AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis dissertation.

Date

Signature

List of abbreviations

Acetyl-CoA	acetyl-coenzyme A
GFP	green fluorescent protein
Nc	effective number of codons
ACS **(Acs)**	acetyl-CoA synthetase
ALD **(Ald)**	aldehyde dehydrogenase
C. glabrata	*Candida glabrata*
CDC	codon deviation coefficient
E. coli	*Escherichia coli*
FBA	flux balance analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
PDC **(Pdc)**	pyruvate decarboxylase
PDH bypass	pyruvate dehydrogenase bypass
RSCU	relative synonymous codon usage
S. cerevisiae	*Saccharomyces cerevisiae*
SNP	single nucleotide polymorphism
S. passalidarum	*Spathaspora passalidarum*
TAG	triacylglycerol
tAI	tRNA adaptation index
TCA cycle	tricarboxylic acid cycle
Y. lipolytica	*Yarrowia lipolytica*

Abstract

Yarrowia lipolytica is an oleaginous budding yeast of biotechnological interest, for instance for the production of single cell oil and organic acids.

Under nitrogen-limited growth conditions, this yeast accumulates high amounts of storage lipid in the form of triacylglycerol and simultaneously excretes citrate as byproduct. For processes aiming at high lipid yields, citrate has to be seen as an undesirable waste of carbon, which could otherwise be channeled towards lipid biosynthesis.

Thus, the flux through the pyruvate dehydrogenase bypass shall be increased, thereby reducing the overflow through the TCA cycle, which in turn would lead to a reduction in citrate secretion. Most importantly, the pyruvate dehydrogenase bypass generates cytosolic acetyl-CoA, which is the precursor for lipid synthesis. This short pathway, composed of a pyruvate decarboxylase, an aldehyde dehydrogenase and an acetyl–CoA synthetase, is essentially inactive in *Yarrowia lipolytica*. Thus, putatively highly active enzymes from other species shall be integrated and overexpressed. Those candidate genes are selected according to their codon usage bias assuming that a highly biased gene is evolutionary more optimized than a rather unbiased one.

Thus, a pyruvate decarboxylase deriving from *Candida glabrata*, a $NADP^+$-dependent aldehyde dehydrogenase from *Saccharomyces cerevisiae* as well as an acetyl-CoA synthetase from *Spathaspora passalidarum* have been selected and suitable expression cassettes for heterologous expression in *Yarrowia lipolytica* have been designed.

Keywords : *Yarrowia lipolytica*, pyruvate dehydrogenase bypass, codon usage bias

Zusammenfassung

Die Sprosshefe *Yarrowia lipolytica* zeichnet sich nicht nur durch die Fähigkeit aus, Fette als Substrate verwerten zu können, sondern auch, diese in Form von Triacylglyceriden (TAG) in großen Mengen in Lipidtröpfchen innerhalb der Zelle einzulagern (20 % der Biomasse und mehr). Diese Eigenschaft ist es, die *Yarrowia lipolytica* interessant für die biotechnologische Produktion von lipid-basierenden Chemikalien macht, wie zum Beispiel Biodiesel.

Unter Stickstoffmangel lagert *Yarrowia lipolytica* TAG ein, produziert aber gleichzeitig auch Citrat, welches als Nebenprodukt sezerniert wird. Aus ökonomischer Sicht stellt dies einen Verlust von Kohlenstoff dar, der auch für die Produktion von Lipiden hätte verwendet werden können. Daher soll der Reaktions-Fluss durch den Pyruvat Dehydrogenase (PDH) Bypass gesteigert werden, wodurch der Überfluss des Krebs-Zyklus gesenkt und die Citratsekretion gemindert werden soll. Abgesehen davon wird durch den Pyruvate Dehydrogenase Bypass cytosolisches Acetyl-CoA gebildet, welches der Grundbaustein für die Lipidbiosynthese ist. Dieser kurze Stoffwechselpfad, bestehend aus einer Pyruvat Decarboxylase, einer Aldehyd Dehydrogenase und einer Acetyl-CoA Synthetase, ist in *Yarrowia lipolytica* nahezu inaktiv. Daher sollen hochaktive Enzyme aus anderen Spezies in *Yarrowia lipolytica* integriert und überexprimiert werden.

Die entsprechenden Kandidaten-Gene wurden auf Basis ihres *Codon Usage Bias* (Präferenz für bestimmte synonyme Codons) unter der Annahme ausgewählt, dass ein hoher Grad an *Codon Usage Bias* auf ein hohes Maß an evolutionärer Optimierung hinweist. Schließlich wurden eine Pyruvat Decarboxylase von *Candida glabrata*, eine $NADP^+$ abhängige Aldehyd Dehydrogenase aus *Saccharomyces cerevisiae* und eine Acetyl-CoA Synthetase aus *Spathaspora passalidarum* ausgewählt und passende Expressionskassetten designed.

Stichworte: *Yarrowia lipolytica*, Pyruvat Dehydrogenase Bypass, Codon Usage Bias

Contents

1 Introduction **1**

1.1 *Yarrowia lipolytica*, an unconventional yeast 1

1.2 A suitable cell factory for lipid-based compounds 2

1.3 Codon usage bias 6

1.3.1 Causes and consequences of codon usage bias 6

1.3.2 Quantifying codon usage bias 9

1.3.3 tAI- the tRNA site of bias 11

1.3.4 Effective number of codons (Nc) 13

1.3.5 Codon deviation coefficient 13

2 Hypothesis and Aim **15**

3 Materials and Methods **16**

3.1 *In silico* analysis 16

3.1.1 Flux balance analysis 16

3.1.2 Calculation of codon usage bias 17

3.2 Cloning 18

3.2.1 Isolation of genomic DNA 19

3.2.2 Polymerase chain reaction (PCR) 19

3.2.3 Overlap-extension PCR 23

3.2.4 Restriction digest 23

3.2.5 Agarose gel electrophoresis 24

3.2.6 Purification of PCR products 25

3.2.7 Plasmid isolation 25

3.2.8 Gibson assembly 25

3.2.9 *Escherichia coli* heat shock transformation 27

4 Results **28**

4.1 Flux balance analysis reveals that the pyruvate dehydrogenase bypass is essentially inactive in *Yarrowia lipolytica* 28

4.2 Implementation of codon usage bias and flux analysis in Python 30

4.3 Codon usage bias reflects physiology 35

4.4 Codon usage bias as estimate of reaction flux values is limited to orthologous genes 37

4.5 Finding suitable codon-biased candidate genes 40

4.5.1 The pyruvate decarboxylase (Pdc1) of *Candida glabrata* shows a strong bias along with other fermentative yeasts 40

4.5.2 The aldehyde dehydrogenase (Ald6) of *Saccharomyces cerevisiae* is $NADP^+$ dependent and highly biased 42

4.5.3 *Spathaspora passalidarum´s ACS1* is highly biased but the localisation of the enzyme is unclear 43

4.6 The tRNA pool of *Yarrowia lipolytica* differs significantly from other *Saccharomycetales* 49

4.7 Construction of expression cassettes 50

5 Discussion **55**

5.1 The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a source of orthologous enzymes, but can be misleading 55

5.2 A selection process based on different codon usage bias indices - the more biased the better ? 56

5.3 The cloning strategy aims to optimize the potential yield, not to verify the impact of codon usage bias on reaction flux 59

5.4 Potential side-effects of enhancing the PDH bypass 59

5.5 Outlook 61

6 Bibliography **62**

7 Appendix **72**

7.1 Composition of Buffer and Media 72

7.2 Devices and Materials 75

1 Introduction

This work deals with metabolic engineering of *Yarrowia lipolytica's* pyruvate dehydrogenase bypass. This pathway, which is composed of a pyruvate decarboxylase, an acetaldehyde dehydrogenase and an acetyl-CoA synthetase, provides a source for cytosolic acetyl-CoA produced from pyruvate. By integrating orthologous enzymes from other species according to their the codon usage bias, the flux through this pathway shall be increased to enhance the cytosolic acetyl-CoA pool for improved lipid production.
The following paragraphs will give an introduction to *Yarrowia lipolytica* as cell factory for lipid-based chemicals by means of metabolic engineering, as well as about the potential of codon usage bias for detecting evolutionary optimized enzymes.

1.1 *Yarrowia lipolytica*, an unconventional yeast

The kingdom of fungi unites different eukaryotic living forms, heterogenous in their growth behaviour and cultivation. One of them are yeasts, still being a phylogenetic diverse group composed of two subphyla : The Ascomycota and Basidiomycota. Within the first group, *Yarrowia lipolytica* belongs to the group of budding yeasts alongside with *Saccharomyces* and *Candida* species. It occurs as food contaminant in dairy products but also in nature: Thus, it can be also isolated from oil contaminated soil, which is rich in hydrocarbon species, as it is able to feed from hydrocarbons as sole carbon source, hence the name 'lipolytica'. Furthermore, it also grows well on glycerol and glucose, whereas it is not able to utilize sucrose as it lacks invertase [1]. In contrast to *Saccharomyces cerevisiae*, it is an obligate aerobic yeast. Notably, growth and cell physiology is heavily influenced by environmental conditions such as nitrogen and carbon source, oxygen concentration, pH and temperature. For instance, *Yarrowia lipolytica* forms mycelia cells in neutral pH, whereas it grows as budding yeast at acidic pHs [1]. Furthermore the secretion of citric acid during cultivation is greatly

dependent on nitrogen availability [2, 3]. Given that optimal growth temperatures are below 35 °C and oxygen is indispensable, it is considered apathogenic, which makes it a convenient laboratory yeast e.g. for studying peroxisome and lipid droplet biogenesis [4]. From a molecular biological perspective, the handling was further eased by the elucidation of the genome sequence of CLIB122 which was published 2004 alongside other frequently applied yeasts in Nature [5]. Its genome of 20.5 Mbp distributed over six chromosomes is bigger, than that of *Saccharomyces cerevisiae*. The authors also found out that the amino acid similarity of orthologous proteins compared to other budding yeasts is only approximately 50 %, fostering the high genetic diversity among yeasts [5]. The corresponding 47.9 kbp mitochondrial DNA sequence of this strain - CLIB122 - was already published by Kerscher *et al.* in 2001 [6]. A second strain, PO1f, a derivate of (CLIB89) was sequenced in 2015 by Liu and Alper [7]. The genome of PO1f is highly similar to CLIB122 with the exception of a loss of four ORFs encoding for proteins putatively involved in strand repair and recombination. Alongside these sequencing efforts, protocols for genetic manipulation have been developed, e.g. expression plasmids [8], hybrid promotors [9] and synthetic terminators [10] for effective overproduction of proteins, making *Yarrowia lipolytica* a suitable model organism. Although *Saccharomyces cerevisiae* is probably the most popular yeast model organism for cell biology in general, *Yarrowia lipolytica* seems to be predestined as a model organism for research on lipid metabolism, as *Yarrowia lipolytica* does not only feed on hydrocarbons, but also stores substantial amounts of lipids as triacylglycerol (TAG) in cytosolic lipid bodies.

1.2 A suitable cell factory for lipid-based compounds

There are many different oleaginous species from different kingdoms, which would be obvious choices for lipid production hosts. Although several plant species are able to accumulate large amounts of different lipids in their vacuoles, their cultivation is not as convenient as a microbial based production system. Albeit bacterial production systems would clearly be best considering economic feasibility, it is noteworthy that in most bacterial species, the lipid composition is different from eukaryotes. Mostly, polyhydroxyalkanoic acids take the place of TAG as internal carbon stock, although

there are some actinobacteria such as *Streptomyces*, that store excess carbon in the form of TAG [11]. Therefore, yeasts seem to be a more promising production system, as the composition of the lipids resembles the need of the food and biodiesel industry. As already mentioned, *Yarrowia lipolytica* is an attractive production platform for lipid-based compounds and chemicals. Depending on the cultivation conditions such as nature of nitrogen and carbon sources, lipids (mainly TAG) stored in lipid bodies can make up 20 % of cell dry weight (biomass) and even more, hence *Yarrowia lipolytica* is also called an oleaginous organism. On the biochemical level, one of the difference between oleaginous yeasts and non-oleaginous yeasts (such as *Saccharomyces cerevisiae*) is that their is twice as much citrate present in the mitochondria of oleaginous yeasts than in non-oleaginous ones. Citrate is further transported into the cytoplasm, where it is converted into oxaloacetate and acetyl-CoA by the ATP-citrate lyase - an enzyme absent in non-oleaginous yeasts [12].

There is a range of industrially attractive lipid-based compounds such as waxes, polyketides, steroids and also biodiesel (fatty acid methyl ester), which are targets for metabolic engineering [13]. Due to this economic value many strategies have been applied to further increase the intracellular level of lipids (Figure 1.1).

On the one hand, lipid content can be elevated by bioprocess engineering, and on the other by metabolic engineering:

As mentioned before, *Yarrowia lipolytica* is amenable to changes in the environment such as temperature, nitrogen and oxygen supply. It is important to note that lipid accumulation does not happen under optimal conditions, but under non-optimal conditions, hence under nitrogen limitation and simultaneous carbon excess. Since nitrogen is lacking for protein biosynthesis and growth, carbon is channelled towards fatty acid biosynthesis. Therefore, cultivation processes need to be well balanced between nitrogen limitation and growth conditions in order to optimize lipid yield (M. Kavščček, personal communication). In contrast to changing cultivation parameters, internal flux distributions can be redirected by metabolic engineering, for instance by knocking out competing pathways or by overexpression or heterologous expression of certain genes in order to channel carbon flux towards lipid production (Figure 1.1).

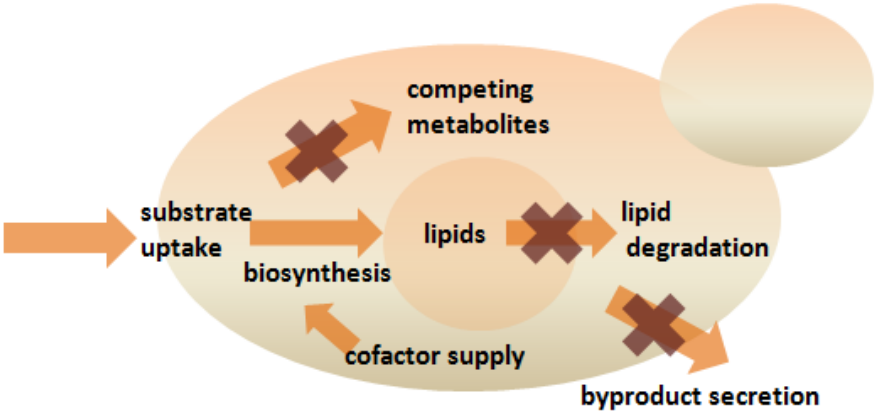


Fig. 1.1 – Different metabolic engineering strategies for lipid overproduction in *Yarrowia lipolytica*. Often, genetic interventions such as deletion or (heterologous) overexpression of genes are combined.

Since metabolic engineering requires in-depth knowledge of the interplay of certain pathways, it is often difficult to find suitable metabolic engineering targets. To support this process, genome scale models, stoichiometric matrices of reactions and metabolites within a cell, linked to the respective genes, have been applied. Two genome-scale models of *Yarrowia lipolytica* have been published, both in 2012. Pan *et al.* [14] reconstructed a network, iYL619_PCP, which accounts for 619 genes, 843 metabolites and 1,142 reactions. It predicts the ability to grow on a range of substrates, while being incapable of predicting dynamic growth, like exponential and stationary phase transition. iNL895, on the other hand, is based on a reconstruction of *Saccharomyces cerevisiae* and adapted according to *Yarrowia lipolytica* CLIB122 [15] genome annotation. Similarly, this model is able to correctly predict growth on many substrates, focussing on the flux of fatty acid related reactions. In addition to those models, another genome-scale network reconstruction has been built, which is able to predict the stationary growth phase correctly using dynamic FBA (M. Kavšček, personal communication). As mentioned so far, there are different strategies to achieve a high lipid overproduction by genetic interventions. Some of which are illustrated in Figure 1.1 and listed in Table 1.1.

