in the metabolite mass balances. The net fluxes for the Fab-producing and control strains growing under different oxygenation conditions are shown in Figure 4. The net fluxes and their standard deviations are included in Additional file 4.

The most prominent feature, as already indicated by the METAFoR analysis (Table 3), was the similarity in flux estimates between corresponding Fab-producing and control strains datasets. Nevertheless, as expected, clear differences were observed when

comparing flux patterns corresponding to different oxygenation set points. In general terms, the metabolic adaptation from oxidative towards respiro-fermentative growth was accompanied by complex changes of carbon flux throughout the whole central carbon metabolism, as previously described in other yeasts (*S. cerevisiae* [37, 38], *P. anomala* [39, 40]).

Glycolytic and Pentose Phosphate Pathway (PPP) fluxes

The METAFoR analysis showed that in fully aerobic conditions up to 50–39 % of phosphoenolpyruvate (Pep) was originated from the pentose phosphate pool. In contrast, the fraction of Pep from the pentose phosphates assuming a maximal contribution of PPP was clearly lower under hypoxic conditions, only about 15 %.

Moreover, the ^{13}C-MFA results (shown in Figure 4) indicate that this decrease of the relative PPP flux was the result of both an increased glycolytic flux, and a decrease in the specific flux through the oxidative branch of the PPP. As previously observed in *S. cerevisiae* [38], the glycolytic flux increased progressively as the oxygen availability decreased.

As already inferred from the macroscopic data, fluxes through some fermentative pathways were increased upon adaptation from respirative to respiro-fermentative metabolism, particularly the fluxes towards the formation of ethanol and arabitol. Production of arabitol had a clear impact on the flux ratios and on the distribution of fluxes through the PPP: The fraction of pentose phosphates showing the reversible action of a transketolase reaction was generally high (> 60 % in all cultivations), with no clear trend, whereas the fraction of pentose phosphates showing the reversible action of a transaldolase clearly decreased at low oxygen availability. Overall, the ^{13}C-MFA results showed that under normoxic conditions there was an important net contribution of the PPP to glucose catabolism (flux of PPP intermediates to triose phosphates). In contrast,

as oxygen availability was decreased and, particularly, when arabitol was produced, this contribution was clearly reduced or, even some of the reactions of the non-oxidative PPP branch showed inverted directionality under hypoxic conditions.

Fluxes around the pyruvate node, intercompartmental transport and Tricarboxylic Acids (TCA) cycle

Following the METAFoR analysis, distinct flux changes were observed for the different pathways utilizing pyruvate (Pyr), that are, pyruvate carboxylase (anaplerosis), pyruvate decarboxylase (fermentative pathways, pyruvate dehydrogenase by-pass), and pyruvate dehydrogenase (direct import of Pyr to the mitochondria). The relative anaplerotic flux (the anaplerotic flux ratio defined as the fraction of mitochondrial oxalocetate Oaa_{mit} molecules originating from Pep) was around 44–41 % under normoxic and hypoxic conditions, while in oxygen-limiting conditions was slightly lower, 35–32 %. Hence, pyruvate carboxylase seemed to be the major anaplerotic reaction under all conditions, and it was accompanied by a substantial transport of carbon (cytosolic oxalocetate Oaa_{cyt} and/or other TCA cycle intermediates) from cytosol to mitochondria. Nevertheless, an important relative carbon efflux from the mitochondria may be occurring, as only 66–63 % of Oaa_{cyt} appears to be directly synthesised from Pep under oxygen-limiting and hypoxic conditions. As previously observed for *P. pastoris* [36] and other yeasts (e.g. *S. cerevisiae* and *P. stipitis*, [38, 41]), cells growing aerobically in glucose-limited chemostats show a bidirectional transport of Oaa and/or other TCA cycle intermediates across the mitochondrial membrane. However, calculation of flux ratios defining the fraction of Oaa_{mit} from Oaa_{cyt} and, Oaa_{cyt} from Pep under normoxic conditions was not possible. The labelling patterns of cytosolic and mitochondrial Oaa pools are accessible through the observation of aspartate labelling patterns (shown to be synthesised from Oaa_{cyt} in yeast in previous studies [42]) and glutamate (synthesised from mitochondrial 2-oxoglutarate, and therefore accessing to Oaa_{mit} labelling patterns). Strikingly, the

fraction of intact C2-C3 bonds of Oaa, which is often used to calculate these flux ratios, where approximately equal for Oaa_{cyt} and Oaa_{mit} (that is, the fraction of intact Cα-Cβ bonds in Asp/Thr and Glu were equal as revealed by the *f*-values of Asp, Thr and Glu, Additional file 5). This could be explained by an extremely fast exchange between cytosolic and mitochondrial pools of TCA cycle intermediates (near equilibrium), resulting in identical labelling patterns in the amino acids synthesised from such pools. However, this possibility can be excluded, as the Asp-Cβ, Thr-Cβ and Glu-Cα labelling patterns under normoxic conditions were not identical (Additional file 5). Notably, the reversibility of the interconversion of cytosolic Oaa to other cytosolic TCA cycle intermediates (defined here as the Oaa_{cyt} interconversion to fumarate ratio) was clearly higher under normoxic conditions, suggesting that aspartate, Oaa, and malate might be participating in a redox shuttle (e.g. malate-Asp and/or malatee-Oaa shuttles) for translocation of NADH across the mitochondrial membrane, as described by Bakker et al [43].

It is worth noting that, although the flux through the PDH bypass was not considered in our metabolic model (see Materials and Methods for explanation), its activity should not be totally excluded since Crabtree negative yeasts are reported to have activity on this pathway [44]: In contrast to *S. cerevisiae*, which does not synthesise carnitine *de novo* (essential for carnitine acetyltransferase-mediated transport of cytosolic AcCoA to the mitochondria [45]), a complete carnitine biosynthesis pathway has been characterised in *Candida albicans*, and the corresponding 4 genes have been identified [46]. Interestingly, the *P. pastoris* genome contains putative homologues to these genes [16, 47]. Moreover, it should be mentioned that a 20-fold change in relative expression levels of the *S. cerevisiae ACS1* homolog encoding the mitochondrial AcCoA synthetase essential for the contribution of mitochondrial PDH

bypass to the formation of mitochondrial AcCoA, was observed when comparing normoxic *vs.* hypoxic conditions. Also, a significant downregulation in expression levels of *CAT2* and *YAT2* homologues involved in carnitine transport to mitochondria was observed.

Overall, carbon flux distributions at the pyruvate branching point (see Figure 5) clearly show the shift from respiratory to respiro-fermentative metabolism: fluxes through the pyruvate dehydrogenase pathway decreased when decreasing oxygen availability, whereas the flux through the pyruvate decarboxylase pathway increased, reflecting the production of ethanol. Also, the anaplerotic flux though the pyruvate carboxylase pathway drastically decreased when oxygen availability was reduced: Under fully aerobic conditions the fraction of carbon flux to the TCA cycle through this pathway was about 29 % (calculated only as a net transfer of Oaa_{cyt} across the mitochondrial membrane), whereas under oxygen-limiting and hypoxic conditions exchange fluxes decreased, resulting in lower net carbon fluxes into the TCA cycle (16–18 % and 12 % in oxygen-limiting and hypoxic conditions, respectively).

Remarkably, although variations in the carbon flux distribution around the pyruvate branch and TCA cycle activity were observed, mitochondrial transporters such as *DIC1*, *OAC1*, *SFC1* showed no significant change on the transcriptional level. Only *YIA6*, which is involved in NAD^+ transport into the mitochondria (and has a disputed role as Pyr transporter), was down regulated under hypoxic conditions. As already mentioned before, the carnitine transporters *CAT2* and *YAT2* were also down regulated under such conditions.

Limitation in oxygen availability reduced the respirative net carbon flux through the TCA cycle (that is, the net flux of α-ketoglutarate through the TCA cycle to Oaa_{mit}, Figure 4). Nevertheless, the fraction of the net carbon flux in the TCA cycle

corresponding to the respirative carbon flux from α-ketoglutarate (or relative TCA cycle activity) remained between 60–70 %, that is to say, the relative anaplerotic fluxes were between 30–40 %. Similar results have been previously observed in *S. cerevisiae* under aerobic conditions [38]. No significant contribution for malic enzyme flux could be observed (Table 3), so the pyruvate carboxylase pathway was the only anaplerotic supply to the TCA cycle.

As discussed above, a good qualitative correlation between both transcriptional and proteomic levels and, the corresponding *in vivo* fluxes in the different oxygenation conditions was observed for the glycolysis, fermentative pathways and in the TCA cycle (Figure 4). Notably, transcriptional levels of *TDH3* under hypoxic conditions were 2.7 fold higher in the Fab-producing strain than in the control strain. *TDH3* codes for glyceraldehyde-3-P dehydrogenase, an enzyme playing an integral role in glycolysis. Interestingly, Tdh3p has also been recognized to be involved in the initiation of apoptosis in *S. cerevisiae* [48]. Furthermore, it has been also found in the yeast cell wall [49], suggesting that *TDH3* may indeed code for a multifunctional protein.

Exceptions were *ZWF1* and *GND2*, coding for enzymes involved in the oxidative branch of the PPP, which did not appear to be significantly regulated, in spite of the strong variation observed in the metabolic fluxes through this pathway. This was also the case for the *GUT1* and *GUT2* genes, coding for enzymes involved in glycerol formation, which were down regulated, whereas the flux to glycerol excretion was very low or even zero under all oxygenation conditions.

Other cellular processes: Lipid metabolism and membrane biogenesis

Lipid metabolism has successively become a focus of attention for linking many important pathways to its intermediate substrates. Sterols for example, long time relegated as passive metabolites modulating the membrane structure, play an

important role as primary sensors of environmental stresses by adjusting the fluidity of the plasma membrane to such perturbations [50, 51]. The cellular lipid metabolism is highly susceptible to hypoxia since the biosynthesis routes for the most essential lipids - ergosterol and fatty acids – require molecular oxygen. Anaerobically grown *S. cerevisiae* has an impaired ability to produce ergosterol, which consequently has to be added to the medium together with Tween80 as a source for unsaturated fatty acids [52, 53]. In this study the medium was not supplemented with either of these reagents since our chemostat cultures were run under severe oxygen limitation but not under strictly anaerobic conditions. As a consequence and due to the lack of exogenous ergosterol uptake, we observed increased transcript levels for a number of enzymes that catalyze oxygen-consuming reactions of the ergosterol pathway (*ERG1, ERG3, ERG5, ERG11* and *ERG25*) which may be upregulated upon hypoxia for compensation of intermediate substrate deficit (see Figure 6A). *NCP1*, a cytochrome P450 reductase that is reported to be uniformly regulated with the key enzyme *ERG11*, also demonstrated increased mRNA levels under hypoxia. It has to be stated that the transcription of *ERG25* was not only induced by oxygen scarcity but also under recombinant protein-producing conditions.

Besides ergosterol as one of the major plasma membrane components in yeast, sphingolipids represent another class of lipids with considerable importance, since they interact with ergosterol to form small platforms (“rafts”) in the cell membrane. Along with the changes in the transcription pattern of the ergosterol pathway, we also observed a hypoxic induction of four sphingolipid synthesis genes (*SUR2, SCS7, DES1*, and *SLD1*) under hypoxic conditions (see Figure 6B for a schematic overview). Interestingly, similar as demonstrated for ergosterol, all of these enzymes need molecular oxygen as substrate, unlike others whose mRNA levels remained constant

(*LAC1, LAG1*, sphingolipid C9-methyltransferase). *SLD1* (Δ8 sphingolipid desaturase), *DES1* (Δ4 sphingolipid desaturase) and the C9-methyltransferase were recently identified and characterized by Ternes and co-workers to be responsible for the synthesis of glucosylceramides (GlcCer) in *P. pastoris* [54]. In a similar study, a *P. pastoris* mutant strain deficient in the endogenous *DES1* gene was not able to produce glucosylceramides either [55]. This is in strong contrast to *S. cerevisiae*, who lacks this common class of sphingolipids.

LAC1 and *LAG1* are the only ceramide synthases in *P. pastoris*. If the cells produced more ceramides (forming the backbone of sphingolipids) at low oxygen levels, at least one of these two genes would be upregulated. Since this is not the case, we speculate that the cell activates the oxygen-dependent enzymes in order to sustain the sphingolipid metabolism and growth in spite of oxygen scarcity.

YPC1, a ceramidase, might not fit in this picture since it also has a reverse, but minor ceramide synthase activity. Its mRNA level was increased to a similar extent (2.2-fold) as *LCB5*, responsible for the phosphorylation of long chain bases in a side-branch reaction of the sphingolipid metabolism. Also *CSH1*, a probable catalytic subunit of a mannosylinositol phosphorylceramide (MIPC) synthase, was upregulated under low oxygen. However, the genes encoding the subunits of serine-palmitoyltransferase (*LCB1, LCB2*), the rate-determining step of sphingolipid synthesis in *S. cerevisiae*, were not induced, and if we exclude any posttranscriptional regulations, this hypothesis about the upregulation of oxygen-consuming reactions might be true.

