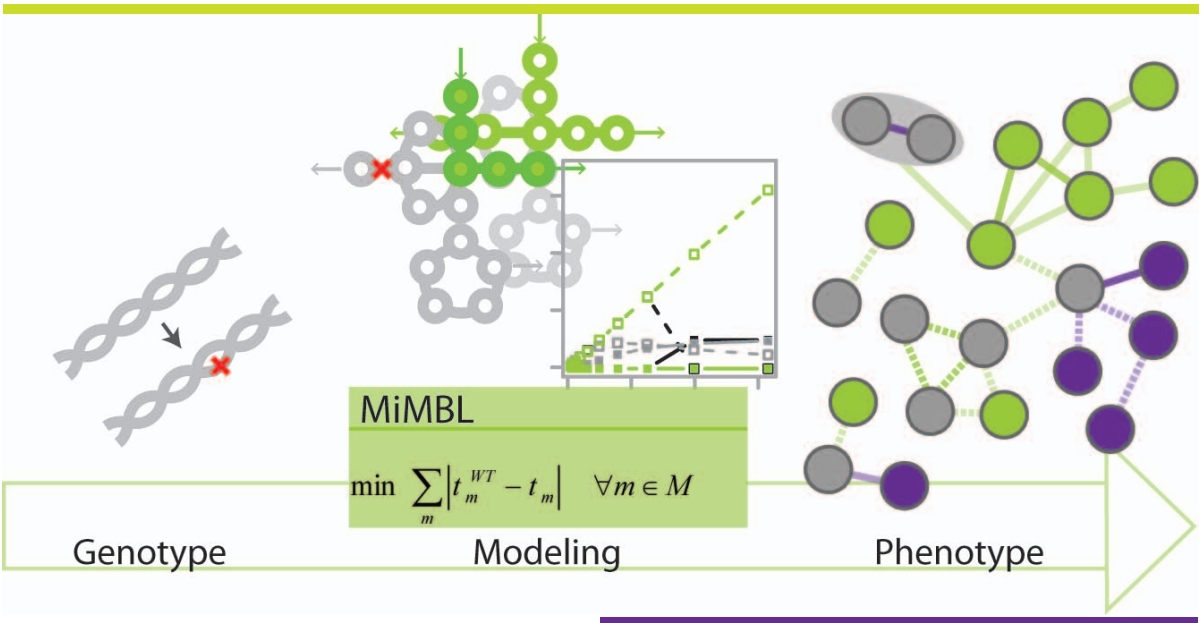
Investigating genotype-phenotype relationships in Saccharomyces cerevisiae metabolic network through stoichiometric modeling

Brochado, Ana Rita; Mortensen, Uffe Hasbro; Patil, Kiran Raosaheb

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Investigating genotype-phenotype relationships in *Saccharomyces cerevisiae* metabolic network through stoichiometric modeling



Ana Rita Brochado
Ph.D. Thesis
April 2012

Center for Microbial Biotechnology
Department of Systems Biology

Technical University of Denmark

Investigating genotype-phenotype relationships in *Saccharomyces cerevisiae* metabolic network through stoichiometric modeling

Ana Rita Brochado

Ph.D. Thesis

April 2012

Center for Microbial Biotechnology
Department of Systems Biology

Address: Center for Microbial Biotechnology
Department of Systems Biology
Technical University of Denmark
Building 223
DK-2800 Kgs. Lyngby
Denmark
Phone: +45 4525 2525
Fax: +45 4588 4148
Web: www.cmb.dtu.dk

Print: J&R Frydenberg A/S
København
September 2012

ISBN: 978-87-91494-13-0

Technical University of Denmark

E-mail: anaritagb@gmail.com

Preface

The work described in this Ph.D. thesis was carried out at the Center for Microbial Biotechnology, Technical University of Denmark, from March 2008 to March 2012, under the joint supervision of Associate Professor Uffe H. Mortensen and Kiran Raosaheb Patil, Group Leader at the European Molecular Biology Laboratory in Heidelberg, Germany (EMBL). From September 2010 to the end of the Ph.D. I have been a visiting Ph.D. student at the Patil group, Computational Biology Unit, EMBL. The Portuguese governmental institution responsible for financing and evaluating the scientific research activities, Fundação para a Ciência e a Tecnologia, ensured financial support throughout the 4 years (ref. SFRH/BD/41230/2007). Travel expenses and conferences have been additionally supported by Otto Mønsteds Fund (Denmark) and SYSINBIO – Systems Biology as a Driver for Industrial Biotechnology, Coordination and Support Action (call FP7-KBBE-2007-1).

The multidisciplinary nature of my Ph.D., which will become apparent to the reader from the beginning of this thesis, was a feature that immediately cached my attention when Kiran firstly proposed it to me in 2007. My Ph.D. started with a very appealing project for improving a yeast vanillin cell factory through metabolic engineering. Despite the successful outcome of this project, I recognized that many aspects which are considered as “fundamental research” assume a determinant role in applied situations. It became clear to me that if we really aim at engineering the metabolism of a cell, there are still a vast amount of “fundamental principles” that we need to uncover, so we achieve a level of knowledge that enables fully controlled manipulation of metabolic features. Therefore, the last project I accomplished for my Ph.D. embraces a much more basic question; rather than engineering a cell, I would like to be able to better predict cellular fate in the metabolic context. While running between the pipettes at the lab bench and the C++ programs I started to develop in my computer, I realized that being able to understand the mathematical aspects underlying model formulation, while having the opportunity to perform the experiments which tackle the hypothesis thereby derived, enables a unique sensation of freedom and inspiration. Experimental and modeling approaches to study biology may demand very distinct cognitive skills, but they complement each other in such rewarding manner, that, to my experience throughout the last 4 years, I can say it was totally worth. At last, I hope I was able to transcribe my enthusiasm into words while writing this thesis.

Ana Rita Brochado, April 2012

Heidelberg. Germany.

Acknowledgments

For accomplishing the multifaceted venture that was my Ph.D., I was very fortunate to have Prof. Uffe H. Mortensen and Kiran R. Patil as my supervisors, to whom I gladly thank for leading me through this experience. Without Uffe's support, who introduced me to molecular biology and yeast genetics, it would have been impossible for me to learn the experimental skills essential to accomplish the work herein presented. Thank you for all valuable discussions and for welcoming me in your laboratory, where there was always people willing to help, not only by providing useful scientific advice, but also by making it a very enjoyable working environment at the bench. To Kiran, who shared with me his experience in fermentation technology in such a tireless way (even if I almost accidently burned his hand while trying to fix a reactor at eleven in the night!), who taught me most of what I learned about modeling of metabolism and for being my mentor, I am especially grateful. Thank you for uncountable valuable discussions and for always supporting me, even when nothing seemed to make sense. Thank you for accepting me as your Ph.D. student in Denmark and also for "bringing" me along when you took a step forward in your carrier by moving to EMBL in Germany.

I thank all our previous and current group members, especially Louise Mølgaard, who was at CMB since the first day I arrived and shared with me much of her scientific knowledge, but also many good moments in the lab and in our office. Aleksej Zelezniak, who started his Ph.D. soon after me and also moved to EMBL, for our valuable scientific discussions and for sharing "a beer" anytime things got hard. To Sergej Andrejev, our scientific programmer, who continuously saved me of drowning in my own programs while sharing the office with me for more than 1.5 years at EMBL. I also want to thank him for teaching me one of the most relevant and truthful sentences I have recently heard. After I questioned him whether the simple program we had just made was optimal or not, he replied, "Optimal?!?! No! It is optimistic!!".

Thanks to the determined and always motivated master's students I was fortunate to co-supervise: Cláudia Matos, Ana Machás, Małgorzata E. Bejtka, Tomasz Boruta, Emauela Monteiro, Joanna Guzicka and Justyna Nocoń.

Thanks to the secretariat team at CMB and EMBL for all administrative support, as well as I thank the research technicians team at CMB for helping with all laboratory assistance I requested. Thanks to Manuel Quirós, Ana Paula Oliveria, Zita Soons, Gary Male, Olga Ponomarova, Lars Poulsen, Line Due Buron, Bo Salomonsen, Martin Engelhard Kogle, Sujata Sohoni, Jakob

Blæsbjerg Nielsen, Tomas Stručko, Erica Valentini, Birger L. Møller, Jørgen Hansen, Martin Beck and Wolfgang Huber for valuable scientific discussions.

To my friends, especially Arnau Montagud, Sarah Lieder, Tilman Plass and Ana Machás, a very big thanks for everything! Last but not least, I thank my family, in particular my mother for her unconditional support.

Synopsis

Genome-scale metabolic models are increasingly used for simulation of cellular phenotypes as reflected in their metabolic physiology, such as growth rate, metabolite secretion and usage of metabolic pathways. In particular, they offer a privileged opportunity to investigate genotype to phenotype relationships for gene deletion mutants, as the deletion of a gene can generally be translated to the absence of the corresponding reaction in the metabolic network. Experimental evidence for the predictive power of genome-scale metabolic models, comprising the metabolic network and the modeling approach, is commonly limited to gene essentiality analysis. Extrapolation of the current modeling tools for covering the deletion of multiple genes, which are relevant from both fundamental and applied perspective, is largely unclear. The work herein presented focused on the experimental evaluation and development of new tools for prediction of cellular phenotypes from a metabolic physiology perspective, *i.e.* prediction of metabolic phenotypes. The baker's yeast *Saccharomyces cerevisiae* was chosen to carry out the work presented in this thesis, motivated by the availability of extensive knowledge about its physiology, its suitability for industrial applications and a vast collection of molecular biology, as well as of high-throughput data generation technologies.

Experimental assessment of the predictive power of the yeast genome-scale metabolic network and established modeling approaches was done in the context of metabolic engineering. The yeast genome-scale metabolic model was used to design a strategy for increasing heterologous vanillin production in *S. cerevisiae* by applying OptGene, a simulation framework for finding gene deletions towards maximization of productivity. The choice of minimization of metabolic adjustment (MoMA) over flux balance analysis (FBA) for maximization of growth as biological objective highly influenced the simulation results for strain design. Up to 2-fold improvement of vanillin production was observed upon physiological characterization of the mutants obtained following MoMA biological objective, confirming that metabolic proximity towards the reference strain can successfully describe the behavior of metabolic networks upon single gene deletion.

In a following study, the metabolic flux through the vanillin biosynthetic pathway was found to be limited by the activity of one of the pathway enzymes. Moreover, such limitation was exclusively observed in one of the mutants previously obtained through *in silico* guided metabolic engineering, for which cofactor and precursor availability was increased. This result demonstrates the fundamental importance of accounting for the entire metabolic network while modulating the metabolic flux of a single pathway. In this context, regulatory circuits often

modulate the expression of several genes or even entire pathways, rendering regulatory genes attractive targets for metabolic engineering. Integration of regulatory information with genome-scale models is still in its infancy, partially due to incomplete knowledge and characterization of regulation mechanisms. Within this work, *HAP4*, a transcription factor involved in glucose repression, was overexpressed in the vanillin producing yeast, based on previous observations that it leads to mutants with increased respiratory capacity. As a consequence of higher availability of ATP, the production of vanillin increased by 30%. In addition to other studies, this example confirms that manipulation of regulatory targets, in this case by gene overexpression, can cause the re-adjustment of fluxes distributed around important metabolic branches. Furthermore, this study demonstrates not only the need for development of new computational tools accounting for regulatory targets for metabolic engineering, but also the need for better understanding the relationship between gene transcription changes and the resulting alteration of metabolic fluxes.

Predicting the phenotype of perturbed metabolic networks is essential for using stoichiometric models for metabolic engineering, as well as to infer gene functions from the phenotype of deleted mutants. Cumulative genetic perturbations embody a large potential for both applied and fundamental research, as they might reveal functional association between the involved genes. The accuracy of phenotype prediction of mutants carrying multiple genetic perturbations is currently poorly assessed, mainly due to lack of large-scale systematic experimental studies. To this end, the recently available map of genetic interactions for *S. cerevisiae* offers the possibility for extending the validation of stoichiometric models to include the accuracy of metabolic phenotype prediction of double deletion mutants, as well as the accuracy of prediction of genetic interactions. The last part of this thesis presents a computational tool for simulation of metabolic phenotypes, termed **Mi**nimization of **M**etabolites **B**a**l**ance – MiMBl, which focuses on the formulation of objective functions, particularly of those concerning simulation of perturbed metabolic networks. Numerous established objective functions were found to be sensitive to the stoichiometric representation of the biochemical reactions, especially visible for cumulative gene deletions. MiMBl allows formulating objective functions describing established and newly suggested biological principles in a robust manner. MiMBl was used to predict and analyze the yeast genetic interaction network and it largely outperformed previously established simulation algorithms for predicting the phenotype of perturbed metabolic networks, namely lMoMA. Moreover, MiMBl and FBA were used in combination for interpreting the metabolic rerouting induced by double gene deletions, illustrating how different

biological principles can be used to provide useful mechanistic insights about operation modes of metabolic networks.

Overall, the work presented in this thesis demonstrates how genome-scale metabolic models can be used for exploring genotype to phenotype relationships, from both experimental and theoretical perspectives.

Resumo em Português

Os modelos metabólicos à escala genómica são cada vez mais utilizados para simulação de fenótipos na perspectiva de fisiologia microbiana, nomeadamente para a simulação de taxa de crescimento, secreção de metabolitos e utilização de vias metabólicas (aqui denominado "fenótipo metabólico"). Em particular, estes modelos proporcionam uma oportunidade privilegiada para investigar a relação genótipo-fenótipo em mutantes com genes eliminados, dado que a supressão de um gene pode ser facilmente traduzida para a ausência da reacção correspondente na rede metabólica. A avaliação do poder preditivo dos modelos metabólicos à escala genómica é actualmente basea, na sua maioria, em evidências experimentais relativas à análise de genes essenciais. A extrapolação das ferramentas de modelação actuais para cobrir a previsão de fenótipos resultantes de eliminação de múltiplos genes, interessantes tanto numa perspectiva fundamental como aplicada, permance, em grande parte, por explorar. O trabalho aqui apresentado centrou-se na avaliação experimental e no desenvolvimento de novas ferramentas computacionais para previsão de fenótipos numa perspectiva de fisiologia metabólica, ou seja, previsão de fenótipos metabólicos. Dada a disponibilidade de amplo conhecimento sobre a sua fisiologia, a sua adequação para aplicações industriais e um vasto conjunto de tecnologias disponíveis para biologia molecular e geração de dados à escala genómica, motivou a escolha de *Saccharomyces cerevisiae* para este estudo.

A capacidade preditiva da rede metabólica à escala genómica da levedura foi experimentalmente availada neste projecto no contexto de engenharia metabólica, assumindo a minimização de ajuste metabólico (MoMA) de mutantes com um gene eliminado em relação a estirpe de referência. Este modelo foi utilizado para projectar uma estratégia de melhoramento de produção heteróloga de vanilina em *S. cerevisiae*, que após implementada resultou no dobro da produção. O algoritmo OptGene foi utilizado para seleccionar os genes a eliminar, tendo em conta a maximização de produtividade. Estes resultados confirmam a proximidade metabólica de mutantes após a eliminação de um gene em relação a estirpe de referência como princípio de simulação adequado.

Num segundo estudo, constatou-se que o fluxo metabólico da via biossintética da vanilina estava condicionado pela actividade de uma das enzimas da via. Tal limitação foi exclusivamente observada num dos mutantes previamente obtidos com base no modelo metabólico à escala genómica, realçando a importância de ter em conta a rede metabólica na sua totalidade, mesmo quando o fluxo de uma única via metabólica está em questão. Neste contexto, os genes

internevientes em cascatas de regulação são também alvos atractivos para a engenharia metabólica, já que estes afectam a expressão de vários genes, ou mesmo de vias metabólicas completas. Na realização deste trabalho a produção de vanilina foi intensificada pela superexpressão do gene *HAP4*, um factor de transcrição envolvido nos circuitos regulatórios para a repressão por glucose em *S. cereviaise*. Embora a inclusão de circuitos de regulação em modelos metabólicos esteja ainda no início do seu desenvolvimento, este resultado evidencía os potenciais benefícios da inclusão de proteínas reguladoras na modelação de metabolismo, nomeadamemte como alvos plausíveis para engenharia metabólica.

A previsão de fenótipos resultantes de múltiplas perturbações genéticas é de grande interesse não só para investigação aplicada, mas também para a investigação científica de carácter fundamental, já que os efeitos cumulativos da eliminação de múltiplos genes podem revelar associações funcionais entre os genes envolvidos. A última parte deste trabalho descreve uma nova ferramenta computacional desenvolvida para formulação de funções objectivo a utilizar na simulação de fenótipos metabólicos, particularmente dedicada à simulação de fenótipos após eliminação de um ou mais genes. O novo algorítmo foi demoniado ***Mi****nimization of* ***M****etabolites* ***B****a****l****ance* – MiMBl. O MiMBl foi utilizado para prever e analisar a recentemente publicada rede de interacções genéticas de levedura, tendo em grande parte superado os algoritmos existentes. Adicionalmente, o MiMBl foi utilizado em combinação com um algoritmo ja estabelecido (FBA) para a interpretação do reencaminhamento metabólico induzido pela eleminação de dois genes, ilustrando como diferentes princípios biológicos podem ser utilizados para fornecer informações mecanísticas sobre modos de operação de redes metabólicas.

Dansk Sammenfatning

Genom-skalerede metabolske modeller bliver hyppigere og hyppigere brugt til at simulere cellulære fænotyper, for eksempel til beskrivelse af den metabolske fysiologi, vækst rate, metabolit sekretion og brug af metabolske reaktionsveje. Især tilbyder de en attraktiv mulighed for at undersøge sammenhængen mellem genotype og fænotype for mutanter, hvor et gen er blevet elimineret, idet denne mangel ofte vil resultere i fraværet en specifik reaktion i cellens metabolske netværk. Eksperimentel evidens til støtte for genom-skalerede modellers evne til at forudsige metabolske ændringer er hovedsageligt begrænset til at forudsige om et gen er essentielt eller ej. En udvidelse af dette koncept til at omfatte, hvad der sker, hvis flere gener elimineres kan i øjeblikket ikke gøres særlig præcist, selv om sådanne forudsigelser ville være særdeles relevante for projekter indenfor både grund- og anvendt forskning. Forskningsarbejdet, der præsenteres i denne rapport, fokuserer på eksperimental evaluering og udvikling af nye værktøjer, som muliggør at forudsige cellers fænotype ud fra et metabolsk fysiologisk perspektiv, det vil sige, til forudsigelse af fænotyper. Bagegær, *Saccharomyces cerevisiae*, blev valgt som modelorganisme for dette stadium, idet den er industriel relevant og idet der er akkumuleret en enorm mængde viden om dens fysiologi, og idet der er udviklet mange nyttige genetiske, molekylærbiologiske og "high throughput" værktøjer.

Eksperimentel validering af genom-skalerbare netværks og modellers evne til at forudsige fænotyper blev udført med fokus på "metabolic engineering" opgaver. Den genom-skalerbare metabolske model for gær blev brugt til at designe en strategi for heterolog produktion af vanillin i *S. cerevisiae*. Ved at følge en *in silico* designet strategi som blev foreslået af computerprogrammet OptGene og algoritmen "minimization of metabolic adjustment" (MoMA) som målfunktion, blev der opnået en fordobling af udbyttet. I et efterfølgende studium blev det påvist, at det metabolske flux gennem vanillin syntesevejen var begrænset af aktiviteten af et af enzymerne i syntesevejen. Denne begrænsning blev udelukkende set i en af de mutanter, der blev skabt som funktion af den *in silico* guidede strategi nævnt ovenfor, hvilket viser nødvendigheden af at gøre rede for hele det metabolske netværk, når fluxet igennem en enkelt biosyntesvej skal ændres. I denne sammenhæng er regulatoriske gener attraktive mål for "metabolic engineering", fordi de kan ændre udtrykket af mange gener eller hele biosyntesveje via en enkelt genetisk ændring. I denne rapport bliver det demonstreret, hvordan vanillin produktionen yderligere kan forøges ved at overudtrykke genet *HAP4*, der koder for en transkriptionsfaktor involveret i glukose repression. Selvom inkludering af regulatoriske kredsløb

i metabolske modeller stadig er i sin barndom, så viser dette resultat, at disse med fordel kan udnyttes i ”metabolic engineering” strategier. At kunne forudsige de fænotyper, der vil opstå som funktion af kumulerende mutationer vil kunne bidrage til at forstå sammenhængen mellem mange gener og deres funktioner, og har derfor et stort potentiale som kan bruges både i grund- og i anvendt forskning. Den sidste del af denne rapport angriber denne problemstilling og beskriver udviklingen af et nyt computerbasere beregningsværktøj, der kan simulere metabolske fænotyper, kaldt MiMBL, og som fokuserer på at formulere målfunktioner, især dem der indvirker på forskudte metabolske netværk. MiMBL blev brugt til at forudse og analysere det samlede netværk af genetiske interaktioner i gær som for nylig er blevet offentliggjort, og det viste sig, at forudsigelserne ved brug af MiMBL generelt var bedre end dem, der blev opnået med andre lignende værktøjer som for eksempel lMoMA. MiMBL og FBA blev også brugt i kombination til at fortolke ændringer i det metabolske kredsløb induceret forårsaget af deletionen af to gener, hvilket illustrerer hvordan forskellige biologiske principper kan blive brugt til at opnå brugbar mekanistisk indsigt i hvordan det metabolske netværk fungerer.