Table 1.1 – Metabolic engineering strategies

Authors	Genetic interventions	Rationale
Tai *et al.* [16]	*ACC1* and *DGA1* overexpression	increasing biosynthesis
Qiao *et al.*. [17]	Δ*scd*, overexpression of *ACC1* and *DGA1*	increasing biosynthesis deletion of feedback inhibitor
Blazek *et al.* [18]	*AMPD*,*ACL*,*MAE*,*DGA1* and *DGA2* overproduction, Δ*pex10*,Δ*mfe1*	increasing biosynthesis reducing lipid degradation
Lazar *et al.* [19]	*HXK1* overexpression, overexpression of *SUC2* from *S.cerevisiae* in Δ*pox1-6* Δ*tgl4*, Δ*dga2* and Δ*gpd1* overexpression	improved substrate uptake consumption of new substrates improved lipid biosynthesis, reducing lipid degradation
Zhao *et al.* [20]	overexpression of *INU1* from *Kluyveromyces marxianus* overexpression	uptake of new substrates
Beopoulos *et al.* [21]	Δ*gut2*,Δ*pox1-6*	reducing lipid degradation reduction of competing metabolites
Wang *et al.* [22]	overexpression of *PYC* from *Pichia guilliermondii*, *ACL* overexpression	reduction of byproduct secretion increase of lipid biosynthesis
Wang *et al.* [23]	heterologous overexpression of *XYNs*	uptake of new substrates

Abbreviations : *ACC* : acetyl-CoA carboxylase, *DGA* : diacylglyceroacyltransferase, *SCD* : delta-9stearoyl-CoA desaturase, *ACL*: ATP-citrate lyase, *AMPD* : adenosine monophosphate deaminase, *MAE*: malic enzyme, *PEX* : peroxisome biogenesis factor, *MFE* : hydroxyacyl-CoA dehydrogenase, *HXK*: hexokinase, *SUC*: Invertase, *TGL4*: triacylglycerol lipase, *GPD* : glycerol-3-phosphate dehydrogenase, *INU* : exo-inulinase, *GUT* : glycerol-3-phosphate dehydrogenase, *POX* : acyl-coenzyme A oxidase, *PYC*: pyruvate carboxylase, *XYN* : xylanase

1.3 Codon usage bias

Genetic information is encoded on chromosomal and extrachromosmal DNA. Most of the DNA is transcribed into different sorts of RNA, such as mRNA which is translated into proteins, as well as rRNA, tRNA and other small regulatory RNAs. The information for amino acids is encoded in from of nucleotide triplets, so-called codons. Excluding codons for termination of protein biosynthesis, 61 codons encode for 20 amino acids. Moreover, with the exception of methionine and tryptophan, amino acids are encoded by more than one codon and up to six so-called synonymous codons for arginine, serine and leucine. Due to this redundancy, it is also called 'degenerate genetic code'.

Figure 1.2 illustrates the genetic code, which is true for most of chromosomal DNA. It is however slightly different for extrachromosomal DNA, such as mitochondrial DNA. Some organisms also encode a variation of cysteine, namely selenocysteine. Another exception to this standard code is the CTG, a default codon for leucine, which encodes a serine in certain *Candida* species [24]. Another layer of complexity is added by the fact that not all codons have exact corresponding tRNA isoacceptors, but interact with several codons. Also, the number of tRNA genes is varying by codons and is greatly different even between strains within a species (*E. coli* K12 : 86, *E. coli* O157H7 : 101, according to the genomic tRNA database [1].

Finally, synonymous codons are not evenly distributed within genes and intergenic sequences, creating a *codon usage bias* with some synonymous codons being more frequently used than others.

1.3.1 Causes and consequences of codon usage bias

One widely accepted theory on the evolution of codon bias is the selection-mutation-drift theory. It states that inter- and intragenic DNA sequences are subject to not only selection, but also a mutational bias, which may counteract selection and even drift, as it happens that in (finite) populations, certain mutations get enriched while others

[1] `http://gtrnadb.ucsc.edu/`, as accessed on 2015-05-05

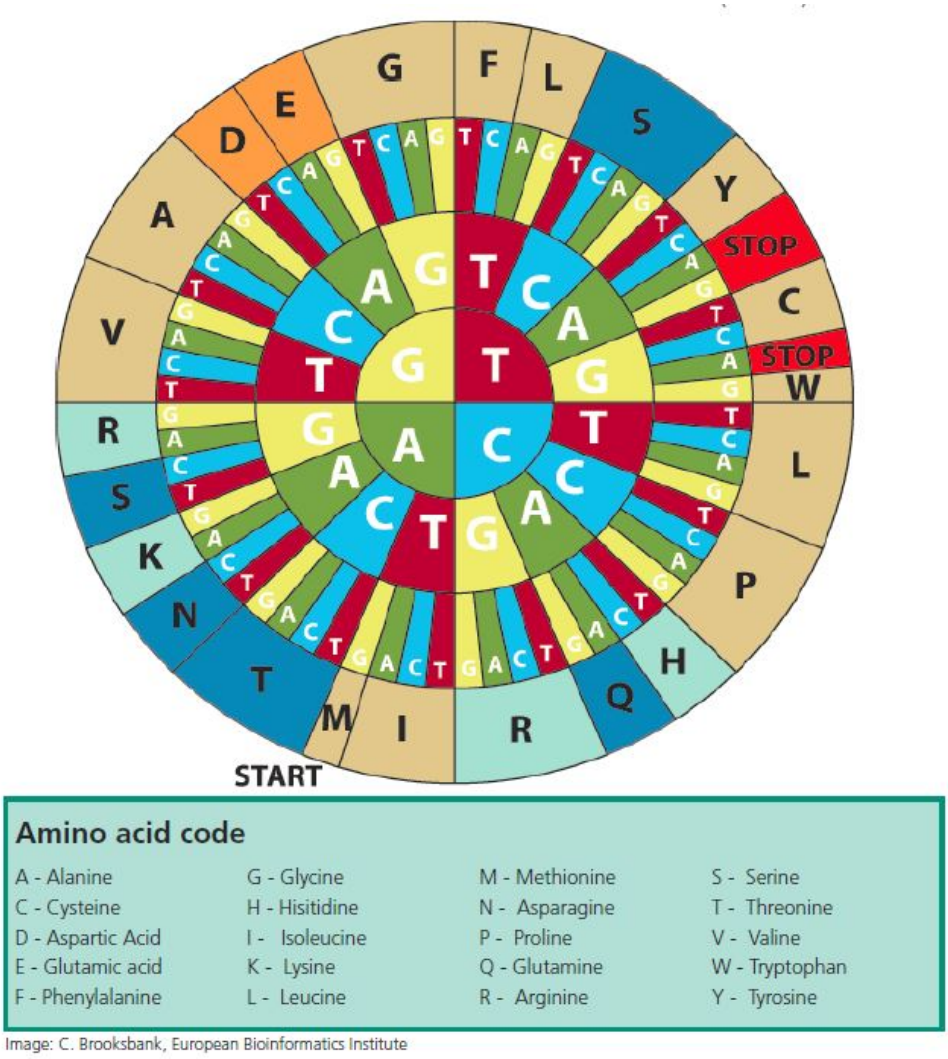


Fig. 1.2 – The genetic code is redundant and many amino acids are encoded by more than one (synonymous) codons. The wheel is read from the inside- out (5' to 3', so that the amino acid methionine is encoded by ATG.

are fading out [25]. Finally, those evolutionary forces taken together tailor genes to the respective needs, for example to the tRNA pool of a cell. Sharp *et al.* [26] for instance pointed out, that the bacterial generation time is correlated with the degree of bias.

Since mutations in synonymous codons do not change the amino acid sequence of a protein, they are called 'silent'. On the contrary, Zuckerkandl and Pauling [27] reasoned that molecules should be seen as *documents of molecular history* and that, furthermore, information from 'isosemantic heterozygosity' at the DNA level is lost when translated into proteins. Since the exchange of a nucleotide might alter the mRNA structure, synonymous single nucleotide polymorphisms (sSNPs) influence the interaction of mRNA and proteins during protein biogenesis. In eukaryotes, the mRNA needs to be processed and exported from the nucleus into the cytoplasm in order to get translated into a protein. During this process, the (pre-) mRNA molecule needs to interact with export proteins, small regulatory RNAs as well as with the splicing machinery. Many diseases have already been linked to sSNPs such as cystic fibrosis and phenylketonuria [28]. Moreover not only the mRNA maturation process is affected, but also its stability and half life as well as translation initiation, elongation efficiency and protein folding in general [29, 30, 31]. Considering that protein folding is a (partly) cotranslational process, the mRNA structure is a pacemaker for proper elongation and folding: Yang *et al.* developed a mathematical model of protein translation, predicting, that the higher the expression and conservation level of a protein is, the slower is the translation as a matter of accuracy [32]. As only a properly folded protein may act at its full activity, even synonymous changes have an impact on protein activity. Thus even synonymous mutations may lead to a reduced fitness of a mutant. Agashe *et al.* [33] varied the codon composition of the formaldehyde activating enzyme (*FAE*) of *Methylobacterium extorquens*, an essential enzyme for growth on methanol. When comparing the different mutants (for frequent or rare codons etc.), they found that even codon-optimized variants show lower specific activity and growth of the mutants, respectively. They found that the observed loss of fitness was not associated with the production of misfolded aggregates but rather with Fae of insufficient activity *per se* . Moreover they found, that once overexpressed, the growth rate was restored. A more systematic study on the effect of codon composition on fidelity of the protein has been

undertaken by Kudla *et al.* by which one codon after another was altered in the green fluorescent protein (GFP) and the fluorescence level was measured [34]. Surprisingly, they found a correlation of fluorescence with mRNA structure rather than with codon bias. Thus, they concluded that codon bias is not a determinant for protein expression levels.
This finding however contradicts the popular assumption that the adaptation of codon bias to frequent codons would increase the protein expression levels. This is especially important whenever proteins should be expressed heterologously in an industrial setting.

It is generally accepted, that codon usage bias influences the protein expression level. This is exploited in codon optimization for the recombinant expression of proteins. There is a myriad of success stories, by which the protein expression levels could be increased by adapting the codon composition of the recombinant gene to the expression host [35] which also goes in hand with computational tools being developed for this purpose [36]. Still, not all of these attempts are crowned with success, as already mentioned previously. Lanza *et al.* [37] suspected that some attempts of codon optimization fail, due to a shortcoming in the underlying rationale. As mentioned previously, codon usage is also biased towards the tRNA (gene) pool of an organism. This pool of tRNAs does change over time and adapts to the physiological needs during a cultivation, e.g. exponential and stationary phase. By developing a condition-dependent codon usage matrix, taking into account dynamic fluctuations of available tRNAs, they could increase the catalytic activity of the target enzyme, catechol 1,2-dioxygenase gene deriving from *Acinetobacter baylyi* expressed in *Saccharomyces cerevisiae*.

1.3.2 Quantifying codon usage bias

Since codon usage bias is of interest for both fundamental science (e.g. convergent evolution of parasitic/viral genomes to host genomes) and applied science, for instance for overexpression of recombinant proteins, many different indices are being developed to describe and quantify codon usage bias of a gene or a species [38]. Basically these methods can be divided into two different approaches:

On the one side, the definition of 'optimality' of a codon is based on the frequency in highly expressed genes. The most popular is the *Codon Adaptation Index (CAI)*. Closely related indices are *Codon Usage Bias Index (CBI)* and *Relative Codon Adaptation Index (rCAI)*, which differ slightly in the weighting of codon frequencies. Further, as genes are also evolved towards the tRNA pool, there exist also tRNA based indices, one of them is called *tRNA Adaptation Index (tAI)*. While the calculated values of this group of indices are well-funded on experimental data, it is also their biggest weakness. Since the values arise from comparison with a reference set, they become intrinsically biased by the choice of protein considered as reference. Friberg *et al.* were closer investigating this shortcoming by comparing the correlation with CAI, which is parametrized with different reference protein sets and protein expression data, which reached a maximum of 0.6 [39]. Its noteworthy that especially the choice of ribosomal proteins for reference might be highly misleading. Hershberg and Petrov [40] compared the preference of codons between ribosomal and non-ribosomal genes and found that in *Ehrlichia ruminantium* the favourite codon for alanine is GCT, whereas for non-ribosomal genes its GCA.

On the other side, other indices are based on statistical measures. They calculate the degree of deviation from an assumed uniform usage of synonymous codons within a gene, hence the degree of bias in the codon composition of a gene. One of the simplest is the so-called *relative synonymous codon usage* (RSCU), proposed by [41].

$$RSCU_{\mathrm{i,j}} = \frac{x_{\mathrm{i,j}}}{\frac{1}{n_{\mathrm{i}}} * \sum_{j=1}^{n_{\mathrm{i}}} x_{\mathrm{i,j}}} \tag{1.1}$$

RSCU is the ratio of number of occurrences of a codon within a sequence and the expected occurrence, when synonymous codons are used equally frequent. In Equation 1.1, $\mathrm{x_{i,j}}$ is the number of occurrences of the j^{th} codon for the i^{th} amino acid, which is encoded by $\mathrm{n_i}$ synonymous codons.

Those indices have been generally less applied than the former type of indices since the biological foundation is more difficult to understand as they do not make use of mRNA and protein abundance data. The advantage of those methods thus lies in their 'objectivity', as they are independent of experimental data and potential experimental

errors. Although these methods have proven to correlate with protein expression, Wang *et al.* [42] observed that different indices might give contradicting results. They compared a statistical method (Nc, which will be discussed in the following subsection) and a reference-based method, CAI, and tried to infer the "optimal" codons in bacterial genomes. However they found out, that for the amino acid alanine the most favoured codon is GCC in case the statistical method was applied and GCA for the CAI index. As discussed already, the use of ribosomal proteins to include in the reference set, was identified as the most likely reason for this discrepancy [40]. To overcome the weaknesses of the single indices applied all separately, O'Neill et al. developed an iterative algorithm for identification of 'optimal codons' and integrated both tAI and CAI to reflect mutational and translational bias [43].

In this work, the tAI, Nc and CDC indices were selected as indices of choice. The tAI was chosen because it belongs to the reference-based kind of indices, which offers a good estimate of bias arising from translational optimization. Furthermore this index does not make use of experimental data, but genomic sequence data only. To complement and maybe also catch transcriptional bias in the sequence of the gene, also two statistical indices, the *effective number of codon* (Nc) index, and the *coefficient of deviation* (CDC) index are chosen as their underlying formalism, are exclusive and not derivative to each other. These three indices will now be introduced in greater detail.

1.3.3 tAI- the tRNA site of bias

During translation, tRNA anticodons match with the respective codons on the mRNA to give rise to a polypeptide chain. To optimize this process, the codon composition is adapted to the tRNA pool, since lack of tRNAs may lead to unwelcome ribosome stalling or misfolded proteins. Ran *et al.* compared the genomes of 80 bacterial species and found that there is a positive correlation between copy number of tRNA-anticodons, overall codon bias and growth rate [44]. Translation *per se* is a complicated process given that cotranslational folding and translation must be well-timed. Zouh *et al.* for example observed that for some regions in the 3C protease of the foot-and-mouth disease virus (FMDV), less frequent tRNAs are used during the transition from

one structural 'unit' to another [45].

This correlation however is constrained by different factors. Firstly, tRNA modifications as catalysed by tRNA-dependent adenosine deaminases and tRNA-dependent uridine methyltransferases. Furthermore, wobble base at position 34 of the tRNA loop, that does not stick to Watson and Crick base-pairing [46]. Moreover tRNA expression levels at different growth phases [37]. There are different theories whether or not these features are positively contributing to translation fidelity [47]. One states, that that codon is preferred which is recognized by the most tRNA isoacceptors, thus by exact matches and 'wobble' matches. An opposing theory suggests that the 'best' codons are those with the most frequent exact isoacceptors. In detail, the formalism, as proposed by dos Reis *et al.* [48] is outlined below.

$$W_{\mathrm{i}} = \sum_{j=1}^{n_{\mathrm{i}}} (1 - s_{\mathrm{ij}}) * tGCN_{\mathrm{i,j}} \tag{1.2}$$

The absolute adaptiveness $\mathrm{W_i}$ (Equation 1.2) is defined by the sum of the number of tRNA gene copy number (tGCN) of the j^{th} tRNA that recognizes the i^{th} codon and a selectivity constraint $\mathrm{s_i}$, taking into account that due to the 34^{th} wobble position, also imperfect matches are accepted. From this absolute adaptiveness value $\mathrm{W_i}$, the relative adaptiveness is calculated by dividing this value by the maximal possible adaptiveness possible for this codon.

$$tAI_{\mathrm{g}} = (\prod (w_{\mathrm{i_{kg}}}))^{1/l_{\mathrm{g}}} \tag{1.3}$$

tAIg estimates the degree of adaptation of a gene g to a given genomic tRNA pool (Equation 1.3). It is the geometric mean of the relative adaptiveness values of its codons k of the gene g, whereas also the length l of the gene is taken into account.

1.3.4 Effective number of codons (Nc)

This index, called effective number of codons (Nc), was first proposed by Wright in 1990 [49]. It is a complex weighting system calculating the relative frequency of each codon, thereby also considering the *codon family*. For instance alanine is encoded by six synonymous codons (fourfold codon family), whereas aspartic acid by two (twofold). Since it was first published, it has been extended and some conceptional problems have been solved [50]. Its current version, published by Sun *et al.* [51] is presented in Equation 1.4. The more codon-biased a gene is, the smaller is the Nc value, with smallest possible value being 21 and the greatest 60, so that it correlates negatively with protein abundances.

$$N_{\mathrm{c}} = N_{\mathrm{s}} + \frac{K_2 \sum_j^{K_2} n_{\mathrm{j}}}{\sum_{j=1}^{K_2} (n_{\mathrm{j}} F_{\mathrm{CFj}})} + \frac{K_3 \sum_j^{K_3} n_{\mathrm{j}}}{\sum_{j=1}^{K_3} (n_{\mathrm{j}} F_{\mathrm{CFj}})} + \frac{K_4 \sum_j^{K_4} n_{\mathrm{j}}}{\sum_{j=1}^{K_4} (n_{\mathrm{j}} F_{\mathrm{CFj}})} \tag{1.4}$$

Most simplified, according to Equation 1.4, the gene-specific Nc value is calculated as sum of codons, weighted by the respective codon family (2,3,4).