A mentionable induction upon hypoxic growth was further noted for the genes *PDR16* (2.8-fold), *YPL206C/PGC1* (3.5-fold) and *PDX3* (5-fold). Pdr16p is a phosphatidylinositol transfer protein of the Sec14p family (involved in protein

504 secretion) and was attributed a role in altering the lipid composition of the plasma
505 membrane [56]. *PGC1* regulates the phosphatidylglycerol (PG) content by
506 degradation of PG to diacylglycerol (DAG) via a phospholipase C activity, as recently
507 reported [57]. *PDX3* is closely related with sterol, fatty acid and cytochrome content
508 in yeast cells [58]. The gene product Pdx3p, a pyridoxamine phosphate oxidase, was
509 additionally shown to have a significantly higher abundance at low oxygen (see
510 proteome data).
511 *OLE1*, a hypoxic gene encoding a key enzyme (Δ-9 fatty acid desaturase) in the
512 synthesis of unsaturated fatty acids, was also significantly expressed under hypoxic
513 conditions. Along with this result, the breakdown of fatty acids was highly impaired
514 by the strong down regulation of the oxygen-dependent β-oxidation pathway (*FAA1*,
515 *FAA2, POX1, ECI1, FOX2, POT1* and *SPS19*), and the genes required for
516 peroxisomal division and metabolite transport, *PEX11* and *ANT1*, respectively.
517 Since alterations in the sterol-sphingolipid balance not only result in defects in the
518 physical properties of membranes, but also affect strongly related events like cell
519 signalling [59] or secretory transport to the cell surface [60, 61], we speculate that this
520 imbalance might somehow also influence recombinant protein secretion.
521 Sphingolipids or sterols, similar to secretory proteins, are first synthesized in the ER
522 and then further processed in the Golgi apparatus, where they are targeted for the
523 proper distribution to the cell surface. In the trans-Golgi network, where secretory
524 vesicles segregate to exchange cargo proteins and lipids between membrane-bound
525 organelles [62, 63], lipid rafts play a pivotal role as sorting point [64]. Pma1 for
526 example, an H^+-ATPase and abundant plasma membrane protein, is misrouted to the
527 vacuole in a mutant strain of *S. cerevisiae* with an inability in sphingolipid acyl chain
528 elongation [65]. Gap1, a general amino acid permease, meets a similar fate in the

absence of sphingolipid neosynthesis and is also sorted to the vacuole instead of the plasma membrane [66]. The heterologous Fab is a soluble protein and probably not directly affected by alterations in the sphingolipid metabolism as it is not targeted to the plasma membrane, however, the degradation of other integral membrane proteins might favour recombinant protein secretion.

Recent experiments in our lab (data not shown) demonstrated that Tween80, a non-ionic surfactant, stimulated the production of the recombinant extracellular Fab fragment considerably. A similar result was also obtained by Apte-Deshpande and co-workers [67] for a recombinant *P. pastoris* strain producing a human growth hormone upon methanol induction. This stimulating effect of surfactants has further been observed in other host organisms including filamentous fungi and bacteria, where authors speculate about a possible correlation with (1) changes in the electrochemical membrane gradients by an altered Na^+/K^+ ratio [68] and (2) an altered membrane stability, even in transport vesicles, leading to enhanced membrane fusion [69]. These explanations would support our hypothesis about a contribution of membrane properties to improved protein secretion.

Stress responses

In this study, *P. pastoris* was exposed to severe oxygen limitation as environmental stress factor. As a consequence and considering also the additional burden of heterologous protein production, the cellular stress responses were highly activated under hypoxic growth conditions. There was a considerable upregulation of genes involved in oxidative (*SVF1, UBA4, AHP1, TSA1, GAD1, NCE103* and *OXR1*) and osmotic stress responses (*AGP2, RRD1, PBS2* and *CAB3*), and of those encoding molecular chaperones (*HSP104, HSP42, HSP31, ZPR1, LHS1*, and genes from the Hsp40/DnaJ family: *XDJ1, SIS1* and *SCJ1*). Many chaperones are catalysts during the protein folding process and provide for a quality checkpoint so that only correctly

555 folded polypeptides are released into the secretory pathway. *TSA1*, one of the genes
556 with the strongest overexpression (almost 6-fold) and with an increased abundance at
557 the proteome level under hypoxia, is a peroxidase under normal conditions and only
558 shifts to its chaperone function in response to stress [33]. It belongs to the so-called
559 'moonlighting proteins' that have multiple functions, which can vary as a
560 consequence of changes in their proximate environment. *TSA1* functions as
561 antioxidant on actively translating ribosomes and thereby maintains the integrity of
562 the translation apparatus. But it was also shown to suppress thermal aggregation by
563 binding to unfolded proteins. Stresses and other events that disrupt/overload the ER
564 folding mechanism can cause accumulation of such unprocessed polypeptides and
565 provoke the unfolded protein response (UPR) (reviewed in [70]). It was recently
566 shown that antibody heavy chain fragments (which may remain partly unfolded) were
567 retained within organelles of the secretory pathway in a recombinant *P. pastoris*
568 strain, indicating a major bottleneck in the secretion process [71]. In the same study,
569 overexpression of heterologous *HAC1*, the transcriptional activator of UPR in *S.*
570 *cerevisiae*, clearly induced the UPR-regulated genes *KAR2*/BiP and *PDI1*. While the
571 mRNA levels of *HAC1* and *PDI1* were significantly increased by hypoxia in our
572 work, the transcription of *KAR2* was not affected. However, we observed a significant
573 induction of *LHS1*, a co-chaperone of *KAR2* and likely to be the *KAR2* nucleotide
574 exchange factor. It is possible that the *S. cerevisiae*-derived *HAC1* exerts a slightly
575 different function in *P. pastoris* than the homologous one by regulating other UPR
576 related genes [72]. In this context, also *IRE1*, a gene that senses misfolded proteins in
577 the ER through interaction with *KAR2* and activation of *HAC1* was not induced.
578 Interestingly, UPR was also reported to be triggered upon lipid deprivation in order to
579 coordinate membrane synthesis in *S. cerevisiae*, in which *HAC1* plays a role in

mediating phospholipid biosynthesis [73]. This finding could be reasonably linked to the observed changes in the lipid balance during oxygen scarcity and may also explain a different regulation of UPR related genes.

ERO1, which is well-known for its crucial role in protein disulfide bond formation and redox homeostasis in the ER, was strongly activated under severe hypoxia (6.5-fold). It interacts with PDI to initiate the transfer of oxidizing equivalents to folding proteins [74]. In the presence of oxygen, *ERO1* generates H_2O_2 while in conditions of severe hypoxia other compounds (i.e. FAD) serve as electron acceptors for *ERO1* thus reducing the accumulation of reactive oxygen species in cells with heavy loads of protein thiols in their secretory pathway [75].

The overexpression of the UPR genes *PDI1, ERO1* and *HAC1* were previously reported as helper factors for enhanced recombinant protein secretion [2], which provides the presumption that their heavy induction upon hypoxia could benefit protein expression in a similar way or even stronger by a more synergistic effect.

Another gene attracting attention due to its involvement in the translocation of soluble secretory proteins and insertion of membrane proteins into the ER membrane was *WSC4*, encoding a cell wall integrity and stress response component with a transmembrane receptor activity. Unlike in a study by Kimata and co-workers [76] where *WSC4* was mentioned to be downregulated by UPR in *S. cerevisiae*, it was significantly induced under low oxygen conditions in our study, thus indicating a form of induction other than unfolded proteins.

It is also worthwhile mentioning that the stress related genes *NCE103*, a carbonic anhydrase, and *PFK3*, encoding a gamma subunit of the 6-phosphofructokinase complex, were considerably up regulated not only in hypoxia but also in the Fab producing strain, pointing at a certain protein-expression related regulation. PFK3 was

recently reported as a novel form of the hetero-oligomeric enzyme 6-phosphofructokinase in *P. pastoris*, but with no similarity to classic PFK subunits [77]. In the same study it was shown that the gamma-subunit tightly regulates the glucose metabolism by fine-tuning PFK activity via AMP and ATP, providing a rapid adaptation to perturbations in the energy balance during environmental changes, like in the case of hypoxia. The enhanced transcript levels in the Fab expressing strain can be explained by the additional energy demand during protein producing conditions. *NCE103* was regulated similarly to *PFK3* in response to oxygen limiting and protein-producing conditions. *NCE103* was described by Cleves et al. [78] to encode for a protein that is a substrate of a non-conventional secretion pathway in yeasts. In a later study a *S. cerevisiae* strain deleted in *NCE103* showed a growth-defect phenotype in the presence of oxygen and enhanced sensitivity to H_2O_2, thus suggesting a protective function against by-products of cellular respiration [79]. This contradicts our results where *NCE103* was highly induced under hypoxic conditions. Clark and co-workers [80] also questioned the proposed antioxidant activity of *NCE103* and detected a functional carbonic anhydrase activity instead. They even suggested a correlation between *NCE103* activity and the supply of sufficient bicarbonate for lipid biosynthetic processes. This hypothesis might explain the higher *NCE103* activity under oxygen scarcity. Its elevated induction in the protein producing strain could be explained by the anti-oxidative effect on ER-oxidative stress derived from UPR. More studies will be necessary to elucidate the proposed function in non conventional protein export.

Validation by quantitative real-time PCR (qRT-PCR)

We performed qRT-PCR for 1 reference gene and 14 target genes, which we selected according to their relevance for this study. We validated five genes that were differentially expressed between control and expressing strain under hypoxic

conditions. We further picked six genes from the central carbon metabolism and compared their transcript levels between normoxic and hypoxic conditions in the expressing strain. We also performed qRT-PCR for the genes encoding the light and the heavy chain of the Fab fragment under hypoxic and normoxic conditions. Since we could not observe any differences when comparing the heavy chain transcripts between producing and control strain in either of the oxygen conditions in the microarrays, we also included samples from the non-producing strain. As suitable reference genes we chose β-actin (*ACT1*). The results are highlighted in Table 4 and indicate a good correlation between microarray and qRT-PCR data. In case of the Fab genes we were able to demonstrate not only the absence of transcripts in the control strain, but also a 1.6 fold change when comparing mRNA levels between hypoxic and normoxic samples, which is similar to the 1.8 fold change seen for the light chain (see Table 4).

Conclusions

In this study we demonstrated that a systems biology approach can serve as a valuable toolbox to investigate the impact of environmental factors on cellular physiology under recombinant protein-producing conditions. The data obtained facilitated the identification of processes that seem to lead to a specific improvement of Fab expression in *P. pastoris* under hypoxic conditions. Besides the obvious changes in the lipid balance under oxygen scarcity, many other regulated genes were shown to be indirectly linked to lipid metabolism, pointing at a potential target for further rational engineering of the expression system *P. pastoris*.

It has become apparent throughout this study that there exist many contradictory data when comparing some of our regulation patterns with those obtained in *S. cerevisiae*. Regulatory functions that are different between different host organisms and lead to a

specific improvement of expression will be of special interest in the future. Further data integration on a cell model scaffold should allow a deeper understanding of physiological adaptation to hypoxia and a more systematic identification of potential targets for strain engineering.

Methods

Yeast strains and chemostat cultivations

The *Pichia pastoris* strain X-33 pGAPZαA Fab3H6, secreting a the light and heavy chain chains of a human monoclonal antibody Fab fragment under the constitutive GAP promoter and the *S. cerevisiae* alpha-mating factor leader, and its empty-vector control strain were cultivated in a glucose-limited chemostat with a working volume of 1 litre at a dilution rate of 0.1 h^{-1}, as previously described by Baumann and co-workers [13]. In brief, cells were grown at 25 ºC, 700 rpm and pH 5.0 under three different oxygen availability conditions. Oxygen concentration in the inlet gas stream was 21 % at the beginning, and was then stepwise reduced by replacing the air with nitrogen, thereby creating oxygen limited (11 %) and hypoxic (8 %) conditions (ethanol and arabitol production). Samples were taken for each physiological equilibrium condition after 5 residence times, with the exception of the hypoxic set point, where wash-out of the culture was observed after 3.5 residence times and, therefore, samples were taken just after 3 residence times. Different combinations of set-points were carried out for the three independent biological replica in order to avoid adaptive effects (see [7]).

Sampling for DNA microarray- and 2D-DIGE analysis from *P. pastoris* chemostat cultures

For DNA microarray analysis, a 9 mL chemostat sample was directly transferred into a pre-chilled Falcon tube containing 5 mL of a freshly prepared, ice cold 5 % (v/v) phenol solution in absolute ethanol (Sigma Aldrich). After thorough mixing of the cell

suspension, aliquots of 1.5 mL were pelletized by centrifugation at 4°C and immediately stored at −80 ºC.

For 2D-DIGE analysis, the chemostat sample was divided into six 2 mL aliquots. Cells were collected by centrifugation at 4 ºC, then the supernatant was stored at −20 ºC for other analysis, and the pellet was immediately frozen at −80 ºC.

Protein analysis by 2D-DIGE

Total protein extraction, DIGE labelling, first and second dimension of the two-dimensional gels as well as data acquisition, data analysis and spot identification were carried out as described by Dragosits and co-workers [7]. In order to prevent dye-specific bias effects in the protein abundance measurements, every individual protein sample (50 µg) was labelled with both Cy3 and Cy5, also known as dye swap correction. The reference pool (mixture of the equal amounts of all individual samples, yielding 50 µg) was labelled with Cy2. Given that two biological replicas of each condition (normoxic, oxygen limiting and hypoxic) were labelled once by Cy3 and once by Cy5, we run 6 gels per strain and generated 12 measurements for each spot. The ratio of the spot volume of the individual samples (Cy3 or Cy5) and the reference pool (Cy2) was determined to obtain the relative abundance of a protein under each oxygen condition across multiple gels. An average ratio of ≥ 1.5 and a 1-ANOVA *p*-value cut-off of 0.05 were the statistical parameters for the determination of proteins whose abundance was significantly different among two groups (in this case, two oxygen set points). The relative protein abundance profiles of these protein spots were illustrated through PCA and heat map clustering using the R software version 2.6.2 (http://www.R-project.org).

DNA microarrays

The *P. pastoris* DNA microarray used in this study was developed by Graf et al. [72]. RNA extraction, cDNA synthesis and labelling, as well as the microarray hybridizations and data analysis were carried out as reported earlier [72]. All samples were labelled in a dye-swap manner and hybridized against a reference cDNA, which was generated from a pool of cells grown under different culture conditions. Microarray data has been deposited in the MIAMExpress by EBI with the accession number xxx (in progress! will be provided ASAP).

Validation of microarrays by qRT-PCR

From the analysis of the microarray data, we selected 14 candidate genes with a significant fold-change, either between two oxygen setpoints or between two strains, to be validated by qRT-PCR. One reference gene shown to be equally expressed in all samples, β-actin (*ACT1*), was chosen for the relative quantification of expression levels. All primer sequences and primer characteristics (amplicon size, Tm and GC content) are given in Additional file 6.

cDNA generation and primer design

For the generation of first strand cDNA, RNA extractions from normoxic and hypoxic chemostat samples from the control and the expressing strain were subjected to a DNAse I (Invitrogen) treatment prior to reverse transcription with SuperScript®VILO cDNA Synthesis Kit (Invitrogen). All steps were performed following the manufacturer's protocol, starting from 1 µg RNA. cDNAs were finally filled up to 100 µL (1:5 dilution) with DEPC treated water (Invitrogen). Oligonucleotides (purchased from biomers.net) were designed with Primer Select 7.0.0 (DNASTAR) considering an amplicon size of 100 – 200 bp and a T_m of approximately 60 ºC.