Nomenclature

List of Abbreviations

3DSD	3-dehydroshikimate dehydratase
3DSH	3-dehydroshikimate
ACAR	Aryl Carboxilic Acid Reductase
acet	acetate
ATP	Adenosine Triphosphate
BPCY	Biomass Product Coupled Yield
COBRA	Constraint-Based Reconstruction and Analysis
DW	Dry weight
eth	ethanol
FBA	Flux Balance Analysis
GLPK	GNU linear programing kit
gly	glycerol
GRAS	Generally Recognized as Safe
hsOMT	*Homo Sapiens* O-methyltransferase
lMoMA	linear Minimization of Metabolic Adjustment
MiMBl	Minimization of Metabolites Balance
MoMA	Minimization of Metabolic Adjustment
NADPH	Nicotinamide adenine dinucleotide phosphate
NMR	Nuclear Magnetic Resonance
OD	Optical Density
PAC	Protocathechuic acid
PAL	Protocathechuic aldehyde
PCR	Polymerase chain reaction
PPTase	Phosphopantetheinyl transferase
ROOM	Regulatory On/Off Minimization of metabolic flux changes
SAM	S-adenosylmethionine
TCA	Tricarboxylic acid
UDP-glc	UDP-glucose
VG	vanillin β-D-glucoside
WT	Wild Type

List of Symbols

Δ	gene deletion
$\uparrow$	gene overexpression
μ	specific growth rate, h^{-1}
μ_{max}	maximum specific growth rate, h^{-1}
$Y_{S\ Metab}$	Yield of metabolite *Metab* on substrate *S*, mg_{Metab}/g_S
r_{Metab}	Specific uptake or production rate of metabolite *Metab*, $mmol_{Metab}/(g_{DW}.h)$
X	Biomass, g_{DW}/l
R^3	Reward-Risk-Ratio
S	Stoichiometric matrix
v	vector of all fluxes
v_i	flux of reaction i, $mmol/(g_{DW}.h)$
t_m	turnover of metabolite m, $mmol/(g_{DW}.h)$
θ	factor for linear scaling of matrix S
ε	epistasis
ub	upper bound
lb	lower bound
N	set of reactions
M	set of intracellular metabolites
$\alpha_{m,i}$	stoichiometric coefficient of metabolite m in reaction i
min	minimization, in the context of linear optimization
max	maximization, in the context of linear optimization
B	optimal basis matrix
c_i	objective function coefficient of variable v_i
$\overline{c_i}$	reduced cost of variable v_i
c_B	vector of the objective coefficients of basic variables
Θ	Scaling matrix

Table of contents

Preface	**i**
Acknowledgments	**iii**
Synopsis	**v**
Resumo em Português	**ix**
Dansk Sammenfatning	**xi**
Nomenclature	**xiii**
List of Abbreviations	xiii
List of Symbols	xiv
Table of contents	**xv**
Table of Figures	**xix**
Chapter 1: Introduction and Thesis Outline	**1**
List of publications and manuscripts included in this thesis	5
Chapter 2: Systems Biology, Genome Scale Metabolic Modeling and Metabolic Engineering of *S. cerevisiae* – Acts and Facts	**7**
Summary	7
Systems Biology	8
An 'omics' briefing	8
Integrating and interpreting high-throughput data	10
Saccharomyces cerevisiae	10
Briefing on yeast physiology and central carbon metabolism	11
An eukaryotic model organism	12
Industrial applications	14
Metabolic Engineering	15
Metabolic engineering cycle	17
Genome-Scale Metabolic Modeling	20
Briefing on genome-scale metabolic models reconstruction	20
Applications of genome-scale models	21
Predicting the distribution of metabolic fluxes	22
Designing and selecting objective functions for metabolic modeling	25
Metabolic modeling for guiding metabolic engineering	26

Conclusions 29

References 29

Chapter 3: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 1 – Metabolic Modeling for Strain Improvement **39**

Abstract 39

Background 40

Results and Discussion 42

Vanillin β-D-glucoside production in *S. cerevisiae* 42

In silico design 42

Strain construction and characterization 46

Analysis of the experimental results 50

Conclusions and Future Perspectives 55

Methods 56

Model simulations 56

Minimization of metabolite turnover 56

Plasmids and strains 57

Media Composition 59

Batch cultivations 60

Continuous cultivations 60

Off-gas analysis 60

Biomass determination 60

Glucose and external metabolites analysis 61

Acknowledgements 61

References 62

Chapter 4: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 2 – Pathway and Regulatory Engineering **65**

Abstract 65

Background 66

Results and Discussion 68

Different PPTases for ACAR post-translational activation lead to altered vanillin β-D-glucoside production 68

hsOMT catalyses the rate-limiting step for de novo vanillin β-D-glucoside biosynthesis in *S. cerevisiae* 69

Overexpression of hsOMT leads to improved vanillin production in baker's yeast only when complemented with model-guided network engineering 72

Engineering glucose repression in *S. cerevisiae* leads to improved vanillin β-D-glucoside production 73
Overview of metabolic engineering strategies towards improving vanillin β-D-glucoside production 75
Conclusions and Future Perspectives 76
Methods 77
Strains and Plasmids 77
Media composition 82
Microtiter plates batch cultivation 83
Biomass determination 83
Glucose and external metabolites analysis 83
Acknowledgments 84
References 84

Chapter 5: Minimization of Metabolite Balance Explains Genetic Interactions within the Yeast Metabolic Network **87**
Abstract 87
Introduction 88
Results and Discussion 89
Minimization of sum of fluxes 89
Minimization of Metabolites Balance - MiMBl 92
Predicting genetic interactions using MiMBl 95
MiMBl predicts genetic interactions between metabolically distant genes 99
Genetic interactions and isoenzymes 99
Using MiMBl and FBA for understanding genetic interactions 100
Methods 101
Yeast genome-scale metabolic models 101
Normalized lMoMA 102
Minimization of metabolites balance – MiMBl 102
Alternative stoichiometry representations 103
Impact of scaling stoichiometry on the optimal solution – Analytical evidence 104
Genetic interactions – epistasis score 106
Metabolic/network distance 107
Acknowledgements 107
References 107

Chapter 6: Conclusions and Future Perspectives **111**

Supplementary Information **115**

Chapter 2: Systems Biology, Genome Scale Metabolic Modeling and Metabolic Engineering of S. cerevisiae – Acts and Facts 117

Chapter 3: Metabolic Engineering for Vanillin Production in S. cerevisiae. Part 1 – Metabolic Modeling for Strain Improvement 126

Chapter 4: Metabolic Engineering for Vanillin Production in S. cerevisiae. Part 2 – Pathway and Regulatory Engineering 127

Chapter 5: Minimization of Metabolite Balance Explains Genetic Interactions within the Yeast Metabolic Network 128

Supplementary methods 128

Supplementary figures 129

Supplementary tables 134

Supplementary notes 135

References 137

Table of Figures

Figure 1.1:	Genome-scale metabolic modeling for investigating genotype to phenotype relationship.	2
Figure 2.1:	Systems Biology tools for understanding genotype to phenotype relationship.	9
Figure 2.2:	Simplified representation of the main regulatory pathways for glucose repression in *S. cerevisiae*.	12
Figure 2.3:	*Saccharomyces cerevisiae*, an established cell factory for small molecules.	16
Figure 2.4:	Metabolic Engineering cycle.	18
Figure 2.5:	Fundamentals of optimization based approaches for simulation of cellular phenotype.	23
Figure 2.6:	OptKnock and OptGene optimization frameworks for guiding Metabolic Engineering.	27
Figure 3.1:	Schematic representation of the *de novo* VG biosynthetic pathway in *S. cerevisiae.*	40
Figure 3.2:	Comparison of targets predicted by OptGene for improved VG productivity.	44
Figure 3.3:	Vanillin β-D-glucoside yield observed for the reference strain (VG0) and metabolically engineered mutants (VG1-4) in batch cultivations.	47
Figure 3.4:	Vanillin β-D-glucoside yield observed for the reference strain (VG0) and metabolically engineered mutants (VG1-4) in continuous cultivations.	50
Figure 3.5:	Flux variability analysis.	52
Figure 3.6:	Minimum turnover of selected metabolites from the central carbon metabolism and from the VG biosynthetic pathway.	54
Figure 4.1:	*De novo* vanillin β-D-glucoside biosynthesis in *S. cerevisiae*.	66
Figure 4.2:	Impact of heterologous expression of different PPTases in ACAR activation in *S. cerevisiae*.	68
Figure 4.3:	Effect of *ACAR* and *hsOMT* overexpression on vanillin β-D-glucoside production in *S. cerevisiae*.	70
Figure 4.4:	Effect of deletion of *YPRCΔ15 locus* on the growth of strain VG4.	72

Figure 4.5:	Overexpression of hsOMT leads to improved vanillin production only when complemented with model-guided network engineering.	73
Figure 4.6:	Overexpression of *HAP4* increases the production of vanillin β-D-glucoside.	74
Figure 4.7:	Overview of metabolic engineering strategies towards improving vanillin β-D-glucoside production.	75
Figure 5.1:	Minimization of overall intracellular flux leads to divergent predictions for flux distribution when using alternative stoichiometry representations.	90
Figure 5.2:	A toy-model illustrating how, and why, alternative stoichiometry representations influence simulation of minimization of metabolic adjustment by using lMoMA or MiMBl.	91
Figure 5.3:	Stoichiometry representation impacts the design of metabolic engineering strategies depending on the nature of the objective function formulation.	95
Figure 5.4:	Comparison of the performance of MiMBl, lMoMA and FBA for predicting genetic interactions in *S. cerevisiae*.	97
Figure 5.5:	Understanding genetic interactions by using MiMBl.	98

Chapter 1: Introduction and Thesis Outline

Understanding how the cell reads and processes the information maintained within the genome so it displays a given phenotype, is one of the most challenging questions faced in biology. The presence of a gene does not imply that this gene is continuously expressed. Expressing it to constant levels, also does not guarantee that it affects the phenotype equally over time or under different environmental conditions. There are several layers of regulation which allow the cell to translate its genotype into a phenotype. An integrated approach for studying cellular systems is expected to enable a better understanding of how this relationship occurs. Systems biology focuses on elucidating the structure and dynamics of such cellular systems, rather than studying individual cellular components, such as a given gene or protein. A cellular state is characterized by a set of expressed genes, a set of synthesised proteins, a set of available metabolites and so on, yet this does not reveal structural or dynamic cellular organization. Uncovering the complex and coordinated interactions between cellular components is fundamental for understanding functional biological systems. For instance reactions must occur between metabolites to form a metabolic network allowing the utilization of external resources and building of new cells. Many of these reactions are catalyzed by individual proteins (enzymes), involving interaction between proteins and metabolites, but also by protein complexes, implying interaction between proteins. Additionally, important regulatory mechanisms also engage interaction between cellular components, such as allosteric regulation, transcriptional regulation or signal transduction. Dynamic coordination of all cellular processes arising from such interactions ultimately enables the cell to translate its genotype to the phenotype.

Metabolism is arguably the subcellular system which is closest to the phenotype, as observable phenotypic traits, such as growth, have a very tight association with metabolic features, for instance provision of energy and cellular building blocks. Cellular metabolism has been studied for several years within Biochemistry and a comprehensive collection of metabolic reactions have been characterized. The rapid development of molecular biology techniques allowed further characterisation of biochemical reactions by helping to uncover gene-to-enzyme associations. Yet, as much as metabolism can be considered a well characterized subcellular system, in many cases phenotype still cannot be predicted from genotype, especially in cases involving changes in the expression of multiple genes. High confidence prediction of cellular phenotypes, as reflected in their metabolic physiology, by using genome-scale metabolic models is commonly confined to prediction of gene essentiality. With the increasing availability of

molecular biology tools and high-throughput characterization of cellular phenotypes, computational approaches which facilitate a systematic and integrative approach for associating phenotypic traits to genotypic features are in demand.

The aim of my thesis was to investigate genotype to phenotype relationships in a systematic manner, through stoichiometric simulation of metabolic phenotypes by using genome-scale metabolic models. Predicting the behaviour of the metabolic network in the absence of a reaction is considered to be a proxy for predicting the phenotype of the corresponding gene deletion mutant, therefore large-scale metabolic models provide excellent platforms for investigating genotype-phenotype relationships. *Saccharomyces cerevisiae* was chosen as a biological system, because of its status as a eukaryotic model organism, for which there is extensive knowledge and technology available. In the present work, stoichiometric modeling of the yeast genome-scale metabolic network was used for studying genetic interactions, as well as for providing guidance for a metabolic engineering strategy for increasing heterologous vanillin production. Both applications are based on predicting the phenotype based on known genetic modifications (**Fig. 1.1**).

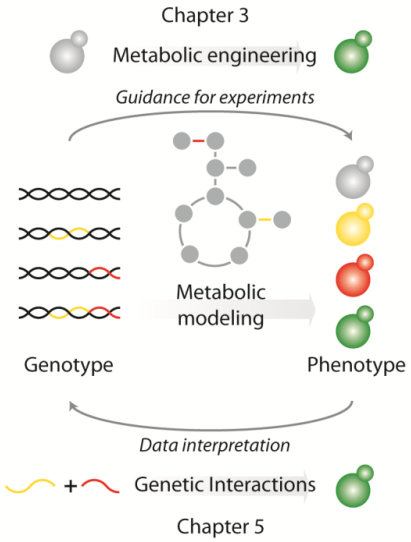


Figure 1.1: Genome-scale metabolic modeling for investigating genotype to phenotype relationships. Different genetic modifications (genotypes) are represented by different colours, yellow or red, as well as their corresponding phenotype. The wild type is represented in gray, while the combination of the yellow and red genetic modifications is represented in green. Stoichiometric modeling was used to provide guidance for metabolic engineering (top arrow) and for analysing genetic interactions, where the simulation results were used to suggest possible modes of operation of the metabolic network, while facilitating a methodical analysis of the functional association between genes (bottom arrow).

For metabolic engineering, prediction capability of a multitude of phenotypic traits is usually desired upon deletion of a gene(s) (reaction(s)), including growth and production yield of the compound of interest. For analysis of genetic interactions, the phenotypic trait most commonly measured is growth, due to the simplicity of its quantitative assessment for a large number of mutants. Although stoichiometric modeling is used in both cases for predicting the phenotype of gene deletion mutants, different uses of the simulation results are illustrated by the different applications. For metabolic engineering, stoichiometric modeling was used to provide guidance for experimental work. For analysing genetic interactions, the simulation results were used to suggest possible modes of operation of the metabolic network, while facilitating a methodical analysis of the functional association between genes (**Fig. 1.1**).

Chapter 2 provides state-of-the-art concepts and a literature survey about Systems Biology, Metabolic Engineering and Genome-Scale Metabolic Modeling.

Chapter 3 illustrates a case study of application of the yeast genome-scale metabolic model towards the improvement of a vanillin cell factory. Vanillin is one of the most widely used flavouring agents, and it is mainly obtained via chemical synthesis from petrol based substrates. Environmental concerns raised the urge for an alternative, and *de novo* vanillin biosynthesis from glucose in *S. cerevisiae* was previously demonstrated. Prior to engineering the cell factory towards increased vanillin production, product toxicity to the host is decreased by supplying the cell factory with an additional enzymatic step catalyzing vanillin glycosylation. A metabolic modelling guided strategy to improve the assembled cell factory is then presented and implemented. The results indicate the relevance of accounting for the entire metabolic network for increasing the flux through the pathway of interest, by the manipulation of genes which are metabolically distant from the target compound.

While taking a global engineering approach for strain improvement, it is assumed that the flux through the pathway of interest is controlled by the availability of precursors and cofactors, rather than by enzyme activity. However, metabolic control resides in a subtle balance between the contributions of both factors and a shift can take place upon altering one of them. Indeed, in **Chapter 4** it is shown that, after global engineering for vanillin production, conversion of pathway intermediates to the final product is limited by the availability of one of the pathway enzymes. Importantly, it is demonstrated that such enzyme limitation arises exclusively when combined with a global engineering approach, supporting the need of a holistic view of metabolism for successful metabolic engineering. The advantages of such integrative perception of cellular processes for metabolic engineering are further explored by engineering the

regulatory circuit for glucose repression, with the goal of simultaneously modulating several metabolic fluxes towards additional vanillin production.

Chapter 5 presents a new computational tool for metabolic modelling, focused on the formulation of objective functions. Several objective functions currently used to describe biological principles for stoichiometric modelling are shown to be susceptible to stoichiometric representation of biochemical reactions. A number of examples are used to illustrate how this affects phenotype simulation results for different purposes, namely for estimating internal flux distribution, as well as for predicting the behaviour of perturbed metabolic networks, *e.g.* by deleting one or more genes. A new approach for formulation of biological principles is presented, Minimization of Metabolites Balance - MiMBl, and several examples are given for formulation of other biological principles using MiMBl methodology. Particular focus is given to prediction of metabolic phenotypes for perturbed networks, because this feature is on the basis of many relevant conclusions and applications driven from genome-scale metabolic modelling. A prominent example is the prediction of genetic interactions, which are revealed from phenotypic behaviour of gene deletion mutants and often emerge from functional relationship between the deleted genes.

List of publications and manuscripts included in this thesis

1. **Brochado AR**, Matos C, Møller BL, Hansen J, Mortensen UH and Patil KR (2010). Improved vanillin production in baker's yeast through *in silico* design. *Microbial Cell Factories*, 9:84.

2. **Brochado AR**, Andrejev S, Maranas CD and Patil KR. Minimization of Metabolite Balance Explains Complex Genetic Interactions within the Yeast Metabolic Network. *Submitted.*

3. **Brochado AR** and Patil KR. Overexpression of O-methyltransferase leads to improved vanillin production in baker's yeast only when complemented with model-guided network engineering. *Submitted.*

Chapter 2: Systems Biology, Genome Scale Metabolic Modeling and Metabolic Engineering of *S. cerevisiae* – Acts and Facts

Summary

A holistic view of the cell is fundamental for gaining insights into genotype to phenotype relationships. Systems Biology is a discipline within Biology, which uses such holistic approach by focusing on the development and application of tools for studying the structure and dynamics of cellular processes. Metabolism is an extensively studied and characterised subcellular system, for which several modeling approaches have been proposed over the last 20 years. Nowadays, stoichiometric modeling of metabolism is done at the genome scale and it has diverse applications, many of them for helping at better characterizing genotype to phenotype relationships. Metabolic Engineering is one of the fields in which the complete understanding of such relationship would have a striking impact, since phenotype prediction based on genotype is fundamental for rationally engineering metabolic networks. This chapter aims at providing the reader with relevant state-of-the-art information concerning Systems Biology, Genome-Scale Metabolic Modeling and Metabolic Engineering. Particular attention is given to the yeast *Saccharomyces cerevisiae*, the eukaryotic model organism used thought the thesis.

Systems Biology

During the last 20 years, Biology went through a fundamental change of focus from individual genes and proteins to cellular processes and interactions between the cellular components [1, 2]. Knowing all the genes, proteins and metabolites (components) existing within the cell at a given stage is not sufficient to understand how the cell operates. How these components interact, what phenotypic traits they trigger and how do they enable the cell to respond to environmental changes, is fundamental to understand a functional system. Such holistic approach to study biological systems has been termed Systems Biology, as suggested by Ideker *et al.* in 2001 [3] and Kitano in 2002 [4]. The aim of Systems Biology is studying "the structure and dynamics of cellular and organismal function". It implies the integration of different data-types resulting from high-throughput measurements of cellular processes in order to understand the functional principles of biological systems [3, 4].

Given its holistic nature, Systems Biology is a very close interplay between Biology, Computational Biology and Technology [1]. Advances in technology are fundamental for automatization of experiments, *e.g.* the synthetic genetic array - SGA [5, 6], as well as for acquiring new types of biological data, ranging from pioneering DNA sequencing [7] to the contemporary challenge of metabolome analysis [8]. Computational biology contributes with tools for modeling and data analysis, primarily to understand the biological principles underplaying the data, as well as for guiding experiments [1, 4]. Biology holds the questions to be asked and experimental expertise, ranging from molecular biology of a simple bacterium to the understanding of brain function or human diseases.

An 'omics' briefing

Systems Biology becomes progressively more quantitative, so as to draw meaningful conclusions from genome-wide/large datasets. The generation and study of such large datasets is generally referred as 'omics' data. Genomics, together with Functional Genomics, focus on identification of genes and entire genome sequences, as well as assignment of gene functions. Recently, the availability of new high-throughput sequencing technologies and consequent decreased price of genome sequencing, lead to the development of Comparative Genomics [9] and Metagenomics [10]. On the path of understanding genotype to phenotype relationship, a comprehensive assessment of gene expression patterns through quantification of RNA molecules is the subject of Tanscriptomics [11–13]. Because regulation occurs at several layers of the *central dogma* (**Fig. 2.1**) [14, 15], such as post-transcriptional modifications and protein degradation, correlation

between gene expression and protein concentration is not always observed [16]. Proteomics focus on the quantification of protein concentrations. Mass spectrometry-based methods have been recently established and, so far, they are the most comprehensive methods for proteome quantification [17, 18]. Metabolomics and Fluxomics are, among the previously mentioned data-types, those that are closer to the cellular phenotype. Metabolomics aims at quantifying all measurable intracellular and extracellular metabolites (*e.g.* carbohydrates, lipids, amino acids), as well as their changes over time or under different environmental/genetic conditions. On the other hand, Fluxomics involves the quantification of rate of the metabolic reactions at the network scale, ultimately leading to a comprehensive characterization of the functional state of the metabolic network. Although different in terms of implied functional information, these two fields are closely related by the fact that they both involve (relative) quantification of (intra)cellular metabolites. Also here, mass-spectroscopy and NMR are the preferred detection methods, often associated with a chromatographic pre-separation of the extracts [19].

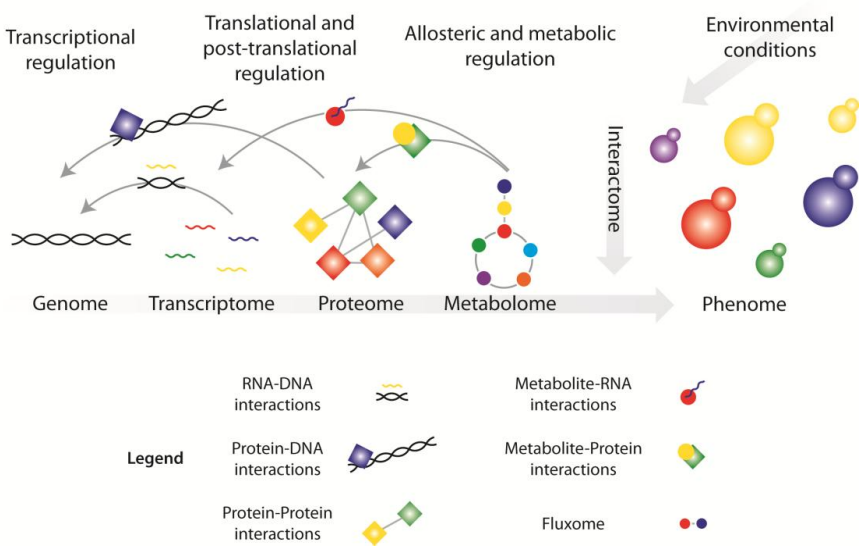


Figure 2.1: Systems Biology tools for understanding genotype to phenotype relationships.