1.3.5 Codon deviation coefficient

Zhang *et al.* [52] developed an index assessing positional codon bias, also called 'background' nucleotide composition (e.g. the GC content at first, second and third position of the codon) and the statistical significance. By applying this algorithm to gene sequences, they could calculate index values which correlate well with protein abundances from different species. Hence, gene expression levels from *Escherichia coli* cultivated in M9 and LB medium were compared with the respective bias values calculated with CAI, Nc and CDC. They found, that the CDC measure outperformed Nc and CAI indices when correlated with M9 culture protein expression data (0.367, 0.187 and 0.288, respectively), albeit the best correlation was discovered for *S.cerevisiae* genes which returned correlation values between 0.6 and 0.675. In this

case Nc and CAI indices perform better than CDC.

$$CDC = 1 - \frac{\sum_{xyz} \pi xyz \times \hat{\pi} xyz}{\sqrt{\sum_{xyz} \pi xyz^2 \times \hat{\pi} xyz^2}} \tag{1.5}$$

The gene-specific CDC (Equation 1.5) compares the observed codon composition πxyz and an expected (uniformly distributed) codon composition $\hat{\pi} xyz$, and calculates the value from the distance between those codon vectors.

2 Hypothesis and Aim

Yarrowia lipolytica is an oleaginous budding yeast which has been successfully applied as production host for lipid-based chemicals and single-cell oil, which can be used for biodiesel production.

In order to further enhance lipid production, cytosolic acetyl-CoA, the major building block for fatty acids, shall be increased. The pyruvate dehydrogenase bypass, composed of a pyruvate decarboxylase, an acetaldehyde dehydrogenase and an acetyl-CoA synthetase, is a short metabolic pathway which is essentially inactive in *Yarrowia lipolytica*.

Thus the aim is to heterologously express (putatively) highly active enzymes to increase flux through this pathway. Selection of orthologous enzymes will be conducted according to their codon usage quantified by different codon usage bias indices. The underlying assumption is, that codon usage bias resembles evolutionary optimization. Thus, the more biased a gene is, the more active is the enzyme.

3 Materials and Methods

In this section methods and tools are described. Detailed lists e.g. of buffer compositions and devices can be found in the Appendix.

3.1 *In silico* analysis

3.1.1 Flux balance analysis

In order to illustrate the relationship between reaction fluxes and codon bias values, flux balance analysis of a genome-scale metabolic model of *Yarrowia lipolytica*, iMK678, was performed using COBRApy [53] (*COnstraints-Based Reconstruction and Analysis* toolbox for Python, in this case Python version 2.7). Standard growth constraints for exchange reactions are listed in Table 3.1. All other lower bounds of exchange reactions were set to zero. The objective function was set to biomass 40 % of TAG (reaction ID : biomass_40). Linear optimization was carried out using a solver from Gurobi,Inc. (free student licence, version 5.6.3.) [1].

[1] Gurobi, Inc. `http://www.gurobi.com/`

Table 3.1 – Growth constraints

Reaction ID	Flux constraint [mmol/(gDW*h)]	Bound
EX_co2_e_	-1000	lower
EX_h2o_e_	-1000	lower
EX_h_e_	-1000	lower
EX_inost_e_ (inositol)	-1000	lower
EX_k_e_	-1000	lower
EX_na1_e_	-1000	lower
EX_nh4_e_	-1000	lower
EX_o2_e_	-1000	lower
EX_so4_e_	-1000	lower
EX_glc_e_(D-glucose)	-3.99	lower
EX_glc_e_	1000	upper
EX_glyc_e_ (glycogen)	0	lower
EX_glyc_e_	1000	upper

3.1.2 Calculation of codon usage bias

For quantification of the codon usage bias, three different codon usage bias indices were used, namely the *effective number of codons* index, the *codon deviation coefficient* and the *tRNA adaptation* index. Gene sequences were downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG) and saved as FASTA sequences in a text file using a custom-made Python module (orthoDB.py). Nc values were calculated using the DAMBE toolbox (version 5.5.9) [2]. The Composition Analysis Toolkit (CAT) version 1.0.[3] was used to calculate CDC values. A Python script was implemented, which allows the parallelization of the calculation on two or more cores (CDC_missing.py). To determine the tRNA-based tAI values, genome sequences were downloaded from NCBI. The genomes were subjected to *tRNAscan-SE*, which predicts

[2] `http://dambe.bio.uottawa.ca/software.asp`, as accessed June 2014
[3] `http://cbrc.kaust.edu.sa/CAT/`, as accessed October 2014

tRNA isoacceptors from the genomic sequence[4]. The bias values were calculated using standard settings. The retrieved tRNA isoacceptor-pool was subsequently used as input for determination of tAI values. The tAI analysis was performed using *CodonR*, written in R[5]. The tutorial R-file was modified for evaluating given input files (e.g. the tRNA data file from tRNAscan-SE and FASTA sequences of gene sequences). The tAI values were calculated in two different ways. The *heterologous* tAI values were determined in order to give an impression on how well the given genes are adapted to the tRNA pool of *Yarrowia lipolytica*. Additionally, *endogenous* values were calculated which give an estimate of the level of adaptation to the original tRNA pool.

Resulting Nc, CDC and tAI values were saved in Microsoft Excel 1997-2003 workbooks and imported into Python for further analysis.

3.2 Cloning

Table 3.2 lists strains and plasmids used as templates and plasmid backbone, respectively.

Table 3.2 – Strains and plasmid

Strain or plasmid	**Source**
Yarrowia lipolytica H222	in-house organism
Saccharomyces cerevisiae CEN.PK	in-house organism
Candida glabrata CBS138	in-house organism
Escherichia coli XL10	in-house organism
Spathaspora passalidarum MYA4345	derived from ATCC
pFA-URA3-hp4d	pFA-URA3 constructed by Gola *et al.*[54] the hp4d promotor was inserted previously

[4] `http://lowelab.ucsc.edu/tRNAscan-SE/`
[5] `http/people.cryst.bbk.ac.uk/~fdosr01/tAI/`

3.2.1 Isolation of genomic DNA

10 mL of overnight culture (30°C 180 rpm) were centrifuged for five minutes at room temperature. The supernatant was removed and the pellet was resuspended in 500 μL of ddH_2O. Subsequently, the resuspension was transferred to a 1.5 mL microcentrifuge tube and centrifuged again. The pellet was resuspended in 200 μL breaking buffer, 200 μL of glass beads and 200 μL of 4 °C cold phenol/chloroform/isoamyl alcohol. The suspension was thoroughly lysed by vortexing for three minutes. Afterwards, 200 μL of TE buffer were added and the cell lysate was mixed by briefly vortexing. The suspension was centrifuged at maximal speed for five minutes at room temperature and the aqueous phase was transferred into a new microcentrifuge tube. 1 mL of absolute ethanol was added and then mixed by inverting the tube. Subsequently, the mixture was centrifuged for three minutes at room temperature and the pellet was resuspended in 400 μL TE buffer. The DNA was precipitated by adding 10 μL of 4 M ammonium acetate (CH_3COONH_4) and 1 mL of absolute ethanol and was mixed by inverting. After another centrifugation step, the supernatant was carefully removed and the microcentrifuge tube was left inverted for drying at room temperature. Finally, the dried pellet was resuspended in 100 μL ddH_2O.

3.2.2 Polymerase chain reaction (PCR)

PCR reactions were carried out using *Herculase* fusion-polymerase. The PCR reactions were assembled on ice as listed in Tables 3.3 and 3.4, respectively, whereas detailed PCR conditions, are listed in the Results. Primer pairs (Table 3.5) were designed with overlapping ends for Gibson assembly. The initial denaturation was set to 8 minutes, followed by 20 - 30 cycles, consisting of denaturation (at 93 °C), annealing (at least 50 °C) and elongation at 72 °C for 2 minutes per kbp. The final elongation time was set to 10 minutes.

Table 3.3 – PCR mix "A"

Compound and stock concentration	**Amount**
PCR Buffer [5x]	to 1x concentration
dNTPs [10 mM]	10 % of final volume
Primer [100 μM]	10 % of final volume
Template	5 % of final volume , at least 5 ng
DMSO	5 % of final volume
Enhancer [15 mM betaine in DMSO]	5 % of final volume
Herculase polymerase	1 % of final volume
ddH_2O	to final volume

Table 3.4 – PCR mix "B"

Compound and stock concentration	**Amount**
PCR Buffer [5x]	to 0.5 x concentration
Janke [55] - PCR Buffer [10x]	to 1 x concentration
dNTPs [10 mM]	10 % of final volume
Primer [100 μM]	8 % of final volume
Template	5 % of final volume ,at least 5 ng
Herculase polymerase	1 % of final volume
ddH_2O	to final volume

Table 3.5 – Primers

Primer sequence 5'-3'	Primer ID
CGCATGCACTAGTCAGCTGATTGGGCGAGCAAAAAGACGA	1_insertion_up_FWD
TTATAAGCTTCAGCTGACTACGTGTGTGCCGATTTGCAAG	1_insertion_up_REV
CTTGCAAATCGGCACACACGTAGTCAGCTGAAGCTTATAA	1_pFa_ura_lox_FWD
AAACGTGAAATACAATTGAGATCGATATAACTTCGTATAA	1_pFa_ura_lox_REV
TTATACGAAGTTATATCGATCTCAATTGTATTTCACGTTT	1_citsProm_FWD
TCAAAGTGTAGCTTAGTCATTGTGTTGATGTGTGTGGGGT	1_citsProm_REV
ACCCCACACACATCAACACAGGATCCATGACTAAGCTACAC TTTGACACTGC	1_Ald6sce_FWD
CTCGGTCCCATCGGTAATCATTTGAAAGATGATACTCTTTA TTTCTAGACAGGTA	1_Ald6sce_REV
AAAGAGTATCATCTTTCAAATGATTACCGATGGGACCGAG	1_insertion_down_FWD
CGGCCGCTCTAGAAGATCTGCAGCTGAAATTTCCCCTGTATGTTG	1_insertion_down_REV
ACTCTTTATTTCTAGACAGTTATATATATATATATATATATATATAAA AAAAAATTAGTGGTGATGGTGATGATGCAACTTAATTCTGACAGCTTTTACTTC	1_ald6_6HIS_synT_REV
TAGGTGACACTATAGAACGCTCTCGGGCAAGTTTTTTTCTCCACA	3_insertion_up_FWD
TTATAAGCTTCAGCTGACTATTGCTGTTCGGGGACTTCCA	3_insertion_up_REV
TGGAAGTCCCCGAACAGCAATAGTCAGCTGAAGCTTATAACTTCGTATAG	3_pFa_ura_lox_FWD
GCTCCAATAACCGCGTTCTTATCGATATAACTTCGTATAATGTATGCTAT	3_pFa_ura_lox_REV
TTATACGAAGTTATATCGATAAGAACGCGGTTATTGGAGC	3_IDH1Prom_FWD

Continued on next page

Table 3.5 – *Continued from previous page*

Primer sequence 5‘ - 3’	**Primer ID**
CCCAAAGTAATCTCAGACAT*ggatcc*TGTGGATGTGGATATGTTTTCGATTA	3_IDHProm_REV
AAAACATATCCACATCCACA*ggatcc*ATGTCTGAGATTACTTTGGGTAGATA	3_PDCcgl_FWD
TCCACGAGGGCAAGGGCTGG*gctagc*TTTGAAAGATGATACT CTTTATTTCTAGACA	3_PDCcgl_REV
CGCTTCCATCAACGCTAAGCAAGAACATCATCACCATCACCAC TAATTTT TTTTTATATATATATATATATATATATATAACTGTCTAGAAATAAAGAGTAT	3_PDCcgl_6HIS_synT_REV
CCATCAACGCTAAGCAAGAACATCATCACCATCACCACTAA	3_PDCcgl_short_REV
AAAGAGTATCATCTTTCAAAGCTAGCCCAGCCCTTGCCCTC GTGGACATAAC	3_insertion_down_FWD
AGGGAGACCGGCAGATCCGCTTTATCCGGGCGGATGCGTC	3_insertion_down_REV

Underlined primer parts overlap with other primers. Nucleotides written in lower case denote restriction enzyme sites.

3.2.3 Overlap-extension PCR

Adjacent expression cassette compounds were designed to have overlapping ends of 20 - 22 base pairs. Thus, the two template fragments were simultaneously used as both template and 'internal' primers. In order to join two fragments, the PCR mix was assembled as outlined in Table 3.6.

Table 3.6 – Overlap extension PCR

Compound and stock concentration	**Amount**
PCR Buffer [5x]	10 μL
dNTPs [10 mM]	5 μL
Smaller template	100 ng
Bigger template	to equimolar amount of smaller template
Herculase polymerase	0.5 μL
ddH_2O	to 48 μL

If necessary, the templates were vacuum concentrated. The overlap extension PCR was carried out at an annealing temperature of 50 °C and the elongation time was set to 2 minutes/kbp (final size). After 15 cycles, 4 μL outer primers (stock concentration : 100 μM) were added during the denaturation step of the 14. or 15. cycle. The PCR was conducted for another 20 cycles (35x in total).

3.2.4 Restriction digest

The restriction digest (Table 3.7) of the plasmid which serves as backbone for further cloning steps was carried out overnight (8 hours at 37 °C followed by 15 minutes at 65 °C).

Table 3.7 – Restriction digest of plasmids

Component	Amount
Restriction buffer	12 μL
Alkaline phosphatase	3 μL
Restriction enzyme	3 μL each
Plasmid	1 - 2 μg
ddH_2O	to 120 μL

Restriction digests of genomic DNA were carried out at 37 °C for 1 - 2 hours. Afterwards, the reaction was stopped by incubating the restriction mix at 65 °C for 20 minutes (Table 3.8).

Table 3.8 – Restriction digest of genomic DNA

Compound and stock concentration	**Amount**
Restriction buffer	1.5 μL
Restriction enzyme	0.2 μL each
Genomic DNA	5 μL
ddH_2O	to 15 μL

3.2.5 Agarose gel electrophoresis

Generally, agarose gels of 0.8 %, spiked with ethidium bromide were used. For verification of PCRs, restriction digests and genomic or plasmid isolations, 1 μL of sample was mixed with 2 μL 6 x loading dye and 9 μL ddH_2O. 110 V were applied for 30 minutes. For preparatory gels (DNA purification from gel), the entire sample volume was mixed accordingly with 6 x loading dye. 90 V were applied for an hour. To measure the concentrations, the *Middle Range* standard ladder (Fermentas) was used. In

order to determine the size of the PCR product more exactly, the *1 kb* DNA standard ladder (Fermentas) was used.

3.2.6 Purification of PCR products

The band of interest was cut out from the preparatory agarose gel and purified using the *GeneJET Gel Extraction* kit. To increase the yield, the agarose gel slice dissolved in binding buffer was applied to the purification column and incubated for 10 minutes at 37 ° C for 10 minutes, briefly centrifuged and the flow-through reapplied on the column. Afterwards, the column was centrifuged for two minutes and then continued with washing steps as described in the manual. The sample was eluted similarly to the first incubation steps. The sample was eluted in in 30 - 50 μL ddH_2O, dependent on the amount of DNA.

3.2.7 Plasmid isolation

Plasmids were isolated using the *GeneJET Plasmid Miniprep* kit. Each step was executed as described in the manual. Plasmids were eluted in ddH_2O and concentrations were inferred from agarose gels.

3.2.8 Gibson assembly

Gibson assembly is a relatively novel alternative to the standard restriction-ligation cloning strategy. The fragments are designed to have overlapping regions at the ends, where they should be joined without the need of many thermocycling steps (Figure 3.1). In contrast to sequential rounds of restriction digests and ligations, respectively, many fragments can be linked at once. Thus, this method was applied in the construction of a synthetic genome of *Mycoplasma genitalium* [56] and was applied since for the rapid assembly of many different fragments, e.g. for constructing of libraries for two-hybrid analysis or for generating circular plasmids in a single step [57, 58].

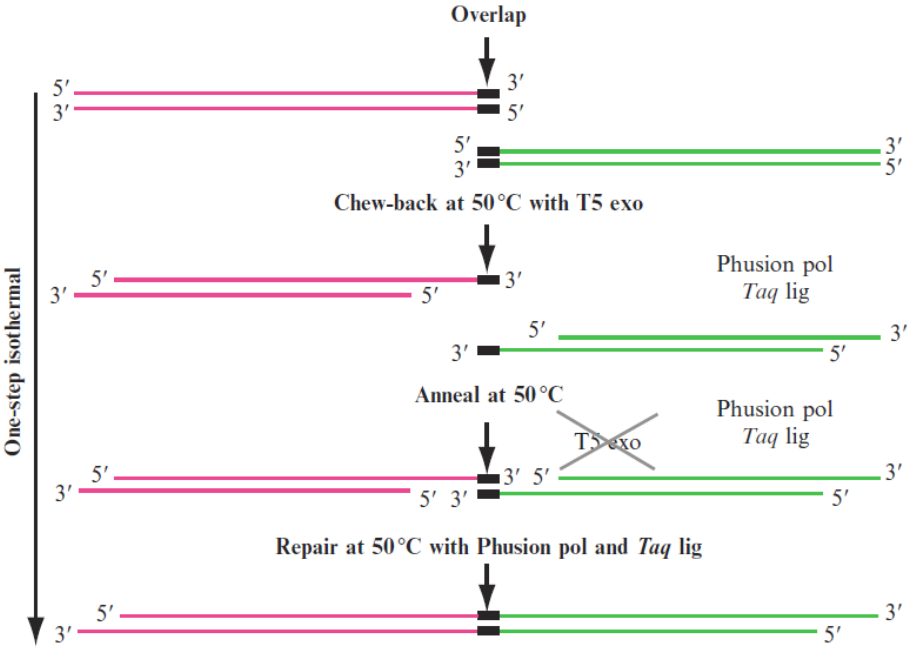


Fig. 3.1 – Isothermal *in vitro* (Gibson) assembly. The overlaps of the fragments are digested by a 5'-exonuclease, which results in single stranded overlaps, that can be joined using a ligase and DNA polymerases. Figure taken from [59].