Primer validation and amplicon purification for standard curve

To guarantee that each primer pair yields a single PCR product of the predicted size, we performed a conventional PCR and confirmed the absence of any primer dimers or

unspecific products on a 2 % (w/v) agarose gel. To additionally check the specificity of the assay, a melt-curve analysis was performed at the end of each PCR assay. An optimized reaction should have a single peak in the melt-curve, corresponding to the single band on the agarose gel. The specific PCR products were purified (Wizard®SV Gel and PCR Clean-Up System, Promega) and quantified on a Nanodrop™3300 (Thermo Scientific). From the concentration and the size of the amplicon, the copy number per µL was determined according to Whelan [81] and decimal dilutions representing $10^7 – 10^4$ copies of target DNA were prepared for the generation of the standard curves.

qRT-PCR assay

Quantitative real-time PCR was carried out in 20 µL reactions using semi-skirted iQ 96-well PCR plates and iQ™SYBR® Green supermix (both from Bio-Rad). Samples were measured in triplicates and standards were measured in duplicates on the iCycler Thermal Cycler (Bio-Rad). A non template control was run in every experiment for each of the primer pairs to avoid detection of unspecific priming. The reactions were incubated at 95 ºC for 5 min to activate the *Taq* polymerase, and then subjected to a three-step cycling protocol including melting (94 ºC, 15 sec), annealing (58 ºC, 15 sec) and extension (72 ºC, 30 sec) for a total of 40 cycles. Each extension was followed by data collection at 72 ºC and a short incubation step at 78 ºC (1 sec) for a second plate read closer to the melting point. After a final extension of 5 min at 72 ºC, we generated a melt-curve profile by data collection during 70 cycles starting at 60 ºC, with 0.5 ºC increments / cycle (1-sec intervals).

Data analysis

The relative gene expression was calculated for each sample with three measurements giving a maximum standard deviation of 10 %. Since the amplification efficiencies of

the target and reference genes were not the same in our experiments, we used the Pfaffl method [82] for the relative quantification of our qRT-PCR results.

Analytical procedures

Cell biomass was monitored by measuring the optical density at 600 nm (OD_{600}). For cellular dry weight, a known volume of cultivation broth was filtered using pre-weighted filters; these were washed with two volumes of distilled water and dried to constant weight at 105 °C for 24 h. Samples for extracellular metabolite analyses were centrifuged at 6,000 rpm for 2 min in a microcentrifuge to remove the cells and subsequently filtered through 0.45 μm-filters (Millipore type HAWP). Glucose, organic acids, ethanol and arabitol were analyzed by HPLC (Series 1050, Hewlett Packard) with an ionic exchange column (Bio-Rad, Aminex HPX-87H). As mobile phase, 15 mM sulphuric acid was used. The metabolites were detected (Detector HP 1047A, Hewlett Packard) and quantified with the Software EmpowerProfor. The exhaust gas of the bioreactor was cooled in a condenser at 2-4 °C (*Frigomix R*, B. Braun Biotech) and dried through a silica gel column. Concentrations of oxygen and carbon dioxide in the exhaust gas of bioreactor cultivations were determined on line with specific sensors (BCP-CO_2 and BCP-O_2, BlueSens, Germany).

Biosynthetically directed fractional (BDF) ^{13}C-labelling

P. pastoris cells were continuously fed with a minimal medium for five residence times until reaching a metabolic steady state, as indicated by a constant cell density and constant oxygen and carbon dioxide concentrations in the bioreactor exhaust gas. Biosynthetically directed fractional ^{13}C labelling (BDF) of cells growing at steady state on a single carbon source has been described elsewhere [36, 41, 83]. After reaching the steady state, 10 % (w/w) of the carbon source in the medium was replaced with uniformly ^{13}C-labelled substrate (^{13}C-labelled glucose, isotopic enrichment of >98 %, from Cortecnet, Voisins le Bretonneux, France). After one

783 residence time, labelled cells were harvested by centrifugation at 4,000 ×g for 10 min,
784 resuspended in 20 mM Tris·HCl (pH 7.6) and centrifuged again. Finally, the washed
785 cell pellets were lyophilized (Benchtop 5L Virtis Sentry). An amount of 100 mg of the
786 lyophilized biomass was resuspended in 6 mL of 6 M HCl and subsequently
787 hydrolyzed in sealed glass tubes at 110 ºC for 21 h. The resulting suspensions were
788 filtered using 0.2 μm-filters (Millex-GP, Millipore) and lyophilized. The lyophilized
789 hydrolysates were dissolved in D_2O for NMR experiments, the pH of the samples
790 being below 1 due to residual HCl.

791 **NMR spectroscopy and metabolic flux ratio (METAFoR) analysis**
792 2D [^{13}C,^{1}H]-COSY spectra were acquired for both aliphatic and aromatic resonances
793 as described [84] at 40 ºC on a Varian Inova spectrometer operating at a ^{1}H resonance
794 frequency of 600 MHz. The spectra were processed using the standard Varian
795 spectrometer software VNMR (version 6.1, C). The program FCAL [85] was used for
796 the integration of ^{13}C-^{13}C scalar fine structures in 2D [^{13}C,^{1}H]-COSY, for the
797 calculation of relative abundances, *f*-values (see Additional file 5), of intact carbon
798 fragments arising from a single carbon source molecule [84], and for the calculation
799 of the resulting flux ratios through several key pathways in central metabolism, as
800 described by Maaheimo [42] and Jouhten [38].
801 As described previously [36, 41, 42, 83-87], the calculation of metabolic flux ratios
802 when using fractional ^{13}C-labelling of amino acids is based on assuming both a
803 metabolic and an isotopomeric steady state. To establish a cost-effective protocol for a
804 larger number of ^{13}C labelling experiments, we fed a chemostat operating in metabolic
805 steady state for the duration of one volume change with the medium containing the
806 ^{13}C-labelled substrates [41, 83] before harvesting the biomass. Then, the fraction of
807 unlabeled biomass produced prior to the start of the supply with ^{13}C-labelled medium

can be calculated following simple wash-out kinetics ([86], see also [36] for additional discussion).

^{13}C-constrained metabolic flux analysis (^{13}C-MFA)

Intracellular metabolic fluxes were determined using ^{13}C-NMR derived flux ratios as additional experimental constraints to solve the MFA system [88]. The biochemical reaction network model (see Additional files 7 and 8) was based on the stoichiometric model of central carbon metabolism formulated for *S. cerevisiae* [38, 42], and *P. pastoris* [36], and adapted conveniently. Briefly, the model included the glycolytic and the pentose phosphate pathways, the TCA cycle and the fermentative pathways, production of glycerol, arabitol, and anabolic fluxes from metabolic intermediates to biosynthesis. The glyoxylate cycle, the PEP carboxykinase and the malic enzyme activity were omitted from the stoichiometric model since the METAFoR data showed that those pathways were either inactive or at basal levels (see [38] for details on the identification of these activities). Separate pools of Pyr, AcCoA and Oaa in the two cellular compartments, cytoplasm and mitochondria, were considered in the metabolic flux ratio analysis. Transports of Pyr and Oaa across the mitochondrial membrane were included in the model but transport of AcCoA, the final step of the cytosolic Pyruvate dehydrogenase (PDH) bypass, was omitted. The potential carbon flux through the PDH bypass was lumped into the flux through the PDH reaction, since the ^{13}C-labelling protocol used in this study does not allow for an assessment of the split flux ratio between these two pathways (that is, given that flux through malic enzyme is essentially zero in our model, labelled Pyr being metabolised through the PDH bypass does not produce labelling patterns in mitochondrial AcCoA that are distinct from those generated when Pyr is channelled through the PDH reaction. Nevertheless, flux through the PDH bypass can not be totally excluded, as discussed in the Results section.

834 The complete model for the calculation of intracellular fluxes, comprised 33
835 (normoxic condition) and 34 (oxygen-limited and hypoxic conditions) metabolic
836 reactions. The measured uptake and excretion rates and the rates of metabolic
837 precursor depletion to biosynthesis, as determined from the composition of *P. pastoris*
838 biomass previously reported for each oxygenation condition [22], were combined with
839 a set of linearly independent equations obtained from METAFoR analysis to render
840 the complete linear system solvable.

841 The determined metabolic flux ratios were used as additional constraints for solving
842 the metabolic network following a ^{13}C constrained flux balancing approach similarly
843 to a previous approach [88]. Using the constraints from the METAFoR analysis, it
844 was not necessary to include redox cofactor mass balances. Cofactor mass balances
845 are sources of errors since the correct balancing requires detailed knowledge of the
846 relative activities of different isoenzymes and the enzyme cofactor specificities on a
847 cell wide scale. The flux ratios considered in the present approach were the following
848 (equations 1 to 4, the reaction numbers are defined in Additional file 8):

849 the fraction of Oaa_{mit} originating from Oaa_{cyt}, that is, Oaa_{cyt} transport into the
850 mitochondria (only applicable to oxygen-limiting and hypoxic conditions):

851 $$a = \frac{x_{23}}{x_{23} + x_{16}} \qquad (1)$$

852 Under normoxic conditions, reaction x_{24} (flux of Oaa_{cyt} into the mitochondria) was
853 calculated as a ***net*** Oaa transport flux across the mitochondrial membrane (that is,
854 Oaa_{cyt} import – Oaa_{mit} export), annotated as x_{24}*. Also, the labelling patterns of Pep
855 instead of Pyr_{cyt} were considered (see Results section) and, therefore, the
856 corresponding anaplerotic flux ratio was defined as:

857 $$b = \frac{x_{23}^*}{x_{23}^* + x_{16}} \quad (2)$$

858 the fraction of Oaa_{cyt} originating from Pyr_{cyt}, that is, the anaplerotic flux ratio (only
859 applicable to oxygen-limiting and hypoxic conditions):

860 $$c = \frac{x_{17}}{x_{17} + x_{24}} \quad (3)$$

861 the fraction of Pep from PPP assuming a maximal contribution of PPP:

862 $$d \geq \frac{x_9 + 2(x_{11}) + 3(x_{10})}{2(x_3) + x_9 + x_{10}} \quad (4)$$

863 The following linear constraint equations (equations 5 to 8) were derived from the
864 flux ratio equations:

865 $$x_{23}(1-a) + x_{16}(-a) = R_a \quad (5)$$

866 $$x_{23}^*(1-b) + x_{16}(-b) = R_b \quad (6)$$

867 $$x_{17}(1-c) + x_{24}(-c) = R_c \quad (7)$$

868 $$x_9(1-d) + x_{11}(2) + x_{10}(3-d) + x_3(-2) \leq R_d \quad (8)$$

869 Equations 5 to 7 were added to the stoichiometric model as a submatrix F, obtaining
870 the complete metabolic model to solve the metabolite mass balances:

871 $$\begin{bmatrix} S \\ F \end{bmatrix} \cdot x = \begin{bmatrix} c \\ 0 \end{bmatrix} \equiv N \cdot x = b \quad (9)$$

872 where S represents the stoichiometric matrix (including input/output reactions), c is a
873 column vector with either 0 for internal reactions or the corresponding value for each one
874 of the input/output rates and x is the vector of fluxes. Solution of the resulting linear
875 system was obtained using the MATLAB function *lsqlin* using equation 8 as an additional
876 constrain. Irreversibility was assumed for several intracellular fluxes and for the depletion

of precursors to biosynthetic reactions, that is, only positive values were allowed for these fluxes (see Additional file 8).

Confidence intervals for the optimized fluxes were calculated up on the determination of their standard deviation using the Fisher Information Matrix approach (FIM) [89] as:

$$\sigma_j = \sqrt{\left(FIM^{-1}\right)_{jj}} \quad (10)$$

Calculation of FIM was performed as:

$$FIM = \sum W^T C^{-1} W \quad (11)$$

where C is the variance-covariance matrix of the measurements (assumed independent) and W is a parameter sensitivity matrix where each element of w_{ij} corresponds to:

$$w_{ij} = \frac{\partial x_i}{\partial p_j} \quad (12)$$

which describes an infinitesimal change of the variable x_i (e.g. a measurement) due to an infinitesimal change in parameter p_j (a flux).

Confidence intervals for the estimated fluxes $\hat{p}_j$ of p_j can be derived [90] from:

$$\hat{p}_j - \sigma_{p_j}.t^{v}_{\alpha/2} < p_j < \hat{p}_j + \sigma_{p_j}.t^{v}_{\alpha/2} \quad (13)$$

where $t^{v}_{\alpha/2}$ corresponds to the Student's t distribution, with v degrees of freedom and α corresponds to the (1-α) confidence interval chosen. All calculations were performed on a PC compatible computer running Matlab ® 7.4 (v2007b) for Windows.

Authors' contributions

KB performed bioreactor cultivations, proteomics and microarray experiments, quantitative real-time PCR, data analysis and interpretation of the results, and drafted the manuscript. MC carried out the ^{13}C-labeling experiments and macroscopic data processing, performed the metabolic flux calculations and participated in

interpretation of results. MD assisted in the design and performance of proteome and microarray experiments. ABG participated in the design and bioinformatic analysis of the microarrays. JS carried out the protein identification by liquid chromatography-tandem mass spectrometry. PJ and HM performed the 2D-NMR and METAFoR analyses and participated in the subsequent interpretation of results. BG helped with the conceptual design of the study and with the interpretation of omics data. JA designed the ^{13}C-constrained MFA approach, and participated in analysis and interpretation of MFA results. DM participated in the overall conceptual and experimental design of this study and interpretation of results. PF participated in the conceptual and experimental design, interpretation of results and in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work has been supported by the European Science Foundation (ESF, program EuroSCOPE), through the Complementary Actions Plan (Project BIO2005-23733-E) of the Spanish Ministry of Science and Education, the Integrated Action HU2005-0001 of the Spanish Ministry of Science and Education, the Austrian Science Fund (FWF), project no. I37-B03, and the Austrian Exchange Service. The Ministry of Innovation and Universities of the Generalitat de Catalunya gave support through contract grant 2005-SGR-00698 and 2009SGR-281, Xarxa de Referència en Biotecnologia and doctoral fellowship for K.B. We would like to thank Klaus Fortschegger for his support with real-time PCR analysis and Philipp Ternes for helpful comments on the sphingolipid metabolism.