More than quantification of metabolites, Fluxomics aims at inferring the rate of reactions between metabolites, thereby tracing what metabolic pathways are active under the given conditions. ^{13}C isotope labeled substrates (*e.g.* glucose) are commonly used, since the use of different metabolic pathways often results in different isotope labeling patterns [20]. This approach is also called metabolic flux analysis (MFA) and it is very useful for elucidating pathway bottlenecks and regulatory mechanisms within metabolic networks [21–23]. While

Transcriptomics, Proteomics and Metabolomics focus on characterizing cellular components and their states, Fluxomics and Phenomics characterize cellular functional states and generally referred as Interactome (**Fig. 2.1**) [24]. The list of comprehensive data collection and analysis continues to expand and more specialized fields are appearing, such as Lipidomics, Localizomics, Glycomics, Phosphoproteomics and others (reviewed in [24, 25]).

Integrating and interpreting high-throughput data

The cellular phenotype results from the interplay of many cellular components, such as several genes, proteins and metabolites [2, 26]. Thus, comprehensive studies for extending our knowledge about cellular function often involve the integration an interpretation of different high-throughput data-types. For example, Usaite *et al.* reported the reconstruction of the yeast *SNF1* kinase regulatory network and its central role in energy metabolism by using a combined analysis of transcriptome, proteome and metabolome [27]. In another example, Nagaraj and colleagues reported the mapping of a human cancer cell line using transcriptome and proteome [28]. In order to cope with increasing high-throughput data generation, computational biology has major tasks to perform, not only the establishment and maintenance of data repositories, but also the provision and update of visualization software for several different types of data, and, most importantly, the development of cutting-edge algorithms for data integration, interpretation and guidance of experiments [24, 29, 30]. Aiming at improving cooperation among different fields in Biology, standardization of the information is an absolute requirement. To this end, the Systems Biology Markup Language (SBML), an XML-based format, was developed for representation of biological networks [31].

Saccharomyces cerevisiae

Saccharomyces cerevisiae is a unicellular eukaryote classified in the kingdom of Fungi. It is a 5-10 µm diameter ovoid-shaped budding yeast, its life cycle alternates between haplophase and diploidphase and both can exist in stable cultures. This yeast has been used for brewing, backing and wine fermentation since 6000 BC. However, it was not until the 19th century that it became object of study, when Louis Pasteur and others firstly associated fermentation with yeast metabolism. The species was named after its ability to ferment sugars - "*Saccharomyces*", meaning sugar fungus, and to produce beer - "*cerevisiae*", from "cerveja" in Portuguese or "cerveza" in Spanish [32, 33].

Briefing on yeast physiology and central carbon metabolism

Much of what is known today about *S. cerevisiae* is due to its biotechnological applications since many years. This yeast naturally produces ethanol and carbon dioxide in aerobic conditions and high sugar concentration, also known as overflow metabolism or Crabtree effect [34]. Although energetically unfavorable, alcoholic fermentation is the main source of energy production in high sugar conditions, while the tricarboxylic acids (TCA) cycle and oxidative phosphorylation are rather less active [34–36]. In aerobic batch conditions, a first exponential phase of growth occurs when glucose is consumed to form biomass and fermentation products, with a biomass yield on glucose of ~3.6 g_{DW}/C-mol [34]. After glucose depletion, the culture typically experiences a short lag phase (diauxic shift) and restarts to grow on the available carbon sources resulting from fermentation, mainly ethanol, although glycerol and acetate are also commonly found. The second exponential growth phase is characterized by a slower growth rate (~0.087 h^{-1} compared to ~0.34 h^{-1} exponential growth rate on glucose), but higher biomass yield, ~10.5 g_{DW}/C-mol, due to the occurrence fully respiratory metabolism [34]. Similarly, fully respiratory growth on glucose is achieved in continuous cultivation (~15 g_{DW}/C-mol), where glucose is kept at low concentration and the dilution rate (and hence growth rate) is below the critical value for onset of fermentation [34, 37]. As in most of the living organisms, central carbon metabolism in yeast provides all the necessary building blocks for biosynthesis of macromolecules: 1) hexose-monophosphates, needed for polysaccharides and pentose-phosphates for nucleic acids; 2) triose-phosphates, needed for glycerol of phospholipids; 3) pyruvate, the main branch point between respiro-fermentatitive metabolism, required for amino acids biosynthesis as well as for 4) Acetyl-CoA, needed for fatty acids; 5) oxaloacetate and 6) 2-ketoglutarate, major building blocks for several amino acids [34].

Glucose metabolism in *S. cerevisiae* has been extensively studied, as glycolysis is usually the pathway carrying the highest flux in the metabolic network [34]. In fact, in the presence of glucose, most of the genes responsible for utilization of alternative carbon sources, as well as for respiration and gluconeogenesis, are repressed due to a phenomenon known as “carbon catabolite repression” or simply “glucose repression” [38–42]. The presence of glucose on the medium is sensed either by membrane-anchored sensor proteins or by hexose transporters (responsible by hexose uptake), which trigger a signaling cascade ultimately responsible for transcriptional repression/activation of several genes [38, 39]. Two signaling pathways work in synergy; glucose induction pathway and glucose repression pathway (**Fig 2.2**).

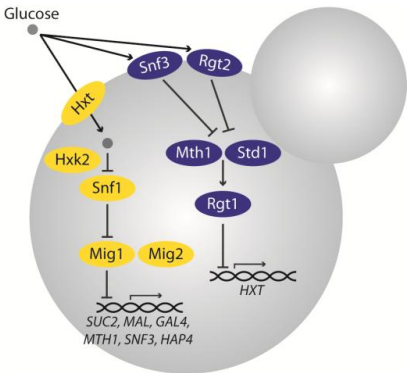


Figure 2.2: Simplified representation of the main regulatory pathways for glucose repression in *S. cerevisiae*. The glucose repression pathway (yellow) is responsible for repressing of genes involved in (or activating) the uptake of alternative carbon sources (*e.g. GAL4, SUC2*), gluconeogenesis, respiration and TCA cycle (*e.g. HAP4*).The glucose induction pathway (blue) is mainly responsible for the de-repression of hexose transporters, *HXT1-HXT16* (Adapted from [43]).

The glucose induction pathway is mainly responsible for the de-repression of hexose transporters, mediated via the membrane receptors Snf3 and Rgt2 leading to Mth1 and Std1 degradation and consequent Rgt1 hyper-phosphorylation, and therefore inability to repress the hexose transporters (*HXT1-HXT16*) [44]. The glucose repression pathway is mainly responsible for repressing genes coding for enzymes involved in the TCA cycle, electron transport chain, gluconeogenesis and consumption of alternative carbon sources. Once inside the cell, glucose is phosphorylated by Hxk2, the Snf1 complex is inactivated and Mig1 is no longer phosphorylated and thereby it represses several genes by binding their promoter regions [44]. Given its similarity to higher eukaryotic systems including humans, as well as its relevance for biotechnology applications, glucose repression in yeast has received much attention, especially the regulatory proteins Snf1, Hxk2, Mig1/Mig2 and the transcriptional activator Hap4 [27, 43–45].

An eukaryotic model organism

A continuously growing and highly interactive scientific community contributed to a great extent to the establishment of *Saccharomyces cerevisiae* as a eukaryotic model organism. One of the most important achievements in yeast genetics was the pioneering possibility to perform genetic manipulations through transformation, during the 1970's [46, 47]. In 1996, *S. cerevisiae* was the first eukaryote with its complete genome sequence available, as a result of an EU-funded consortium [48]. By then, from the ~6000 structurally annotated genes, a small fraction had characterized biological function (30%). Nowadays, the genes with unknown function in yeast range from 10 to 15% [33, 49]. The availability of the genome sequence made possible to

perform large-scale studies, thus yeast became a preferred organism for establishing several high-throughput assays, as well as for libraries construction; gene expression started soon being studied with microarray technology [50], as well as the first protein-protein interactions networks also started to be uncovered [51]. In 2002, a yeast deletion library was constructed, where nearly all the open reading frames were deleted [52]. In 2003, a comprehensive library of green fluorescent tagged proteins (GFP) was assembled in order to study protein cellular localization [53]. During the same year, the first of a series of genome-scale reconstructions of the *S. cerevisiae* metabolic network was published [54]. Again, this was the first of its kind among eukaryotes.

As consequence of a continuous development of novel technological resources, the number of high-throughput studies using yeast continues to increase. One of the most recent examples is the assembly of a comprehensive double gene deletion library and the corresponding genetic interactions, by Constanzo *et al*. [55, 56]. A genetic interaction (also called epistatic interaction) occurs when the combination of gene deletions results in a phenotype different than expected based on the individual gene deletion phenotypes [57–59]. When the observed phenotype for a double deletion mutant is worse than expected, it is said to be a negative or aggravating interaction between the two genes. If the opposite happens, it is a positive or alleviating interaction. Genetic interactions are often the outcome of associated gene functions, and therefore highly contribute to better understand genotype to phenotype relationship. The study by Constanzo *et al*. uncovered the interaction profile of ~75% of the yeast annotated genes. Similar biological processes were found to cluster together and functional cross-connections between all bioprocesses (pleiotropy) were observed. In a subsequent work, Szappanos *et al*. expanded the genetic interactions dataset to include the interaction profiles of 80% of the metabolic genes and performed an integrated data analysis using a genome-scale model of yeast metabolism [60]. They investigated the connection between degree of genetic interaction and pleiotropy within metabolism, a feature also observed in the dataset by Constanzo *et al*.. Ultimately, they introduced a machine-learning method to reconcile empirical interaction data with the model predictions, in order to improve the predictive power of the metabolic model. Within my thesis, the genetic interactions dataset focused on metabolism was used to validate a new algorithm for predicting cellular phenotype using the yeast genome-scale metabolic model, while simultaneously aiming at gaining new insights about modes of operation of the metabolic network.

Profiting from accumulated knowledge over the years, and due to conserved protein sequences and functions across species up to higher eukaryotes, *S. cerevisiae* provides useful insights in several disciplines of Biology and Medicine, ranging from Biochemistry and Evolution to Aging and even complex Neurological Disorders studies [26, 61, 62]. Many of the generated resources useful for yeast research are accessible from dedicated databases and repositories, some of them presented in **Table ST2.1**.

Industrial applications

Since the 1970's, when yeast geneticists introduced direct genetic engineering by transformation, a complete new dimension was added to biotechnology of *S. cerevisiae*. This would allow not only improving current knowledge about yeast products and physiology, but also the possibility to extend the genetic repertoire of the cell factory in order to obtain desired features, such as expanding substrate and product ranges [33]. Together, the practical knowledge about yeast physiology and biochemistry obtained over the years and the genome sequencing project, further enhanced its potential for industrial application. In fact, *S. cerevisiae* presents several advantages for applications in industry: it holds the GRAS (Generally Recognized as Safe) status from the American Food and Drug Administration; its genetic engineering toolbox is highly optimized; extensive knowledge about physiology and biochemistry; relatively easy scaling-up of industrial processes; tolerance to low pH (3 to 5), high sugar and ethanol concentration, thereby decreasing the risk for bacterial contamination; ability to grow anaerobically and aerobically; ability to utilize a wide range of sugars [63].

Nowadays, a broad variety of compounds are obtained from *S. cerevisiae*; not only pharmaceutical proteins, as human insulin, hepatitis and papillomavirus vaccines, but also fine and commodity chemicals and biofuels [19, 64]. While protein production potential by this yeast can be limited by its inability of accurately achieve the desired glycosylation patterns, production of small molecules is continuously gaining popularity among industry and academia [64, 65]. As a matter of fact, the fossil based natural resources are becoming scarce, and at the same time they directly or indirectly provide most of the fuels, fine and commodity chemicals worldwide. Biotechnology presents an attractive alternative, since microorganisms can be used to convert renewable sources, *e.g.* plant biomass, to valuable chemicals [64, 66].

Figure 2.3 shows several examples of added-value compounds which have been, or can potentially be, produced in *S. cerevisiae*. Among the represented compounds, native molecules such as and glycerol and fumarate are currently used as commodity chemicals. Biofules are also

strongly represented; ethanol, native to yeast and by far the dominating biofuel, but also other biomolecules with better physicochemical properties, such as propane-1,2-diol, isobutanol and 2/3-Methyl-1-butanol. The extensive list of fine chemicals contains polyketides (6-methylsalicylic acid), isoprenoids and sesquiterpenes (farnesol, valencene, artemisinic acid), carboxylic acids and aldehydes (vanillin), polyunsaturated fatty acids (PUFAs), among others [67–76]. These are chemically very diverse molecules, typically found as secondary metabolites of fungi, bacteria or plants, and have several applications as pharmaceutical ingredients, flavouring and preserving agents to be used in perfume and food industries [19, 64]. With a currently increasing capability for screening for novel compounds, many other molecules are to come [19, 77]. Despite their chemical diversity, most of the compounds can be obtained from intermediates of the central carbon metabolism, such as pyruvate and acetyl-CoA or from amino acid biosynthesis, such as aromatic amino acids and isoleucine, leucine and valine super family (**Fig. 2.3**). Tuning the distribution of the metabolic fluxes around these key branches of the metabolic network could, in principle, favour the synthesis of all the products obtained from the same native precursor. Additional network rewiring would be required for specific energy and cofactor supply for each of the products. Retrofitting of the metabolic network towards desired cellular properties has been termed Metabolic Engineering [78, 79], as described in detail in the next section.

Metabolic Engineering

Many added-value biomolecules are naturally found in a panoply of microbial or plant hosts. However they are usually present in trace amounts, which impair the economical viability of a potential biotech application [19, 77]. In 1991, J. Bailey and G. Stephanopoulos proposed, in two independent publications, what today is a field in biological sciences, Metabolic Engineering [78, 79]. Metabolic engineering was defined as “the improvement of cellular activities by manipulations of enzymatic, transport and regulatory functions of the cell with the use of recombinant DNA technology” [79]. This way, they propose the retrofitting of the metabolic network towards the desired feature in a rational and targeted manner. Indeed, in order to improve the yield and productivity of the desired product, knowledge-oriented strategies are required often involving 1) improving precursor metabolites and cofactors supply, 2) up-regulation of genes involved in product export routes or 3) redesign of metabolic pathway, so feedback inhibitions and competing pathways are avoided, among others [19].

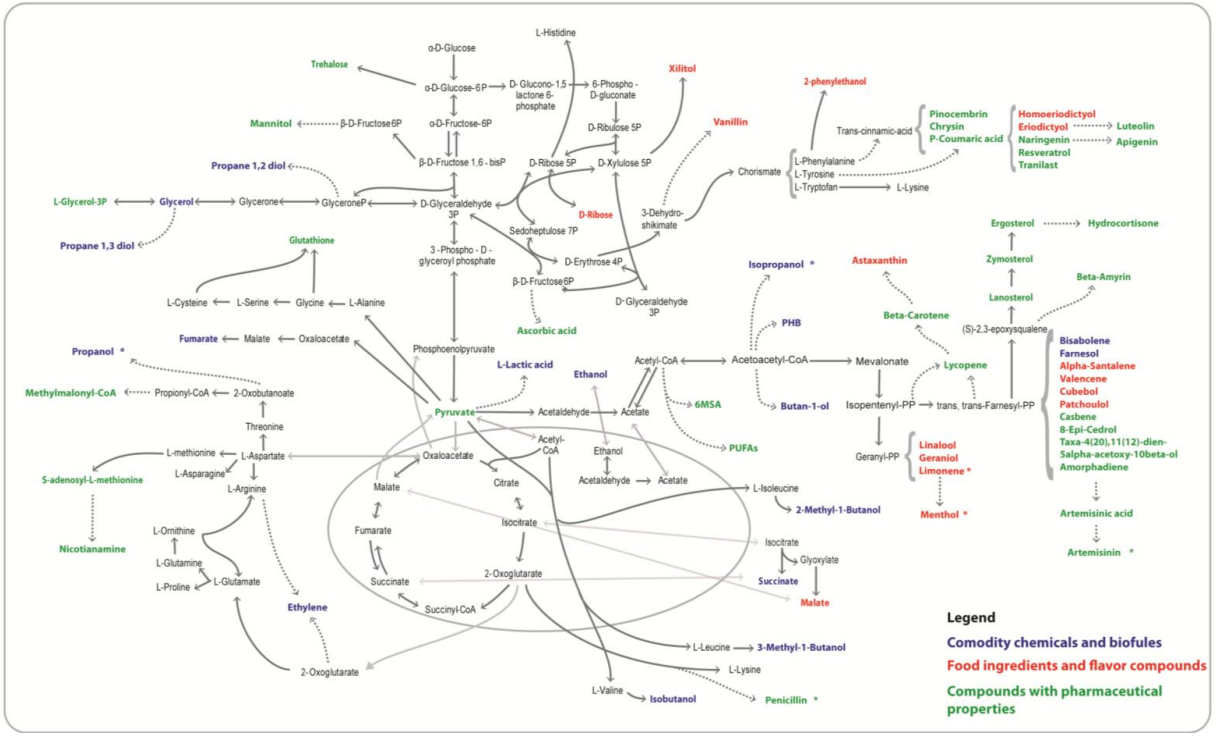


Figure 2.3: *Saccharomyces cerevisiae*, an established cell factory for small molecules. Schematic representation of the *S. cerevisiae* metabolic network containing metabolites and connecting reactions (arrows) relevant for the biosynthesis of added-value small molecules (blue, red and green). Native (pointed at by full arrows) and heterologous (pointed at by dashed arrows) compounds which were reported to be produced in *S. cerevisiae* are presented, as well as potential interesting compounds which were not yet reported in literature (*). The products are classified as commodity chemicals and biofuels (blue), flavouring agents, food ingredients and aroma compounds (red) and molecules with pharmaceutical properties (green). The complete list of compounds is presented in Table ST1.2. This image was kindly prepared by Erica Valentini and Tomasz Boruta.

One of the most prominent examples of successful metabolic engineering is the production of the antimalarial drug precursor, artenimisinic acid, in *S. cerevisiae* [73]. The strategy required the heterologous expression of two enzymes for synthesizing artenimisinic acid from farnesyl pyrophosphate, an intermediate of the mevalonate pathway, essential for obtaining ergosterol. Additionally, de-regulation of the mevalonate pathway was performed by overexpressing several pathway genes, in order to increase the precursor supply for artenimisinic acid [73]. Other examples of successful metabolic engineering are higher alcohols production by *E. coli* [80], sesquiterpenes and biofules production by *S. cerevisiae* [76, 81]. Applications of metabolic engineering are not limited to increase product range and yield. Several attempts for engineering microbes for expanding substrate range are underway, due to the absolute requirement of avoiding resource competition between food and chemicals production [63, 64, 82]. Substrates such as waste cheese whey and agricultural by-products (corn-cob, bagasses and lignocellulosic plant stocks) represent possible options, yet they usually comprise a diversity of sugars which are not naturally metabolized by conventional microorganisms [64, 83]. Thus, several efforts have been put into expanding *S. cerevisiae* range of substrate utilization, *e.g.* cellulose [84], L-arabinose [85], lactose [86], galactose [62] and xylose [87, 88].

Metabolic engineering cycle

The process to obtain microbial cell factories engages a sequence of **design**, **construction** and **analysis**, often applied in a cyclic manner (**Fig. 2.4**) [89]. Over the last decade, advances in Synthetic and Systems Biology generated new resources and technologies extremely useful for metabolic engineering [19, 90]. Nowadays, the concept of metabolic engineering for strain improvement is expanded from “rational metabolic engineering” to include “evolutionary engineering” and “reverse metabolic engineering”, with implication on different stages of the strain improvement process [19, 90].

The first stage while establishing a cell factory is the decision about what workhorse to select. Several factors must be considered, such as whether the host is, or not, a native producer, its response to different substrates, toxicity of the compound and pathway intermediates, availability of genetic engineering tools, availability of vast knowledge about physiology and cell biology (important for further engineering) and suitability to the fermentation process [77].

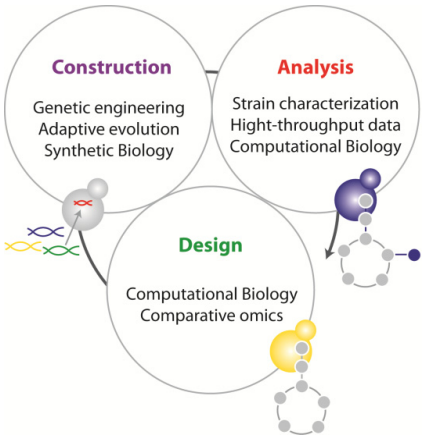


Figure 2.4: Metabolic engineering cycle. Improving cell factories involves three consecutive steps, design, construction and analysis. Each circle corresponds to one step. Design employs computational tools and comparative 'omics' in order to find target for genetic manipulation. Construction employs genetic and protein engineering, as well as synthetic biology tools for introducing the desired modifications. Adaptive evolution can be used as a global approach to evolve the cell factory under the appropriate selective pressure, so the desired phenotype is favored. Analysis employs strain characterization and systems biology tools for obtaining a complete readout of the engineered phenotype. Computational biology plays a fundamental role, often applied for integrated data analysis.