The fragments were mixed according to Equation 3.1. At least 100 ng of the smallest fragment were used. The sample was vacuum concentrated, resuspended in 5 μL ddH_2O and added to 15 μL assembly master mix. The isothermal assembly was carried out at 50 °C for 60 minutes in a thermocycler. 5 μL of the reaction mix (composition see [59]), putatively containing readily assembled plasmid, were directly used for transformation of *Escherichia coli*.

$$amount_{\text{large fragment}}[ng] = \frac{amount_{\text{small fragment}}[ng] * size_{\text{large fragment}}[kb]}{size_{\text{small framgnet}}[kb]} \tag{3.1}$$

3.2.9 *Escherichia coli* heat shock transformation

Aliquots of chemically competent *Escherichia coli* cells were incubated on ice for half an hour and then aliquoted into 40 μL each. 5 μL of Gibson reaction mix were added and incubated on ice for another half an hour. Subsequently, the cell suspension was incubated at 42 °C for 90 seconds followed by 2 minutes on ice. 700 μL prewarmed LB medium was added to the cell suspension which was incubated for an hour at 37 °C. After spinning down the resuspension for 10 seconds, the supernatant was partially removed, the cell pellet was resuspended in the residual ∼300 μL and plated on ampicillin-containing LB plates. The plates were incubated overnight at 37 °C.

4 Results

In order to find a suitable metabolic engineering target for enhanced lipid production, flux balance analysis of a genome scale model was performed. According to *in silico* flux distributions, the PDH bypass is essentially inactive and overexpressing the enzymes of this pathway seems to be a promising approach to achieve increased cytosolic acetyl-CoA levels. Orthologous enzymes of the pyruvate decarboxylase, the aldehyde dehydrogenase and the acetyl-CoA synthetase were screened for a high degree of codon bias using different codon usage bias indices. Furthermore, in order to examine the applicability of codon usage bias as approximation for enzyme performance, Python scripts and functions were implemented, which allow automated screening for orthologous enzymes in KEGG and linking of *in silico* flux values with different codon bias values. Finally, design and first cloning steps towards an overexpression cassettes of *Saccharomyces cerevisiae's ALD6* and *Candida glabrata's PDC1* are discussed.

4.1 Flux balance analysis reveals that the pyruvate dehydrogenase bypass is essentially inactive in *Yarrowia lipolytica*

Acetyl-CoA is the central building block in lipid biogenesis. In oleaginous yeasts, such as *Yarrowia lipolytica*, *cytosolic* acetyl-CoA is produced by the ATP citrate lyase. This enzyme uses citrate, which is exported from the mitochondrium, and produces acetyl-CoA and oxaloacetate. Non-oleaginous yeasts, like *Saccharomyces cerevisiae* however lack this enzyme and are dependent on the pyruvate dehydrogenase bypass, composed of the pyruvate decarboxylase, the acetaldehyde dehydrogenase and the acetyl-CoA synthetase to produce cytosolic acetyl-CoA. In order to compare the flux rates of the two pathways, a genome-scale metabolic model of *Yarrowia lipolytica* (iMK678) was subjected to flux balance analysis as described in Materials and Methods. It was

found, that the PDH bypass is essentially inactive (Figure 4.1).

This finding suggests that the pyruvate dehydrogenase bypass is a valuable metabolic engineering target. By heterologous overexpression of (putative) highly active enzymes the flux through this pathway might be increased and thus acetyl-CoA levels.

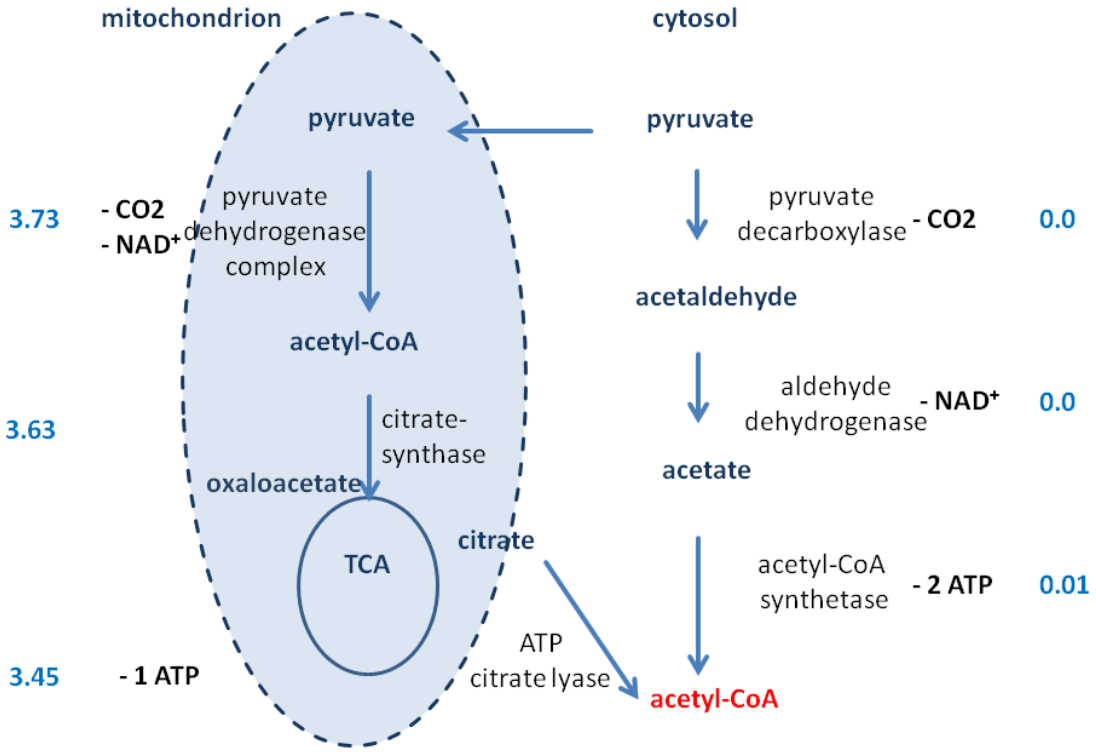


Fig. 4.1 – Production of acetyl-CoA from pyruvate in *Yarrowia lipolytica* mitochondrial PDH complex and cytosolic bypass. There are many different sites within a eukaryotic cell where acetyl-CoA is produced. Two of them are the mitochondrion and the cytosol. Both pathways are similar regarding their cofactor stochiometry with the expection for acetyl-CoA synthetase, which is AMP-forming and thus consumes one ATP more than the ATP citrate lyase. From this point of view, the mitochondrial pathway involving the PDH complex and TCA cycle is preferable over the straight cytosolic pathway. Thus, simulations using the genome-scale model of *Yarrowia lipolytica*, iMK678, also favour the mitochondrial route, whereas the PDH bypass is essentially inactive. Simulation : (light blue) flux values [mmol/(gDW*h)] ; biomass objective function: 40 % lipid in biomass.

4.2 Implementation of codon usage bias and flux analysis in Python

In order to compare different codon usage indices as well as to analyse their correlation with *in silico* reaction flux rates, various scripts and functions were implemented in Python 2.7. In particular, these scripts and functions enable the automated download of nucleotide sequences of orthologous enzymes from KEGG, import of different codon usage values, the subsequent coupling to flux values as derived from flux balance analysis (Table 4.1) and graphical output.

Table 4.1 – Python scripts and functions used for analysis

Python Module/Function	**Description**
orthoDB	The script orthoDB.py contains necessary functions to extract and analyse orthologous enzymes. Input : proteinID (K number) e.g. proteinID = 'K11160' file path to save the FASTA sequences of orthologous énzymes. e.g.filePath = "C:/Users/Lydia/Desktop/" + proteinID + ".txt" Optional Input : QueryTax: in case, only certain taxonomic groups shall be investigated e.g. QueryTax = ["Saccharomycetes"] KeyList: list of species abbreviations from KEGG Output: FASTA sequences of orthologous enzymes saved in a txt-file. Plot of orthologous enzymes vs Nc values (this requires a file of Nc values).
fnProteinOrthos	The function fnProteinOrthos extracts all orthologous enzymes in the KEGG for the given KEGG number Input: proteinID (KEGG number) Output : KeyList : KEGG field keys for given KEGG number

Continued on next page

Table 4.1 – *Continued from previous page*

Python Module/Function	**Description**
mkFASTA	The function mkFASTA uses a key as returned from fnProteinOrthos and returns the FASTA sequence (subfunction : callFASTApage) Input : Key: e.g. yli:YALI0D10131g Output : FASTA sequence
TaxBrowse	The function TaxBrowse uses taxonomic information and returns the respective species abbreviations from KEGG. Input : taxonomic class; "Saccharomycetes" Output : species abbreviations e.g. ["yli","sce",...] Example : KeyList = TaxBrowse(QueryTax)
GeneNCDict	The function GeneNCDict generates a dictionary linking genes and their NC values. Input : Path to an Excel file ('.xls', 97-2003) containing genes and NC values Output : Dictionary of genes and corresponding Nc values Example : GeneNCDict = makeGeneNCList(NCPath)
mkSpeciesDict	The function mkSpeciesDict generates a dictionary of species abbreviations and full species names for all entries in KEGG. Input : void Output : dictionary e.g. dict = "yli" : "Yarrowia lipolytica" Example : dict = mkSpeciesDict()
PickSpecies	The function PickSpecies shortens the Gene-Nc dictionary by picking the given species and deleting the rest. Input: KeyList e.ge ["yli","sce",..] GeneNCDict e.g. dict = YALIXXXX: Nc value Output : shorter Gene-Nc dictionary Example : PickSpecies(KeyList,GeneNCDict)

Continued on next page

Table 4.1 – *Continued from previous page*

Python Module/Function	**Description**
reorgDict	The function reorgDict reorganizes the given Gene-Nc dictionary to the form of dict = abbrev : 'genes' : Nc value,'species' : name Input : Gene-Nc Dictionary, species dictionary (see: mkSpeciesDict) Output: reorganized dictionary Example : dict = reorgDict(GeneNCDict,speciesDict)
mkBoxPlot	The function mkBoxPlot plots the given Nc values by species; the Nc values have been once plotted as boxplots but are now displayed as single points Input: organizedDict (function reorgDict),proteinID e.g. "K01208", taxon: default : "all taxa in KEGG" , both proteinID as well as taxon are necessary settings for the header of the figure Output : figure of Nc plot Example: mkBoxPlot(pdcDict,proteinID,QueryTax)
mkGenomeTable	The function mkGenomeTable imports the codon usage table from KAZUSA and arranges it in dictionaries. Input : KAZUSA link; e.g. `http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=284591&aa=1&style=N` Output : d : Codon : AA : occurrence of codon and codonTable : AA : Codon1 : occurrence of codons,Codon2:occurrence of codons Example : mkGenomeTable(link1)

Continued on next page

Table 4.1 – *Continued from previous page*

Python Module/Function	**Description**
mkRSCUTable	The function RSCUTable calculates the RSCU values for a given codonTable Input : codonTable; codonTable = AA : Codon1 : no. of occurences,CodonX: no. of occurrences Output : RSCU dictionary; AA : Codon1 : RSCU,CodonX: RSCU Output : RSCUTable Example : mkRSCUTable(codonTable)
AnalyseRSCU	The function AnalyseRSCU calculates the RSCU values for every codon of a given gene. Input : GeneName; e.g. "R0040C"; d (dictionary) output from mkGenomeTable myfile: file containing FASTA sequences of genes Output : RSCU table of the gene of interest (RSCUTable-Gene) Example : AnalyseRSCU(GeneName,d,myfile)
CDC_missing	The script CDC_missing calculates the CDC values for given input sequences. CAT1.3.exe is embedded Input: dirPath = /Desktop/CDC/, PathCDC = dirPath + 'CDC.xls' (at the beginning, this file is empty) PathSeq = dirPath + 'infile.txt'; e.g. .txt file of FASTA sequences Output: xls.file containing CDC values
tRNADBScore_main	The script tRNADBScore_main calculates the RSCU values of the tRNA pool of different organisms and returns a dendrogram. Input : Path_trnaFungi = "C:/Users.../trna_Fungi.xls" Output : dendrogram of different species, scored by the similarity of their tRNA anticodons

Continued on next page

Table 4.1 – *Continued from previous page*

Python Module/Function	**Description**
mktrnaDatabase	The function mktrnaDatabase creates a draft tRNA-RSCU-dictionary, with is used as template for sample tRNA pools. Input : r'C:/.../tRNA_Ecoli.xls', KazusaPath ="`http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4952&aa=1&style=N`" (default) Output: tRNA-RSCU dictionary Example : tRNARSCUdictionary = mktrnaDatabase(), default input parameters are used
mkOneLetterCode	The function mkOneLetterCode translates amino acid abbreviations to one letter code, e.g. "Ala" : "A". Input : amino acid as imported from infile. Output : amino acid in one-letter-code Example : mkOneLetterCode(AA)
translateAnticodon	The function translateAnticodon uses the given anticodon and translates it to a codon. Input: anticodon, as imported from the input file. Output: codon. Example : translateAnticodon(anticodon)
PWorthoNcT	The script PWorthoNcT is used to create Figure 4.2. Nc values are plotted along a pathway e.g. glycolysis for different species of interest. Input: NCPath = r'C:/.../YeastPW_NC.xls', QueryTax = ["Saccharomycetes"], KeyList = TaxBrowse(QueryTax) Output: plot of Nc values
main2YLnewT	The script main2YLnewT is used for flux balance analysis links flux values and codon usage bias indices

4.3 Codon usage bias reflects physiology

As mentioned in the Introduction, codon usage bias reflects physiological niches, such as fast and slow-growing organisms. Beyond this, codon usage bias patterns shed light on the impact of a pathway for an organism and thus aid to further discriminate different lifestyles. Figure 4.2 illustrates the distribution of codon bias along different core pathways of different yeasts. Most remarkably, there is a clear difference in codon bias for enzymes of the TCA cycle, which is weaker for fermentative yeasts than for the respiratory yeast *Yarrowia lipolytica*. Furthermore, there is a strong bias of the pyruvate decarboxylase, which is involved in ethanol production, in *Saccharomyces cerevisiae* and *Candida glabrata*, which are two strong fermentative yeasts. This finding is plausible, since fermentative yeasts are not as dependent on the TCA cycle and subsequent oxidative phosphorylation as are respiratory ones, since NAD^+ is regenerated via ethanol production.