References

1. Porro D, Mattanovich D: **Recombinant protein production in yeasts.** *Methods Mol Biol* 2004, **267**:241-258.

2. Gasser B, Sauer M, Maurer M, Stadlmayr G, Mattanovich D: **Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts.** *Appl Environ Microbiol* 2007, **73**(20):6499-6507.

3. Bonander N, Bill R: **Relieving the first bottleneck in the drug discovery pipeline: using array technologies to rationalize membrane protein production.** *Expert Rev Proteomics* 2009, **6**(5):501-505.

4. Mattanovich D, Gasser B, Hohenblum H, Sauer M: **Stress in recombinant protein producing yeasts.** *J Biotechnol* 2004, **113**(1-3):121-135.

5. Knijnenburg T, Daran J, van den Broek M, Daran-Lapujade P, de Winde J, Pronk J, Reinders M, Wessels L: **Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: a quantitative analysis of a compendium of chemostat-based transcriptome data.** *BMC Genomics* 2009, **10**:53.

6. Tai S, Daran-Lapujade P, Walsh M, Pronk J, Daran J: **Acclimation of Saccharomyces cerevisiae to low temperature: a chemostat-based transcriptome analysis.** *Mol Biol Cell* 2007, **18**(12):5100-5112.

7. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P, Mattanovich D: **The effect of temperature on the proteome of recombinant Pichia pastoris.** *J Proteome Res* 2009, **8**(3):1380-1392.

8. Georgiou G, Valax P, Ostermeier M, Horowitz P: **Folding and aggregation of TEM beta-lactamase: analogies with the formation of inclusion bodies in Escherichia coli.** *Protein Sci* 1994, **3**(11):1953-1960.

9. Li Z, Xiong F, Lin Q, d'Anjou M, Daugulis A, Yang D, Hew C: **Low-temperature increases the yield of biologically active herring antifreeze protein in Pichia pastoris.** *Protein Expr Purif* 2001, **21**(3):438-445.

10. Haynes C, Titus E, Cooper A: **Degradation of misfolded proteins prevents ER-derived oxidative stress and cell death.** *Mol Cell* 2004, **15**(5):767-776.

11. Malhotra J, Miao H, Zhang K, Wolfson A, Pennathur S, Pipe S, Kaufman R: **Antioxidants reduce endoplasmic reticulum stress and improve protein secretion.** *Proc Natl Acad Sci U S A* 2008, **105**(47):18525-18530.

12. Gasser B, Saloheimo M, Rinas U, Dragosits M, Rodríguez-Carmona E, Baumann K, Giuliani M, Parrilli E, Branduardi P, Lang C *et al*: **Protein folding and conformational stress in microbial cells producing**

recombinant proteins: a host comparative overview. *Microb Cell Fact* 2008, **7**:11.

13. Baumann K, Maurer M, Dragosits M, Cos O, Ferrer P, Mattanovich D: **Hypoxic fed-batch cultivation of Pichia pastoris increases specific and volumetric productivity of recombinant proteins.** *Biotechnol Bioeng* 2008, **100**(1):177-183.

14. Daly R, Hearn M: **Expression of heterologous proteins in Pichia pastoris: a useful experimental tool in protein engineering and production.** *J Mol Recognit*, **18**(2):119-138.

15. Regenberg B, Grotkjaer T, Winther O, Fausbøll A, Akesson M, Bro C, Hansen L, Brunak S, Nielsen J: **Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae.** *Genome Biol* 2006, **7**(11):R107.

16. Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B: **Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris.** *Microb Cell Fact* 2009, **8**:29.

17. Görg A, Weiss W, Dunn M: **Current two-dimensional electrophoresis technology for proteomics.** *Proteomics* 2004, **4**(12):3665-3685.

18. de Groot M, Daran-Lapujade P, van Breukelen B, Knijnenburg T, de Hulster E, Reinders M, Pronk J, Heck A, Slijper M: **Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes.** *Microbiology* 2007, **153**(Pt 11):3864-3878.

19. Bruckmann A, Hensbergen P, Balog C, Deelder A, Brandt R, Snoek I, Steensma H, van Heusden G: **Proteome analysis of aerobically and anaerobically grown Saccharomyces cerevisiae cells.** *J Proteomics* 2009, **71**(6):662-669.

20. van Dijken J, van den Bosch E, Hermans J, de Miranda L, Scheffers W: **Alcoholic fermentation by 'non-fermentative' yeasts.** *Yeast* 1986, **2**(2):123-127.

21. Gancedo C, Serrano R: **Energy-yielding metabolism**. In: *The Yeasts.* Edited by & JSH, amp, Rose AH, 2nd edn. New York: Academic Press: 205-259.

22. Carnicer M, Baumann K, Töplitz I, Sánchez-Ferrando F, Mattanovich D, Ferrer P, Albiol J: **Macromolecular and elemental composition analysis and extracellular metabolite balances of Pichia pastoris growing at different oxygen levels.** *Microb Cell Fact* 2009, **8**:65.

23. Bellinger Y, Larher F: **A 13C comparative nuclear magnetic resonance study of organic solute production and excretion by the yeasts Hansenula anomala and Saccharomyces cerevisiae in saline media.** *Can J Microbiol* 1988, **34**(5):605-612.

24. Tokuoka K, Ishitani T, Chung W-C: **Accumulation of polyols and sugars in some sugar-tolerant yeasts**. *The Journal of General and Applied Microbiology* 1992, **38**(1):11.

25. Dragosits M, Stadlmann J, Graf A, Gasser B, Maurer M, Sauer M, Kreil D, Altmann F, Mattanovich D: **The response to unfolded protein is involved in osmotolerance of Pichia pastoris.** *BMC Genomics* 2010, **11**(1):207.

26. Passoth V, Fredlund E, Druvefors U, Schnürer J: **Biotechnology, physiology and genetics of the yeast Pichia anomala.** *FEMS Yeast Res* 2006, **6**(1):3-13.

27. Träff K, Jönsson L, Hahn-Hägerdal B: **Putative xylose and arabinose reductases in Saccharomyces cerevisiae.** *Yeast* 2002, **19**(14):1233-1241.

28. Petrash J, Murthy B, Young M, Morris K, Rikimaru L, Griest T, Harter T: **Functional genomic studies of aldo-keto reductases.** *Chem Biol Interact* 2001, **130-132**(1-3):673-683.

29. Wong B, Leeson S, Grindle S, Magee B, Brooks E, Magee P: **D-arabitol metabolism in Candida albicans: construction and analysis of mutants lacking D-arabitol dehydrogenase.** *J Bacteriol* 1995, **177**(11):2971-2976.

30. Ingram J, Wood W: **Enzymatic basis for D-Arabitol production by Saccharomyces Rouxii**. *J Bacteriol* 1965, **89**:1186-1194.

31. Singh J, Kumar D, Ramakrishnan N, Singhal V, Jervis J, Garst J, Slaughter S, DeSantis A, Potts M, Helm R: **Transcriptional response of Saccharomyces cerevisiae to desiccation and rehydration.** *Appl Environ Microbiol* 2005, **71**(12):8752-8763.

32. Zuzuarregui A, Monteoliva L, Gil C, del Olmo M: **Transcriptomic and proteomic approach for understanding the molecular basis of adaptation of Saccharomyces cerevisiae to wine fermentation.** *Appl Environ Microbiol* 2006, **72**(1):836-847.

33. Jang H, Lee K, Chi Y, Jung B, Park S, Park J, Lee J, Lee S, Moon J, Yun J *et al*: **Two enzymes in one; two yeast peroxiredoxins display oxidative stress-dependent switching from a peroxidase to a molecular chaperone function.** *Cell* 2004, **117**(5):625-635.

34. Urban C, Xiong X, Sohn K, Schröppel K, Brunner H, Rupp S: **The moonlighting protein Tsa1p is implicated in oxidative stress response and in cell wall biogenesis in Candida albicans.** *Mol Microbiol* 2005, **57**(5):1318-1341.

35. Kodaki T, Tsuji S, Otani N, Yamamoto D, Rao K, Watanabe S, Tsukatsune M, Makino K: **Differential transcriptional regulation of two distinct S-adenosylmethionine synthetase genes (SAM1 and SAM2) of Saccharomyces cerevisiae.** *Nucleic Acids Res Suppl* 2003(3):303-304.

36. Solà A, Maaheimo H, Ylönen K, Ferrer P, Szyperski T: **Amino acid biosynthesis and metabolic flux profiling of Pichia pastoris.** *Eur J Biochem* 2004, **271**(12):2462-2470.

37. Frick O, Wittmann C: **Characterization of the metabolic shift between oxidative and fermentative growth in Saccharomyces cerevisiae by comparative 13C flux analysis.** *Microb Cell Fact* 2005, **4**:30.

38. Jouhten P, Rintala E, Huuskonen A, Tamminen A, Toivari M, Wiebe M, Ruohonen L, Penttilä M, Maaheimo H: **Oxygen dependence of metabolic fluxes and energy generation of Saccharomyces cerevisiae CEN.PK113-1A.** *BMC Syst Biol* 2008, **2**:60.

39. Fredlund E, Broberg A, Boysen M, Kenne L, Schnürer J: **Metabolite profiles of the biocontrol yeast Pichia anomala J121 grown under oxygen limitation.** *Appl Microbiol Biotechnol* 2004, **64**(3):403-409.

40. Fredlund E, Blank L, Schnürer J, Sauer U, Passoth V: **Oxygen- and glucose-dependent regulation of central carbon metabolism in Pichia anomala.** *Appl Environ Microbiol* 2004, **70**(10):5905-5911.

41. Fiaux J, Cakar Z, Sonderegger M, Wüthrich K, Szyperski T, Sauer U: **Metabolic-flux profiling of the yeasts Saccharomyces cerevisiae and Pichia stipitis.** *Eukaryot Cell* 2003, **2**(1):170-180.

42. Maaheimo H, Fiaux J, Cakar Z, Bailey J, Sauer U, Szyperski T: **Central carbon metabolism of Saccharomyces cerevisiae explored by biosynthetic fractional (13)C labeling of common amino acids.** *Eur J Biochem* 2001, **268**(8):2464-2479.

43. Bakker B, Overkamp K: **Stoichiometry and compartmentation of NADH metabolism in Saccharomyces cerevisiae.** *FEMS Microbiol Rev* 2001, **25**(1):15-37.

44. van Dijken J, Weusthuis R, Pronk J: **Kinetics of growth and sugar consumption in yeasts.** *Antonie Van Leeuwenhoek* 1993, **63**(3-4):343-352.

45. van Roermund C, Hettema E, van den Berg M, Tabak H, Wanders R: **Molecular characterization of carnitine-dependent transport of acetyl-CoA from peroxisomes to mitochondria in Saccharomyces cerevisiae and identification of a plasma membrane carnitine transporter, Agp2p.** *EMBO J* 1999, **18**(21):5843-5852.

46. Strijbis K, van Roermund C, Hardy G, van den Burg J, Bloem K, de Haan J, van Vlies N, Wanders R, Vaz F, Distel B: **Identification and characterization of a complete carnitine biosynthesis pathway in Candida albicans.** *FASEB J* 2009, **23**(8):2349-2359.

47. De Schutter K, Lin Y, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouzé P, Van de Peer Y, Callewaert N: **Genome sequence of the recombinant protein production host Pichia pastoris.** *Nat Biotechnol* 2009, **27**(6):561-566.

48. Berry M, Boulton A: **Glyceraldehyde-3-phosphate dehydrogenase and apoptosis.** *J Neurosci Res* 2000, **60**(2):150-154.

49. Delgado M, O'Connor J, Azorín I, Renau-Piqueras J, Gil M, Gozalbo D: **The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins.** *Microbiology* 2001, **147**(Pt 2):411-417.

50. Beney L, Gervais P: **Influence of the fluidity of the membrane on the response of microorganisms to environmental stresses.** *Appl Microbiol Biotechnol* 2001, **57**(1-2):34-42.

51. Swan T, Watson K: **Stress tolerance in a yeast lipid mutant: membrane lipids influence tolerance to heat and ethanol independently of heat shock proteins and trehalose.** *Can J Microbiol* 1999, **45**(6):472-479.

52. Andreasen A, Stier T: **Anaerobic nutrition of Saccharomyces cerevisiae. I. Ergosterol requirement for growth in a defined medium.** *J Cell Physiol* 1953, **41**(1):23-36.

53. Andreasen A, Stier T: **Anaerobic nutrition of Saccharomyces cerevisiae. II. Unsaturated fatty acid requirement for growth in a defined medium.** *J Cell Physiol* 1954, **43**(3):271-281.

54. Ternes P, Sperling P, Albrecht S, Franke S, Cregg J, Warnecke D, Heinz E: **Identification of fungal sphingolipid C9-methyltransferases by phylogenetic profiling.** *J Biol Chem* 2006, **281**(9):5582-5592.

55. Michaelson L, Zäuner S, Markham J, Haslam R, Desikan R, Mugford S, Albrecht S, Warnecke D, Sperling P, Heinz E *et al*: **Functional characterization of a higher plant sphingolipid Delta4-desaturase: defining the role of sphingosine and sphingosine-1-phosphate in Arabidopsis.** *Plant Physiol* 2009, **149**(1):487-498.

56. van den Hazel H, Pichler H, do Valle Matta M, Leitner E, Goffeau A, Daum G: **PDR16 and PDR17, two homologous genes of Saccharomyces cerevisiae, affect lipid biosynthesis and resistance to multiple drugs.** *J Biol Chem* 1999, **274**(4):1934-1941.

57. Simocková M, Holic R, Tahotná D, Patton-Vogt J, Griac P: **Yeast Pgc1p (YPL206c) controls the amount of phosphatidylglycerol via a phospholipase C-type degradation mechanism.** *J Biol Chem* 2008, **283**(25):17107-17115.

58. Loubbardi A, Marcireau C, Karst F, Guilloton M: **Sterol uptake induced by an impairment of pyridoxal phosphate synthesis in Saccharomyces cerevisiae: cloning and sequencing of the PDX3 gene encoding pyridoxine (pyridoxamine) phosphate oxidase.** *J Bacteriol* 1995, **177**(7):1817-1823.