The design phase might be the first stage of the cycle, when one aims at producing a heterologous compound, or a more advanced step, when the host already produces the desired metabolite and the goal is to improve its productivity [89]. The design of cell factories for heterologous production of valuable compounds benefits from Synthetic Biology tools for assembly of new metabolic pathways [91]. *In silico* tools allowing the simulation of novel pathways, such as the Biochemical Network Integrated Computational Explorer (BNICE), are especially helpful for providing guidance for experimental work [92]. Alternatively, when the aim is to improve an existent pathway, the question is how to retrofit the entire metabolic network so it can support high product yields. Several options can be employed here; one possibility is to take a rational approach involving a genome-scale metabolic model and available algorithms to find gene targets to be manipulated, thereby increasing the production [93, 94]. This approach is discussed in detail in the next section. Otherwise, one can take an organism which has the desired phenotype, carry out different high-throughput analysis and perform reverse engineering to implement the relevant findings on the selected strain. The organism with the desired features might be a native producer, or it may also be found while screening a mutant library, or even a mutant obtained through adaptive evolution [95–98].

The **construction** step is the implementation of the designed strategy. Synthetic biology and protein engineering are the excellence fields, providing the most sophisticated tools for

engineering the cells. Depending on the strategy to be implemented, different tools might be applied for constructing the improved cell factory. Some of them focus on targeted effects, such as control of gene expression and protein activity, while other approaches are rather global, such as adaptive evolution or global transcription machinery engineering [19, 90]. Quantitative control of gene expression levels is the most established approach and it can be achieved in a number of different ways. One option is to control of the gene copy number, which can be accomplished by providing a plasmid or by genomic integration (or deletion) of the target gene. Instead, the native promoter of the target gene can also be swapped with another promoter of interest. Both approaches for controlling gene expression largely benefit from promoter libraries, generated either from native [99] or synthetic promoter regions [100–102]. Finally, riboregulators, such as antisense RNA or short interfering RNAs (siRNA) can also be used to efficiently regulate gene expression by post-transcriptional control [103–105]. The increasing availability of fast cloning techniques, such as the Gateway system [106] or USER cloning [107, 108], highly facilitate the construction step, since they allow the assembly of fairly big DNA constructs, while being relatively cheap and uncomplicated techniques. Rather than controlling gene expression, increasing efficiency of an enzyme, either by increased enzyme activity or selectivity (*e.g.* change of the cofactor usage), constitutes another option for improving cell factories. Protein and enzyme engineering research areas contribute with different options; 1) rational design by site-directed mutagenesis based on existing knowledge and computational design, and 2) directed evolution achieved through random mutagenesis based on error-prone PCR or DNA shuffling [109, 110].

Global approaches are useful if fine-tuning of multiple genes is required. Multiple Gene Promoter Shuffling – MGPS and global Transcription Machinery Engineering – gTME, are two examples [111, 112]. Modulating the spacial organization of enzymes, either by using a protein scaffold or by directly fusing of the proteins of interest, is an alternative option also to be used for increasing the metabolic flux through a given pathway [113, 114]. Al last, adaptive evolution can be used as a global approach to evolve the cell factory, if the appropriate selective pressure can be applied and the desired phenotype is favored [62].

The **analysis** step aims at evaluating the cell factory and it always follows the strain construction and precedes the next design step. In order to assess the quality of the cell factory, fully characterization of the obtained mutant is needed. This usually involves physiological characterization of the strains, if possible, in well controlled conditions [89]. The analysis of mutant phenotypic traits has improved enormously within the last 10 years, since it highly

benefits from Systems Biology tools, such as transcriptome, proteome, metabolome and interactome [19, 90]. If the selected construction strategy implied random genetic mutations, e.g. random mutagenesis or adaptive evolution, genome (re)sequencing will also take part in the analysis [97]. Computational tools very much contribute to the data analysis, since models are helpful for data integration and interpretation, especially when multiple types of data are in use [93]. The algorithm for identification of reporter metabolites is a pioneering example of integration of gene expression data with the genome-scale metabolic network topology in order to find the metabolites around which the most significant transcriptional changes occur [115]. The analysis step is fundamental for a good following strategy design step in the case of further improvement, yet it may be also the last step of the strain improvement, when the desired features are obtained [89].

Genome-Scale Metabolic Modeling

In order to fully characterize a given cellular state, one would profit of knowing all the genes being transcribed, as well as all the proteins and metabolites existing within the cell given certain conditions. Yet, as already mentioned above, how all these components interact is fundamental to fully understand biological systems. To this end, metabolism is a privileged subcellular system, since interactions between enzymes and substrates/products have been extensively collected and characterized over several years [34, 93]. Metabolic pathway reconstruction has been around for reasonably long time, and availability of complete genome sequences, from the late 1990's, prompted the assembly of complete metabolic networks, the so called genome-scale metabolic network reconstructions [93, 116]. Such reconstructions essentially contain a set of stoichiometric equations representing the biochemical reactions of the cell, as well as detailed information about reaction reversibility and cofactor usage. Furthermore, as many of the reactions have known association with genes, these models provide an excellent opportunity for studying genotype to phenotype relationships [93].

Briefing on genome-scale metabolic models reconstruction

Historically, the initial application of genome-scale metabolic models was to provide understanding of the characteristics of microbial pathogens, so the first model available was for *Haemophilus influenza* [117]. Since then, several organisms had their network reconstructed, including *E. coli* [118], among other bacteria, *S. cerevisiae* [54], the first eukaryote, *Arabidopsis Thaliana* [119], the first plant, and even *Homo sapiens* [120]. To date, more than 60 species have

genome-scale metabolic reconstructions available and an especially dedicated knowledgebase – Biochemical Genetic and Genomic (BiGG) – is assembled [121]. Since whole genome sequencing prices continue to decrease and an automated pipeline for metabolic network reconstructions is already available, many more genome scale reconstructions are to come [122]. Kim and co-workers (2011) provide a comprehensive timeline of genome-scale models reconstructed since 1999 until 2011 [116].

Methods for building genome-scale metabolic reconstructions have been established and extensively reviewed [116, 123, 124]. There are three important steps; 1) initial network reconstruction is carried out based on the functional annotated genome sequence followed by 2) extensive literature search for collecting biochemical evidence from several different sources including databases, text books and scientific articles. As much information as possible is collected about enzymes and reactions characteristics, such as EC number, localization, associated gene(/s) and reaction reversibility, so as to obtain a well curated reconstruction. 3) The reconstruction is converted into a mathematical model (described in the next section) that can be analysed through constraint based approaches and validated using experimental data on cell physiology. Although network reconstruction has been extensively standardized, new relevant information about the organisms is constantly arising, especially concerning organisms with extensive research communities, such as *S. cerevisiae* or *E. coli*. Thereby, several updated and, in some cases, manually curated high-quality models are usually available for such organisms. *S. cerevisiae* is an outstanding example with 10 different models available, including a consensus model obtained by a community effort towards standardization of this resource [125, 126].

Applications of genome-scale models

The applications of the genome-scale metabolic models are continuously increasing and they are often conveniently divided into different categories, such as: contextualization of high-throughput data, directing hypothesis-driven discovery, network properties analysis, study of multi species relationships, evolution of metabolic networks and guidance for metabolic engineering. Several reviews focusing on genome scale models applications are recently available [93, 116, 126], thus only few examples will be referred here.

Numerous studies are available where the metabolic network is used to interpret high-throughput data in an integrative fashion [27, 115, 127, 128]. Zelezniak *et al.* used network topology analysis to elucidate transcriptional regulatory signals of Type 2 Diabetes [129]. Other

examples suggest the use of genome-scale models for finding drug targets [130, 131]. Recent applications have appeared suggesting the use of such models for studying microbial interactions [132–134]. At last, genome-scale models are extensively used for designing metabolic engineering strategies [81, 135, 136], also the subject of Chapter 2 of the thesis.

Predicting the distribution of metabolic fluxes

While some applications (*e.g.* network properties or contextualization of high-throughput data) often use stoichiometric models as scaffold for analysis, focusing on the network topology, others make use of the predictive capabilities of the models for estimating cell physiology and internal flux distributions (*e.g.* metabolic engineering). Therefore, these quantitative approaches require a proper mathematical description of the metabolic network and its operation modes.

Already in the 1970's, **Metabolic Control Analysis** aimed at elucidating the parameters responsible for the control of metabolic fluxes [137, 138]. Flux control is determined by the kinetic parameters of the enzymes, as well as by the thermodynamic constraints of each reaction. Therefore, kinetic modeling of metabolism requires extensive knowledge about enzyme parameters, which should, ideally, be collected at conditions similar to those *in vivo*. Unfortunately, this is not the case for most of the available data, mostly because of lack of knowledge for reproducing the intracellular environmental conditions. It was not until 2010 that efforts have been put together to characterize the intracellular pH, as well as the concentration of several salts, so standard conditions for measuring kinetic parameters could be established [139–141]. Yet, dynamic modeling of metabolism would also require information about allosteric regulation and post-translational modifications of the enzymes, which still remain highly uncharacterized [142].

In contrast, **Metabolic Flux Analysis** (or Metabolic Network Analysis) became popular in the 1990's, initially using models describing the central carbon metabolism, and it aims at elucidating the metabolic flux distribution based on reaction stoichiometry, metabolite exchange rates and thermodynamic constraints [21, 126, 143]. This approach for modeling cellular metabolism relies on mass conservation and thermodynamic constrains at a steady state, thereby it does not require extensive knowledge about reaction kinetics. The metabolic model is represented as a system of linear equations (**Fig. 2.5A**). The stoichiometric coefficients are represented in the stoichiometric matrix, where the number of rows corresponds to the number of intracellular metabolites, and each column corresponds to a reaction (**Fig. 2.5B**). Because the metabolic networks usually contain more reactions than metabolites (more linearly independent

variables than constraints), the system is frequently underdetermined and many solutions are possible. There are two major categories for solving such linear system, thus predicting what metabolic pathways are active under certain environmental and genetic conditions: pathway-based approaches and optimization-based approaches [144]. Pathway-based approaches explore the entire reactions space by using one of the four methods currently established: extreme pathways [145], elementary flux modes [146], extreme currents [147] and minimal generators [148] (reviewed in [149, 150]). Given the combinatorial nature of the underlying solution space, application to large-scale metabolic networks, as genome-scale models, is still limited [151]. Therefore, optimization-based approaches, which use the same mass balance and thermodynamic constrains, can be applied in order to find the active flux distribution by using a pre-defined objective function.

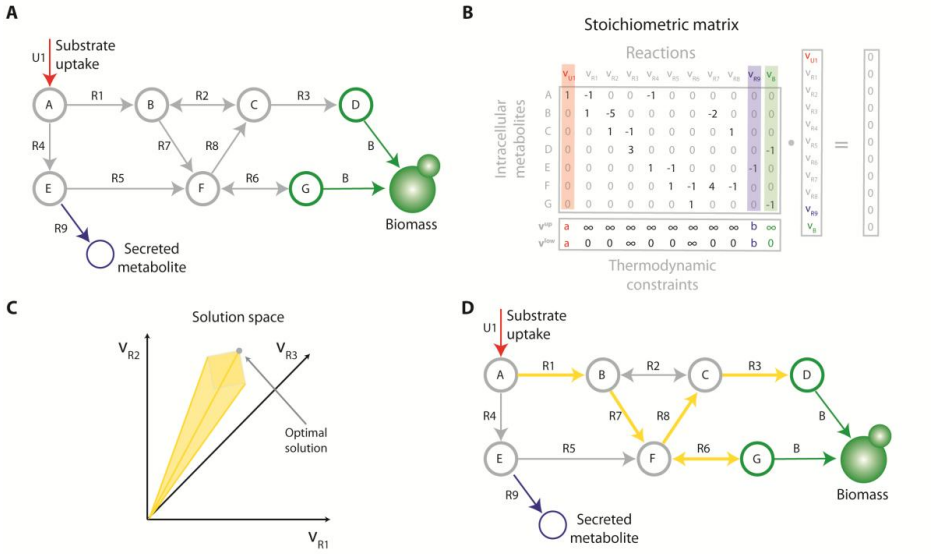


Figure 2.5: **Fundamentals of optimization based approaches for simulation of cellular phenotype. A)** The metabolic network, which contains all stoichiometric relationships between metabolites (circles), as well as reversibility information concerning all the reactions (arrows). U1 represents the reaction for substrate uptake (red). Reaction B (green) connects all the metabolites required for growth. Reaction R9 (purple) represents a metabolite which is known to be secreted (blue circle). **B)** The stoichiometric matrix contains all the stoichiometric coefficients. Each row corresponds to one metabolite, while each column corresponds to one reaction. The steady state assumption after applying mass balance constraints to each intracellular metabolite implies that there is no variation of the metabolites concentration over time (column of zeros). **C)** The stoichiometric matrix, together with the thermodynamic constraints, as well as information about metabolites uptake and secretion constraint the solution space. **D)** An objective function is established and maximization/minimization of the objective function yields to the optimal internal flux distribution (yellow).

Flux Balance Analysis (**FBA**) was initially proposed by Varma and co-workers in 1993, and it was the first optimization-based method established for estimating active flux distributions [152, 153]. The problem was implemented by using a linear programming formulation (LP [154]), and solved for maximizing growth, hypothesised as metabolic **objective function** (**Fig. 2.5D**) [152, 153]. Reactions reversibility information is used for *upper* and *lower* bounds to constraint the flux of the reactions (variables), and the exchange rates between the cell and the extracellular broth, *e.g.* glucose uptake and ethanol production rates, are also taken into account for further constraining the solution space (**Fig. 2.5C**).

$$\begin{aligned} &\max v_{Growth} \\ &s.t. \quad S \cdot v = 0 \\ &\qquad v_i^{lb} \le v_i \le v_i^{ub} \qquad \forall i \in N \\ &\qquad v_{glc} \le a \\ &\qquad v_{eth} \ge b \end{aligned}$$

Where v_{Gorwth} represents the flux through the reaction converting all required metabolites into biomass (biomass equation), S represents the stoichiometric matrix, *v* is the vector containing the fluxes, v_i^{lb} and v_i^{ub} correspond to the lower and upper bounds applied to reaction *i*, v_{glc} represents the glucose (or other substrate) uptake rate, which is experimentally determined (*a*). If no more information is given, *e.g.* experimentally determined fluxes, the upper and lower bounds are defined by the reaction reversibility. For reversible reactions both upper and lower bounds are considered to be unbounded (+/- infinity). In the case of irreversible reactions, the lower bound is set to zero (**Fig. 2.5B**).

This approach proved to be very useful, especially for predicting gene essentiality and, to some extent, the flux distribution for microbes growing under optimal conditions [155, 156]. Nowadays, FBA is vastly used for metabolic modeling and there are several reviews on the method [157, 158]. New methods were established based on FBA, such as dynamic FBA, for estimating metabolic distributions over a time-course [159], rFBA, for including regulatory constraints [160], iFBA, which includes regulatory constraints over a time-course [128], FBAwMC, which includes molecular crowding as additional constraints [161], among many others (reviewed [94]). The COBRA toolbox is a specialized Matlab package for constraint-based reconstruction and analysis of genome scale models [162, 163].

Designing and selecting objective functions for metabolic modeling

While using optimization-based approaches, the outcome of the simulations is totally dependent on the **objective function** to be used, therefore it is very important that the selected objective function is suitable for the problem [154]. The design of objective functions, properly formulating the desired biological principles, has been a topic of discussion over time. Computational tools have been proposed to address this question, *e.g.* ObjFind [164], GrowMatch [165] and BOSS [166]. Moreover, Schuetz and colleagues experimentally tested several hypothesis applicable to *E. coli* growing under different experimental conditions, *e.g.* maximization of growth, maximization of ATP yield or minimization of overall intracellular flux [167]. Their findings suggest that there is no universal objective function which applies to all the conditions and systems, implying that the biological principles governing cellular metabolism under different conditions might also be different.

The principle underlying the objective function first selected for FBA, maximization of growth, implies that the organisms are optimized for maximal growth, coherent with evolutionary theories [152, 153]. On the other hand, many approaches to answer biological questions lay on perturbation of the system and evaluation of the outcome. For instance, a widely spread method for assigning gene function is to delete the gene and look for missing features [52, 55]. Thus, if metabolic modeling is to be used for understanding such mechanisms, other objective functions must be used instead of maximization of fitness, since deleting a gene may impair the assumption of "fully adapted behaviour" [168]. Predicting the behaviour of perturbed metabolic networks is extremely relevant for basic research, as simulation of gene deletions, but also for applied fields, *e.g.* metabolic engineering, where the aim is to select the genetic perturbation leading to specific cellular features [79].

In 2003, Segre and co-workers proposed **Minimization of Metabolic Adjustment** (**MoMA**) as biological principle to simulate the behaviour of perturbed metabolic networks [168]. In contrast to FBA, MoMA relies on the premise that the mutant flux distribution will be as similar as possible to that of the wild-type. It was originally formulated as a quadratic programming problem [169] for minimization of quadratic distance between the flux distribution of the wild-type and the mutant [168]. A linear programming formulation for MoMA has also been proposed, lMoMA, and it is has been shown to have comparable results [162]. MoMA was reported to outperform FBA in some applications, namely related to the prediction of single gene deletion phenotypes and their internal flux distributions [168, 170], metabolic engineering [81] and microbial community analysis [132]. It is important to mention that the quality of the

wild-type flux distribution to be used as reference is crucial for obtaining meaningful results with MoMA. Such reference flux distribution is often obtained by using FBA while, desirably, extra information from ^{13}C based flux analysis can be used to impose additional constraints on the metabolic network. At this stage, minimization of intracellular flux (another objective function) can also be used in order to minimize mathematical artifacts, such as closed loops of fluxes that do not bring net change, and further improve the reference flux distribution [170]. At last, for predicting the phenotype of perturbed metabolic networks, Regulatory On/Off Minimization of metabolic flux changes – ROOM, has also been proposed [171]. With a similar biological hypothesis to MoMA, ROOM aims at minimizing the number of significant flux changes with respect to the wild type, but it applies a Mixed Integer Linear Programming (MILP [154]) formulation, as opposed to minimizing the total distance between the flux vectors.

Metabolic modeling for guiding metabolic engineering

As previously mentioned, retrofitting the metabolic network by rewiring metabolic fluxes is fundamental for obtaining increased production of the desired compound, and thereby sustainable cell factories [19, 90]. However, the high connectivity-degree of metabolic networks, often related to the usage of cofactors, such as ATP and NADH, weakens the potential of intuitive solutions for the problem [172]. Furthermore, an optimized cell factory requires a fine-tuned balance of allocation of cellular resources between the production of biomass and of the desired compound [21, 89, 173]. To this end, genome-scale metabolic models can be applied in order to provide guidance for metabolic engineering, while finding target genes for genetic manipulation. The implementation of such problem is not trivial; if a linear programming routine is applied for maximizing the production of a given metabolite, the optimal solution does not lead to biomass formation in most of the cases, therefore the solutions are often not biologically meaningful. In 2003, the first platform developed towards finding target genes, deletion of which would lead to improved yield of a desired compound, while the biological objective function is still satisfied, was proposed by Burgard *et al*. [174]. **OptKnock** uses an elegant approach for finding the genes that, when deleted, would couple the production of the desired compound to biomass formation. In genome-scale metabolic modeling context, a gene deletion is formulated by setting the upper and lower bounds of the corresponding reaction to zero. By using a bi-level (or nested) optimization, they suggested the maximization of the production of the desired metabolite (design objective), subject to maximization of growth (biological objective, **Fig. 2.6A**). The number of genes to be deleted is a parameter decided prior to simulation and it depends on the product and producing host; deletion of several genes is often

required, due to the robustness of metabolism [172, 175, 176]. Given the number of genes to be deleted, OptKnock is guaranteed to find the targets leading to the highest production (global optimal solution) [174]. Patil *et al.* put forward another platform for finding metabolic engineering targets based on Darwinian evolutionary principles, **OptGene** [177]. They propose the use of a genetic algorithm, which goes through rounds of mutations, crossing and selection in order to find the best candidate, given a fitness function (**Fig. 2.6B**). The fitness function is also of particular novelty within this algorithm, because such framework enables the use of non-linear functions as objective. A proxy for productivity is suggested by the authors, which is obtained by multiplying the growth flux by the flux leading to the product of interest, Biomass-Product Coupled Yield - BPCY. As a genetic algorithm cannot guarantee to find the optimal solution for the problem, the number of generations or the fitness function can be used as stopping criteria for OptGene [177].

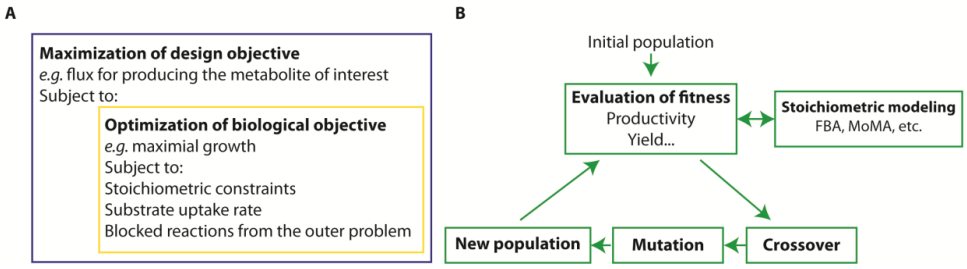


Figure 2.6: OptKnock and OptGene optimization frameworks for guiding Metabolic Engineering. A) Bi-level optimization structure of OptKnock (in [174]). **B)** OptGene schematic overview (adapted from [177]).

Several other algorithms with different/improved characteristics are available nowadays; OptStrain, which finds optimal reactions to be deleted from, as well as to be added to, the metabolic network in order to improve the phenotype [178]. OptForce, which finds reaction targets for deletion or overexpression in order to achieve a desired phenotype [98]. OptORF, which aims at finding optimal metabolic and regulatory perturbations to design metabolic engineering strategies [179]. An open-source software platform for *in silico* metabolic engineering, termed OptFlux, focused on providing the user with some of these frameworks, as well as different options for alternative objective functions, is already available [180]. Alternatively, the latest version of the COBRA Toolbox (v2.0) also includes some of these algorithms [163].