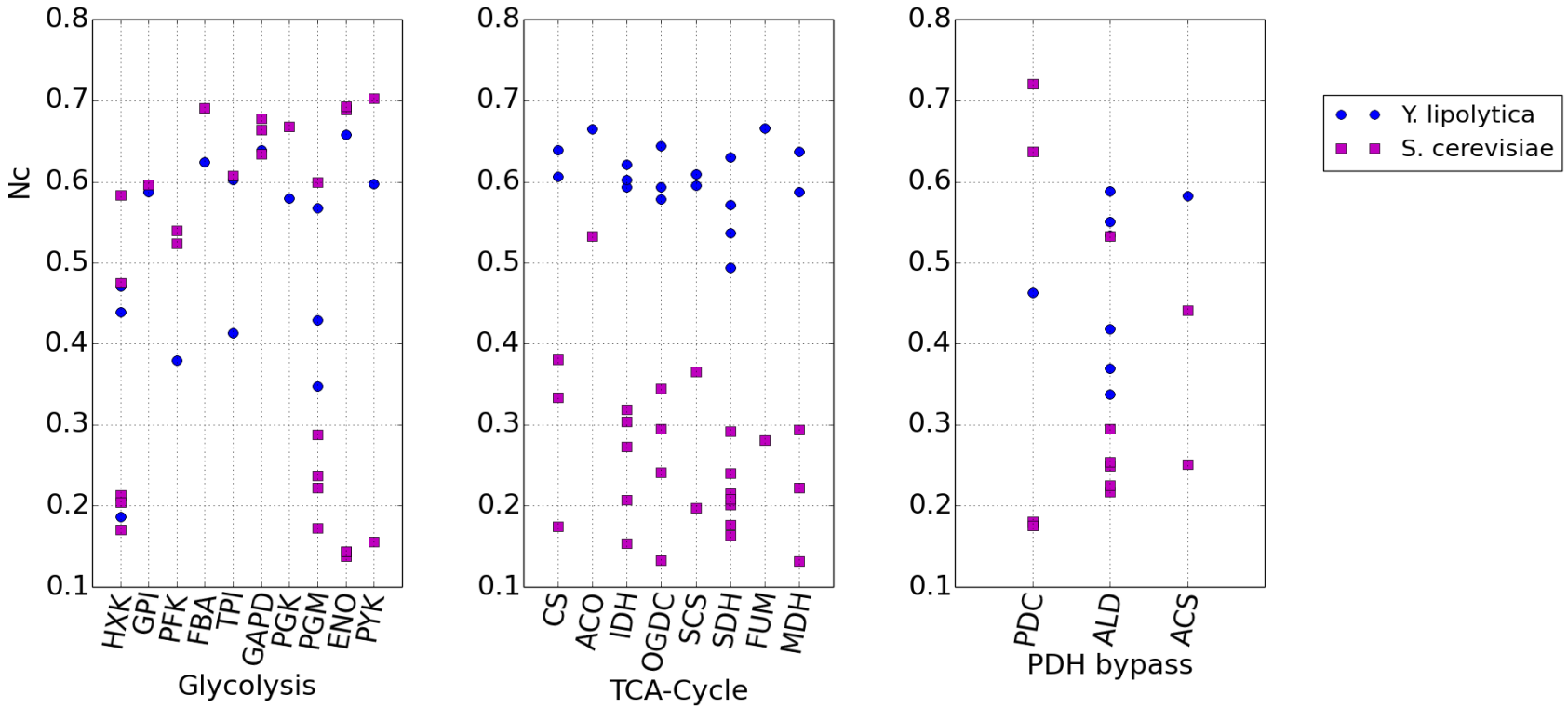
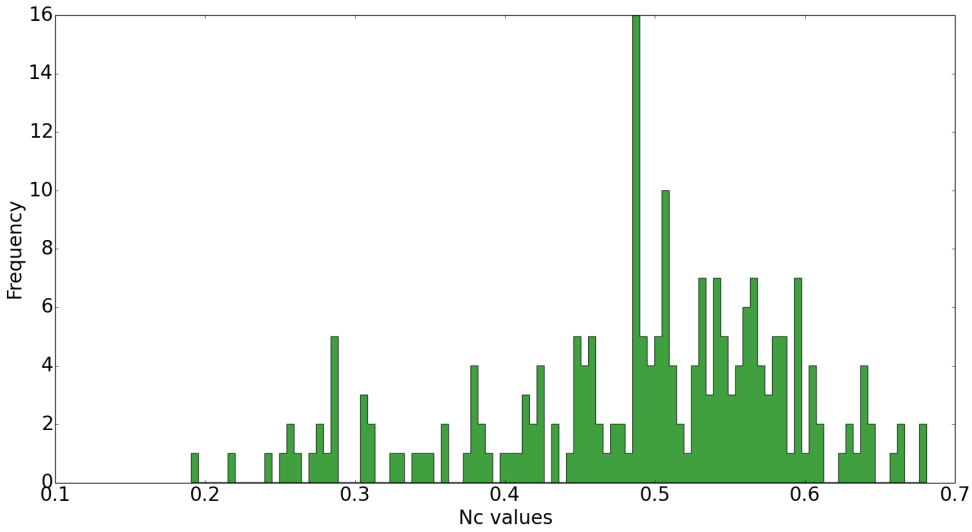


Fig. 4.2 – Comparison of Nc values of different pathways in *Saccharomyces cerevisiae, Yarrowia lipolytica, Candida glabrata* and *Spathaspora passalidarum*. Nc values are normalised to values between 0 and 1, 1 denotes a high degree of bias. There is a clear trend between the yeasts for genes of the TCA cycle, especially between *Y. lipolytica* and *S.cerevisiae* and *C.glabrata*, respectively. Thus, when comparing orthologous enzymes, the TCA cycle is given relatively less importance in strong fermentative yeasts compared to respiratory yeasts. Concerning the PDH bypass, the pyruvate decarboxylase shows a stronger bias indicating the importance of this enzyme for ethanol production. Abbreviations : HXK: hexokinase ,GPI: glucose-6-phosphate isomerase ,PFK: 6-phosphofructokinase, FBA: fructose-bisphosphate aldolase, TPI: triosephosphate isomerase, GAPD: glyceraldehyde 3-phosphate dehydrogenase, PGK: phosphoglycerate kinase, PGM: phosphoglycerate mutase, ENO: enolase, PYK: pyruvate kinase, CS: citrate synthase, ACO: aconitate hydratase, IDH: isocitrate dehydrogenase, OGDC: 2-oxoglutarate dehydrogenase complex, SCS: succinyl-CoA synthetase, SDH: succinate dehydrogenase, FUM: fumarate hydratase, MDH: malate dehydrogenase, PDC: pyruvate decarboxylase, ALD: aldehyde dehydrogenase, ACS: acetyl-CoA synthetase

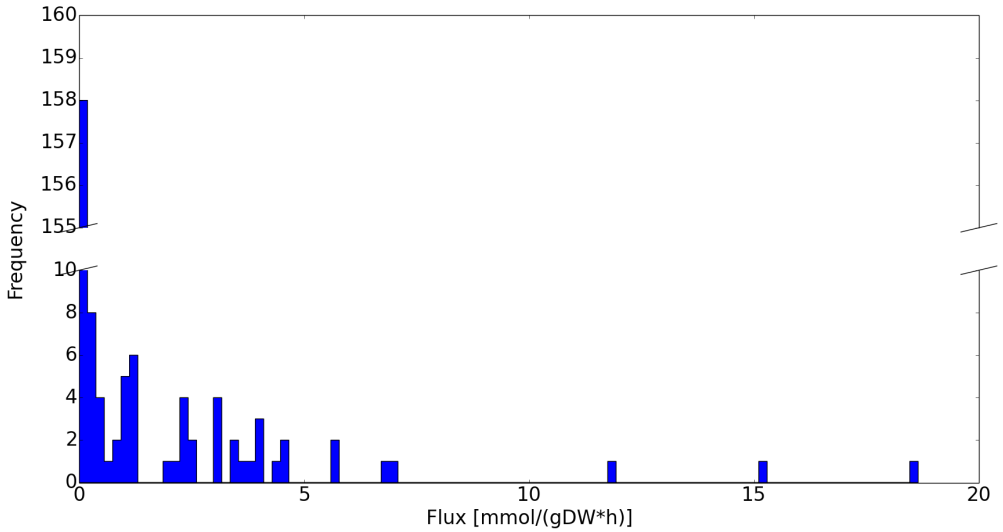
4.4 Codon usage bias as estimate of reaction flux values is limited to orthologous genes

The degree of codon usage bias can be applied to get an impression of the general metabolic capacity. Thus, it should be possible to use this bias values for prediction of fluxes, hence in a *quantitative* way. This implies a correlation of reaction flux and degree of bias. To this purpose, a genome-scale metabolic model of *Yarrowia lipolytica*, iMK678, was subjected to flux balance analysis as mentioned in Materials and Methods, and resulting reaction flux values were compared to codon usage bias values. Despite some highly active core metabolic pathways, such as glycolysis or the pentose phosphate pathway, the majority of fluxes is diminishing (Table 4.3b). Normed Nc values on the other side, range from 0.2 to 0.7, with two peaks at 0.5 (Table 4.3a). Thus, Nc values (or any other index values) cannot be *readily* linked to a given flux, because there obviously exist different *flux levels*. Moreover, assuming identical fluxes through all reactions the genes coding for these reactions should all be equally biased. Figure 4.4 however clearly indicates, that the bias of glycolytic genes of *Yarrowia lipolytica* is not strongly correlated with reaction fluxes.

From a biochemical perspective, Nc and flux values cannot be correlated or compared across different classes of enzymes. For example a highly active oxidoreductase might be able to catalyse a reaction at much higher rates [e.g. mmol/gDW*h)] than does a hydroxynitrile lyase. To conclude, comparison of enzyme activities and their correlation to codon bias is only possible for orthologous enzymes.



(a) Histogram of Nc values



(b) Histogram of flux values

Fig. 4.3 – Distributions of Nc and flux values.
Nc values have been normalised to 0 - 1, 1 indicating a strong bias. Nc values have been assigned to the genome scale model of *Yarrowia lipolytica*, iMK678, and flux balance analysis was performed using standard settings (Materials and Methods). Reactions that do not carry flux have been excluded. These two distributions illustrate, that the codon bias follows a normal distribution, whereas most reaction fluxes are found at rather low rates. There are only few reactions which have high flux rates such as core metabolism, but most fluxes are small. Ultimately, the two figures imply, that Nc and flux rates cannot be directly compared with each other.

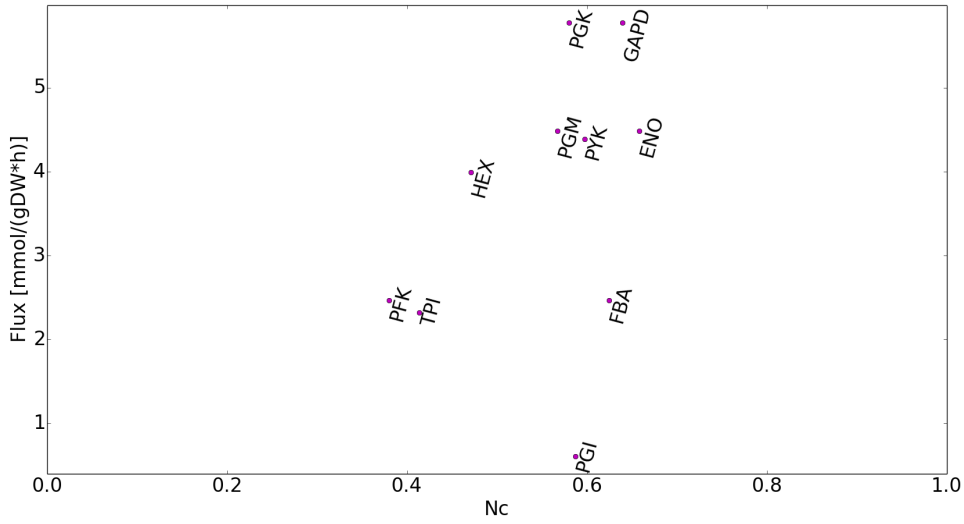


Fig. 4.4 – Flux and bias of glycolytic enzymes of *Yarrowia lipolytica*. Nc and flux values of some glycolytic genes suggest, that Nc values are not dependent on the position of the enzyme within this pathway.

4.5 Finding suitable codon-biased candidate genes

Orthologous candidate genes for the pyruvate decarboxylase, the aldehyde dehydrogenase and the acetyl-CoA synthetase were extracted from KEGG. Subsequently, these genes were subjected to the Nc and CDC calculation, respectively. The resulting values were then ranked from most biased to less biased values in both cases. Those ranks were then summed up, and the smallest sums (highest ranked) genes, were furthermore subjected to tAI calculation.
Whereas Nc and CDC are gene-specific indices, tAI reflects the translational level: on the one hand the endogenous tAI was calculated by using the tRNA pool of the respective organism and on the other also for the tRNA pool of *Yarrowia lipolytica*, herein referred to as heterologous tAI. For this purpose, *CodonR* was used for calculation of the tAI as published by dos Reis *et al.* [48] and was adapted for heterologous or endogenous tAI.

Remarkably, all of those enzymes had their best hits within different yeast species (*Candida glabrata*, *Saccharomyces cerevisiae* and *Spathaspora passalidarum*), although the database search was not restricted to them.

4.5.1 The pyruvate decarboxylase (Pdc1) of *Candida glabrata* shows a strong bias along with other fermentative yeasts

The pyruvate decabroxylase (Pdc) is the first enzyme of the pyruvate dehydrogenase bypass. It catalyzes the conversion of pyruvate to acetaldehyde and thus also determines the flux through this pathway. After the evaluation using different codon usage bias indices, Table 4.2 sums up the most-biased genes as well as *Yarrowia lipolytica´*s endogenous *PDC*. Most of the values are similar. Only *Candida glabrata* stands out from the others regarding the Nc value.

Table 4.2 – Highly biased pyruvate decarboxylases (KEGG number : K01568)

Gene	**Species**	**Nc**	**CDC**	**tAI (endogenous)**	**tAI *(Y.lipolytica)***
YLR044C	*Saccharomyces cerevisiae*	0.72	0.35	0.73	0.43
ZYRO0F01606g	*Zygosaccharamyces rouxii*	0.72	0.35	0.79	0.51
,KAFR_0E02020	*Kazachstania africana*	0.71	0.36	0.72	0.40
CAGL0M07920g	*Candida glabrata*	0.76	0.34	0.70	0.33
NCAS_0D01660	*Naumovozyma castellii*	0.72	0.35	0.74	0.44
YALI0D10131g	*Yarrowia lipolytica*	0.46	0.12	0.41	0.41

CAGL0M07920g encoding for Pdc1 from *Candida glabrata* was characterized by Wang *et al.* 2004 [60]. In the same year, Liu *et al.* [61] published a study on metabolic engineering of the PDH-bypass in this organism. They generated a Pdc1 mutant with *decreased* activity in order to increase the cytosolic pyruvate pool and decrease ethanol secretion, hence carbon spillage.

4.5.2 The aldehyde dehydrogenase (Ald6) of *Saccharomyces cerevisiae* is $NADP^+$ dependent and highly biased

The second enzyme of the PDH bypass converts the toxic acetaldehyde to acetate and liberated protons are transferred to either NAD^+ or $NADP^+$. The latter is a more favourable cofactor to utilise than NAD^+, since $NADPH/H^+$ is required for the reduction steps in fatty acid biosynthesis. The choice of enzyme based on cofactor preference is beneficial in two ways. Not only should the gene be selected based on a strong bias, which is herein used as a proxy for high activity, but the simultaneous generation of $NADP^+$ is also advantageous for lipid biogenesis.
There exist two separate orthologous groups in KEGG, one for (putatively) NAD^+ dependent aldehyde dehydrogenases and another for $NADP^+$ dependent ones. *Saccharomyces cerevisiae*'s aldehyde dehydrogenases Ald5 and Ald6 are reportedly $NADP^+$-dependent [62], while they are misclassified to be NAD^+-dependent in KEGG. Thus the database seems to be unreliable in this case, the research for suitable candidates is restricted to well-characterised, published enzymes. In Table 4.3, the different index values and their values are listed.

Table 4.3 – Aldehyde dehydrogenase isoforms of *Saccharomyces cerevisiae* and *Yarrowia lipolytica*

Gene	Species	Nc	CDC	tAI (endogenous)	tAI *(Y.lipolytica)*
YOR374W (Ald4)	*Saccharomyces cerevisiae*	0.30	0.17	0.50	0.38
YER073W (Ald5)	*Saccharomyces cerevisiae*	0.25	0.11	0.52	0.41
YPL061W (Ald6)	*Saccharomyces cerevisiae*	0.53	0.24	0.61	0.31
YALI0C03025g (similar to Ald4)	*Yarrowia lipolytica*	0.37	0.18	0.45	0.45
YALI0E00264g (similar to Ald4)	*Yarrowia lipolytica*	0.59	0.22	0.52	0.52
YALI0F04444g (similar to Ald5)	*Yarrowia lipolytica*	0.42	0.13	0.43	0.43

Ald6 is for most of the indices the most favourable candidate. This enzyme is furthermore resident in the cytosol [63], which makes it a suitable enzyme for increasing cytosolic acetyl-CoA levels in *Yarrowia lipolytica*.

4.5.3 *Spathaspora passalidarum´s ACS1* is highly biased but the localisation of the enzyme is unclear

The acetyl-CoA synthetase (Acs) is the final step in the cytosolic generation of acetyl-CoA, thereby consuming two molecules ATP.
According to Table 4.4, the six best candidates are originating from only three different organisms, namely *Dechlorosoma suillum*, *Candida tropicalis*, *Spathaspora passalidarum*. Regarding their codon usage index values each of these enzymes might be a valuable candidate, as either Nc, CDC or tAI are ranked very good. *Candida tropicalis*

seems to be the best among them. However, as it is classified as S2 microorganism, which requires special precautions while handling, this organism was neglected. An interesting alternative organism would be *Dechlorosoma suillum*, which is an anaerobic perchlorate-reducing bacterium [64]. Its ability to grow even without oxygen suggests that it makes extensive use of the PDH bypass for acetyl-CoA production, as the mitochondrial PDH complex is inactive under these conditions. *Spathaspora passalidarum* on the other hand is a xylose fermenting yeast [65]. Being a strong fermentative yeast, *Spathaspora passalidarum* also seems to harbour a highly active acetyl-CoA synthetase. Thus, *SPAPADRAFT_135964* (Acs1), the more biased isoform of *Spathaspora passalidarum* was chosen for expression in *Yarrowia lipolytica*.