59. Incardona J, Eaton S: **Cholesterol in signal transduction.** *Curr Opin Cell Biol* 2000, **12**(2):193-203.

60. Bagnat M, Keränen S, Shevchenko A, Simons K: **Lipid rafts function in biosynthetic delivery of proteins to the cell surface in yeast.** *Proc Natl Acad Sci U S A* 2000, **97**(7):3254-3259.

61. Proszynski T, Klemm R, Gravert M, Hsu P, Gloor Y, Wagner J, Kozak K, Grabner H, Walzer K, Bagnat M *et al*: **A genome-wide visual screen reveals a role for sphingolipids and ergosterol in cell surface delivery in yeast.** *Proc Natl Acad Sci U S A* 2005, **102**(50):17981-17986.

62. Barlowe C: **COPII and selective export from the endoplasmic reticulum.** *Biochim Biophys Acta* 1998, **1404**(1-2):67-76.

63. Mellman I, Warren G: **The road taken: past and future foundations of membrane traffic.** *Cell* 2000, **100**(1):99-112.

64. Simons K, Ikonen E: **Functional rafts in cell membranes.** *Nature* 1997, **387**(6633):569-572.

65. Eisenkolb M, Zenzmaier C, Leitner E, Schneiter R: **A specific structural requirement for ergosterol in long-chain fatty acid synthesis mutants important for maintaining raft domains in yeast.** *Mol Biol Cell* 2002, **13**(12):4414-4428.

66. Lauwers E, André B: **Association of yeast transporters with detergent-resistant membranes correlates with their cell-surface location.** *Traffic* 2006, **7**(8):1045-1059.

67. Apte-Deshpande A, Rewanwar S, Kotwal P, Raiker V, Padmanabhan S: **Efficient expression and secretion of recombinant human growth hormone in the methylotrophic yeast Pichia pastoris: potential applications for other proteins.** *Biotechnol Appl Biochem* 2009, **54**(4):197-205.

68. Jacques N, Jacques V, Wolf A, Wittenberger C: **Does an increase in membrane unsaturated fatty acids account for Tween 80 stimulation of glucosyltransferase secretion by Streptococcus salivarius?** *J Gen Microbiol* 1985, **131**(1):67-72.

69. Mukaiyama H, Giga-Hama Y, Tohda H, Takegawa K: **Dextran sodium sulfate enhances secretion of recombinant human transferrin in Schizosaccharomyces pombe.** *Appl Microbiol Biotechnol* 2009, **85**(1):155-164.

70. Kaufman R: **Stress signaling from the lumen of the endoplasmic reticulum: coordination of gene transcriptional and translational controls.** *Genes Dev* 1999, **13**(10):1211-1233.

71. Gasser B, Maurer M, Gach J, Kunert R, Mattanovich D: **Engineering of Pichia pastoris for improved production of antibody fragments.** *Biotechnol Bioeng* 2006, **94**(2):353-361.

72. Graf A, Gasser B, Dragosits M, Sauer M, Leparc G, Tüchler T, Kreil D, Mattanovich D: **Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.** *BMC Genomics* 2008, **9**:390.

73. Cox J, Chapman R, Walter P: **The unfolded protein response coordinates the production of endoplasmic reticulum protein and endoplasmic reticulum membrane.** *Mol Biol Cell* 1997, **8**(9):1805-1814.

74. Frand A, Kaiser C: **The ERO1 gene of yeast is required for oxidation of protein dithiols in the endoplasmic reticulum.** *Mol Cell* 1998, **1**(2):161-170.

75. Gross E, Sevier CS, Heldman N, Vitu E, Bentzur M, Kaiser CA, Thorpe C, Fass D: **Generating disulfides enzymatically: reaction products and electron acceptors of the endoplasmic reticulum thiol oxidase Ero1p**. *Proc Natl Acad Sci U S A* 2006, **103**(2):299-304.

76. Kimata Y, Ishiwata-Kimata Y, Yamada S, Kohno K: **Yeast unfolded protein response pathway regulates expression of genes for anti-oxidative stress and for cell surface proteins.** *Genes Cells* 2006, **11**(1):59-69.

77. Tanneberger K, Kirchberger J, Bär J, Schellenberger W, Rothemund S, Kamprad M, Otto H, Schöneberg T, Edelmann A: **A novel form of 6-phosphofructokinase. Identification and functional relevance of a third type of subunit in Pichia pastoris.** *J Biol Chem* 2007, **282**(32):23687-23697.

78. Cleves A, Cooper D, Barondes S, Kelly R: **A new pathway for protein export in Saccharomyces cerevisiae.** *J Cell Biol* 1996, **133**(5):1017-1026.

79. Götz R, Gnann A, Zimmermann F: **Deletion of the carbonic anhydrase-like gene NCE103 of the yeast Saccharomyces cerevisiae causes an oxygen-sensitive growth defect.** *Yeast* 1999, **15**(10A):855-864.

80. Clark D, Rowlett R, Coleman J, Klessig D: **Complementation of the yeast deletion mutant DeltaNCE103 by members of the beta class of carbonic anhydrases is dependent on carbonic anhydrase activity rather than on antioxidant activity.** *Biochem J* 2004, **379**(Pt 3):609-615.

81. Whelan J, Russell N, Whelan M: **A method for the absolute quantification of cDNA using real-time PCR.** *J Immunol Methods* 2003, **278**(1-2):261-269.

82. Pfaffl M: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Res* 2001, **29**(9):e45.

83. Sauer U, Hatzimanikatis V, Bailey J, Hochuli M, Szyperski T, Wüthrich K: **Metabolic fluxes in riboflavin-producing Bacillus subtilis.** *Nat Biotechnol* 1997, **15**(5):448-452.

84. Szyperski T: **Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.** *Eur J Biochem* 1995, **232**(2):433-448.

85. Szyperski T, Glaser R, Hochuli M, Fiaux J, Sauer U, Bailey J, Wüthrich K: **Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional 13C labeling and two-dimensional NMR spectroscopy.** *Metab Eng* 1999, **1**(3):189-197.

86. Szyperski T: **13C-NMR, MS and metabolic flux balancing in biotechnology research.** *Q Rev Biophys* 1998, **31**(1):41-106.

87. Sauer U, Lasko D, Fiaux J, Hochuli M, Glaser R, Szyperski T, Wüthrich K, Bailey J: **Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism.** *J Bacteriol* 1999, **181**(21):6679-6688.

88. Fischer E, Zamboni N, Sauer U: **High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints.** *Anal Biochem* 2004, **325**(2):308-316.

89. Faller D, Klingmueller U, Timmer J: **Simulation methods for optimal experimental design in systems biology**. *Simulation* 2003(79):9.

90. Press W, Flannery B, Teukolsky S, Vetterling W: **Numerical Recipes Example Book (C++), The Art of Scientific Computing**, 2nd edition edn. Cambridge: Cambridge University Press; 2002.

91. Pitkänen J, Aristidou A, Salusjärvi L, Ruohonen L, Penttilä M: **Metabolic flux analysis of xylose metabolism in recombinant Saccharomyces cerevisiae using continuous culture.** *Metab Eng* 2003, **5**(1):16-31.

92. Stephanopoulos GN, Aristidou AA, Nielsen J: **Metabolic Engineering—Principles and Methodologies**. New York: Academic Press; 1998.

Figures

Figure 1 – Venn diagram

Venn diagram illustrating the relationship of up- and downregulated annotated genes ($p \leq 0.05$, no FC) in the control (C) and expressing (E) strain under hypoxic conditions. The intersections reflect equally regulated genes among both data sets.

Figure 2 – Percentage distribution of regulated genes (cut-off $p \leq 0.05$) to their respective GO biological process term(s)

Upregulated (purple), downregulated (green) and unregulated genes (white) in response to hypoxia were classified into their GO functional group(s) and illustrated as relative numbers summing up 100 %. A: Control strain, B: Fab expressing strain.

Figure 3 – Principal component analysis (PCA) and heat map of proteome data

A: Principal component analysis of the proteome data presented in a 2D graph with components 1 and 2 displayed on the two axes. PCA projection demonstrates that the maximum variability in the dataset occurs between normoxia and hypoxia, independent of the strain genetic background.

B: Heat map presentation of a hierarchical cluster of the 45 proteins that show significantly different ($p \leq 0.05$) abundances in both strains (Fab expressing (E) and control (C) strain) at different oxygen concentrations in the inlet gas (20, 10 or 8). The green colour represents relatively low expression and pink colour represents relatively high expression levels.

Figure 4 - Metabolic flux distributions in *P. pastoris* Fab-expressing and control strains under different oxygenation conditions

Relative net flux distributions of *P. pastoris* X-33/pGAPαA_Fab and X-33/pGAPαA in glucose-limited chemostats at a D = 0.1 h^{-1} under different oxygenation conditions. Fluxes are shown as relative fluxes normalised to the specific glucose uptake rate (expressed as mmol glucose g^{-1} DCW h^{-1}) in the corresponding experiment. The specific glucose uptake rates corresponding to the different oxygenation conditions and strains are given at the top of the figure. The fluxes for each reaction in the network corresponding to 21 %, 11 % and 8 % oxygen in the bioreactor inlet gas are given from top to bottom; the flux values from the Fab-producing strain are shown on the left and those from the corresponding control strain on the right. The transport of Oaa across the mitochondrial membrane under normoxic conditions is given as a single net influx value. Fluxes with SD values are provided in the Additional file 4. Arrows indicate higher (red) or lower (green) mRNA levels (T) and protein abundances (P) during hypoxia compared to normoxia. The corresponding gene/protein names are displayed above the arrows.

Figure 5 - Fractional distributions of carbon fluxes in metabolic branching points derived from ^{13}C-MFA.

Fractional distributions of carbon fluxes. A: the glucose-6-P flux split to glycolysis and PPP B: the pyruvate branching point, and C: the TCA cycle, in *P. pastoris* Fab-producing (E) and control (C) strains growing in glucose-limited chemostats at D = 0.1 h^{-1}, in 21 %, 11 % and 8 % oxygen in the chemostat inlet gas.

Figure 6 – Scheme of pathways involved in lipid metabolism

Schematic overview of the discussed pathways involved in lipid metabolism. Colour code: red = upregulated genes; green = downregulated genes; and blue = un-regulated genes under hypoxic conditions (FC ≥ 1.5 and $p \leq 0.05$) A: Sphingolipid metabolism in the yeast *P. pastoris* adapted from [55]. MIPS = mannosyl-inositol-phosphorylceramide, GlcCer = glucosylceramides B: Outline of the post-squalene ergosterol biosynthetic pathway, dashed arrows indicate no specification of intermediates

Tables

Table 1 – Microarray statistics

Microarray statistics including all genes that passed the adjusted *p*-value cut-off ≤ 0.05. Comparisons between oxygen set-points within one strain (e.g. C 8/21) or between strains at a certain oxygen set-point (E/C 8) are given. C = control strain; E = expressing strain

3904 annotated sequences, *p* ≥ 0.05, no FC				
comparison	**threshold passed**	**up**	**down**	**% regulated**
C 8/21	656	357	299	16.80
C 11/21	504	261	243	12.91
C 8/11	4	2	2	0.10
E 8/21	649	342	307	16.62
E 11/21	349	172	177	8.94
E 8/11	157	62	95	4.02
E/C 8	6	4	2	0.15
E/C 11	0	0	0	0.00
E/C 21	3	2	1	0.08

Table 2 – Strain-dependent gene regulation

List of differently regulated genes ($p \leq 0.05$ and FC > 1.5) in both strains under hypoxic conditions. E/C 8 up corresponds to the upregulated genes in the Fab expressing strain compared to the control, while the downregulated genes are referred to as E/C 8 down.

Gene Name	***p*-value**	FC	Description
E/C 8 up			
NCE103	2.39E-02	2.27	***Carbonic anhydrase, involved in a non-classical protein export pathway***
PFK3	1.06E-02	2.53	***Pichia pastoris 6-phosphofructokinase gamma-subunit***
ERG25	3.29E-03	2.79	***Required in the ergosterol biosynthesis pathway***
AQR1	3.99E-02	3.18	***Multidrug transporter of the major facilitator superfamily***
E/C 8 down			
YCT1	3.20E-04	- 9.23	***High-affinity cysteine-specific transporter***
FLR1	4.08E-02	- 2.23	***Plasma membrane multidrug transporter***

Table 3 – Metabolic flux ratio (METAFoR) analysis results

Origins of metabolic intermediates during growth of *P. pastoris* in glucose-limited chemostat ^{13}C-labelled cultures at D = 0.1 h^{-1} at different fractions of oxygen in the chemostat inlet gas. Values for both control and Fab producing strains are given. Ratios highlighted in grey have been used as constraints for metabolic flux analysis (n.a., not applicable; n.d., not determined. See main text for explanation).

% Fraction of total pool	Expressing Strain			Control Strain		
	21% O_2	11% O_2	8% O_2	21% O_2	11% O_2	8% O_2
Pep from PPs, upper bound	50 ± 9	23 ± 6	15 ± 7	39 ± 9	32 ± 8	15 ± 6
R5P from T3P and S7P (transketolase)	71 ± 2	78 ± 2	70 ± 2	66 ± 2	70 ± 2	62 ± 2
R5P from E4P (transaldolase)	44 ± 2	24 ± 2	23 ± 2	40 ± 2	29 ± 2	24 ± 2
Ser originating from Gly and C1-unit	61 ± 4	68 ± 4	68 ± 4	62 ± 4	69 ± 4	72 ± 4
Gly originating from CO2 and C1-unit	10 ± 4	13 ± 3	13 ± 3	6 ± 4	12 ± 3	10 ± 3
PEP originating from OAA-cyt (PEPck)	0 ± 4	0 ± 8	0 ± 10	2 ± 5	0 ± 10	5 ± 11
OAA-mit originating from PEP	44 ± 2	32 ± 2	44 ± 2	42 ± 2	35 ± 2	41 ± 2
OAA-mit originating from OAA-cyt	na	43 ± 3	55 ± 3	na	44 ± 3	51 ± 3
OAA-cyt originating from PEP	na	63 ± 3	64 ± 4	na	66 ± 3	66 ± 4
OAA-cyt reversibly converted to FUM	63 ± 11	7 ± 5	10 ± 5	63 ± 11	11 ± 4	9 ± 4
Flux through malic enzyme, upper bound	1 ± 4	nd	nd	1 ± 6	nd	nd
Flux through malic enzyme, lower bound	1 ± 2	nd	nd	0 ± 3	nd	nd

Table 4 – qRT-PCR results

Quantitative real-time PCR results compared with microarray data. Standard deviations derive from triplicate measurements; all numbers reflect relative gene expression to a reference gene (*ACT1*).