Even though several *in silico* tools for guidance of metabolic engineering are available, obtaining improved cellular features through rational design is still a challenge. For instance, though optimization problems are guaranteed to provide an optimal solution, this does not imply

uniqueness [154]. Given the existence of several alternative pathways within metabolism, some of these pathways may satisfy the same requirements for the objective function, thereby they are also alternative solutions. Identifying the uniqueness of a solution can be accomplished by, *e.g.* maximizing and minimizing all fluxes under the given set of constraints (flux variability analysis), while identifying all optimal solutions may be accomplished by implementing a MILP routine or by elementary mode analysis [146, 176, 181]. Flux variability analysis should always be taken into account while using genome-scale models, especially for metabolic engineering, in order to guarantee that the selected targets will lead to improved cellular features, independently of the uniqueness of the solution. RobustKnock is an extension of OptKnock and it aims at finding optimal set of gene deletions for improved production, while guaranteeing that non-unique solutions (solutions presenting alternative pathways) are avoided [182].

Fine-tuning of metabolic networks often involves the reduction, suppression and/or amplification of fluxes. While deletion of a reaction is a binary problem, therefore relatively simple for *in silico* formulation, reduction or amplification of fluxes implies knowing the amplitude of the variation. More importantly, once a reaction if found as a target for amplification (or reduction), *e.g.* by OptForce, a quantitative relationship between reaction amplification (or reduction) and gene overexpression (or downregulation) should be known in order to accurately predict the effect of the genetic manipulation [98]. Because of complex mechanisms of regulation, the attempts to verify linear correlations between gene expression and flux changes have failed so far, suggesting the prevalence of non-linear relationships [183]. Further characterization of the cellular regulatory mechanisms is expected to contribute to better characterize such non-linear relationships. Moreover, this would allow more accurate quantitative prediction of genetic manipulations, useful for metabolic engineering, but also for expanding our understanding of genotype to phenotype relationships.

At last, incorporation of regulatory constraints in metabolic models is expected to greatly influence the outcome of predictions. As mentioned before, some frameworks for including transcriptional networks within phenotype simulations are already available [160], as well as a simulation platform for including regulatory genes as targets for metabolic engineering, OptORF [179]. Despite providing better predictive power in some cases [160], these frameworks are limited due to lack of extensive knowledge about regulatory mechanisms from a qualitative, but mostly from a quantitative perspective. Regulatory networks mostly present non-linear behaviour and, also here, finding a solution for combining non-linear approaches with the established linear optimization tools would be of great interest for the scientific community.

Conclusions

Metabolism is one of the most well characterized subcellular systems, as biochemical reactions occurring within the cell have been studied and assembled in metabolic pathways over the years. A large fraction the genes coding for enzymes, referred as metabolic genes, are functionally annotated. Yet, prediction of phenotype based on genotype is, in many cases, still not possible. A holistic approach of cellular systems is needed to completely understand how phenotype arises from the genotype, and genome-scale stoichiometric models represent a suitable computational framework for simulating and studying such relationships. Better characterization of biological features, such as regulation, are expected to expand the range of application and quality of genome-scale models.

Availability of extensive knowledge and dedicated technology for genetic manipulation and high-throughput data generation conferred *S. cerevisiae* its current status of a eukaryotic model organism. Additionally, this yeast is largely used for production of food and beverages and, thereby, it is well established among industrial applications. Combined, these attributes make *S. cerevisiae* a preferred host for metabolic engineering. This chapter illustrated the current applications of yeast as a model organism and as cell factory. The importance of applying stoichiometric genome-scale modeling for metabolic engineering was also emphasized.

References

1. Snyder M, Gallagher JEG: **Systems biology from a yeast omics perspective.** *FEBS Letters* 2009, **583**:3895-9.

2. Chuang H-Y, Hofree M, Ideker T: **A decade of systems biology.** *Annual Review of Cell and Developmental Biology* 2010, **26**:721-44.

3. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology.** *Annual Review of Genomics and Human Genetics* 2001, **2**:343-72.

4. Kitano H: **Systems biology: a brief overview.** *Science* 2002, **295**:1662-4.

5. Tong AH, Evangelista M, Parsons AB, et al.: **Systematic genetic analysis with ordered arrays of yeast deletion mutants.** *Science* 2001, **294**:2364-8.

6. Hin A, Tong Y, Lesage G, et al.: **Global Mapping of the Yeast Genetic Interaction Network**. *Science* 2004, **303**:808-813.

7. Sanger F, Nicklen S, Coulson A: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**:5463-5467.

8. Reaves ML, Rabinowitz JD: **Metabolomics in systems microbiology.** *Current Opinion in Biotechnology* 2011, **22**:17-25.

9. Rubin GM: **Comparative Genomics of the Eukaryotes**. *Science* 2000, **287**:2204-2215.

10. Tringe SG, Mering C von, Kobayashi A, et al.: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**:554-7.

11. Mockler TC, Chan S, Sundaresan A, et al.: **Applications of DNA tiling arrays for whole-genome analysis.** *Genomics* 2005, **85**:1-15.

12. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Research* 2008, **18**:1509-17.

13. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature Reviews Genetics* 2009, **10**:57-63.

14. Crick FC: **On protein synthesis**. *Symposia of the Society for Experimental Biology* 1958, **12**:138-63.

15. Crick FC: **Central Dogma of M olecular Biology**. *Nature* 1970, **227**:561-3.

16. Straub L: **Beyond the Transcripts : What Controls Protein Variation ?** *PLoS Biology* 2011, **9**:9-10.

17. Gstaiger M, Aebersold R: **Applying mass spectrometry-based proteomics to genetics, genomics and network biology.** *Nature Reviews Genetics* 2009, **10**:617-27.

18. Sabidó E, Selevsek N, Aebersold R: **Mass spectrometry-based proteomics for systems biology.** *Current Opinion in Biotechnology* 2011:1-7.

19. Kim I-K, Roldão A, Siewers V, Nielsen J: **A systems-level approach for metabolic engineering of yeast cell factories.** *FEMS Yeast Research* 2012, **12**:228-48.

20. Wiechert W, Möllney M, Isermann N, Wurzel M, Graaf AA De: **Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems.** *Biotechnology and Bioengineering* 1999, **66**:69-85.

21. Christensen B, Nielsen J: **Metabolic network analysis. A powerful tool in metabolic engineering.** *Advances in Biochemical Engineering* 2000, **66**:209-31.

22. Nielsen J: **It Is All about Metabolic Fluxes**. *Journal of Bacteriology* 2003, **185**:7031-5.

23. Sauer U: **Metabolic networks in motion: 13C-based flux analysis.** *Molecular Systems Biology* 2006, **2**:62.

24. Gedela S: **Bioinformatics for Omics Data**. *Methods in Molecular Biology* 2011, **719**:399-414.

25. Joyce AR, Palsson BØ: **The model organism as a system: integrating "omics" data sets.** *Nature Reviews Molecular Cell Biology* 2006, **7**:198-210.

26. Khurana V, Lindquist S: **Modeling neurodegeneration in Saccharomyces cerevisiae: why cook with baker's yeast?** *Nature Reviews Neuroscience* 2010, **11**:436-49.

27. Usaite R, Jewett MC, Oliveira AP, et al.: **Reconstruction of the yeast Snf1 kinase regulatory network reveals its role as a global energy regulator.** *Molecular Systems Biology* 2009, **5**:319.

28. Nagaraj N, Wisniewski JR, Geiger T, et al.: **Deep proteome and transcriptome mapping of a human cancer cell line.** *Molecular Systems Biology* 2011, **7**:548.

29. Gehlenborg N, O'Donoghue SI, Baliga NS, et al.: **Visualization of omics data for systems biology.** *Nature Methods* 2010, **7**:S56-68.

30. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H: **Software for systems biology: from tools to integrated platforms.** *Nature Reviews Genetics* 2011, **12**:821-32.

31. Hucka M, Finney a., Sauro HM, et al.: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**:524-531.

32. Feldmann H: *Yeast Molecular and Cell Biology*. Wiley Blackwell; 2009.

33. Roberts IN, Oliver SG: **The yin and yang of yeast: biodiversity research and systems biology as complementary forces driving innovation in biotechnology.** *Biotechnology Letters* 2011, **33**:477-87.

34. Fraenkel DG: *Yeast Intermediary Metabolism*. Cold Spring Harbor Laboratory Press; 2011.

35. Deken RH De: **The Crabtree effect: a regulatory system in yeast.** *Journal of General Microbiology* 1966, **44**:149-56.

36. Postma E, Verduyn C, Scheffers WA, Dijken JP van: **Enzymic Analysis of the Crabtree Effect in Glucose-Limited Chemostat Cultures of Saccharomyces cerevisiae**. *Applied and Environmental Microbiology* 1989, **55**:468-77.

37. Nielsen J, Villadsen J, Liden G: *Bioreaction Engineering Principles*. Second Edi. Kluwer Academy/Plenum; 2002.

38. Gancedo JM: **Yeast carbon catabolite repression.** *Microbiology and Molecular Biology Reviews* 1998, **62**:334-61.

39. Carlson M: **Glucose repression in yeast**. *Current Opinion in Microbiology* 1999, **2**:202-207.

40. Johnston M: **Feasting , fasting and fermenting. Glucose sensing in yeast and other cells**. *Trends in Genetics* 1999, **15**:29-33.

41. Rolland F, Winderickx J, Thevelein JM: **Glucose-sensing and -signalling mechanisms in yeast.** *FEMS Yeast Research* 2002, **2**:183-201.

42. Santangelo GM: **Glucose Signaling in Saccharomyces cerevisiae Glucose Signaling in Saccharomyces cerevisiae**. *Microbiology and Molecular Biology Reviews* 2006, **70**:253-82.

43. Christensen TS, Oliveira AP, Nielsen J: **Reconstruction and logical modeling of glucose repression signaling pathways in Saccharomyces cerevisiae.** *BMC Systems Biology* 2009, **3**:7.

44. Westergaard SL, Oliveira AP, Bro C, Olsson L: **A Systems Biology Approach to Study Glucose Repression in the Yeast Saccharomyces cerevisiae**. *Biotechnology* 2007, **96**:134-145.

45. Maris AJ van, Bakker BM, Brandt M, et al.: **Modulating the distribution of fluxes among respiration and fermentation by overexpression of HAP4 in Saccharomyces cerevisiae.** *FEMS Yeast Research* 2001, **1**:139-49.

46. Beggs JD: **Transformation of yeast by a replicating hybrid plasmid**. *Nature* 1978, **275**:104-9.

47. Hinnen a, Hicks JB, Fink GR: **Transformation of yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 1978, **75**:1929-33.

48. Goffeau A, Barrell BG, Bussey H, et al.: **Life with 6000 Genes**. *Science* 1996, **274**.

49. Botstein D, Fink GR: **Yeast: an experimental organism for 21st Century biology.** *Genetics* 2011, **189**:695-704.

50. DeRisi JL: **Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale**. *Science* 1997, **278**:680-686.

51. Uetz P, Giot L, Cagney G, et al.: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403**:623-7.

52. Giaever G, Chu AM, Ni L, et al.: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418**:387-91.

53. Huh W-K, Falvo JV, Gerke LC, et al.: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425**:686-91.

54. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13**:244-53.

55. Costanzo M, Baryshnikova A, Bellay J, et al.: **The genetic landscape of a cell**. *Science* 2010, **327**:425-31.

56. Baryshnikova A, Costanzo M, Kim Y, et al.: **Quantitative analysis of fitness and genetic interactions in yeast on a genome scale.** *Nature Methods* 2010, **7**:1017-24.

57. Bendert A, Pringle JR: **Use of a Screen for Synthetic Lethal and Multicopy Suppressee Mutants To Identify Two New Genes Involved in Morphogenesis in Saccharomyces cerevisiae**. *Molecular and Cell Biology* 1991, **11**:1295-1305.

58. Avery L, Wasserman S: **Ordering gene function: the interpretation of epistasis in regulatory hierarchies.** *Trends in Genetics* 1992, **8**:312-6.

59. Boone C, Bussey H, Andrews BJ: **Exploring genetic interactions and networks with yeast.** *Nature Reviews Genetics* 2007, **8**:437-49.

60. Szappanos B, Kovács K, Szamecz B, et al.: **An integrated approach to characterize genetic interaction networks in yeast metabolism.** *Nature Genetics* 2011, **43**:656-62.

61. VanderSluis B, Bellay J, Musso G, et al.: **Genetic interactions reveal the evolutionary trajectories of duplicate genes.** *Molecular Systems Biology* 2010, **6**:429.

62. Hong K-K, Vongsangnak W, Vemuri GN, Nielsen J: **Unravelling evolutionary strategies of yeast for improving galactose utilization through integrated systems level analysis.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:12179-84.

63. Nevoigt E: **Progress in metabolic engineering of Saccharomyces cerevisiae.** *Microbiology and Molecular Biology Reviews* 2008, **72**:379-412.

64. Hong K-K, Nielsen J: **Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries.** *Cellular and Molecular Life Siciences* 2012, **69**.

65. Porro D, Branduardi P: **Yeast cell factory: fishing for the best one or engineering it?** *Microbial Cell Factories* 2009, **8**:51.

66. Yan Y, Liao JC: **Engineering metabolic systems for production of advanced fuels.** *Journal of Industrial Microbiology & Biotechnology* 2009, **36**:471-9.

67. Wattanachaisaereekul S, Lantz AE, Nielsen ML, Andresson OS, Nielsen J: **Optimization of Heterologous Production of the Polyketide 6-MSA in Saccharomyces cerevisiae**. *Biotechnology and Bioengineering* 2007, **97**:893-900.

68. Chemler J a, Yan Y, Koffas M a G: **Biosynthesis of isoprenoids, polyunsaturated fatty acids and flavonoids in Saccharomyces cerevisiae.** *Microbial Cell Factories* 2006, **5**:20.

69. Steen EJ, Chan R, Prasad N, et al.: **Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol.** *Microbial Cell Factories* 2008, **7**:36.

70. Overkamp KM, Bakker BM, Kötter P, et al.: **Metabolic Engineering of Glycerol Production in Saccharomyces cerevisiae**. *Applied and Environmental Microbiology* 2002, **68**:2814-21.

71. Lee SK, Chou H, Ham TS, Lee TS, Keasling JD: **Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels.** *Current Opinion in Biotechnology* 2008, **19**:556-63.

72. Hansen EH, Møller BL, Kock GR, et al.: **De novo biosynthesis of vanillin in fission yeast (Schizosaccharomyces pombe) and baker's yeast (Saccharomyces cerevisiae).** *Applied and Environmental Microbiology* 2009, **75**:2765-74.

73. Ro D-K, Paradise EM, Ouellet M, et al.: **Production of the antimalarial drug precursor artemisinic acid in engineered yeast.** *Nature* 2006, **440**:940-3.

74. Møller K, Nielsen KF, Asadollahi MA, et al.: **Production of Plant Sesquiterpenes in Saccharomyces cerevisiae : Effect of ERG9 Repression on Sesquiterpene Biosynthesis**. *Biotechnology* 2008, **99**:666-677.

75. Tavares S, Grotkjær T, Obsen T, et al.: **Metabolic engineering of Saccharomyces cerevisiae for production of Eicosapentaenoic Acid, using a novel delta5-Desaturase from Paramecium tetraurelia.** *Applied and Environmental Microbiology* 2011, **77**:1854-61.

76. Lee W, Dasilva N a: **Application of sequential integration for metabolic engineering of 1,2-propanediol production in yeast.** *Metabolic Engineering* 2006, **8**:58-65.

77. Keasling JD: **Manufacturing molecules through metabolic engineering.** *Science* 2010, **330**:1355-8.

78. Stephanopoulos G, Vallino JJ: **Network Rigidity and Metabolic Engineering in in Metabolite Overproduction**. *Science* 1991, **252**:1675-1681.

79. Bailey JE: **Toward a science of metabolic engineering.** *Science* 1991, **252**:1668-75.

80. Atsumi S, Hanai T, Liao JC: **Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels.** *Nature* 2008, **451**:86-9.

81. Asadollahi M, Maury J, Patil KR, et al.: **Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering.** *Metabolic Engineering* 2009, **11**:328-34.

82. Lynd LR, Laser MS, Bransby D, et al.: **How biotech can transform biofuels.** *Nature Biotechnology* 2008, **26**:169-72.

83. Elkins JG, Raman B, Keller M: **Engineered microbial systems for enhanced conversion of lignocellulosic biomass.** *Current Opinion in Biotechnology* 2010, **21**:657-62.

84. Haan R Den, Rose SH, Lynd LR, Zyl WH van: **Hydrolysis and fermentation of amorphous cellulose by recombinant Saccharomyces cerevisiae.** *Metabolic Engineering* 2007, **9**:87-94.

85. Wisselink HW, Toirkens MJ, Rosario Franco Berriel M del, et al.: **Engineering of Saccharomyces cerevisiae for efficient anaerobic alcoholic fermentation of L-arabinose.** *Applied and Environmental Microbiology* 2007, **73**:4881-91.

86. Becerra M, Dias Prado S, Rodriıguez-Belmonte E, Cerdan ME, Gonzalez Siso MI: **Metabolic engineering for direct lactose utilization by Saccharomyces cerevisiae**. *Biotechnology Letters* 2002, **24**:1391-1396.

87. Katahira S, Fujita Y, Mizuike A, Fukuda H, Kondo A: **Construction of a Xylan-Fermenting Yeast Strain through Codisplay of Xylanolytic Enzymes on the Surface of Xylose-Utilizing Saccharomyces cerevisiae Cells**. *Applied and Environmental Microbiology* 2004, **70**:5407-5414.

88. Wohlbach DJ, Kuo A, Sato TK, et al.: **Comparative genomics of xylose-fermenting fungi for enhanced biofuel production**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:13212-7.

89. Nielsen J: **Metabolic engineering**. *Applied Microbiology and Biotechnology* 2001, **55**:263-283.

90. Lee JW, Kim TY, Jang Y-S, Choi S, Lee SY: **Systems metabolic engineering for chemicals and materials.** *Trends in Biotechnology* 2011, **29**:370-8.

91. Prather KLJ, Martin CH: **De novo biosynthetic pathways: rational design of microbial chemical factories.** *Current Opinion in Biotechnology* 2008, **19**:468-74.

92. Hatzimanikatis V, Li C, Ionita J a, et al.: **Exploring the diversity of complex metabolic networks.** *Bioinformatics* 2005, **21**:1603-9.

93. Oberhardt MA, Palsson BØ, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Molecular Systems Biology* 2009, **5**:320.

94. Lewis NE, Nagarajan H, Palsson BO: **Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods**. *Nature Reviews Microbiology* 2012.

95. Lee K-S, Hong M-E, Jung S-C, et al.: **Improved galactose fermentation of Saccharomyces cerevisiae through inverse metabolic engineering.** *Biotechnology and Bioengineering* 2011, **108**:621-31.

96. Wright J, Bellissimi E, Hulster E de, et al.: **Batch and continuous culture-based selection strategies for acetic acid tolerance in xylose-fermenting Saccharomyces cerevisiae.** *FEMS Yeast Research* 2011, **11**:299-306.

97. Oud B, Maris AJ a van, Daran J-M, Pronk JT: **Genome-wide analytical approaches for reverse metabolic engineering of industrially relevant phenotypes in yeast.** *FEMS Yeast Research* 2012, **12**:183-96.

98. Ranganathan S, Suthers PF, Maranas CD: **OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.** *PLoS Computational Biology* 2010, **6**:e1000744.

99. Partow S, Siewers V, Bjørn S, Nielsen J, Maury J: **Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae**. *Yeast* 2010, **27**:955-964.

100. Nevoigt E, Kohnke J, Fischer CR, et al.: **Engineering of promoter replacement cassettes for fine-tuning of gene expression in Saccharomyces cerevisiae.** *Applied and Environmental Microbiology* 2006, **72**:5266-73.

101. Gertz J, Siggia ED, Cohen B a: **Analysis of combinatorial cis-regulation in synthetic and genomic promoters.** *Nature* 2009, **457**:215-8.

102. Tyo K, Nevoigt E, Stephanopoulos G: **Directed Evolution of Promoters and Tandem Gene Arrays for Customizing RNA Synthesis Rates and Regulation**. In *Methods in Enzymology*. edited by Voigt C Elsevier Inc.; 2011:135-58.

103. Isaacs FJ, Dwyer DJ, Ding C, et al.: **Engineered riboregulators enable post-transcriptional control of gene expression.** *Nature Biotechnology* 2004, **22**:841-7.

104. Isaacs FJ, Dwyer DJ, Collins JJ: **RNA synthetic biology.** *Nature Biotechnology* 2006, **24**:545-54.

105. Kötter P, Weigand JE, Meyer B, Entian K-D, Suess B: **A fast and efficient translational control system for conditional expression of yeast genes.** *Nucleic Acids Research* 2009, **37**:e120.

106. Alberti S, Gitler AD, Lindquist S: **A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae**. *Yeast* 2007, **24**:913-919.

107. Nour-Eldin HH, Hansen BG, Nørholm MHH, Jensen JK, Halkier B a: **Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments.** *Nucleic Acids Research* 2006, **34**:e122.

108. Geu-Flores F, Nour-Eldin HH, Nielsen MT, Halkier B a: **USER fusion: a rapid and efficient method for simultaneous fusion and cloning of multiple PCR products.** *Nucleic Acids Research* 2007, **35**:e55.

109. Saven JG: **Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins.** *Current Opinion in Chemical Biology* 2011, **15**:452-7.

110. Brustad EM, Arnold FH: **Optimizing non-natural protein function with directed evolution.** *Current Opinion in Chemical Biology* 2011, **15**:201-10.

111. Lu C, Jeffries T: **Shuffling of promoters for multiple genes to optimize xylose fermentation in an engineered Saccharomyces cerevisiae strain.** *Applied and Environmental Microbiology* 2007, **73**:6072-7.

112. Alper H, Stephanopoulos G: **Global transcription machinery engineering: a new approach for improving cellular phenotype.** *Metabolic Engineering* 2007, **9**:258-67.

113. Dueber JE, Wu GC, Malmirchegini GR, et al.: **Synthetic protein scaffolds provide modular control over metabolic flux.** *Nature Biotechnology* 2009, **27**:753-9.