Table 4.4 – Highly biased acetyl - CoA snythetases (KEGG number : K01895)

Gene	**Species**	**Nc**	**CDC**	**tAI (endogenous)**	**tAI (*Y.lipolytica*)**
Dsui_2003	*Dechlorosoma suillum*	0.72	0.20	0.86	0.32
Dsui_2187	*Dechlorosoma suillum*	0.70	0.17	0.72	0.33
CTRG_04729	*Candida tropicalis*	0.65	0.25	0.78	0.43
CTRG_02032	*Candida tropicalis*	0.62	0.26	0.73	0.41
SPAPADRAFT_135964	*Spathaspora passalidarum*	0.61	0.27	0.63	0.43
SPAPADRAFT_130864	*Spathaspora passalidarum*	0.61	0.23	0.62	0.38
YALI0F05962g	*Yarrowia lipolytica*	0.58	0.16	0.64	0.64

Little is known about *Spathaspora passalidarum* or its acety-CoA synthetase. In *Sac-*

charomyces cerevisiae, Acs1 and Acs2 were investigated by Takahashi *et al.* [66]. The two enzymes also partly reside in the nucleus, involved in lysine acetylation. Since this works aims to manipulate a cytosolic pathway, it is crucial to ensure that also Acs1 of *Spathaspora passalidarum* is acting in the cytosol. Thus, *Spathaspora passalidarum*'s Acs1 (*SPAPADRAFT_135964*) was subjected to localisation prediction programs and compared to the *Saccharomyces cerevisiae* Acs1 amino acid sequence. Thus the amino acid sequence was subjected to targetP [1] (Table 4.5) :

Table 4.5 – Subcellular localisation prediciton output (targetP)

Name	Length	mTP	SP	other	Loc	RC
ACS1	676	0.127	0.043	0.890	_	2

According to *targetP* evaluation, *ACS1* does not possess neither a secretory nor a mitochondrial signal sequence and localises "elsewhere" (see: Loc : _ with decent reliability [RC = 2 on a scale from 1 - 5 whereas 1 denotes high reliability of the prediction).

To make sure, that Acs1 is not a peroxisomal enzyme the amino acid sequence was also submitted to the PTS1 predictor (using fungal settings). [2].

Table 4.6 – subcellular localisation prediciton output (PTS1)

Name	C terminus	Score	Profile	S_ppt (non-accessibility)	S_ppt (accessibility)	P (false positive)	Prediction classification
ACS1	SQIIDVVKTSRK	-35.579	-20.814	-2.838	-11.927	20.81 %	Not targeted

As this enzyme is also not targeted to the peroxisome, it might be still resident in either nucleus or the cytoplasm as these localisations have not been explicitly specified

[1] `http://www.cbs.dtu.dk/cgi`
[2] `http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp`

in the bioinformatics softwares used so far. According to *YLoc* [3] , it has a probability of 86.92 % to be cytosolic and only 13.0 % probability to be nuclear.

Taken these three *in silico* predictions together, Acs1 (*SPAPADRAFT_135964*) seems to be cytosolic.

Apart from the (experimentally unresolved) localisation, there is another complication, namely, that it might be feedback regulated. Starai *et al.* [67] observed that the reversible acetylation of a lysine at the C-terminal end of Acs of *Salmonella enterica* renders this enzyme inactive. Tucker *et al.* [68] also investigated those regions for acetylation for Acs of *Streptomyces lividans* and *Salmonella enterica*. Although both enzymes share a common stretch of amino acids in the C-terminus, the Acs of *Streptomyces lividans* is not acetylated (Figure 4.6). By generating chimeric sequences, consisting of a combination of *Streptomyces lividans*' and *Salmonella enterica*'s Acs, they found that residues upstream and downstream of this acetylation site influence lysine acetylation. In order to find out, if Acs1 from *Spathaspora passalidarum* is regulated in a similar manner, it was compared to the orthologs of *Streptomyces lividans* and *Salmonella enterica* using BlastP. This resulted in a query coverage of 98 % and 49 % identity. Most notably, both sequences share a C-terminal conserved region, including the acetylation site. However, as it remains unclear whether Acs1 of *S. passalidarum* is acetylated and thus feedback inhibited, a mutant *ACS1* devoid of this acetylation motif will be cloned.

Summing up, before *ACS1* of *Spathaspora passalidarum* can be subjected to heterologous overexpression in *Yarrowia lipolytica*, the intracellular localisation of Acs1 needs to be clarified by expression of a GFP-tagged version. Furthermore the putative acetylation site needs to be disrupted and the activity compared to the original Acs1.

[3] `http://abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi`

Fig. 4.5 – Multiple Alignment of N-terminal region of *S.passalidarum´s ACS1 (SPAPADRAFT_135964)* and *S.cerevisiae's ACS1 (YAL054C)* using ClustalW 2.1 EMBL-EBI. Although there is a significant similarity at around 300 - 360. amino acids, there is only little similarity within the first 100 amino acids.

Fig. 4.6 – Multiple Alignment of C-terminal region of *Spathaspora passalidarum´s ACS1 (SPAPADRAFT_135964)*, *Salmonella enterica´s ACS (STM4275)* and *Streptomyces lividans´ ACS (SLIV_20470)* using ClustalW 2.1 (EMBL-EBI). Despite the overall similarity is low, there is a conserved domain (blue frame), which contains a lysine that is acetylated and thus the enzyme is inactivated- which happens only in *Salmonella enterica*, not in *Streptomyces lividans*. The neighbouring amino acids upstream and downstream are dissimilar between those three organisms, leaving unclear whether or not this site is acetylated in *Spathaspora passalidarum*.

4.6 The tRNA pool of *Yarrowia lipolytica* differs significantly from other *Saccharomycetales*

In general, Nc values reflect well the overall impact of cellular pathways within an organism (Figure 4.2). Most interestingly, a clear physiological imprint becomes visible as the Nc values of the TCA cycle have a distinct tendency for *S. cerevisiae* and *Y. lipolytica*, which are strong fermentative and respiratory, respectively. Thus, this clear distinction between fermentative and respiratory phenotypes agrees well with the findings of Man *et al.* [69], who investigated the impact of the tRNA pool on basic physiological characteristics in yeasts. Another important aspect in choosing candidate genes is the similarity of tRNA gene copy number composition between organism of origin and expression host (Figure 4.7). Here, however, *Y. lipolytica* is the most dissimilar species when compared to *S. cerevisiae*, *C. glabrata* and *S. passalidarum*, although still belonging to the same taxonomic clade. This dissimilarity might arise from the phylogenetic distance between the rest of the other *Saccharomycetales* as well as the high similarity in between those species. When reconsidering the genes of choice, it is remarkable, that the heterologous tAI values never fall below 70 % of the tAI value of *Yarrowia lipolytica´*s counterpart. The significance of tRNA adaptation in heterologous expression will be further examined in the Discussion.

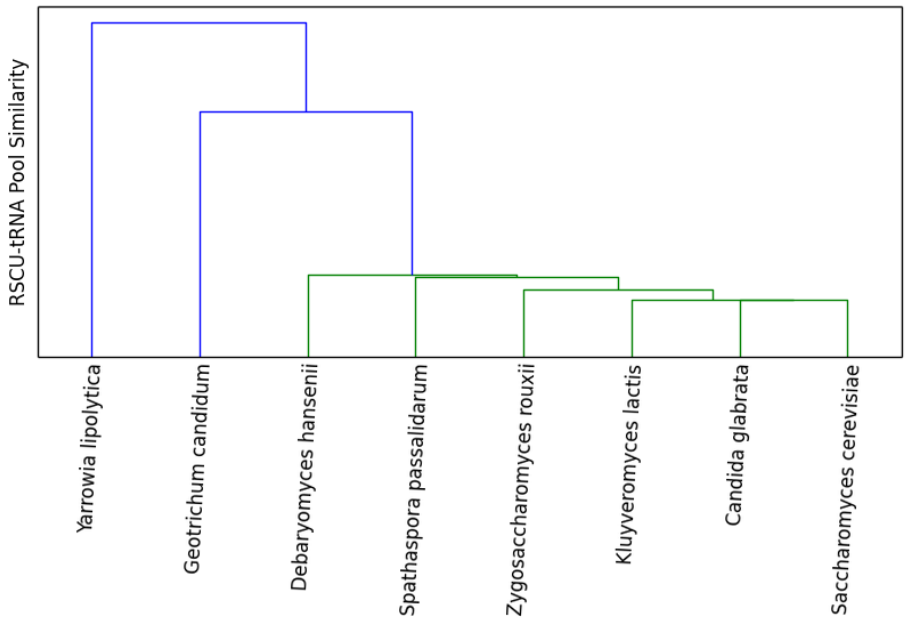


Fig. 4.7 – Cluster of various yeasts according to similarity in tRNA copy number composition. The genome sequences of different genes were subjected to tRNAscan-SE to retrieve tRNA genes. The tRNA isoacceptors were weighted using RSCU (Relative Synonymous Codon Usage). Among the *Saccharomycetaceae*, *Saccharomyces cerevisiae* and *Candida glabrata* are most similar in their tRNA gene composition, whereas *Yarrowia lipolytica* is comparably different and closer to *Geotrichum candidum*. Froissard *et al.* [70] reported a relationship between lipid composition and pylogenetic distance in different yeast species. Similarly, the tRNA pool of *Yarrowia lipolytica* is more similar to *Geotrichum candidum* another species of the *Dipodascaceae* family, than to post-WGD (whole genome duplication) clade yeasts such as *S. cerevisiae* and *C. glabrata*.

4.7 Construction of expression cassettes

To increase cytosolic acetyl-CoA levels in *Yarrowia lipolytica*, the pyruvate decarboxylase from *Candida glabrata (CAGL0M07920g)*, the $NADP^+$-dependent aldehyde dehydrogenase (*YPL061W*) deriving from *Saccharomyces cerevisiae* and the acetyl-CoA synthetase of *SPAPADRAFT_135964* of *Spathaspora passalidarum* will be over-

expressed under strong native promotors of the TCA cycle. Since overexpression-plasmids are not as stably maintained in *Yarrowia lipolytica* as in *Escherichia coli*, these overexpression cassettes will be integrated close to glycolytic genes, hence at highly transcribed regions (Figure 4.8).

Each of the fragments (Table 4.7) has been designed so that it is overlapping by 20 - 22 base pairs to the neighbouring fragment, to join them using isothermal assembly (Gibson cloning) without the need of (many) different restriction sites and ligation steps.

Table 4.7 – Components of the expression cassettes

Fragment ID	**Locus**
ALD6	*Saccharomyces cerevisiae YPL061W* Chromosome XVI 432588..434090
Insertion site upstream (for *ALD6*)	*Yarrowia lipolytica* Chromosome VI 2249325..2250074
Insertion site downstream (for *ALD6*)	*Yarrowia lipolytica* Chromosome VI 2250075..2250824
Promotor of citrate synthase (for *ALD6*	*Yarrowia lipolytica* Chromosome V 65192..64592
PDC1	*Candida glabrata CAGL0M07920g* Chromosome XIII 789467..791161
Insertion site upstream (for *PDC1*)	*Yarrowia lipolytica* Chromosome VI 2250750..2250001
Insertion site downstream (for *PDC1*)	*Yarrowia lipolytica* Chromosome VI 2250000..2249251
Promotor of isocitrate dehydrogenase (for *PDC1*)	*Yarrowia lipolytica* Chromosome V 569160..568561
URA3	*Yarrowia lipolytica*'s orotidine-5'-phosphate decarboxylase

The single fragments were generated and amplified using PCR as described in Ma-

terials and Methods. Two different buffer systems "A" and "B" as well as different annealing temperatures have been tested. The different PCR setups for the fragments are summarized in Table 4.8.

Table 4.8 – PCR conditions

Fragment	**Buffer**	**Tm**	**Elongation Time**	**Template**
ALD_HISterm	A	52	3	CEN.PK cut with BamHI
ALD_overlap	A	52	3	purified ALD_HISterm
PDC	B	50	3,50	uncut *Candida glabrata*
PDC_HISterm	B	55	3,50	purified PDC
PDC_overlap	B	55	3,50	purified PDC_HisTerm
PDC_overlap	B	55	3,50	purified PDC_HisTerm
promotors	A	52	2	*Yarrowia lipolytica* cut with BamHI
insertion sites (upstream and downstream)	A	52	2	*Yarrowia lipolytica* cut with XbaI

Gibson cloning was performed as noted in Materials and Methods, whereas 150 ng of the smallest fragment have been used and the rest the other fragments were added to equimolar amounts.

Gibson assembly failed twice, yielding either no or negative colonies, for the overexpression of *ALD6* suggesting that six different fragments are exceeding the capability of the Gibson assembly for this sample. Thus, some fragments were merged using overlap PCR (as described in Materials and Methods). Thus for both *ALD6* and *PDC1*, two fragments have been generated: one "upper fragment" composed of the upstream insertion site, the *URA3* marker and a promotor, as well as a "downstream fragment" consisting of the gene (e.g. "PDC1_overlap") and the downstream insertion site. Despite the overlap PCRs for the upstream fragments were quite efficient, each downstream fragment had low yield and did not yield significant product in a

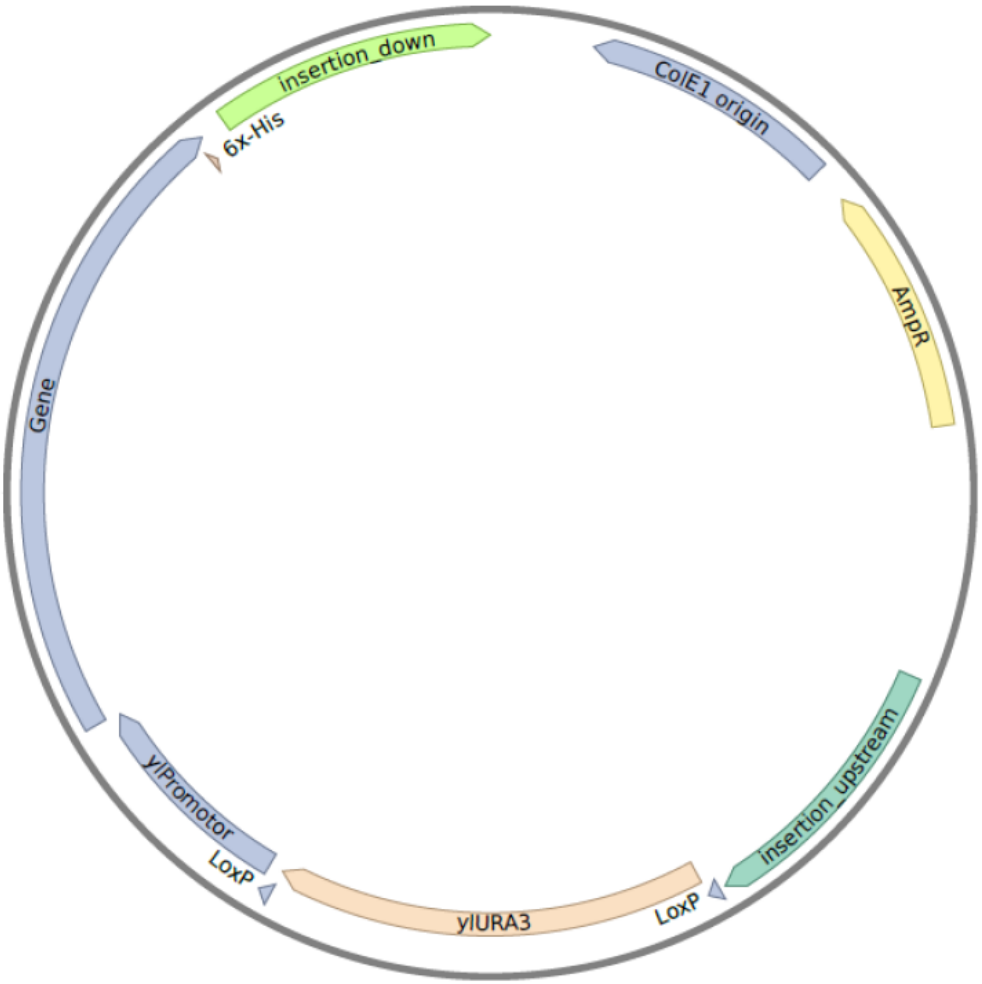


Fig. 4.8 – Standard expression cassette embedded in a pFA-URA3-hp4d plasmid backbone. At either end of the expression cassette is an upstream and downstream region which is the targeted insertion site in *Yarrowia lipolytica*. A *URA3* auxotrophic marker is included to screen for integrated cassettes. This marker is flanked by IoxP sites in order to remove the marker from the genome once proper integration has been confirmed. The gene, either *ALD6* or *PDC1* is preceded by a strong promotor deriving from genes of the TCA cycle in *Yarrowia lipolytica*. A polyhistidine tag is added to the gene for protein analysis later on as well as a strong synthetic terminator region. At the ends of the expression cassettes are cut sites for restriction enzymes (EcoRI and HindIII) in order to integrate this linear cassette in the pFA-URA3-hp4d. Thus it can be stably amplified in *Escherichia coli*. For integration, the cassette will be cut out from the vector and the linear fragment will be transformed into *Yarrowia lipolytica*

subsequent standard PCR using the fused fragment. A possible reason for this lack of success might be the overlapping region between the C-terminal end of the gene, which is the AT-rich terminator (18% GC), and the N-terminus of the downstream insertion site (50 - 60 % GC). This unequal GC distribution might lead to an inefficient hybridisation of primer and template during PCR. Thus, an approximately 20 bp linker was designed to bridge the gap in GC content between the synthetic terminator and the downstream insertion site.

Further experimental steps such as fusion of upstream -, downstream fragment and vector backbone using Gibson assembly and/or overlap extension PCR are envisaged.

5 Discussion

In this work, flux through a short pathway in *Yarrowia lipolytica*, the pyruvate dehydrogenase bypass, shall be increased by heterologously expressing putative highly active enzymes from other species, chosen based on their degree of codon usage bias.