Gene product	accession nr	8/21 E Arrays	p-value	8/21 E qPCR	stdev
MDH1	PIPA02244	**-1.61**	4.90E-03	**-2.03**	± 0.212
FUM1	PIPA02844	**-2.07**	2.23E-02	**-1.42**	± 0.121
YDL124W	PIPA01263	**2.58**	3.00E-04	**1.52**	± 0.139
RKI1	PIPA02895	**2.44**	2.50E-03	**1.93**	± 0.091
CDC19	PIPA00751	**5.24**	8.24E-10	**3.71**	± 0.189
TDH3	PIPA02510	**4.34**	5.44E-06	**2.03**	± 0.151
Fab Hc		**1.65**	8.80E-02	**1.68**	± 0.064
Fab Lc		**2.07**	8.30E-05	**1.80**	± 0.019
		E/C 8 Arrays	**p-value**	**E/C 8 qPCR**	**stdev**
NCE103	PIPA03864	**2.21**	2.30E-02	**1.71**	± 0.087
PFK3	PIPA09969	**2.53**	1.20E-02	**3.21**	± 0.128
ERG25	PIPA00945	**2.79**	3.00E-03	**2.21**	± 0.093
AQR1	PIPA04502	**3.18**	3.90E-02	**4.12**	± 0.188
YCT1	PIPA00376	**-9.19**	3.90E-04	**-5.03**	± 0.322
FLR1	PIPA02458	**-2.22**	4.10E-02	**-1.78**	± 0.161
Fab Hc		**2.96**	1.75E-01	**10.31**	± 0.397

Additional files

Additional file 1 – Design 2D DIGE Gels

An example of the experimental design for the acquisition of statistical data on differences between untreated (normoxia (20.9 %)) and treated (oxygen limitation (10.9 %) and hypoxia (8.4 %)) samples labelled with Cy Dyes (GE Healthcare).

Additional file 2 – 2D DIGE Gels

Representative gel image from a 2D gel electrophoresis experiment with proteins obtained from the Fab expressing strain. We identified 45 out of 81 proteins with a different expression pattern when comparing high and low oxygen experiments. Green spots show proteins downregulated and pink ones show those upregulated under hypoxia.

Additional file 3 – Identified protein spots

List of the 45 identified proteins with different abundances comparing normoxic and hypoxic conditions in the *P. pastoris* expressing and control strain. Proteins were identified by MALDI-TOF MS and grouped into 6 different biological processes. The protein name, short name and accession number, theoretical Mw and pI are reported together with the percentage of peptide coverage and number of identified peptides. Average ratios and 1-ANOVA (DeCyder) are given and only not indicated where no spot could be matched.

Additional file 4 – Metabolic fluxes

Metabolic fluxes in the central carbon metabolism of *P. pastoris* Fab-producing and control strain in glucose-limited chemostats at a D = 0.1 h^{-1}, in different oxygenation conditions. The standard deviations of each net flux are given.

Additional file 5 – Relative abundances of intact carbon fragments in proteinogenic amino acids

Relative abundances of intact C2 and C3 fragments (*f*-values) in proteinogenic amino acids describing the conservation of carbon chain fragments in *P. pastoris* Fab-producing and control strains growing in glucose-limited chemostats at a D = 0.1 h^{-1}, in different oxygenation conditions.

Additional file 6 – qRT-PCR primers used in this study.
Table showing the primer sequences of the genes analyzed by qRT-PCR and characteristics of the corresponding amplicons. Calculated copy numbers and the dilution factors in order to obtain 10^9 copies μl^{-1} are indicated (see main text for explanation).

Additional file 7 – Metabolic network model of the central carbon metabolism of *P. pastoris*
Bioreaction network model of the central carbon metabolism of *P. pastoris* used in the ^{13}C-metabolic flux analysis for the determination of net fluxes under the different oxygenation conditions. Fluxes are represented as net fluxes and the directions of the arrows indicate the directions of the positive net fluxes. The metabolites consumed or produced by extracellular fluxes (shown as dashed arrows) are denoted with (E).

Additional file 8 – Stoichiometric model of the central carbon metabolism of *P. pastoris*
Reactions in the stoichiometric model of the central carbon metabolism of *P. pastoris* applied in the ^{13}C-MFA determination of the metabolic fluxes under different oxygenation conditions; it also includes anabolic reactions from metabolic intermediates to biosynthesis, transport reactions across the mitochondrial membrane and uptake and excretion reactions. Note that O_2, CO_2, energy and redox cofactor mass balances were not included in the mass balance constraints in ^{13}C-MFA.

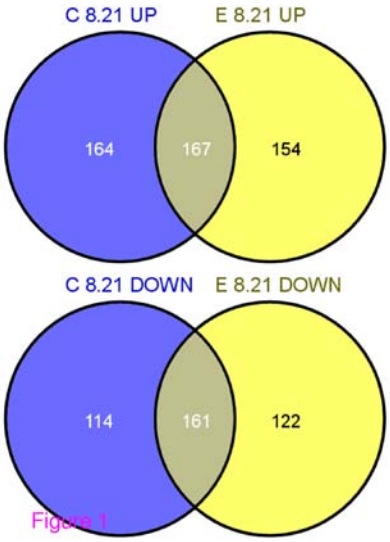


Figure 1

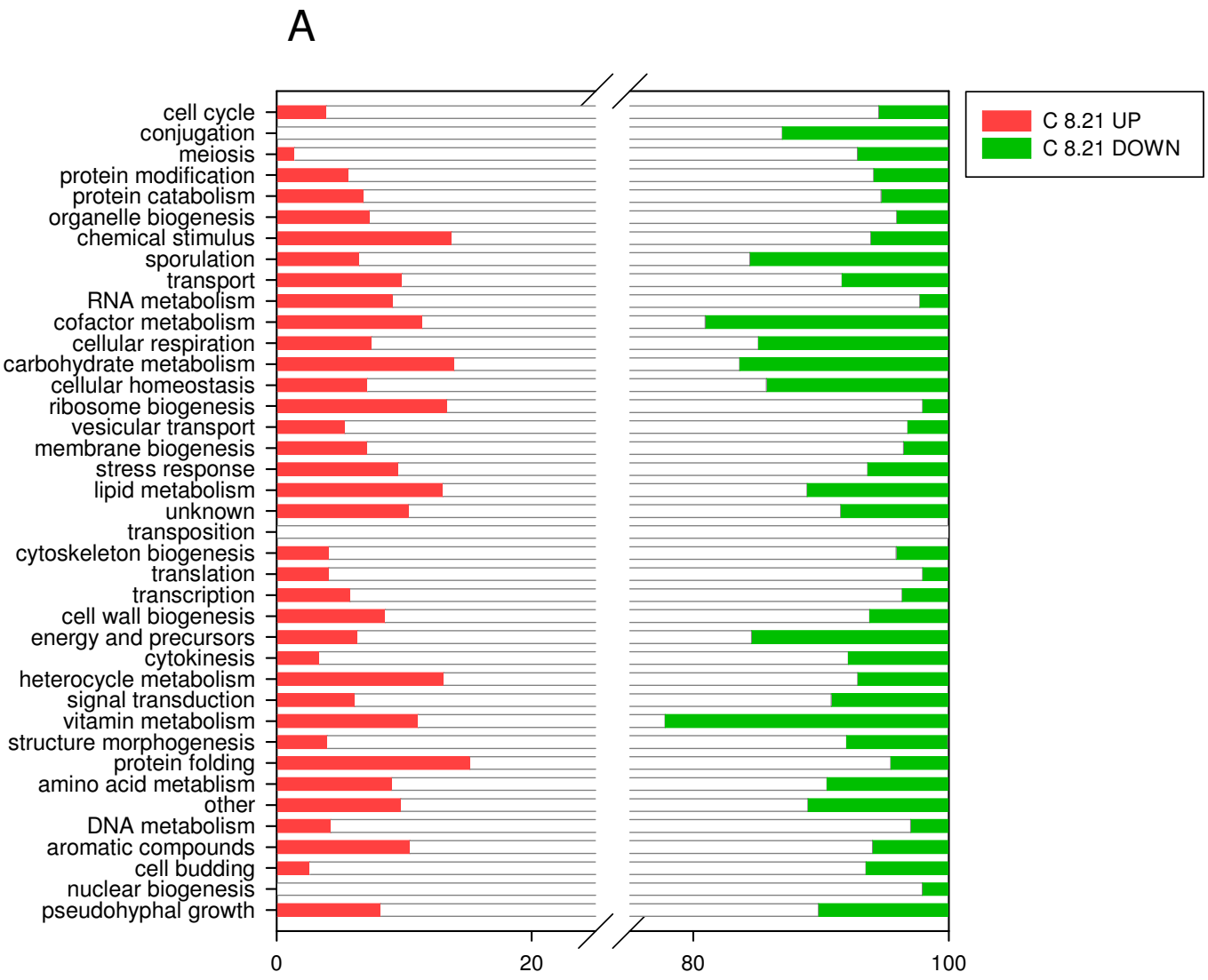

A
cell cycle
conjugation
meiosis
protein modification
protein catabolism
organelle biogenesis
chemical stimulus
sporulation
transport
RNA metabolism
cofactor metabolism
cellular respiration
carbohydrate metabolism
cellular homeostasis
ribosome biogenesis
vesicular transport
membrane biogenesis
stress response
lipid metabolism
unknown
transposition
cytoskeleton biogenesis
translation
transcription
cell wall biogenesis
energy and precursors
cytokinesis
heterocycle metabolism
signal transduction
vitamin metabolism
structure morphogenesis
protein folding
amino acid metablism
other
DNA metabolism
aromatic compounds
cell budding
nuclear biogenesis
pseudohyphal growth
0
20
80
100
C 8.21 UP
C 8.21 DOWN

Figure 2

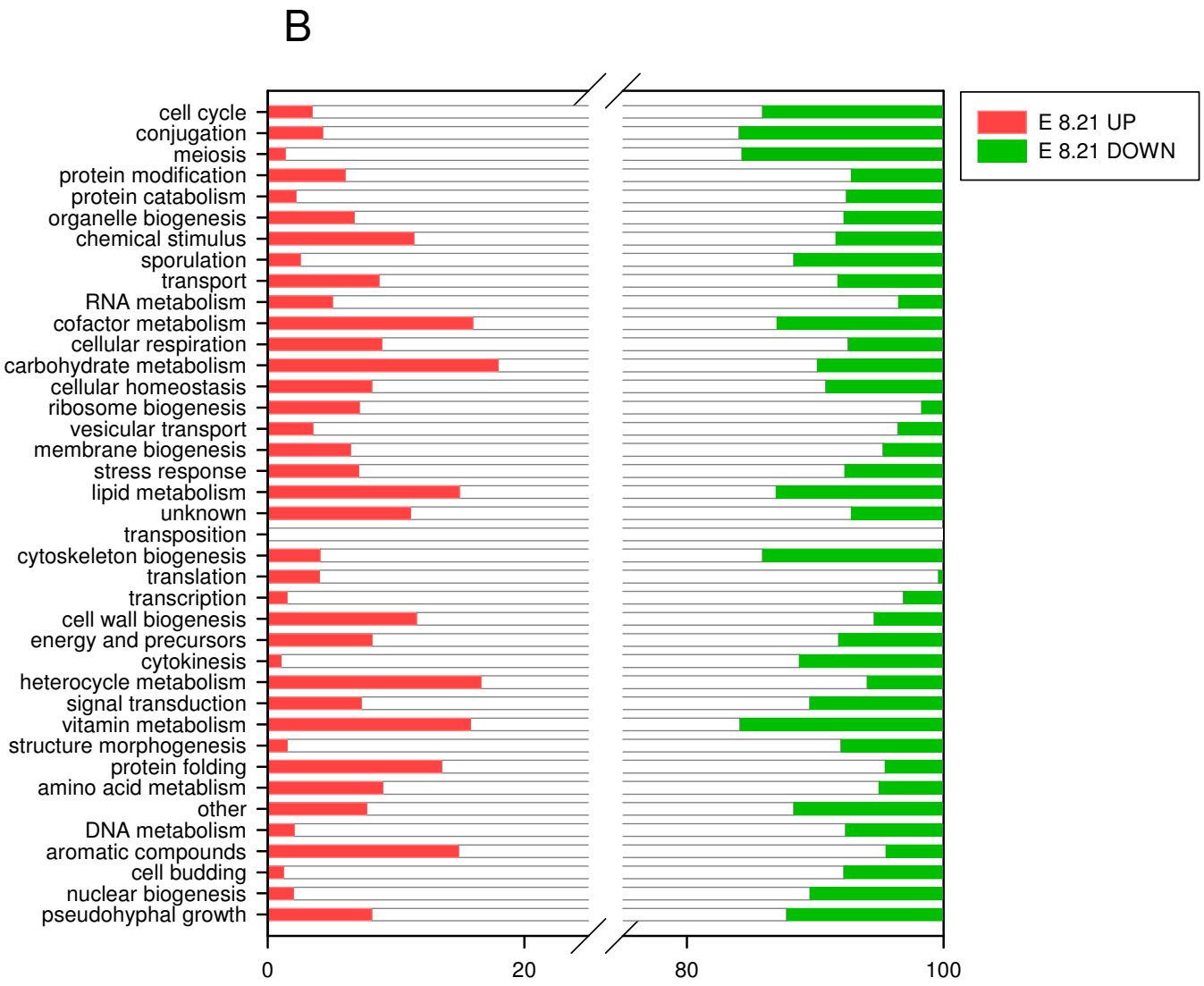

B
cell cycle
conjugation
meiosis
protein modification
protein catabolism
organelle biogenesis
chemical stimulus
sporulation
transport
RNA metabolism
cofactor metabolism
cellular respiration
carbohydrate metabolism
cellular homeostasis
ribosome biogenesis
vesicular transport
membrane biogenesis
stress response
lipid metabolism
unknown
transposition
cytoskeleton biogenesis
translation
transcription
cell wall biogenesis
energy and precursors
cytokinesis
heterocycle metabolism
signal transduction
vitamin metabolism
structure morphogenesis
protein folding
amino acid metablism
other
DNA metabolism
aromatic compounds
cell budding
nuclear biogenesis
pseudohyphal growth
0
20
80
100
E 8.21 UP
E 8.21 DOWN