114. Albertsen L, Chen Y, Bach LS, et al.: **Diversion of flux toward sesquiterpene production in Saccharomyces cerevisiae by fusion of host and heterologous enzymes.** *Applied and Environmental Microbiology* 2011, **77**:1033-40.

115. Patil KR, Nielsen J: **Uncovering transcriptional regulation of metabolism by using metabolic network topology.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:2685-9.

116. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY: **Recent advances in reconstruction and applications of genome-scale metabolic models**. *Current Opinion in Biotechnology* 2011:1-7.

117. Edwards JS, Palsson BO: **Systems properties of the Haemophilus influenzae Rd metabolic genotype.** *The Journal of Biological Chemistry* 1999, **274**:17410-6.

118. Edwards JS, Palsson BO: **The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:5528-33.

119. Oliveira Dal'Molin CG de, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK: **AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis.** *Plant Physiology* 2010, **152**:579-89.

120. Duarte NC, Becker S a, Jamshidi N, et al.: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:1777-82.

121. Schellenberger J, Park JO, Conrad TM, Palsson BØ: **BiGG : a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC Bioinformatics* 2010, **11**.

122. Henry CS, DeJongh M, Best A a, et al.: **High-throughput generation, optimization and analysis of genome-scale metabolic models.** *Nature Biotechnology* 2010, **28**:977-82.

123. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ: **Reconstruction of biochemical networks in microorganisms.** *Nature Reviews Microbiology* 2009, **7**:129-43.

124. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nature Protocols* 2010, **5**:93-121.

125. Herrgård MJ, Swainston N, Dobson P, et al.: **A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.** *Nature Biotechnology* 2008, **26**:1155-60.

126. Osterlund T, Nookaew I, Nielsen J: **Fifteen years of large scale metabolic modeling of yeast: Developments and impacts.** *Biotechnology Advances* 2011.

127. Cakir T, Patil KR, Onsan ZIi, et al.: **Integration of metabolome data with metabolic networks reveals reporter reactions.** *Molecular Systems Biology* 2006, **2**:50.

128. Covert MW, Xiao N, Chen TJ, Karr JR: **Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli.** *Bioinformatics* 2008, **24**:2044-50.

129. Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR: **Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes.** *PLoS Computational Biology* 2010, **6**:e1000729.

130. Kim TY, Kim HU, Lee SY: **Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks.** *Metabolic Engineering* 2010, **12**:105-11.

131. Folger O, Jerby L, Frezza C, et al.: **Predicting selective drug targets in cancer through metabolic networks.** *Molecular Systems Biology* 2011, **7**.

132. Wintermute EH, Silver P: **Emergent cooperation in microbial metabolism.** *Molecular Systems Biology* 2010, **6**:407.

133. Klitgord N, Segrè D: **Environments that Induce Synthetic Microbial Ecosystems**. *Ecosystems* 2010, **6**.

134. Zomorrodi AR, Maranas CD: **OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities.** *PLoS Computational Biology* 2012, **8**:e1002363.

135. Choi HS, Lee SY, Kim TY, Woo HM: **In silico identification of gene amplification targets for improvement of lycopene production.** *Applied and Environmental Microbiology* 2010, **76**:3097-105.

136. Lee SJ, Lee D-yup, Kim TY, et al.: **Metabolic Engineering of Escherichia coli for Enhanced Production of Succinic Acid , Based on Genome Comparison and In Silico Gene Knockout Simulation**. *Applied and Environmental Microbiology* 2005, **71**:7880-7.

137. Kacser H, Burns JA: **The control of flux**. *Symposia of the Society for Experimental Biology* 1973, **27**:65-104.

138. Heinrich R, Rapoport TA: **Linear Steady-State Treatment of Enzymatic Chains. General properties, control and effector strength.** *European Jornal of Biochemistry* 1974, **42**:97-105.

139. Rohwer JM: **Kinetic modeling of plant metabolic pathways**. *Journal of Experimental Botany* 2012:1-18.

140. Gombert a K, Nielsen J: **Mathematical modeling of metabolism.** *Current Opinion in BiotechnologyCurrent opinion in biotechnology* 2000, **11**:180-6.

141. Eunen K van, Bouwman J, Daran-Lapujade P, et al.: **Measuring enzyme activities under standardized in vivo-like conditions for systems biology.** *FEBS Journal* 2010, **277**:749-60.

142. Oliveira AP, Sauer U: **The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism.** *FEMS Yeast Research* 2012, **12**:104-17.

143. Gulik WM van, Heijnen JJ: **A metabolic network stoichiometry analysis of microbial growth and product formation.** *Biotechnology and Bioengineering* 1995, **48**:681-98.

144. Reed JL, Senger RS, Antoniewicz MR, Young JD: **Computational approaches in metabolic engineering.** *Journal of Biomedicine & Biotechnology* 2010, **2010**:207414.

145. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.** *Journal of Theoretical Biology* 2000, **203**:229-48.

146. Schuster S, Dandekar T, Fell D a: **Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering.** *Trends in Biotechnology* 1999, **17**:53-60.

147. Clarke BL: **Stoichiometric network analysis.** *Cell Biophysics* 1988, **12**:237-53.

148. Urbanczik R, Wagner C: **An improved algorithm for stoichiometric network analysis: theory and applications.** *Bioinformatics* 2005, **21**:1203-10.

149. Papin J a, Stelling J, Price ND, et al.: **Comparison of network-based pathway analysis methods.** *Trends in Biotechnology* 2004, **22**:400-5.

150. Llaneras F, Picó J: **Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators.** *Journal of Biomedicine & Biotechnology* 2010:753904.

151. Klamt S, Stelling J: **Combinatorial complexity of pathwat analysis in metabolic netwroks**. *Molecular Biology Reports* 2002, **29**:223-36.

152. Varma A, Boesch BW, Palsson BØ: **Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates.** *Applied and Environmental Microbiology* 1993, **59**:2465-73.

153. Varma A, Palsson BØ: **Metabolic Flux Balancing: Basic Concepts, Scientific and Pratical Use**. *Bio/Technology* 1994, **12**:994-8.

154. Bertsimas D, Tsitsiklis JN: *Introduction to Linear Optimization*. First Edit. Belmont, Massachusetts: Athena Scientific; 1997.

155. Ibarra RU, Edwards JS, Palsson BO: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth**. *Nature* 2002, **420**:20-23.

156. Famili I, Förster J, Nielsen J, Palsson BØ: **Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:13134-9.

157. Kauffman KJ, Prakash P, Edwards JS: **Advances in flux balance analysis**. *Current Opinion in Biotechnology* 2003, **14**:491-496.

158. Price ND, Reed JL, Palsson BØ: **Genome-scale models of microbial cells: evaluating the consequences of constraints.** *Nature Reviews Microbiology* 2004, **2**:886-97.

159. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli.** *Biophysical Journal* 2002, **83**:1331-40.

160. Covert MW, Palsson BØ: **Transcriptional regulation in constraints-based metabolic models of Escherichia coli.** *The Journal of Biological Chemistry* 2002, **277**:28058-64.

161. Beg QK, Vazquez A, Ernst J, et al.: **Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:12663-8.

162. Becker S a, Feist AM, Mo ML, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nature Protocols* 2007, **2**:727-38.

163. Schellenberger J, Que R, Fleming RMT, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.** *Nature Protocols* 2011, **6**:1290-307.

164. Burgard AP, Maranas CD: **Optimization-based framework for inferring and testing hypothesized metabolic objective functions.** *Biotechnology and Bioengineering* 2003, **82**:670-7.

165. Kumar VS, Maranas CD: **GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.** *PLoS Computational Biology* 2009, **5**:e1000308.

166. Gianchandani EP, Oberhardt M a, Burgard AP, Maranas CD, Papin J a: **Predicting biological system objectives de novo from internal state measurements.** *BMC Bioinformatics* 2008, **9**:43.

167. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.** *Molecular Systems Biology* 2007, **3**:119.

168. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:15112-15117.

169. Boyd SP, Vandenberghe L: *Convex Optimization*. Cambridge University Press; 2004.

170. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in Saccharomyces cerevisiae**. *Genome Research* 2005, **15**:1421-30.

171. Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:7695-700.

172. Blank LM, Kuepfer L, Sauer U: **Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast.** *Genome Biology* 2005, **6**:R49.

173. Stephanopoulos GN, Aristidou AA, Nielsen J: *Metabolic Engineering*. First Edit. San Diego, California: Academic Press; 1998.

174. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnology and Bioengineering* 2003, **84**:647-57.

175. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási a L: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-4.

176. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED: **Metabolic network structure determines key aspects of functionality and regulation.** *Nature* 2002, **420**:190-3.

177. Patil KR, Rocha I, Förster J, Nielsen J: **Evolutionary programming as a platform for in silico metabolic engineering.** *BMC Bioinformatics* 2005, **6**:308.

178. Pharkya P, Burgard AP, Maranas CD: **OptStrain: a computational framework for redesign of microbial production systems.** *Genome Research* 2004, **14**:2367-76.

179. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.** *BMC Systems Biology* 2010, **4**:53.

180. Rocha I, Maia P, Evangelista P, et al.: **OptFlux: an open-source software platform for in silico metabolic engineering.** *BMC Systems Biology* 2010, **4**:45.

181. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metabolic Engineering* 2003, **5**:264-276.

182. Tepper N, Shlomi T: **Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways.** *Bioinformatics* 2010, **26**:536-43.

183. Daran-Lapujade P, Rossell S, Gulik WM van, et al.: **The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:15753-8.

Chapter 3: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 1 – Metabolic Modeling for Strain Improvement

This chapter is based on the publication:

"Improved vanillin production in baker's yeast through in silico design". **Ana Rita Brochado**, Claudia Matos, Birger L Møller, Jørgen Hansen, Uffe H Mortensen, Kiran Raosaheb Patil. *Microbial Cell Factories* 9:84 (2010)

Abstract

Vanillin is one of the most widely used flavoring agents, originally obtained from cured seed pods of the vanilla orchid *Vanilla planifolia*. Currently vanillin is mostly produced *via* chemical synthesis. A *de novo* synthetic pathway for heterologous vanillin production from glucose has recently been implemented in baker's yeast, *Saccharomyces cerevisiae*. In this study we aimed at engineering this vanillin cell factory towards improved productivity and thereby at developing an attractive alternative to chemical synthesis.

Expression of a glycosyltransferase from *Arabidopsis thaliana* in the vanillin producing *S. cerevisiae* strain served to decrease product toxicity. An *in silico* metabolic engineering strategy of this vanillin glucoside producing strain was designed using a set of stoichiometric modeling tools applied to the yeast genome-scale metabolic network. Two targets (*PDC1* and *GDH1*) were selected for experimental verification resulting in four engineered strains. Three of the mutants showed up to 1.5 fold higher vanillin β-D-glucoside yield in batch mode, while continuous culture of the *Δpdc1* mutant showed a 2-fold productivity improvement. This mutant presented a 5-fold improvement in free vanillin production compared to the previous work on *de novo* vanillin biosynthesis in baker's yeast.

Use of constraints corresponding to different physiological states was found to greatly influence the target predictions given minimization of metabolic adjustment (MoMA) as biological objective function. *In vivo* verification of the targets, selected based on their predicted metabolic adjustment, successfully led to overproducing strains. Overall, we propose and demonstrate a framework for *in silico* design and target selection for improving microbial cell factories.

Background

Vanillin is a plant secondary metabolite and the main constituent of natural vanilla - one of the most important flavoring agents. The annual market for vanillin exceeds 16,000 tons, although only 0.25% of this originates from cured seed pods of the vanilla orchid, *Vanilla planifolia*. The remaining demand for vanillin is fulfilled by chemical synthesis from lignin or fossil hydrocarbons, in particular guaiacol [1]. Sustainable and environmental friendly routes have been proposed for obtaining vanillin through bioconversion of eugenol and ferulic acid by bacteria or fungi [2–4]. Alternatively, an attractive option was recently reported by *Hansen et al* (2009), who demonstrated *de novo* vanillin biosynthesis from glucose in baker's and fission yeasts [5]. The native metabolic precursor for this *de novo* pathway is 3-dehydroshikimate (3-DHS), an intermediate of the shikimate pathway for aromatic amino acids biosynthesis. To convert 3-dehrydroshikimate into vanillin, four genes encoding the required four enzymatic activities were obtained from different organisms, *Podospora pausiceta*, *Nocardia sp.*, *Escherichia coli* and *Homo sapiens* (**Fig. 1**) [5]. Inspired by the fact that metabolic engineering has been successfully applied to improve the yield of e.g. sesquiterpenes [6], ethanol [7, 8], artemisinic acid [9] and succinic acid [10] production in *Saccharomyces cerevisiae,* we hypothesized that vanillin production could also be improved by implementing a metabolic engineering strategy [11]. An immense collection of systems biology tools, in addition to well-established technologies for genetic manipulation, renders *S. cerevisiae* a very amenable organism for metabolic engineering [12–14].

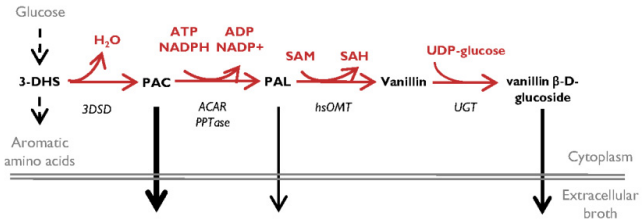


Figure 3.1: Schematic representation of the de novo VG biosynthetic pathway in *S. cerevisiae* (as designed by Hansen et al [5])**.** Metabolites are shown in black, enzymes are shown in black and in italic, cofactors and additional precursors are shown in red. Reactions catalyzed by heterologously introduced enzymes are shown in red. Reactions converting glucose to aromatic amino acids are represented by dashed black arrows. Metabolite secretion is represented by solid black arrows where relative thickness corresponds to relative extracellular accumulation. 3-DSH stands for 3-dedhydroshikimate, PAC stands for protocathechuic acid, PAL stands for protocatechuic aldehyde, SAM stands for S-adenosylmethionine. 3DSD stands for 3-dedhydroshikimate dehydratase, ACAR stands for aryl carboxylic acid reductase, PPTase stands for phosphopantetheinyl transferase, hsOMT stands for O-methyltransferase, and UGT stands for UDP-glycosyltransferase. Adapted from Hansen et al. [5].

The aim of this study was to design and construct an improved S. *cerevisiae* vanillin cell factory guided by genome-scale metabolic modeling. *In silico* metabolic engineering algorithms were used to identify target reactions in the metabolic network, knockout of which would lead to improved vanillin production. A set of knockouts that maximizes the flux towards a desired metabolite must be searched for, while the overall flux distribution is determined by the cellular objective function (e.g. maximization of biomass yield). This problem was formulated by Burgard *et al.* (2003) as a bi-level optimization algorithm termed OptKnock [15]. OptKnock can be applied in case of a linear design objective, such as maximizing flux towards a desired metabolite [15]. Optimization of non-linear objective functions, such as productivity, is also of great interest for a variety of metabolic engineering problems. OptGene, an extension of OptKnock, allows maximization of non-linear objective functions, while at the same time accounting for non-linear constraints on the metabolic network [16].

The most widely used approach for calculating flux distribution is flux balance analysis (FBA), where a given flux (or a linear combination of chosen fluxes) is used as the objective function [17]. For microorganisms, biomass maximization is generally accepted as a cellular objective function while simulating flux distributions [18, 19]. FBA has been successfully applied to predict gene essentiality [20, 21], end point of adaptive evolution experiments [22] and optimal metabolic states under given environmental conditions [19]. However, a mutant strain that is not subjected to evolutionary pressure might have a disturbed metabolic network and the principle of optimality for growth may not be prevailing. To address this question, the algorithm Minimization of Metabolic Adjustment (MoMA) has been suggested by Segrè *et al.* (2002), where it is advocated that the cellular objective for a mutant strain is to minimize its metabolic distance from the wild type flux distribution [23]. Mathematically, metabolic adjustment of a mutant is defined as the Euclidean distance between the reference and the mutant flux distribution vectors. In MoMA approach, it is crucial to have a physiologically meaningful wild type flux distribution, as it will strongly influence the predicted phenotype [23]. Within this study, the *S. cerevisiae* genome-scale stoichiometric model iFF708 [21] was used to identify and select target reactions for maximizing vanillin production by using OptGene [16]. MoMA [23] was used as the biological objective function with wild type flux distributions spanning three major modes of yeast metabolic physiology. The model-based metabolic engineering strategy was tested experimentally by strain construction and characterization. These research efforts resulted in three mutant yeast strains with significantly increased vanillin production.

Results and Discussion

Vanillin β-D-glucoside production in S. cerevisiae

Vanillin is toxic to many living organisms. In case of *S. cerevisiae*, growth defect is significant with concentrations as low as 0.5 g/l [5]. Tackling the problem of vanillin toxicity is therefore an important pre-requisite for building an economically viable vanillin cell factory. An elegant solution is glycosylation of vanillin, which is observed in the natural producer *Vanilla planifolia* [25, 26]. This strategy was successfully implemented by Hansen *et al.* (2009) in *Schizosaccharomyces pombe* [5]. The glycosylation step implies reducing the maximum theoretical yield from 486 mg_{van}/g_{glc} changes to 293 mg_{Van}/g_{glc}. Nevertheless, given the toxicity and low solubility of vanillin, reaching high titers is not a favourable option. Extracellular concentration of vanillin β-D-glucoside (VG) up to 25 g/l has been shown not to affect growth, therefore it is more suitable for commercial production.

In *S. pombe,* heterologous expression of *UGT72E2*, a gene encoding a plant family 1 glycosyltransferase from *Arabidopsis thaliana,* resulted in 80% conversion of vanillin to vanillin β-D-glucoside [5]. Within this study, *UGT72E2* was expressed in the vanillin producing *S. cerevisiae* strain VAN286, obtained from Hansen *et al*. [5]. The resulting strain, VG0, produced significantly more VG than vanillin (> 100 mg/l versus < 7 mg/l), indicating efficient conversion of vanillin into VG.

In silico design

The yeast genome-scale metabolic model iFF708 [21] was used to provide guidance for a metabolic engineering strategy for improving the vanillin β-D-glucoside cell factory. The existence of large number of alternative flux routes (pathways) within the metabolic network requires the use of experimental constraints in order to obtain physiologically meaningful flux distributions. This is even more so in the case of MoMA, where predictions of mutant flux distributions will be highly dependent on the solution provided for the wild type (or reference) flux distribution [23]. Different metabolic states of *S. cerevisiae* are characterized by different nutrient uptake rates, different metabolite production rates and different biomass yields. Therefore, it is to be expected that different constraints (corresponding to different metabolic states) may lead to different suggestions for the metabolic engineering targets.

Reference metabolic states

S. cerevisiae uses two major pathways for generating energy necessary for growth: glycolysis and oxidative phosphorylation (respiration). In the presence of oxygen and high glucose concentration, typically observed in batch conditions, glycolysis is highly active and ethanol formation is observed, also called overflow metabolism. The co-existence of fermentation and respiration leads to significantly lower biomass yield than purely respiratory metabolism, due to carbon channeling towards ethanol [26, 27]. However, if glucose concentration (or glucose uptake rate) is bellow a critical threshold, the flux through glycolysis is reduced, most of the energy is obtained from respiration and higher biomass yield on glucose is observed [26, 28, 29].

Different metabolic scenarios spanning different yeast life-styles were used for obtaining the reference flux distribution for MoMA simulations. Reference 1, representing fully respiratory metabolism, is characterized by no ethanol formation and low glucose uptake rate. Reference 2 was simulated for respiro-fermentative metabolism, characterized by high glucose uptake rates, alcoholic fermentation and less active respiration. Reference 3 corresponds to VG0 flux distribution predicted by using FBA and physiological constraints obtained in this study from chemostat cultivations at a dilution rate of 0.1 h^{-1}. This reference was introduced in order to have a biological meaningful flux distribution of the reference strain concerning the production of vanillin related compounds, as the vanillin biosynthetic pathway is predicted to have no flux is while maximizing growth, unless constrained.

Assessment of in silico predictions

Simulations were performed using OptGene for maximization of Biomass Product Coupled Yield (BPCY), a proxy for productivity [16], by deletion of up to six reactions. Improved vanillin β-D-glucoside production was not predicted when using maximization of biomass production as biological objective, while the use of MoMA [23] suggested interesting targets even after a single reaction deletion. Among a variety of possible target reactions, biomass and vanillin β-D-glucoside yield are generally related with an inverted trend; when predicted biomass yield is high, predicted product yield is modest and vice-versa (**Fig. 3.2A**).

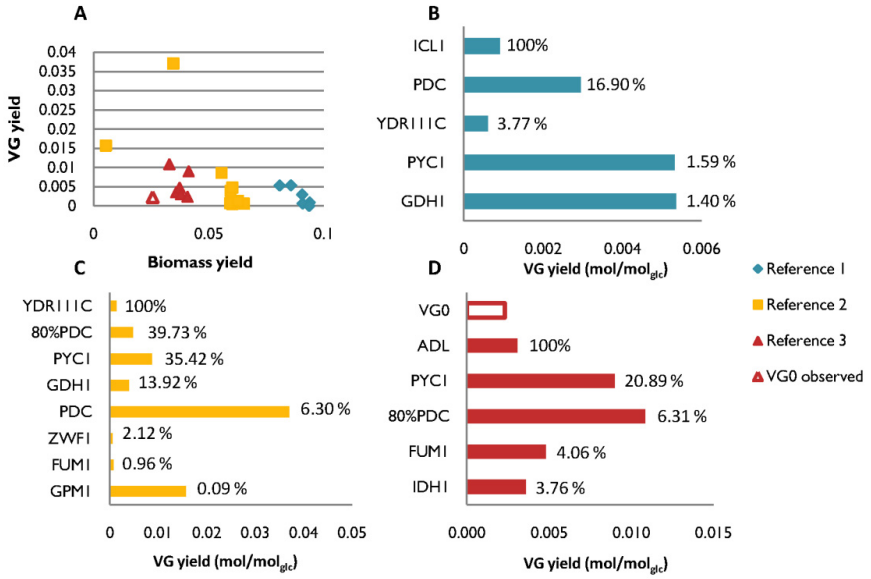


Figure 3.2: Comparison of targets predicted by OptGene for improved VG productivity. A – Biomass versus VG yield is represented for each knockout mutant phenotype obtained after OptGene simulation using three different reference flux distributions for MoMA. Experimental yields observed for VG0 are represented by the red empty triangle and bar. **B/C/D** – The predicted VG yield (mol/mol$_{glc}$) obtained for each knockout mutant after OptGene simulation using Reference 1/2/3 is given by the length of the coloured bars. For each reference flux distribution, the R^3 score was estimated for each of the mutants was calculated and normalized to the mutant presenting the highest value: MA_{mut}/MA_{max}*100. 100% represents the mutant with highest R^3 score for a given flux distribution. Candidate 80%PDC is not a knockout *in silico* mutant, rather its PDC reaction is constrained to 80% of the upper bound.