5.1 The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a source of orthologous enzymes, but can be misleading

Orthologous enzymes of the pyruvate decarboxylase, the aldehyde dehydrogenase as well as the acetyl-CoA synthetase were screened using the Kyoto Encyclopedia of Genes and Genomes (KEGG), as this database sorts its entries in groups of orthologous proteins ('K numbers'), hence enzymes catalysing the same reaction in different organisms. Although this organisation is perfectly fitting the goal of this work, this database has some drawbacks. For some species little is known besides the genome sequence, so that these K numbers are assigned by homology search. Thus, there are entries in KEGG that have no K number assigned if the homology to other enzymes is low. Hence, there is a general risk of overseeing possible candidate genes if the screening solely relies on KEGG. This can easily be prevented by also searching in NCBI (Gene database) for the respective enzymes. Another shortcoming of databases in general is that entries might be simply wrong. For example, the aldehyde dehydrogenases 4 to 6 of *Saccharomyces cerevisiae* have been experimentally verified to be $NADP^+$ dependent [62]. Nevertheless, these isoforms have been assigned K numbers for NAD^+ dependency, although the description of these enzymes states $NADP^+$ dependency. Since this implies that database entries for cofactor dependencies are misleading, the screening in such cases has to be verified with (published) experimental data.

Unsurprisingly, although the screening was not restricted to yeasts, all three enzymes (Pdc, Ald and Acs) were found to be among the most biased in different yeast species. On the whole, since non-oleaginous yeasts lack an ATP citrate lyase, they have to rely on this pathway for direct production of acetyl-CoA in the cytoplasm. Furthermore, the pyruvate decarboxylase is involved in the production of ethanol as it produces acetaldehyde, the substrate for the alcohol dehydrogenase, making Crabtree-positive yeasts such as *Saccharomyces cerevisiae* plausible sources.

5.2 A selection process based on different codon usage bias indices - the more biased the better ?

The key assumption underlying the screening procedure is that the highly active enzymes are more intensely subjected to evolutionary selection pressures (such as mutation and drift) than the less active ones.

There are two ways to estimate evolutionary forces. Different codon usage bias indices calculate the current state of bias resembling an evolutionary endpoint, while more elaborate models determine codon substitution rates as approximation for evolutionary strengths exhibited on an enzyme.

In this work, orthologous enzymes have been ranked based on their codon usage bias values from CDC, Nc and tAI indices, hence an evolutionary time point as mentioned previously.
The usage of different indices should assure, that the screen is not mislead by an intrinsic bias of only one method. Apart from these indices, evolutionary (optimization) forces can also be measured in rates of non-synonymous and synonymous substitution, respectively. Du *et al.* [71], for example, employed a complex phylogenetic mutation-selection likelihood model to investigate the evolutionary forces on rhodopsins in mammals and discovered that evolutionary substitution rates vary significantly between motifs of the protein, e.g. low rates at well-conserved sites of splicing enhancers.

Here, the general assumption is that more biased genes might be more optimized in terms of kinetic properties and overall fitness such as solubility, half life and stability,

because they are under higher evolutionary selective pressure. This in turn implies that it is not only the amino acid sequence which contributes to the properties of a protein but also the choice synonymous codons. Gorochowski *et al.* [72] pointed out that the translation process has to be well-paced, mediated by mRNA structure and different tRNA isoacceptors, to allow proteins to be properly folded. Thus, it is not surprising, that even *slowing down* the translation speed is helpful for heterologous protein expression, especially when (eukaryotic) multidomain proteins are expressed in bacterial expression hosts [73, 74]. This however implies, that the highest biased enzymes in an organism might not automatically be the 'best'. Moreover although kinetic properties might remain unchanged, because the amino acid sequence is unchanged, the folding of heterologously expressed proteins is altered - thus the specific activity - as the tRNA copy number and furthermore the tRNA isoacceptor concentrations might be different in expression hosts and in the native organism [75].

In this work, the tAI was applied as a proxy for expressibility of a given orthologous protein in *Yarrowia lipolytica*. It was found, that all of the candidate genes have smaller tAI values than *Yarrowia lipolytica*'s intrinsic ones. This would disqualify these orthologs from being expressed in *Yarrowia lipolytica*, as it might be either weakly expressed or aggregate due to an apparent lack of adaptation to local tRNA isoacceptor levels. Nevertheless, abovementioned findings also imply that an extremely high tAI value is also not beneficial, as translation might progress too fast for a protein to fold properly.
Thus the tAI values should rather be related to the intrinsic gene, and might not even be disadvantageous if they are comparably lower. Although proteins tend to aggregate if the adaptation to the tRNA pool is too low, it is difficult to set a threshold for tAI values, as this depends on other aspects, such as the host organism, the heterologous protein and the extent of overexpression.

This work aimed to optimise the potential yield of acetyl-CoA, not to investigate whether the key hypothesis of evolutionary optimization is actually valid.

Ma *et al.* [76], however tested exactly this hypothesis - though the present work is not based on this paper - by integrating and overexpressing genes for succinate overproduction in *Escherichia coli* under exact the same promotors. They found

an exponential correlation of Nc values and succinate production, with higher bias resulting in a higher yield. On the other hand, they also compared Km values for glycolytic genes. Even though they seem to support the reaction flux - Nc correlation in some cases, there is no clear correlation and some are even contradicting. For example, *Escherichia coli*'s pyruvate kinase (Pyk) has a similar Nc value like *Bacillus subtilis*'s Pyk, although the latter one has a Km value almost twice as high. However, *Corynebacterium glutamicum´*s Pyk indeed is stronger biased and also has a Km value lower than that of *Bacillus subtilis*. It seems only logical that codon usage bias indices would rather resemble enzymes optimized for stability, conveyed by the composition of synonymous codons influenced by the tRNA pool as mentioned previously, rather than enzyme catalysis stemming from distinct amino acids, hence evolutionary selection of non-synonymous codons. Ma et al. [76] restricted their comparison to Km values, although maximal velocity of catalysis (Vmax) as well as the turnover (kcat) are also major kinetic parameters. The lack of clear correlation between flux and Km values is an evidence that synonymous codon usage apparently has little impact on enzyme activity with regard to the kinetic properties, but is influencing the *specific* enzyme activity and modulation of translation and mRNA stability which ultimately result in a *properly* folded protein.

A myriad of studies have been undertaken comparing the kinetic properties of different enzymes, such as between mutant and wild type. However, there is an apparent lack in studies of orthologous proteins, expressed and studied under exactly the same conditions. In the abovementioned work, only Km values have been compared but not Vmax and kcat, which would eventually correlate better with observed fluxes than Km values.

5.3 The cloning strategy aims to optimize the potential yield, not to verify the impact of codon usage bias on reaction flux

When looking at the cloning strategy in detail, it becomes obvious, that this work aims to enhance the acetyl-CoA pool best as possible in the first place, not proving this general evolutionary hypothesis: The genes will be inserted in different places of the genome, in the neighbourhood of highly expressed (glycolytic) genes where the chromatin structure seems to be more accessible for transcription factors. Additionally these genes have different strong promotors - from genes coding for TCA cycle enzymes. This should lead to a high expression of the target genes but should also 'redirect' the respective transcription factors from the TCA cycle, so that less citrate would be produced at that site and eventually excreted. A further advantage of choosing different promotors is the reduced risk of recombination between identical (promotor) sequences, if the sequence is only duplicated in the genome (one for the native gene and one for heterologous expression) than when this promotor sequence would appear many times within the genome.

5.4 Potential side-effects of enhancing the PDH bypass

As mentioned before, flux through the PDH bypass will be increased so that more acetyl-CoA is produced which may in turn also increase lipid biogenesis, as this is the key building block. However, there are also some risks associated with increasing activity of this pathway. Firstly, the pyruvate decarboxylase is a highly active enzyme in fermentative yeasts, since this enzyme is also involved in the biosynthesis of ethanol.

Is there a chance of *Yarrowia lipolytica* getting Crabtree-positive if a pyruvate decarboxylase is overexpressed?

VanUrk *et al.* [77] and Hagman *et al.* [78] investigated the different physiological responses to glucose excess in Crabtree-positive and Crabtree-negative yeasts. They found, that glucose uptake is increased during excess of glucose in Crabtree-positive yeasts, which subsequently leads to an *intracellular* excess of glucose which is metabolised to pyruvate and further to ethanol. Contrarily, Crabtree-negative yeasts, use the substrate excess to accumulate biomass (thus storage metabolites). Moreover, the activity of the pyruvate dehydrogenase bypass is different between Crabtree-positive and - negative yeasts. While ethanol-producing yeasts exhibit an active pyruvate decarboxylase, its activity is relatively low in Crabtree-negative yeasts.

Summing up, *Yarrowia lipolytica* is only prone to turn Crabtree-positive if the level of pyruvate in the cell is high and the pyruvate decarboxylase is excessively more active than the subsequent aldehyde dehydrogenase. Thus it seems unlikely that *Yarrowia lipolytica* secretes ethanol solely by genetic engineering of the PDH-bypass. Still, mutants will also be checked for ethanol secretion as experimental verification.

Within the PDH bypass acetaldehyde, which is produced by the pyruvate decarboxylase is further metabolised to acetate. As acetaldehyde is toxic, there is a risk of creating a lethal phenotype in case the aldehyde dehydrogenase (or any other enzyme) fails to further convert this metabolite. Lastly, acetate is metabolized to acetyl-CoA by the acetyl-CoA synthetase. As already pointed out in the Results, this enzyme might also be feedback regulated, so that excess of acetyl-CoA leads to downregulation of this enzyme. Thus, as already noted in the Results, an acetylation site mutant of *ACS* is envisaged, to avoid negative feedback regulation. Moreover, the envisaged additional cytosolic acetate and acetyl-CoA,respectively might not be completely used for lipid biogenesis but is transported into the nucleus, where it might lead to global histone-acetylation, hence global transcription control. High level of acetyl-CoA in the cytoplasm or the nucleus are known to have an impact on the transcription of autophagy-related genes [79], which is repressed upon histone hyperacetylation, as well as cell cycle progression by inducing transcription of the *CLN3* cyclin [80] in *Saccharomyces cerevisiae*.

5.5 Outlook

Most of the present work was dedicated to explore the applicability of different codon usage indices and the setup of Python scripts. Therefore the original goal to generate a mutant strain having a heterologous PDH bypass could not be achieved, though some cloning steps towards an expression cassette have been undertaken. Thus, single mutants as well as mutant strains containing the complete heterologous pathway will be generated and tested for protein expression, citrate secretion as well as lipid production. Finally, future research work will also be dedicated to combine resulting valuable mutants with others (e.g. for enhanced substrate uptake) in order to optimally redirect cellular resources towards lipid biogenesis.

6 Bibliography

[1] Coelho, M.A.Z., Amaral, P.F.F., Belo, I.: Yarrowia lipolytica : an industrial workhorse. Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology, 930–944 (2010)

[2] Fickers, P., Nicaud, J.M., Gaillardin, C., Destain, J., Thonart, P.: Carbon and nitrogen sources modulate lipase production in the yeast Yarrowia lipolytica. Journal of Applied Microbiology **94**, 742–749 (2004). doi:10.1111/j.1365-2672.2004.02190.x

[3] Silva, L.V., Tavares, C.B., Amaral, P.F.F., Coelho, M.A.Z.: Production of Citric Acid by Yarrowia lipolytica in Different Crude Glycerol Concentrations and in Different Nitrogen Sources. Chemical Engineering Transactions **27**, 199–204 (2012)

[4] Nicaud, J.M.: Yarrowia lipolytica. Yeast **29**(10), 409–418 (2012). doi:10.1002/yea

[5] Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.-M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.-F., Straub, M.-L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., Souciet, J.-L.: Genome evolution in yeasts. Nature **430**(6995), 35–44 (2004). doi:10.1038/nature02579

[6] Kerscher, S., Durstewitz, G., Casaregola, S., Gaillardin, C., Brandt, U.: The complete mitochondrial genome of Yarrowia lipolytica. Comparative and Functional Genomics **2**(2), 80–90 (2001). doi:10.1002/cfg.72

[7] Liu, L., Alper, H.S.: Draft Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica PO1f, a Commonly Used Metabolic Engineering Host. Genome Announcements **2**(4), 00652–14 (2014). doi:10.1128/genomeA.00652-14

[8] Juretzek, T., Le Dall, M., Mauersberger, S., Gaillardin, C., Barth, G., Nicaud, J.M.: Vectors for gene expression and amplification in the yeast Yarrowia lipolytica. Yeast **18**(2), 97–113 (2001). doi:10.1002/1097-0061(20010130)18:2<97::AID-YEA652>3.0.CO;2-U

[9] Blazeck, J., Liu, L., Redden, H., Alper, H.S.: Tuning gene expression in Yarrowia lipolytica by a hybrid promoter approach. Applied and Environmental Microbiology **77**(22), 7905–14 (2011). doi:10.1128/AEM.05763-11

[10] Curran KA, Morse NJ, Markham KA, Wagman AM, Gupta A, A.H.: Short, Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. ACS Synthetic Biology **4**(7), 824–3 (2015). doi:10.1021/sb5003357

[11] Liang, M.H., Jiang, J.G.: Advancing oleaginous microorganisms to produce lipid via metabolic engineering technology. Progress in Lipid Research **52**(4), 395–408 (2013). doi:10.1016/j.plipres.2013.05.002

[12] Ageitos, J.M., Vallejo, J.A., Veiga-Crespo, P., Villa, T.G.: Oily yeasts as oleaginous cell factories. Applied Microbiology and Biotechnology **90**(4), 1219–1227 (2011). doi:10.1007/s00253-011-3200-z

[13] Nielsen, J.: Synthetic Biology for Engineering Acetyl Coenzyme A Metabolism in Yeast **5**(6), 14–16 (2014). doi:10.1128/mBio.02153-14.Copyright

[14] Pan, P., Hua, Q.: Reconstruction and in silico analysis of metabolic network for an oleaginous yeast, Yarrowia lipolytica. PLoS ONE **7**(12), 51535 (2012). doi:10.1371/journal.pone.0051535

[15] Loira, N., Dulermo, T., Nicaud, J.-M., Sherman, D.: A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica. BMC Systems Biology **6**(1), 35 (2012). doi:10.1186/1752-0509-6-35

[16] Tai, M., Stephanopoulos, G.: Engineering the push and pull of lipid biosynthesis in oleaginous yeast Yarrowia lipolytica for biofuel production. Metabolic engineering **15**, 1–9 (2013). doi:10.1016/j.ymben.2012.08.007

[17] Qiao, K., Imam Abidi, S.H., Liu, H., Zhang, H., Chakraborty, S., Watson, N., Kumaran Ajikumar, P., Stephanopoulos, G.: Engineering lipid overproduction in the oleaginous yeast Yarrowia lipolytica. Metabolic Engineering **29**, 56–65 (2015). doi:10.1016/j.ymben.2015.02.005

[18] Blazeck, J., Hill, A., Liu, L., Knight, R., Miller, J., Pan, A., Otoupal, P., Alper, H.S.: Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production. Nature Communications **5**, 3131 (2014). doi:10.1038/ncomms4131

[19] Lazar, Z., Dulermo, T., Neuvéglise, C., Crutz-Le Coq, A.-M., Nicaud, J.-M.: Hexokinase—A limiting factor in lipid production from fructose in Yarrowia lipolytica. Metabolic Engineering **26**, 89–99 (2014). doi:10.1016/j.ymben.2014.09.008

[20] Zhao, C.H., Cui, W., Liu, X.Y., Chi, Z.M., Madzak, C.: Expression of inulinase gene in the oleaginous yeast Yarrowia lipolytica and single cell oil production from inulin-containingmaterials. Metabolic Engineering **12**(6), 510–517 (2010). doi:10.1016/j.ymben.2010.09.001

[21] Beopoulos, A., Mrozova, Z., Thevenieau, F., Le Dall, M.-T., Hapala, I., Papanikolaou, S., Chardot, T., Nicaud, J.-M.: Control of lipid accumulation in the yeast Yarrowia lipolytica. Applied and Environmental Microbiology **74**(24), 7779–89 (2008). doi:10.1128/AEM.01412-08

[22] Wang, G.-Y., Zhang, Y., Chi, Z., Liu, G.-L., Wang, Z.-P., Chi, Z.-M.: Role of pyruvate carboxylase in accumulation of intracellular lipid of the oleaginous yeast Yarrowia lipolytica ACA-DC 50109. Applied Microbiology and Biotechnology **99**(4), 1637–1645 (2015). doi:10.1007/s00253-014-6236-z

[23] Wang, W., Wei, H., Alahuhta, M., Chen, X., Hyman, D., Johnson, D.K., Zhang, M., Himmel, M.E.: Heterologous Expression of Xylanase Enzymes in Lipogenic Yeast Yarrowia lipolytica. PLoS ONE **9**(12), 111443 (2014). doi:10.1371/journal.pone.0111443

[24] Tuite, M.F., Santos, M.A.S.: Codon reassignment in Candida species: An evolutionary conundrum. Biochimie **78**(11-12), 993–999 (1996). doi:10.1016/S0300-9084(97)86722-3