Figure 3

A

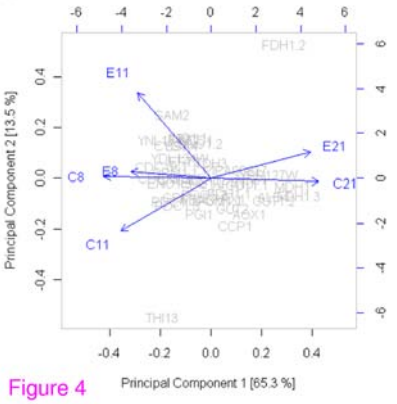

-6
-4
-2
0
2
4
6
FDH1.2
E11
SAM2
E21
C8
E8
C21
PGI1
CCP1
C11
THI13
-0.4
-0.2
0.0
0.2
0.4
Principal Component 2 [13.5 %]
Principal Component 1 [65.3 %]

Figure 4

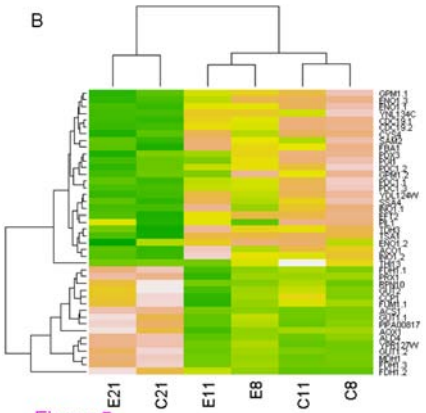


Figure 5

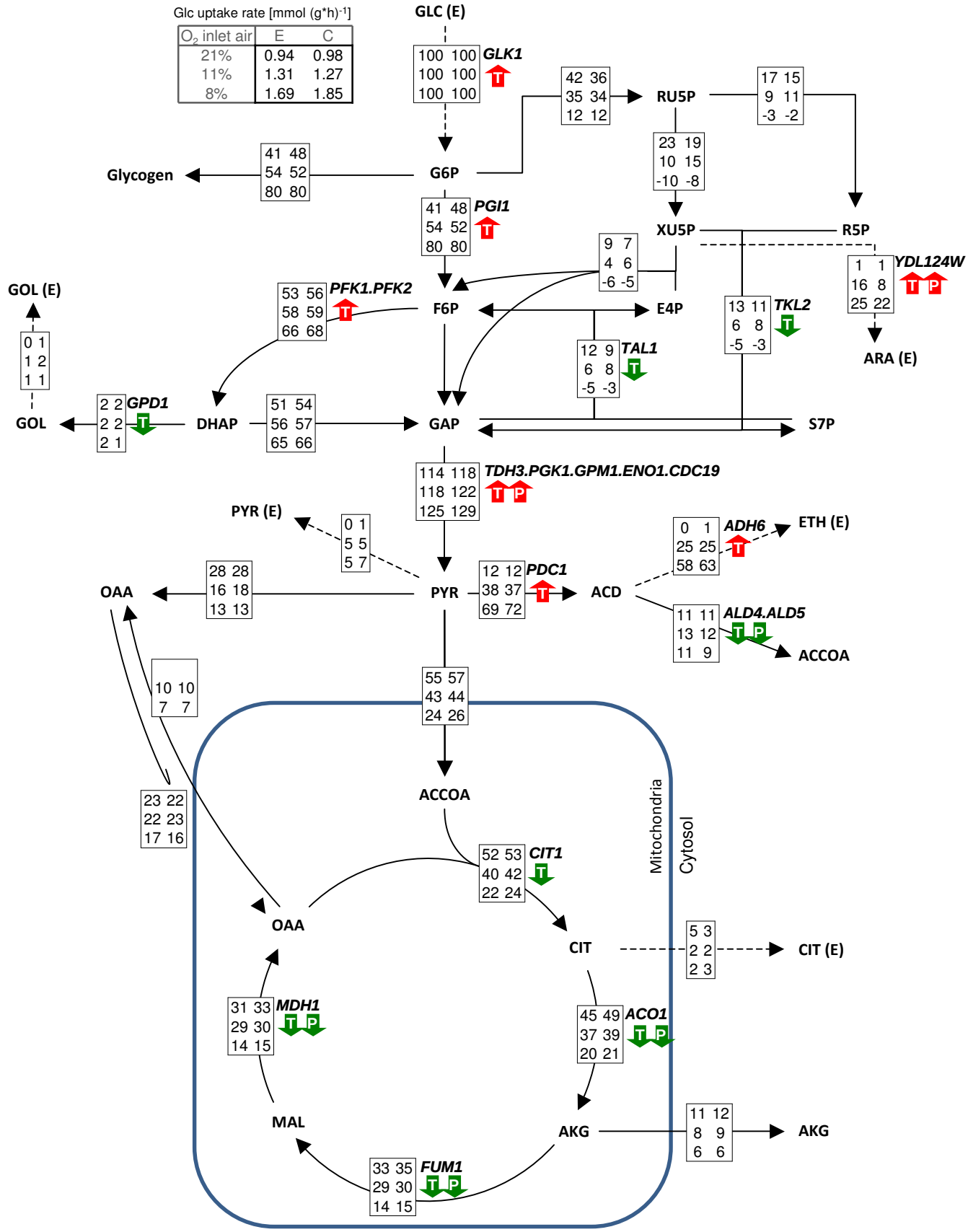

Glc uptake rate [mmol (g*h)-1]
O2 inlet air | E | C
21% | 0.94 | 0.98
11% | 1.31 | 1.27
8% | 1.69 | 1.85
GLC (E)
100 100
100 100
100 100
GLK1
T
42 36
35 34
12 12
RU5P
17 15
9 11
-3 -2
23 19
10 15
-10 -8
Glycogen
41 48
54 52
80 80
G6P
41 48
54 52
80 80
PGI1
T
XU5P
R5P
9 7
4 6
-6 -5
1 1
16 8
25 22
YDL124W
T P
ARA (E)
GOL (E)
53 56
58 59
66 68
PFK1.PFK2
T
F6P
E4P
13 11
6 8
-5 -3
TKL2
T
12 9
6 8
-5 -3
TAL1
T
0 1
1 2
1 1
GOL
2 2
2 2
2 1
GPD1
T
DHAP
51 54
56 57
65 66
GAP
S7P
114 118
118 122
125 129
TDH3.PGK1.GPM1.ENO1.CDC19
T P
PYR (E)
0 1
5 5
5 7
0 1
25 25
58 63
ADH6
T
ETH (E)
OAA
28 28
16 18
13 13
PYR
12 12
38 37
69 72
PDC1
T
ACD
11 11
13 12
11 9
ALD4.ALD5
T P
ACCOA
10 10
7 7
55 57
43 44
24 26
23 22
22 23
17 16
ACCOA
Mitochondria
Cytosol
52 53
40 42
22 24
CIT1
T
OAA
CIT
5 3
2 2
2 3
CIT (E)
31 33
29 30
14 15
MDH1
T P
45 49
37 39
20 21
ACO1
T P
MAL
AKG
11 12
8 9
6 6
AKG
33 35
29 30
14 15
FUM1
T P

Figure 6

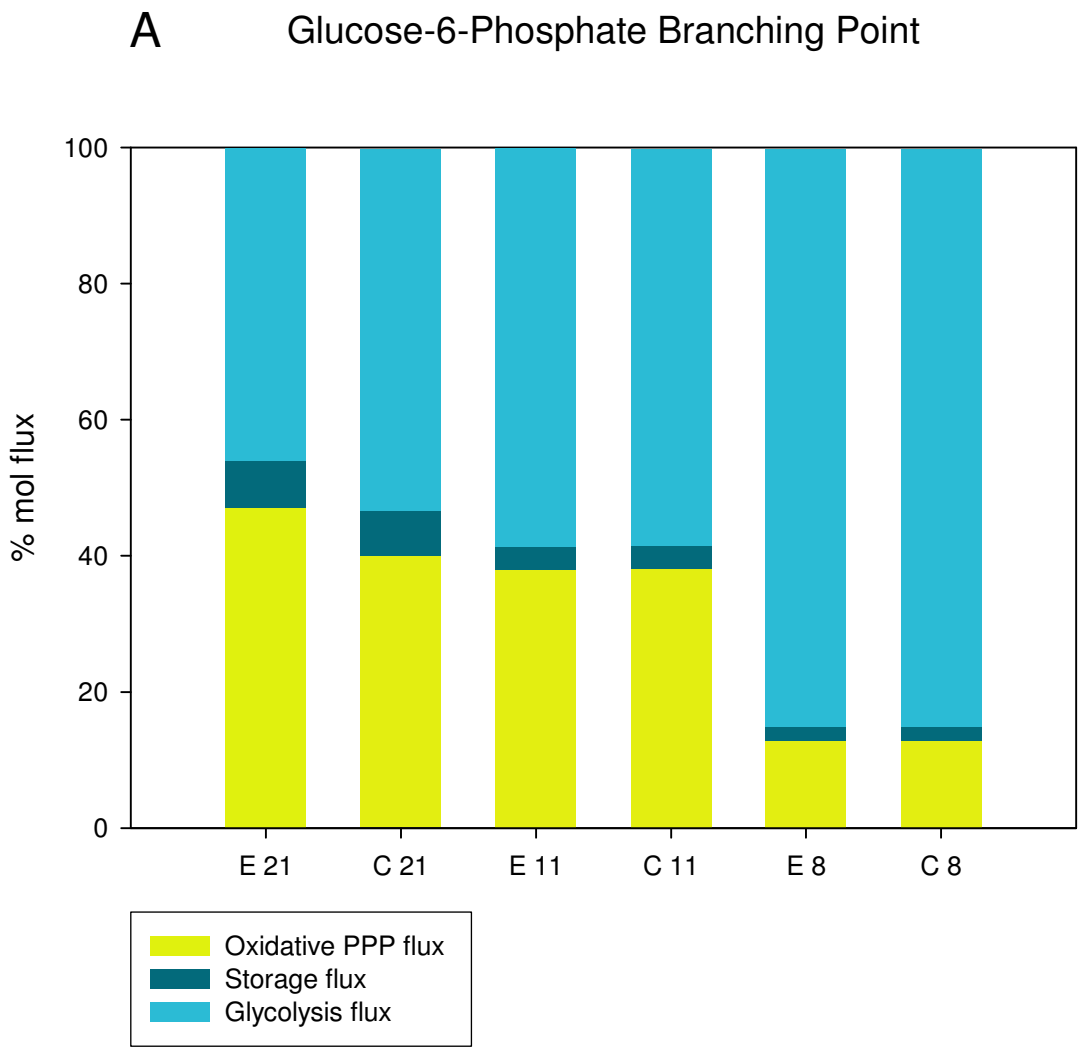

A
Glucose-6-Phosphate Branching Point
% mol flux
100
80
60
40
20
0
E 21
C 21
E 11
C 11
E 8
C 8
Oxidative PPP flux
Storage flux
Glycolysis flux

Figure 7

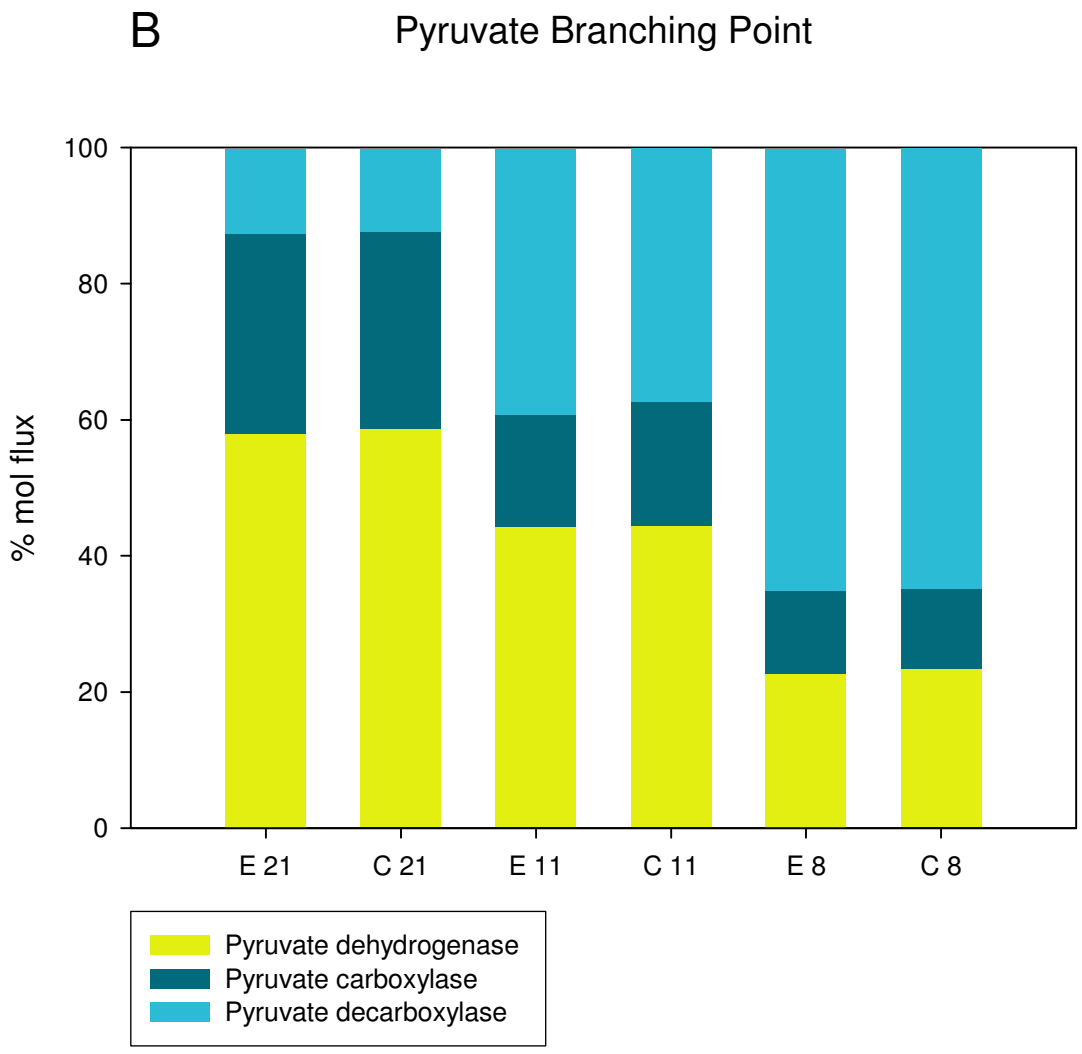

B
Pyruvate Branching Point
% mol flux
100
80
60
40
20
0
E 21
C 21
E 11
C 11
E 8
C 8
Pyruvate dehydrogenase
Pyruvate carboxylase
Pyruvate decarboxylase