Target selection for experimental validation must strike a good balance between improved vanillin β-D-glucoside production and a reasonable prediction for biomass yield. Herein, metabolic adjustment (as defined by Segrè *et al.*, 2002 [23]) was used as an additional factor to rank each of the candidate target sets. Our hypothesis is that smaller metabolic adjustments are more likely to be achieved *in vivo* than large adjustments. This was taken into account by introducing a new metric for ranking *in silico* predicted mutants, *viz.*, the Reward-Risk-Ratio (R^3), defined as the ratio between BPCY (reward) and metabolic adjustment (risk).

$$R^3 = \frac{BPCY}{metabolic\ adjustment}$$

Depending on the reference flux distribution used for MoMA, different reactions were suggested for deletion (**Fig. 3.2 B-D**). A number of reactions, for example GDH1, were identified while using more than one reference flux distributions. Yet they are differently ranked using R^3, demonstrating that the best targets in one physiological scenario are not necessarily the best targets in another. The results of the prediction analysis suggest that VG production is favoured

at those physiological conditions leading to respiro-fermentative metabolism as compared to exclusively respiratory metabolism. Diverting ethanol flux towards formation of VG is predicted to compel less metabolic adjustment than diverting flux from biomass constituents. Indeed, several candidates found using the references 2 and 3 are predicted to have lower ethanol formation than the reference. Furthermore, the biomass yield predicted for all the suggested mutants using reference 3 is slightly higher than for the reference, again at the expense of ethanol formation.

Target selection

The targets for *in vivo* implementation were selected based on their R^3 score under different simulation conditions in combination with manual evaluation of their suitability for gene deletion. Literature and database search included the possible existence of iso-enzymes, experimentally observed single gene deletion phenotypes and the assessment of the importance of regulatory links to other processes [30]. Based on this analysis, two gene candidates, *PDC1* (Pyruvate decarboxylase) and *GDH1* (Glutamate dehydrogenase), were selected for strain construction and characterization.

The Gdh1 catalyzed reaction was predicted for deletion when using respiratory or respiro-fermentative metabolism, therefore it was selected for further work, despite its prediction for large metabolic adjustment. *GDH1* encodes an NADPH-dependent glutamate dehydrogenase involved in ammonium metabolism through glutamate biosynthesis, which is reported to provide 85% of the cellular nitrogen sources [31]. Ammonium metabolism has been extensively studied in *S. cerevisiae* and in particular, deletion of *GDH1* was previously used as a metabolic engineering strategy for improving ethanol and sesquiterpenes production [6, 8, 32]. Even though it was not a priority target considering our R^3 score due to high metabolic adjustment, prior work reported that deletion of *GDH1* leads to increased NADPH pool, which may favour the conversion of PAC to PAL by ACAR.

Pyruvate is a key metabolite in *S. cerevisiae* metabolism and the branch point between respiratory and fermentative metabolism. Pyruvate decarboxylases (PDCs) have a crucial significance for fermentation, since they catalyze the conversion from pyruvate to acetaldehyde, an intermediate towards ethanol formation [33]. The *S. cerevisiae* genome harbours three PDC structural genes (*PDC1, 5* and *6*), one regulatory gene (*PDC2*) and two other genes with potential contribution towards PDC activity (*PDC3* and *4*) [33, 34]. Complete suppression of pyruvate decarboxylase activity (*pdc1Δ, pdc5Δ & pdc6Δ*) creates a mutant which is unable to grow on glucose as sole carbon source [34, 35]. By *in silico* analysis, PDC was found as a target to increase

the formation of VG considering both respiratory and respiro-fermentative reference flux distributions for MoMA, but not when VG0 data was used to constrain the network. Furthermore, complete absence of pyruvate decarboxylase activity under highly fermentative conditions (as reference 3) was predicted to result in zero growth, as lack of PDC activity would require a very large metabolic adjustment. This *in silico* observation is in good agreement with the experimental observation that a PDC negative mutants are not able to grow while utilizing glucose [34, 35]. Remains to be clarified to what extent partial reduction of PDC activity, *e.g.* by deletion of one of the structural genes would influence VG production. *In silico* analysis was carried out by constraining PDC flux to 80% of the flux observed in the reference strain. This simulation predicted positive growth, as well as higher VG production in comparison to other targets such as *GDH1* (**Fig. 3.2C and D**). Among the three PDC structural genes, *PDC1* was selected for deletion *in vivo* as there is experimental evidence that its removal results in ~30% reduction of total pyruvate decarboxylase activity [35].

Strain construction and characterization

Following the selection of the two target genes, two single gene deletion mutants, *gdh1*Δ and *pdc1Δ*, were constructed from strain VG0, resulting in strains VG1 and VG2, respectively. To test whether simultaneous deletion of *PDC1* and *GDH1* would have a positive synergistic effect on VG accumulation, mutant VG3, with both deletions, was also obtained. The strains were initially characterized in batch cultures in well-controlled bioreactors, using minimal medium.

Table 3.1: Physiological parameters for the reference and metabolically engineered strains in batch cultivation.

Strains	Engineered Genotype	μ_{max} [a]	$Y_{S\,X}$ [b]	$Y_{S\,EtOh}$ [c]	$Y_{S\,gly}$ [d]
VG0		0.14	0.10	0.23	0.05
VG1	*gdh1Δ*	0.10	0.07	0.25	0.03
VG2	*pdc1Δ*	0.20	0.14	0.23	0.07
VG3	*pdc1 Δgdh1Δ*	0.11	0.10	0.27	0.05
VG4	*pdc1 Δgdh1Δ ↑GDH2*	0.17	0.17	0.25	0.07

[a]Maximum specific growth rate, h^{-1}.
[b]Overall yield of biomass on substrate, $g_{DW}.g_{glc}^{-1}$
[c]Overall yield of ethanol on substrate, $g_{eth}.g_{glc}^{-1}$
[d]Overall yield of glycerol on substrate, $g_{Gly}.g_{glc}^{-1}$
The biomass overall yield in glucose ($Y_{S\,X}$) was calculated based on all the biomass obtained after glucose and ethanol consumption. Ethanol and glycerol yields were calculated based on production only in the glucose consumption phase.

Mutant VG2 (*pdc1Δ*) showed an overall increased fitness compared to the reference strain (VG0) as documented by its 43% higher maximum specific growth rate and 40% higher biomass yield on glucose (Y_{SX}, **Table 3.1**). On the other hand, mutant VG1 (*gdh1*Δ) showed reduced strain

fitness, illustrated by reduced μ_{max} and Y_{SX}. The adverse fitness effects of strain VG1 were partially relieved by deletion of *PDC1* (strain VG3), as verified by slightly improved μ_{max} and Y_{SX} (**Table 3.1**).

In agreement with previous reports, our results confirm the deleterious effect of *GDH1* deletion, most likely due to a reduced nitrogen assimilation rate [36]. In the absence of *GDH1*, the GS-GOGAT system (co-action of two enzymes, a glutamate synthase, *GLT1,* and a glutamine synthase, *GLN1*) and the glutamate dehydrogenase, coded by *GDH2*, are responsible for ammonia assimilation [37–39]. Both alternatives use NADH instead of NADPH, thus explaining the high metabolic adjustment predicted for the *GDH1 in silico* mutants (**Fig. 3.2B** and **C**). In comparison with Gdh1, lower activity has been reported for both of the alternative systems. Consequently, overexpression of the enzymes involved in the alternative pathways is a required step in order to recover the cellular fitness [40]. We proceeded experimentally with *GDH2* overexpression as the use of the GS-GOGAT system has the disadvantage of using an important cellular resource – ATP. The resulting strain VG4 (*pdc1 Δgdh1Δ* ↑*GDH2*) showed a significantly improved cellular fitness compared to VG3 with similar μ_{max} and Y_{SX} to those observed for VG2 and better than those obtained with VG0. At these experimental conditions, none of the introduced mutations seem to affect ethanol production, except *GDH1* deletion, which induces a slight increase in the substrate yield of ethanol (**Table 3.1**).

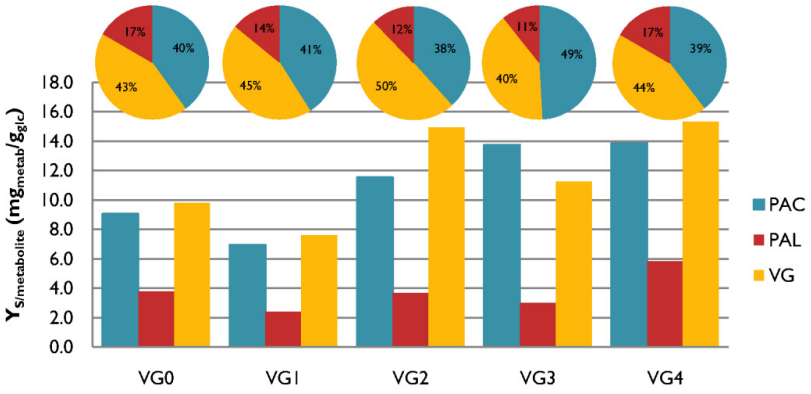


Figure 3.3: Vanillin β-D-glucoside yield observed for the reference strain (VG0) and metabolically engineered mutants (VG1-4) in batch cultivations. Substrate overall yield for vanillin β-D-glucoside ($Y_{S\ VG}$, mg_{VG}/g_{glc}), protocatechuic acid ($Y_{S\ PAC}$, mg_{PAC}/g_{glc}) and protocatechuic aldehyde ($Y_{S\ PAL}$, mg_{PAL}/g_{glc}) obtained for the reference and engineered strains in batch culture. Pie charts are presented to illustrate relative distribution of PAC, PAL and VG for each strain.

Strain VG1 showed the lowest total yield of vanillin related compounds, most likely as consequence of the general decreased fitness observed of this strain (**Fig. 3.3**). The other three engineered strains displayed better performance than VG0 concerning VG production (**Fig. 3.3**). Single deletion of *PDC1* (strain VG2) increased VG yield by 52%, as well as the total yield of all vanillin related compounds by 30%. Despite the adverse effect of *GDH1* deletion, double deletion of *PDC1* and *GDH1* (strain VG3) yielded 15% increase in VG production. Strain VG4 has over 55% increased VG yield. Accumulation of several intermediates from the vanillin pathway was observed for all mutants, particularly protocatechuic acid (PAC) and protocatechuic aldehyde (PAL), which accounted for more than 50% of the total products formed (**Fig. 3.3**).

Respiratory vs fermentative metabolism for VG production

The extent to which VG production is affected by the presence or absence of alcoholic fermentation is still unknown. In chemostat cultivation, glucose concentration is kept at low levels and, below the critical dilution rate, glucose uptake rate is low enough to ensure exclusive respiratory metabolism, thus no ethanol is formed [26, 28, 29].

VG0, the reference strain, and the best performing mutants in batch cultivations (VG2 and VG4) were selected for further characterization in glucose-limited chemostat cultures at a dilution rate of 0.1 h^{-1}. VG2 showed increased fitness as demonstrated by a higher biomass yield (Y_{SX}) accompanied by a slightly decreased glycerol yield, as compared to VG0 (**Table 3.2**). Under these conditions, VG2 shows 40% higher VG yield on glucose than VG0. In contrast, VG4 displayed a considerable decrease in Y_{SX}, while ethanol and glycerol production were significantly increased (**Table 3.2**). The VG production by this strain remained very similar to that of the reference strain.

The lower glucose uptake observed for VG2 decreases the overflow metabolism, and subsequently smaller rates of ethanol, glycerol and acetate formation by this strain (**Table 3.2**). The opposite trend is verified for mutant VG4, indicating higher overflow metabolism. Accordingly, a significantly increased production of ethanol (1.5-fold) and glycerol (4-fold) was observed (**Table 3.2**).

Table 3.2: Physiological parameters for the reference and metabolically engineered strains in chemostat cultivation at dilution rate 0.1 h^{-1}.

Strain	**VG0**	**VG2**	**VG4**
Engineered Genotype		*pdc1Δ*	*pdc1 Δgdh1Δ ↑GDH2*
$Y_{S\,X}$ ($g_{DW}.g_{glc}^{-1}$)	0.151 ± 0.008	0.159 ± 0.003	0.115 ± 0.005
$Y_{S\,eth}$ ($g_{eth}.g_{glc}^{-1}$)	0.283 ± 0.005	0.290 ± 0.008	0.32 ± 0.01
$Y_{S\,gly}$ ($mg_{gly}.g_{glc}^{-1}$)	3.4 ± 0.3	2.1 ± 0.4	11 ± 9
$Y_{S\,acet}$ ($mg_{acet}.g_{glc}^{-1}$)	7.3 ± 0.3	7.6 ± 0.3	4 ± 2
$Y_{S\,PAC}$ ($mg_{PAC}.g_{glc}^{-1}$)	27 ± 2	23.5 ± 0.3	15.5 ± 0.3
$Y_{S\,PAL}$ ($mg_{PAL}.g_{glc}^{-1}$)	7 ± 2	9 ± 2	7.2 ± 0.2
$Y_{S\,VG}$ ($mg_{VG}.g_{glc}^{-1}$)	4 ± 1	5.6 ± 1	4.16 ± 0.07
r_s($mmol_{glc}.g_{DW}^{-1}.h^{-1}$)	3.9 ± 0.2	3.5 ± 0.3	4.8 ± 0.2
r_{eth}($mmol_{eth}.g_{DW}^{-1}.h^{-1}$)	4.3 ± 0.3	3.9 ± 0.3	6.1 ± 0.4
r_{gly}($mmol_{gly}.g_{DW}^{-1}.h^{-1}$)	0.026 ± 0.003	0.014 ± 0.004	0.10 ± 0.08
r_{acet}($mmol_{acet}.g_{DW}^{-1}.h^{-1}$)	0.086 ± 0.006	0.079 ± 0.006	0.05 ± 0.003
r_{PAC}($mmnol_{PAC}.g_{DW}^{-1}.h^{-1}$)	0.12 ± 0.01	0.10 ± 0.01	0.087 ± 0.03
r_{PAL}($mmol_{PAL}.g_{DW}^{-1}.h^{-1}$)	0.036 ± 0.008	0.040 ± 0.007	0.045 ± 0.003
r_{VG}($mmol_{VG}.g_{DW}^{-1}.h^{-1}$)	0.009 ± 0.003	0.011 ± 0.002	0.012 ± 0.001

Substrate yields for biomass (X), ethanol (eth), glycerol (Gly), acetate (Ace), protocathechuic acid (PAC), protocathechuic aldehyde (PAL) and vanillin β-D-glucoside (VG) on glucose (glc) are represented by $Y_{S\,X\,(or\ metabolite)}$. Specific glucose uptake rate is represented by r_S. Specific production rate for ethanol, glycerol, acetate, PAC, PAL and VG are represented by $r_{metabolite}$. Triplicates of all chemostats were performed.

VG2 exhibits a higher conversion of PAC into the products of the later steps in the vanillin pathway (**Fig. 3.4A**). Strain VG4, despite the severe impact of the overflow metabolism on its biomass yield on glucose, displays the highest conversion of PAC into downstream metabolites, particularly PAL (**Fig. 3.4A**). As hypothesized during target selection, this mutant is likely to have higher availability of NADPH, due to the engineered reduced demand for this cofactor for ammonium metabolism.

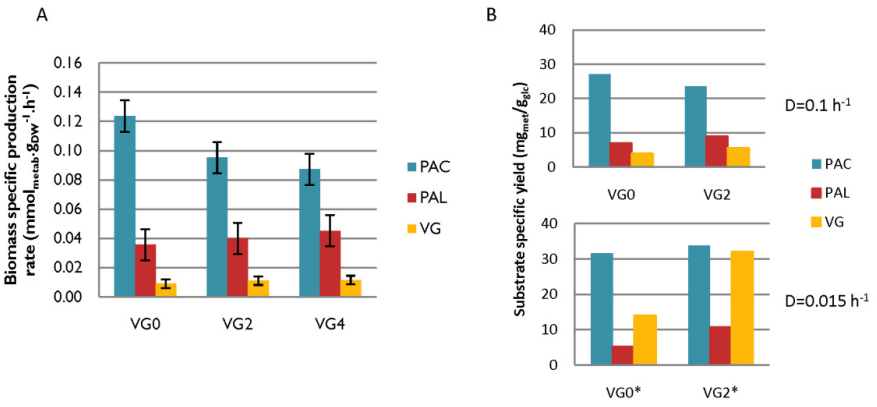


Figure 3.4: Vanillin β-D-glucoside yield observed for the reference strain (VG0) and metabolically engineered mutants (VG1-4) in continuous cultivations. A - Biomass specific production rate ($mg_{metab}.g_{dw}^{-1}.h^{-1}$) for protocatechuic acid (PAC), protocatechuic aldehyde (PAL) and VG in glucose limited chemostat cultivation at dilution rate of 0.1h^{-1}. **B** - Substrate specific yield ($Y_{S\ Metab}$, , mg_{metab}/g_{glc}) for PAC, PAL and VG for strains VG0 and VG2 in glucose limited chemostat cultivation at different dilution rates – 0.1 h^{-1}(Top) and 0.015 h^{-1}(*, Bottom).

The predominantly respiro-fermentative metabolism observed for all engineered strains implies that their critical dilution rate, indicative of respiratory capacity, is very low. By gradually decreasing the dilution rate in a glucose-limited chemostat, we verified that the critical dilution rate for both VG2 and VG0 was below 0.015 h^{-1}. Such a low respiratory capacity could be the result of combined effects of product/by-products toxicity and the background of the strain. By employing this experimental setup, ethanol production rate was as low as 0.130 $mmol_{eth}/(g_{gw}.h)$ for VG0 and 0.065 $mmol_{eth}/(g_{gw}.h)$ for VG2, while the rate of glucose uptake was 0.37 and 0.30 $mmol_{glc}g_{gw}^{-1}.h^{-1}$, respectively. The VG concentration in the fermentation broth of VG2 was 500 mg/l, (32 mg_{VG}/g_{glc}), a two-fold increase compared to that of VG0, which produced 250 mg/l (15 mg_{VG}/g_{glc}) (**Fig. 3.4B**). These results confirm that VG2 has better respiratory capacity than the reference strain VG0 and lowering the overflow metabolism results in higher VG yield. For both strains, the observed conversion of PAL into VG is more efficient at low dilution rates, suggesting that low overflow metabolism is linked to increased precursor and/or cofactors supply, enabling higher VG productivity.

Analysis of the experimental results

In an attempt to better understand the metabolic flux changes at the whole network level, and therefore the basis for the observed improved vanillin β-D-glucoside production, the flux phenotypes of VG0, VG2 and VG4 were simulated by using FBA [17]. The experimental results

obtained from the chemostat cultures, *i.e.* glucose uptake rate and biomass, ethanol, acetate, glycerol and CO_2 production rates, were added as constraints to the metabolic model. The bounds for reaction GDH1 were set to zero. As deleting *PDC1* does not imply that pyruvate decarboxylase activity will be zero, the choice of constraints for the *PDC1* deletion is not straightforward. To circumvent this issue, and to account for different glucose uptake rates of the mutants, the upper bound for PDC flux was identified in each condition. Subsequently, the phenotypes were simulated with upper bound for PDC constrained to 80% of the previously found upper bound.

Flux variability analysis

The flux distributions obtained with FBA are guaranteed to be optimal, but not necessarily unique due to the existence of a large number of alternative routes. This renders the transformation of the experimentally determined levels of the products obtained into intracellular fluxes, a difficult task. Nevertheless, stoichiometric simulations provide an estimate of the possible range of flux values for every reaction in the network. Fluxes which are unique will have the same maximum and minimum possible values. The flux ranges of all reactions of VG2 and VG4 were calculated and compared with those of VG0, resulting in different categories as illustrated in **Fig. 5**. The first category consists of those reactions for which flux is infeasible at steady-state, *i.e.* blocked reactions. Among the remaining reactions (~570), only 50 have unique flux values for all the strains. Almost all of these reactions belong to category *a*, *i.e.* with no change between the reference strain and the tested mutants. As chemostat experiments at the same dilution rate were used for all strains, the flux of reactions directly coupled to biomass biosynthesis also fall in this category. Examples include reactions from lipid, nucleotide and amino acid metabolism.