[25] Bulmer, M.: The selection-mutation-drift theory of synonymous codon usage. Genetics **129**(3), 897–907 (1991). doi:10.1002/yea.320070702

[26] Sharp, P.M., Emery, L.R., Zeng, K.: Forces that influence the evolution of codon bias. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **365**(1544), 1203–1212 (2010). doi:10.1098/rstb.2009.0305

[27] Zuckerkandl, E., Pauling, L.: Molecules as Documents of Evolutionary History. Journal of theoretical biology **8**(2), 357–366 (1965). doi:10.1016/0022-5193(65)90083-4

[28] Bali, V., Bebok, Z.: Decoding mechanisms by which silent codon changes influence protein biogenesis and function. The International Journal of Biochemistry & Cell Biology **64**, 58–74 (2015). doi:10.1016/j.biocel.2015.03.011

[29] Hunt, R.C., Simhadri, V.L., Iandoli, M., Sauna, Z.E., Kimchi-Sarfaty, C.: Exposing synonymous mutations. Trends in Genetics **30**(7), 308–321 (2014). doi:10.1016/j.tig.2014.04.006

[30] Shabalina, S.A., Spiridonov, N.A., Kashina, A.: Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research **41**(4), 2073–94 (2013). doi:10.1093/nar/gks1205

[31] Hurst, L.D.: Molecular genetics: The sound of silence. Nature **471**(7340), 582–583 (2011). doi:10.1038/471582a

[32] Yang, J.R., Chen, X., Zhang, J.: Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biology **12**(7), 1–14 (2014).

doi:10.1371/journal.pbio.1001910

[33] Agashe, D., Martinez-Gomez, N.C., Drummond, D.A., Marx, C.J.: Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Molecular Bology and Evolution **30**(3), 549–60 (2013). doi:10.1093/molbev/mss273

[34] Kudla, G., Murray, A., Tollervey, D., Plotkin, J.: Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science **324**(April), 255–258 (2009)

[35] Elena, C., Ravasi, P., Castelli, M.E., Peirú, S., Menzella, H.G.: Expression of codon optimized genes in microbial systems: current industrial applications and perspectives. Frontiers in Microbiology **5**(2), 21 (2014). doi:10.3389/fmicb.2014.00021

[36] Chung, B.K.-S., Lee, D.-Y.: Computational codon optimization of synthetic gene for protein expression. BMC Systems Biology **6**(1), 134 (2012). doi:10.1186/1752-0509-6-134

[37] Lanza, A.M., Curran, K.a., Rey, L.G., Alper, H.S.: A condition-specific codon optimization approach for improved heterologous gene expression in Saccharomyces cerevisiae. BMC Systems Biology **8**(1), 33 (2014). doi:10.1186/1752-0509-8-33

[38] Roth, A., Anisimova, M., Cannarozzi, G.M.: Measuring codon usage bias. In: Cannarozzi, G.M., Schneider, A. (eds.) Codon Evolution - Mechanisms and Models, pp. 189–217. Oxford University Press, Oxford (2012)

[39] Friberg, M., von Rohr, P., Gonnet, G.: Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae. Yeast **21**(13), 1083–93 (2004). doi:10.1002/yea.1150

[40] Hershberg, R., Petrov, D.: On the Limitations of Using Ribosomal Genes as References for the Study of Codon Usage: A Rebuttal. PLoS ONE **7**(12) (2012). doi:10.1371/journal.pone.0049060

[41] Sharp, P.M., Tuohy, T.M.F., Mosurskil, K.R.: Codon usage in yeast : cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids

Research **14**(13), 5125–5143 (1986)

[42] Wang, B., Shao, Z.Q., Xu, Y., Liu, J., Liu, Y., Hang, Y.Y., Chen, J.Q.: Optimal codon identities in bacteria: Implications from the conflicting results of two different methods. PLoS ONE **6**(7) (2011). doi:10.1371/journal.pone.0022714

[43] O'Neill, P.K., Or, M., Erill, I.: scnRCA: A Novel Method to Detect Consistent Patterns of Translational Selection in Mutationally-Biased Genomes. PLoS ONE **8**(10) (2013). doi:10.1371/journal.pone.0076177

[44] Ran, W., Higgs, P.G.: Contributions of Speed and Accuracy to Translational Selection in Bacteria. PLoS ONE **7**(12) (2012). doi:10.1371/journal.pone.0051652

[45] Zhou, J.-H., You, Y.-N., Chen, H.-T., Zhang, J., Ma, L.-N., Ding, Y.-Z., Pejsak, Z., Liu, Y.-S.: The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus. Infection, Genetics and Evolution **16**, 270–274 (2013). doi:10.1016/j.meegid.2013.02.017

[46] Novoa, E.M., Pavon-Eternod, M., Pan, T., Ribas De Pouplana, L.: A role for tRNA modifications in genome structure and codon usage. Cell **149**(1), 202–213 (2012). doi:10.1016/j.cell.2012.01.050

[47] Rocha, E.P.C.: Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Research **14**(11), 2279–2286 (2004). doi:10.1101/gr.2896904

[48] dos Reis, M., Savva, R., Wernisch, L.: Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research **32**(17), 5036–44 (2004). doi:10.1093/nar/gkh834

[49] Wright, F.: The 'effective number of codons' used in a gene. Gene **87**, 23–29 (1990)

[50] Banerjee, T., Gupta, S.K., Ghosh, T.C.: Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene". Biochemical and Biophysical Research Communications **330**(4), 1015–1018 (2005). doi:10.1016/j.bbrc.2005.02.150

[51] Sun, X., Yang, Q., Xia, X.: An Improved Implementation of Effective Number of Codons (Nc). Molecular Bology and Evolution **30**(1), 191–196 (2012). doi:10.1093/molbev/mss201

[52] Zhang, Z., Li, J., Cui, P., Ding, F., Li, A., Townsend, J.P., Yu, J.: Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics **13**(1), 43 (2012). doi:10.1186/1471-2105-13-43

[53] Ebrahim, A., Lerman, J., Palsson, B.O., Hyduke, D.R.: COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology **7**(1), 74 (2013). doi:10.1186/1752-0509-7-74

[54] Gola, S., Martin, R., Walther, A., Dünkler, A., Wendland, J.: New modules for PCR-based gene targeting in Candida albicans: Rapid and efficient gene targeting using 100 bp of flanking homology region. Yeast **20**(16), 1339–1347 (2003). doi:10.1002/yea.1044

[55] Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., Knop, M.: A versatile toolbox for PCR-based tagging of yeast genes: New fluorescent proteins, more markers and promoter substitution cassettes. Yeast **21**(11), 947–962 (2004). doi:10.1002/yea.1142

[56] Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison, C.A., Smith, H.O.: Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science **319**(5867), 1215–1220 (2008). doi:10.1126/science.1151721

[57] Merryman, C., Gibson, D.G.: Methods and applications for assembling large DNA constructs. Metabolic Engineering **14**(3), 196–204 (2012). doi:10.1016/j.ymben.2012.02.005

[58] Ellis, T., Adie, T., Baldwin, G.S.: DNA assembly for synthetic biology: from parts to pathways and beyond. Integrative Biology **3**(2), 109–118 (2011). doi:10.1039/c0ib00070a

[59] Gibson, D.G.: Enzymatic assembly of overlapping DNA fragments. Methods in Enzymology **498**, 349–361 (2011). doi:10.1016/B978-0-12-385120-8.00015-2

[60] Wang, Q., He, P., Shen, A., Jiang, N., Wang, Q.: Purification, Characterization, Cloning and Expression of Pyruvate Decarboxylase from Torulopsis glabrata IFO005. Journal of Biochemistry **455**, 447–455 (2004). doi:10.1093/jb/mvh

[61] Liu, L.M., Li, Y., Li, H.Z., Chen, J.: Manipulating the pyruvate dehydrogenase bypass of a multi-vitamin auxotrophic yeast Torulopsis glabrata enhanced pyruvate production. Letters in Applied Microbiology **39**(2), 199–206 (2004). doi:10.1111/j.1472-765X.2004.01563.x

[62] Saint-Prix, F., Bönquist, L., Dequin, S.: Functional analysis of the ALD gene family of Saccharomyces cerevisiae during anaerobic growth on glucose: The NADP+-dependent Ald6p and Ald5p isoforms play a major role in acetate formation. Microbiology **150**(7), 2209–2220 (2004). doi:10.1099/mic.0.26999-0

[63] Meaden, P.G., Dickinson, F.M., Mifsud, A., Tessier, W.: The ALD6 Gene of Saccharomyces cerevisiae Encodes a Cytosolic , Mg 2 + -Activated Acetaldehyde Dehydrogenase. Yeast **1327**, 1319–1327 (1997)

[64] Byrne-Bailey, K.G., Coates, J.D.: Complete genome sequence of the anaerobic perchlorate-reducing bacterium Azospira suillum strain PS. Journal of Bacteriology **194**(10), 2767–2768 (2012). doi:10.1128/JB.00124-12

[65] Hou, X.: Anaerobic xylose fermentation by Spathaspora passalidarum. Applied Microbiology and Biotechnology **94**(1), 205–214 (2012). doi:10.1007/s00253-011-3694-4

[66] Takahashi, H., McCaffery, J.M., Irizarry, R.A., Boeke, J.D.: Nucleocytosolic Acetyl-Coenzyme A Synthetase Is Required for Histone Acetylation and Global Transcription. Molecular Cell **23**(2), 207–217 (2006). doi:10.1016/j.molcel.2006.05.040

[67] Starai, V.J., Gardner, J.G., Escalante-Semerena, J.C.: Residue Leu-641 of acetyl-CoA synthetase is critical for the acetylation of residue Lys-609 by the protein acetyltransferase enzyme of Salmonella enterica. Journal of Biological Chemistry **280**(28), 26200–26205 (2005). doi:10.1074/jbc.M504863200

[68] Tucker, A.C., Escalante-Semerena, J.C.: Determinants within the C-terminal domain of Streptomyces lividans acetyl-CoA synthetase that block acetylation of its active site lysine in vitro by the protein acetyltransferase (Pat) enzyme. PLoS ONE **9**(6) (2014). doi:10.1371/journal.pone.0099817

[69] Man, O., Pilpel, Y.: Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nature Genetics **39**(3), 415–421 (2007). doi:10.1038/ng1967

[70] Froissard, M., Canonge, M., Pouteaux, M., Cintrat, B., Mohand-Oumoussa, S., Guillouet, S.E., Chardot, T., Jacques, N., Casaregola, S.: Lipids containing medium-chain fatty acids are specific to post-whole genome duplication Saccharomycotina yeasts. BMC Evolutionary Biology **15**(1), 1–16 (2015). doi:10.1186/s12862-015-0369-2

[71] Du, J., Dungan, S.Z., Sabouhanian, A., Chang, B.S.: Selection on synonymous codons in mammalian rhodopsins: a possible role in optimizing translational processes. BMC Evolutionary Biology **14**(1), 96 (2014). doi:10.1186/1471-2148-14-96

[72] Gorochowski, T.E., Ignatova, Z., Bovenberg, R.A.L., Roubos, J.A.: Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. Nucleic Acids Research **43**(6), 3022–3032 (2015). doi:10.1093/nar/gkv199

[73] Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C., Barral, J.M.: Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency. Journal of Molecular Biology **396**(5), 1310–1318 (2010). doi:10.1016/j.jmb.2009.12.042

[74] Hess, A.-K., Saffert, P., Liebeton, K., Ignatova, Z.: Optimization of Translation Profiles Enhances Protein Expression and Solubility. PLoS ONE **10**(5), 0127039 (2015). doi:10.1371/journal.pone.0127039

[75] Zhu, J., Lin, J.-L., Palomec, L., Wheeldon, I.: Microbial host selection affects intracellular localization and activity of alcohol-O-acetyltransferase. Microbial Cell Factories **14**(1), 1–10 (2015). doi:10.1186/s12934-015-0221-9

[76] Ma, X., Zhang, X., Wang, B., Mao, Y., Wang, Z., Chen, T., Zhao, X.: Engineering microorganisms based on molecular evolutionary analysis: a succinate production case study. Evolutionary Applications **7**(8), 913–20 (2014). doi:10.1111/eva.12186

[77] Van Urk, H., Voll, W.S.L., Scheffers, W.A., Van Dijken, J.P.: Transient-state analysis of metabolic fluxes in Crabtree-positive and crabtree-negative yeasts. Applied and Environmental Microbiology **56**(1), 281–287 (1990)

[78] Hagman, A., Säll, T., Piškur, J.: Analysis on yeast short-term Crabtree effect and its origin. FEBS Journal **281**, (2014). doi:10.1111/febs.13019

[79] Schroeder, S., Pendl, T., Zimmermann, A., Eisenberg, T., Carmona-Gutierrez, D., Ruckenstuhl, C., Mariño, G., Pietrocola, F., Harger, A., Magnes, C., Sinner, F., Pieber, T.R., Dengjel, J., Sigrist, S.J., Kroemer, G., Madeo, F.: Acetyl-coenzyme A: A metabolic master regulator of autophagy and longevity. Autophagy **10**(7), 1335–1337 (2014). doi:10.4161/auto.28919

[80] Shi, L., Tu, B.P.: Acetyl-CoA induces transcription of the key G1 cyclin CLN3 to promote entry into the cell division cycle in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America **110**(18), 7318–23 (2013). doi:10.1073/pnas.1302490110

7 Appendix

7.1 Composition of Buffer and Media

If not declared otherwise, all buffers are stored at room temperature. Media and agar plates are stored at 4 °C.

Table 7.1 – Agarose Running Buffer TBE

Compound	**Amount**
TRIS	54 g
Boric acid	27.5 g
0.5 M EDTA pH 8	20 mL
dH_2O	to 5 L

Table 7.2 – YPD liquid medium (final volume : 800 mL)

Compound	**Amount**
Yeast extract	8 g
Glucose (monohydrate)	17.6 g
Peptone	16 g
dH_2O	to 800 mL

Glucose was autoclaved seperately from the other components. After autoclaving, it was added to the mix and the volume was readjusted to 800 mL with sterile dH_2O.

Table 7.3 – LB liquid medium (final volume : 800 mL)

Compound	Amount
Trypton	8 g
Yeast extract	4 g
NaCl	4 g
dH_2O	to 800 mL

For plasmid selection, ampicillin was added 1 : 1000 (stock concentration : 100 mg/mL).

Agar Plates: add 16 g agar / 800 mL medium and autoclave.

Table 7.4 – Breaking Buffer (DNA isolation)

Compound	Amount
Triton X-100 (stock concentration 10 %)	2 % v/v
SDS (stock concentration : 20 %)	1 % v/v
NaCl	100 mM
Tris-HCl	100 mM pH 8
EDTA	1 mM pH 8

Table 7.5 – 10 x TE-buffer (DNA isolation)

Compound	**Amount**
1M Tris-HCl pH 8.0 (stock concentration 100 mM)	100 mL
0.5M EDTA pH 8.0 (stock concentration 10 mM)	20 mL
ddH2O	880 mL

7.2 Devices and Materials

Table 7.6 – Chemicals

Chemical	Specifications	Supplier
Agarose		Biozyme LE Agarose Art No. 840004
NaCl	$\geq$ 88.5 % p.a. ACS ISO	Roth; Art. No. 3957.1
Agar		Bacto^{TM}Agar (BD); Ref. No. 214030
Peptone		Bacto^{TM}Peptone(BD); Ref. No. 211820
Yeast extract		Bacto^{TM}Yeast extract (BD); Ref. No. 212720
α - D(+) - Glucose monohydrate		Roth; Art No. 6887.2
SDS	Sodium dodecyl sulfate, suitable for electrophoresis, for molecular biology, $\geq$ 98.5% (GC)	Sigma-Aldrich, Art No. L3771
TRIS	Pufferan $\geq$ 99.9% p.a.	Roth, Art No. 4855.3
Triton X-100		Sigma -Aldrich, Art No. T-8787
Boric acid	$\geq$ 99,8 %, p.a., ACS, ISO	Roth
EDTA	Triplex	Merck, Art No. 1.08418.1000
ddH_2O	Water for chromatography, Art No. 1.15333.1000	LiChrosolv
Polymerase (and PCR components)	Herculase	Agilent

Table 7.7 – Laboratory Equipment

Device	**Specifications**	**Supplier**
Agarose chamber and power supply	Biorad Powerpac 3000	Biorad
Analytical scale	max: 220g, min: 10 mg d:0.1 mg	Kern ABJ
SpeedVac	SpeedVac conecentrator Refidgerated condensation trap RT400	Savant
SpeedVac- pump	VAP5	Vacuubrand
Centrifuge (large scale)	Eppendorf centrifuge 5810 R rotor: Eppendorf A-4-81 max : 4 * 1.4 kg, 4000 rpm buket: 58100	Servo Lab
Centrifuge (small scale)	Eppendorf 5415 R rotor: max : 13200 rpm, 24 * 3.75 g FL031 F45-24-11	Servo Lab
Pipettes	100 - 1000 μL, 20 - 200 μL, 2 - 10 μL Pipetman	Gilson
Pipette	0.1 - 2.5 μL	Eppendorf
Heater	Thermomixer comfort	Eppendorf
Heater	HTM130(L)	HLC
Vortex		IKA
Thermocycler	GeneAmp RCR system 9700	A&B Applied Biosystems
Thermocycler	PTC-200	Peltier