Figure 8

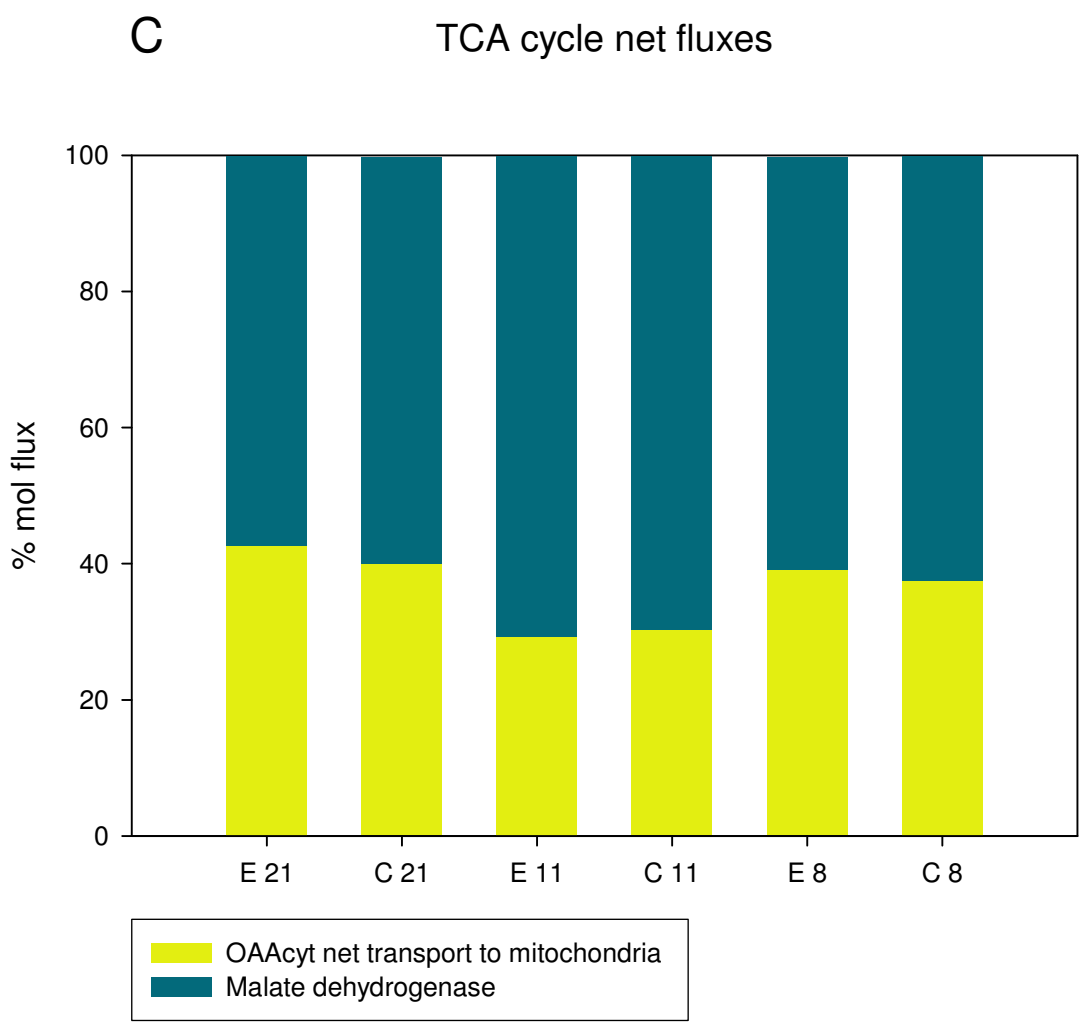

C
TCA cycle net fluxes
% mol flux
100
80
60
40
20
0
E 21
C 21
E 11
C 11
E 8
C 8
OAAcyt net transport to mitochondria
Malate dehydrogenase

Figure 9

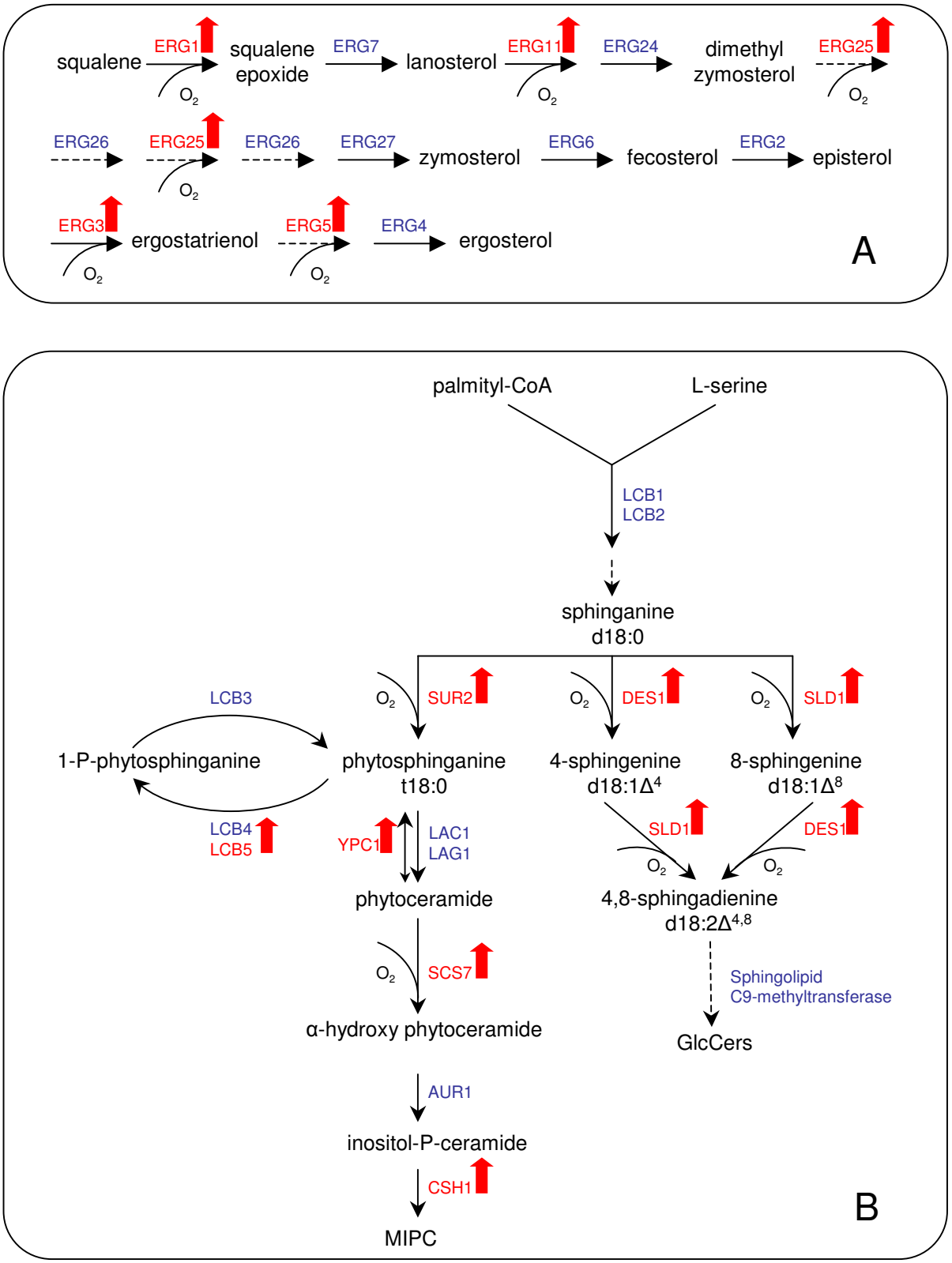

squalene
ERG1
O2
squalene epoxide
ERG7
lanosterol
ERG11
O2
ERG24
dimethyl zymosterol
ERG25
O2
ERG26
ERG25
O2
ERG26
ERG27
zymosterol
ERG6
fecosterol
ERG2
episterol
ERG3
O2
ergostatrienol
ERG5
O2
ERG4
ergosterol
A
palmityl-CoA
L-serine
LCB1
LCB2
sphinganine
d18:0
LCB3
O2
SUR2
O2
DES1
O2
SLD1
1-P-phytosphinganine
phytosphinganine
t18:0
4-sphingenine
d18:1Δ4
8-sphingenine
d18:1Δ8
LCB4
LCB5
YPC1
LAC1
LAG1
SLD1
O2
DES1
O2
phytoceramide
4,8-sphingadienine
d18:2Δ4,8
O2
SCS7
Sphingolipid
C9-methyltransferase
α-hydroxy phytoceramide
GlcCers
AUR1
inositol-P-ceramide
CSH1
MIPC
B

Figure 10

Additional files provided with this submission:

Additional file 1: Additional file 1.doc, 31K
http://www.biomedcentral.com/imedia/5259995693882010/supp1.doc
Additional file 2: Additional file 2.png, 950K
http://www.biomedcentral.com/imedia/8261575553882010/supp2.png
Additional file 3: Additional file 3.doc, 157K
http://www.biomedcentral.com/imedia/1471352014388201/supp3.doc
Additional file 4: Additional file 4.xls, 23K
http://www.biomedcentral.com/imedia/1310907439388201/supp4.xls
Additional file 5: Additional file 5.xls, 34K
http://www.biomedcentral.com/imedia/1788828927388201/supp5.xls
Additional file 6: Additional file 6.doc, 84K
http://www.biomedcentral.com/imedia/1605522227388201/supp6.doc
Additional file 7: Additional file 7.ppt, 93K
http://www.biomedcentral.com/imedia/1851943689388201/supp7.ppt
Additional file 8: Additional file 8.doc, 32K
http://www.biomedcentral.com/imedia/6759484388201014/supp8.doc

7 Curriculum Vitae

Alexandra Bettina Graf

Date of birth: 04. July 1974

Place of birth: Mödling, Austria

Education

2007 - 2010 PhD Thesis - *Towards Pichia pastoris Systems Biotechnology: Genome Sequence, Expression Microarrays, and Genome-Scale Metabolic Model*; University of Natural Resources and Life Sciences Vienna; Supervisor: Ao Prof. DI Dr Diethard Mattanovich.

2007 Diploma (DI (FH)) graduation with great distinction; Diploma Thesis - *Building a Bioinformatics Pipeline for the Design of Pichia pastoris Whole Genome Microarrays*; University for Applied Sciences FH-Campus Wien; Supervisor - Ao Prof. DI Dr Diethard Mattanovich.

2003 - 2007 Bioengineering course at the University for Applied Sciences FH-Campus Wien, Austria. Specialization in Bioinformatics.

1998 Diploma (MSc) graduation with great distinction; Diploma Thesis - *The Evolution of the Zonienwoud under Human Influence*; Free University Brussels; Supervisor - Prof. Nico Koedam.

1996 - 1998 Master program in Human Ecology at the Free University Brussels (VUB); Belgium.

1993 - 1997 First degree in Biology at the Uiversity of Vienna; Austria.

1988 - 1993 A-Levels at the Advanced Institute for Economic Studies, Biedermannsdorf, Austria.

Publications

Graf A, Dragosits M, Gasser B, Mattanovich D. *Yeast systems biotechnology for the production of heterologous proteins.* FEMS Yeast Res. 2009 May; 9(3):335-48

Graf A, Gasser B, Dragosits M, Sauer M, Leparc GG, Tüchler T, Kreil DP, Mattanovich D. *Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays.* BMC Genomics 2008, August **9**:390

Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, Mattanovich D, Lee SY. Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. Biotechnology Journal 2010, Jul;5(7):705-15.

Tüchler T, Velez G, Graf A, Kreil DP. *BibGlimpse: The case for a light-weight reprint manager in distributed literature research.* BMC Bioinformatics 2008, October **9**:406

Mattanovich D, Graf A, Stadlmann J, Dragosits M, Redl A, Maurer M, Kleinheinz M, Sauer M, Altmann F, Gasser B. *Genome, secretome and glucose transport highlight unique features of the protein production host. Pichia pastoris.* Microbial Cell Factories 2009, June **8**:29

Mattanovich D, Callewaert N, Rouzé P, Lin YC, Graf A, Redl A, Tiels P, Gasser B, De Schutter K. *Open access to sequence: Browsing the Pichia pastoris genome.* Microbial Cell Factories 2009, October **8**:53

Dragosits M, Stadlmann J, Graf A, Gasser B, Maurer M, Sauer M, Kreil DP, Altmann F, Mattanovich D. *The response to unfolded protein is involved in osmotolerance of Pichia pastoris.* BMC Genomics 2010, March **11**:207

Baumann K, Carnicer M, Dragosits M, Graf AB,Stadlmann J, Jouhten P, Maaheimo H, Gasser B, Albiol J, Mattanovich D, Ferrer P. *A multi-level study of recombinant Pichia pastoris in different oxygen conditions as knowledge base for strain improvement.* **(submitted)**

Baumann K, Branduardi P, Dato L, Dragosits M, Ferrer P, Frascotti G, Graf AB, Mattanovich D, Porro D. *The impact of oxygen on the transcriptome of recombinant S. cerevisiae and P. pastoris a comparative analysis.* **(in preparation)**

Work Experience

2001 - 2006 LEM Norma GmbH, A-2345 Brunn a. Geb., Austria; change to the IT Department with responsibilities in project management, network administration and help desk.

2000 - 2001 LEM Norma GmbH, A-2345 Brunn a. Geb., Austria; executive assistant, responsibilities in marketing projects.

1998 - 1999 Austropersonal, 1060 Vienna, Austria; Web designer

Language Skills

German: first language

English: excellent spoken and written (CAE: A)

French: basic