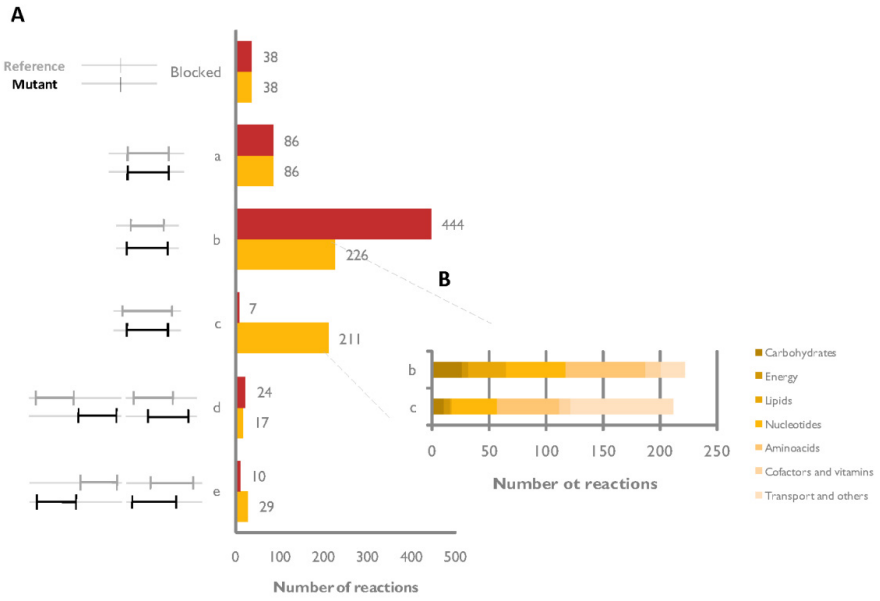


Figure 3.5: Flux variability analysis. Reactions were classified based on the comparison of their flux variability range between the reference VG0 and the mutants VG2 and VG4. **A -** The scheme on the left-hand side illustrates the flux variability ranges defining the six different categories (Blocked and *a* to *e*). Flux variability range for the reference strain (VG0) is represented in gray, and the mutant in black. The distribution of VG2 and VG4 reactions among different categories are presented in the bar chart, yellow and red, respectively. **B -** The reactions from mutant VG2 belonging to categories *b* and *c* are further classified accordingly to their metabolic function.

Categories *b* and *c* contain the majority of the reactions (~70%) for VG2 and VG4, yet the distribution among the two categories is not alike for the two mutants. For the strain VG4, most of the reactions fall in category *b*, a larger variability range than observed for the reference strain, as expected due to its higher glucose uptake rate. In the case of VG2, there is an even distribution of reactions between the categories *b* and c (reactions with smaller variability range than the reference). VG2 exhibits a decreased glucose uptake rate as compared to the reference, therefore reactions from glycolysis belong to category c. However, as its biomass yield is similar to the reference, equivalent quantity of energy must be available. In strain VG2, several reactions from the tricarboxylic acid (TCA) cycle and gluconeogenesis fall in category b, reflecting its improved respiratory capacity. A more active respiration in this strain is further confirmed by the *in silico* predicted increased oxygen uptake rate. Even though this reaction is found in category *c* (with decreased flux range), its lower bound is higher in VG2 than in VG0.

The remaining reactions present flux variability ranges with partial or no overlap between the reference strain and the mutants. These reactions were grouped in category *d*, where the

mutant flux upper bound is higher than the reference flux upper bound, and category *e* where the mutant flux lower bound is lower than the reference flux lower bound. Together, categories *d* and *e* comprise the reactions with the clearest differences between the mutants and the reference. These include reactions related to product formation (e.g. biosynthesis of S-adenosylmethionine – the methyl-group donor in the vanillin pathway) and glucose uptake (which was experimentally determined), as well as reactions from the ammonia metabolism.

Metabolite-centric analysis

Despite the increased VG production, strain VG2 exhibits a decreased flux through the aromatic amino acid biosynthesis pathway from which the vanillin precursor is derived. Likewise, this strain shows a reduced total flux through the VG pathway. The same trend was found for the VG4 strain, implying that the metabolic network is being adjusted for increased PAL and VG production at the expense of a reduction of the total carbon flow into aromatic amino acids until the VG biosynthesis branch. In fact, the reaction after which the production is increased is the conversion from PAC to PAL. This reaction uses NADPH and ATP, two of the most highly connected metabolites and cofactors that are competed for by growth requirement.

To systematically explore the usage of cofactors and other metabolites in the engineered strains, the turnover of these metabolites can be calculated by summing all the fluxes which are producing (or consuming) them [41]. As the unique values of all the fluxes are unknown due to alternative optima inherent to FBA of metabolic networks, the minimum metabolite turnover was calculated by solving a linear programming (LP) problem for the *minimization of metabolite turnover* (see **Methods** for details). This LP formulation guarantees to find the minimum turnover of a given metabolite that ensures the observed phenotype. The direction of optimization, *i.e.* minimization, not only avoids the unbound cyclic fluxes around the metabolite under question, but also confers with the biological hypothesis of minimal resource allocation by the cell in terms of enzyme expression. Minimum metabolite turnover denotes how much flux needs to pass through a given metabolite, although the distribution of this flux among possible reactions may not be unique in all cases. Nevertheless, the turnover calculated in this way provides a lower bound on the flux through a metabolite that can be used in complementation with flux variability analysis.

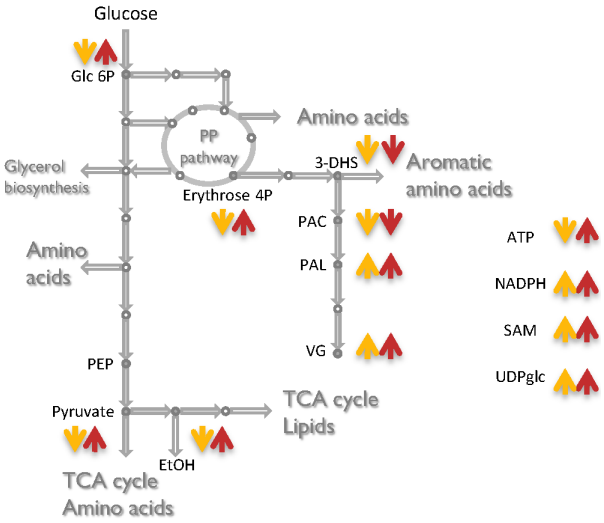


Figure 3.6: Minimum turnover of selected metabolites from the central carbon metabolism and from the VG biosynthetic pathway (including cofactors). Metabolites from the central carbon metabolism: glucose-6-phosphate, erythrose-4-phosphate, pyruvate and ethanol; Metabolites from the VG biosynthetic pathway (ATP, NADPH, SAM and UDP-glucose). Metabolites for which minimum turnover was calculated are represented by filled circles, metabolites for which no minimum turnover was calculated are represented by open rings. Reactions are represented as arrows. Qualitative variation of the minimum turnovers relatively to the reference (VG0) is shown by the arrows next to each metabolite; yellow corresponds to VG2 while red corresponds to VG4.

Besides NADPH and ATP, the minimum turnover of some other relevant metabolites was also calculated (**Fig. 3.6**). The minimum turnover for PAC is lower in the strains VG2 and VG4 than for VG0; while for PAL and VG the opposite trend is verified, in agreement with the flux variability analysis. An increase in the glucose uptake rate will result in an increase in glycolysis and pentose phosphate pathway, which is reflected in the increased minimum turnover of pyruvate and erythrose-4-phosphate in the VG4 strain. On the other hand, the VG2 strain exhibits a decreased glucose uptake rate and consequently less flux through glycolysis and pentose phosphate pathway. The same trends apply to ATP, implying that most ATP available in strain VG0 is being produced from glycolysis. NADPH, S-adenosylmethionine and UDP-glucose minimum turnovers are increased for both simulated phenotypes, reflecting the increased flux through ACAR, hsOMT and UGT leading to PAL and VG. The analysis above provides insight into the intracellular flux changes and pinpoints metabolites that play a role in the engineered strains.

Conclusions and Future Perspectives

The framework presented in this study comprises of i) *in silico* target prediction that accounts for the available physiological information; ii) systematic ranking of the targets based on the predicted metabolic adjustments; iii) *in vivo* verification through genetic engineering and fermentation; and iv) reaction/metabolite centric analysis of the experimental results. Our results demonstrate the applicability of *in silico* modeling tools for overproduction of a product from a multistep heterologous pathway in a eukaryotic system. Three vanillin β-D-glucoside overproducing strains were successfully designed and constructed during this study. Low dilution rate continuous cultivation of an overproducer resulted in notably high titer of vanillin β-D-glucoside – 500 mg/l, 5-fold higher than the 45 mg/l reported by Hansen and co-workers [5].

The yeast genome-scale model was used throughout this study, from the strain design to the analysis of the physiological data resulting from the fermentation studies of the constructed mutants. The *in silico* strategy used herein revealed the sensitivity of target predictions towards the choice biological objective function, as well as towards the flux distribution used for reference of simulation while using MoMA. To this end, it was crucial to use basic physiological knowledge for simulating different relevant yeast phenotypes. *In silico* analysis of fluxes variability and metabolites turnover of the overproducing strains suggested that the modification imposed on the metabolic network led to increased cofactor availability for vanillin β-D-glucoside production.

The observed discrepancy between model predictions and experimental observations concerning the production yield of vanillin β-D-glucoside (experimental ~3 *versus* predicted 10 $mmol_{VG}/mol_{glc}$ in batch conditions) is most likely linked to limited kinetic and regulatory information. The need for accounting for regulatory information is even more apparent when considering an isoenzyme as metabolic engineering target. Quantitative prediction of flux distributions following downregulation (similar to deletion of an isoenzyme) or overexpression of genes is still in its infancy. Advances towards integration of regulation in metabolic models require additional knowledge on the relationship between changes in gene expression and metabolic fluxes, as well as flux simulation tools that can integrate such information. Moreover regulation at both transcriptional and metabolic level is of particular relevance for vanillin production as the shikimate pathway for aromatic amino acids biosynthesis is tightly regulated in yeast [42]. As an example, Luttik *et al.* were able to increase the total flux through this

pathway by 4.5-fold in *S. cerevisiae* through alleviation of the DAHP synthase feed-back inhibition mechanisms [43].

Due to considerable accumulation of intermediates of the vanillin pathway observed in this study, identification and overexpression of an eventual rate limiting enzyme within the biosynthetic pathway may serve to enhance conversion of pathway intermediates into the final product. Further enhancing the respiratory capacity of the vanillin producing strains is also expected to increase vanillin β-D-glucoside production, as illustrated within this study.

Methods

Model simulations

Five new reactions were introduced in the *Saccharomyces cerevisiae* stoichiometric model [21] to convert 3-dehydroshikimate, a natural intermediate in aromatic amino acids biosynthesis, into vanillin β-D-glucoside (VG). Furthermore all the intermediates in the pathway were allowed to be secreted, based on the experimental evidence. FBA simulations were performed using linear programming library GLPK (ftp://ftp.gnu.org/gnu/glpk/), while MoMA simulations were performed using quadratic programming library OOQP [44].

Strategy design to improve VG production in *S. cerevisiae* was performed by using OptGene with Biomass-Product Coupled Yield (BPCY) as design objective function [16].

Minimization of metabolite turnover

Metabolite turnover or flux-sum is the sum of all incoming or outgoing fluxes around a particular metabolite under pseudo-steady state conditions [42,46]. Let Φ_i denote metabolite turnover of metabolite i and its mathematical definition is given by $\Phi_i = \frac{1}{2}\sum_k |S_{ik} v_k|$, where S_{ik} represents the stoichiometric coefficient of metabolite *i* in reaction *k* and v_k is the flux of reaction *k*. By calculating the sum of all absolute flux values through (in and out) a given metabolite eliminates further concern about reactions reversibility. Furthermore, given the steady state assumption, the metabolite turnover will be half of the calculated sum.

Due to existence of alternative optimal FBA solutions within genome-scale models, *minimum of metabolite turnover* was calculated, given predetermined exchange fluxes (including growth), using a linear programming formulation as follows:

$$\min \Phi_i$$
$$s.t.$$
$$\sum_j S_{ij} v_j = 0$$
$$\alpha_j \le v_j \le \beta_j \text{ for } \alpha_j \text{ and } \beta_j \in \Re \text{ and } \alpha_j \le \beta_j \text{ (Flux capacity constraints, including uptake and secretion reactions)}$$

Plasmids and strains

The strain *Saccharomyces cerevisiae* VAN286 obtained by Hansen *et al.* was used as the background strain in this work. In order to produce vanillin, this strain must be transformed with a plasmid containing the gene *EntD* from *Escherichia coli*, coding for a PPTase. This enzyme is indispensable for post-translational activation by phosphopantetheinylation of ACAR in *S. cerevisiae* [5]. All the strains constructed during this study (**Table 3.3**) were transformed with the plasmid containing *EntD* prior to cultivation in order to produce VG.

Cloning UGT72E2 in S. cerevisiae VAN286

The integrative plasmid pARB021 containing the gene UGT72E2 from *Arabidopsis thaliana* coding for a UDP-glucosyltransferase was obtained by replacing the *URA3* marker by the *HIS3* marker in plasmid pJH665 with the restriction enzyme *XmaI*. Restriction enzymes and buffers from New England Biolabs were used and the conditions for restriction followed manufacture instructions. The *HIS3* gene was PCR amplified using the primers His3_Fw and His3_Rev (**Table ST3.1**), the plasmid pWJ1213 [47] was used as template DNA and amplification was achieved using Phusion™ Hot Start High-Fidelity DNA Polymerase (Finnzymes Oy, Espoo, Finland). GFX™ PCR DNA and Gel Band Purification Kit (GE Healthcare) were used for DNA purifications and ligation was performed with T4 DNA ligase (New England Biolabs). The plasmid was treated with Antarctic Phosphate (New England Biolabs) in order to avoid recirculation. The ligation mixture was deactivated and transformed into chemo-competent *DH5α E. coli* cells. Ampicillin resistance was used as *E. coli* selection marker and plasmid extraction was performed using a GenElute HP Plamsid Miniprep Kit (Sigma-Aldrich). The plasmid ARB021 was verified by restriction analysis and sequencing of the PCR amplified *HIS3* marker with the primers MarkSeq_Fw and MarkSeq_Rev from **Table ST3.1** (StarSEQ, Mainz, Germany). The correct plasmid was then restricted with *SphI* and transformed into the yeast *TPI1* promoter locus of VAN286, thus

creating the strain VG0, producing VG from glucose. High efficiency yeast transformation method was used to construct the yeast strains [48].

Table 3.3: Yeast strains and plasmids used in this study.

Yeast Strain	Relevant genotype	Reference
VAN286	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 PTPI1::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]	Hansen *et al.* 2009 [5]
VG0	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 PTPI1::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]::UGT72E2 [*HIS3*]	This study
VG1	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 PTPI1::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]::UGT72E2 [*HIS3*] *gdh1*	This study
VG2	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 PTPI1::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]::UGT72E2 [*HIS3*] *pdc1*	This study
VG3	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 PTPI1::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]::UGT72E2 [*HIS3*] *pdc1 gdh11*	This study
VG4	*MAT*a *his3*D1 *leu2*D0 *met15*D0 *ura3*D0 *adh6*::*LEU2 bgl1*::KanMX4 P_{TPI1}::3DSD [AurC]::HsOMT [NatMX]::ACAR [HphMX]::UGT72E2 [*HIS3*] *pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2*	This study

Plasmid	Gene content	Plasmid type	Selection marker	Reference
pJH589[a]	*EndD* (*E. coli*)	*CEN*-ARS (*S. cerevisiae*)	*URA3*	Hansen *et al.* [5]
pJH665[a]	*UGT72E2 (Arabidopsis thaliana)*	*CEN*-ARS (*S. cerevisiae*)	*URA3*	Hansen *et al.* [5]
pARB021	*UGT72E2 (Arabidopsis thaliana)*	Integration (*S. cerevisiae*)	*HIS3*	This work
pWJ1042[b]	Recyclable *URA3 (Kluyveromyces lactis)*	*CEN*-ARS (*S. cerevisiae*)		Reid *et al.* [46]
pPGK1-GDH2[c]	P_{PGK1}-*GDH2* (*S. cerevisiae*)	Integration (*S. cerevisiae*)	*KanMX3*	Nissen *et al*. [8]

Model guided strain construction

PDC1 and *GDH1* gene deletions, as well as *GDH2* overexpression were achieved by gene targeting through homologous recombination of bipartite PCR fragments, using *URA3* gene from *Kluyveromyces lactis* as a marker [49]. The marker was flanked by direct repeats that allowed restoring of uracil auxotrophy by plating the cells in agar medium containing 5-Fluoroorotic acid (5-FOA) after each genetic manipulation [50].

The primers used for amplifying the up and downstream regions flanking the *PDC1* and *GDH1* gene (approximately 500 bp each) are listed in **Table ST3.1**, as well as the primers used to amplify *K. lactis URA3* flanked by direct repeats from the plasmid pWJ1042 [46]. Strain VG1 was obtained by deleting the gene *GDH1* in the strain VG0. Strain VG2 was obtained by deleting the gene *PDC1* also in the strain VG0. Strain VG3 was obtained by deleting gene *GDH1* in the strain VG2. The deletions were verified by analytical PCR using the primers PDC1_Ver_FW and PDC1_Ver_REV for *PDC1* deletion, and GDH1_Ver_FW and GDH1_Ver_REV for *GDH1* deletion (**Table ST3.1**). Strain VG4 was obtained from VG3 by swapping the native *GDH2* promoter by the strong constitutive promoter of the gene *PGK1*, as previously reported by Nissen *et al*. [8]. A 500 bp fragment upstream the *GDH2* open reading frame (ORF) used for homologous recombination was obtained from VG3 genomic DNA with the primers GDH2(UP)_Fw and GDH2(UP)_Rev (**Table ST3.1**). The downstream fragment used for homologous recombination was amplified from the plasmid pPGK1-GDH2 [8] with the primers PGK1_GDH2(Dw)_Fw and PGK1_GDH2(Dw)_Rev listed in **Table ST3.1**. 1479 bp of the *PGK1* promoter region were used to substitute the *GDH2* original promoter, while the initial 500 bp of the *GDH2* OFR were used to ensure accurate targeting. The promoter swapping was verified by analytical PCR with the primers PGK1verif and Gdh2verif (**Table ST3.1**), amplifying 420 bp of the *PGK1* promoter to 1300 bp of *GDH2*.

Media Composition

A defined minimal medium as described by Verduyn *et al.* (1992) with 20 g/l glucose as sole carbon source was used for cell cultivation [51]. The medium composition used for batch and continuous cultivation in well controlled bioreactors was as follows: 5.0 g/l $(NH_4)_2SO_4$, 3.0 g/l KH_2PO_4, 0.5 g/l Mg_2SO_4, 2.0 ml/l trace metal solution, 1.0 ml/l vitamins solution, 0.05 ml/l antifoam 204 (Sigma-Aldrich A-8311) and 80 mg/l L-methionine. Trace metal solution contained 3 g/L $FeSO_4.7H_2O$, 4.5 g/L $ZnSO_4.7H_2O$, 4.5 g/L $CaCl_2.6H_2O$, 0.84 g/L $MnCl_2.2H_2O$, 0.3 g/L $CoCl_2.6H_2O$, 0.3 g/L $CuSO_4.5H_2O$, 0.4 g/L $NaMoO_4.2H_2O$, 1 g/L H_3BO_3, 0.1 g/L KI and 15 g/L $Na_2EDTA.2H_2O$. Vitamins solution included 50 mg/l d-biotin, 200 mg/l *para*-amino benzoic acid, 1.0 g/l nicotinic acid, 1.0 g/l Ca-pantothenate, 1.0 g/l pyridoxine HCL, 1.0 g/l thiamine HCl and 25 mg/l m-inositol.

The pH was adjusted to 5 by addition of 2N NaOH prior to autoclavation, the glucose was autoclaved separately and methionine and vitamins solutions were sterile filtered (0.2 µm pore-size Ministart®-Plus Sartorius AG, Geottingen, Germany) and added after autoclavation.

Batch cultivations

Batch cultivations were executed in well-controlled, aerobic, 2.2 B Braun Biotech Biostat B fermentation systems with a working volume of 2 L (Sartorius AG, Geottingen, Germany). Proper mixing conditions were ensured by two disk-turbine impellers rotating at 800 RPM and 4 baffles. The pH was automatically controlled at 5 by addition of 2N NaOH. The temperature was kept constant at 30 ^{o}C. The air flow rate was 1 vvm (volume air per volume of broth per minute).

Prior to inoculation, 100 ml precultures were cultivated in 500 ml baffled shake-flasks at 30 ^{o}C until $OD_{600\ nm}$ 5 in an orbital shaker (150 RPM). Minimal medium as described above was used to grow the precultures with 20 g/l glucose. The bioreactors were inoculated to an initial $OD_{600\ nm}$ ranging from 0.5 to 0.7.

Continuous cultivations

Aerobic, carbon limited continuous cultivations were carried out in 2.2 B Braun Biotech Biostat B fermentation systems (as described above for batch cultivations) with a constant working volume of 1.5 L. The temperature was kept at 30 ^{o}C, the pH was maintained at 5 by addition of 2N NaOH, the stirring speed was 600 RPM and the air flow was 1 vvm. Minimal medium with 20 g/l glucose was used to feed the bioreactors at a constant dilution rate of 0.1 h^{-1}. The volume was kept constant at 1.5 l by controlling the level of broth inside the vessel. Steady state conditions were assumed after at least 5 residence times and CO_2 and biomass concentrations were constant.

Off-gas analysis

For both cultivation modes (batch and continuous), off-gas passed through a condenser to minimize evaporation loss during the fermentation and filter sterilized before carbon dioxide and oxygen were quantified in a Brüel & Kjær 1308 acoustic gas analyzer (Brüel & Kjær, Nærum, Denmark).

Biomass determination

Samples were maintained at 4 ^{o}C post sampling and the biomass concentration was monitored by optical density at 600 nm ($OD_{600\ nm}$) and dry cell weight. $OD_{600\ nm}$ was measured throughout all the fermentation in a Shimadzu UV mini 1240 spectrophotometer (Shimadzu Europe GmbH, Duidberg, Germany). The samples were diluted with distilled water in order to obtain

measurements in the linear range of 0 to 0.6 $OD_{600\ nm}$. Dry cell weight was determined by filtering a known volume of fermentation broth with pre-dried 0.45 µm pore-size nitrocellulose filters (Sartorius AG, Geottingen, Germany), which were subsequently washed with a 3x sample volume 0.9 % NaCl saline solution. The filters were then dried for 20 minutes at 150 W in a microwave oven and kept in a desiccator while cooling for at least 2 h. The filters where finally weighted using an analytical balance.

Glucose and external metabolites analysis

The fermentation samples were immediately filtered using a 0.45 µm pore-size syringe-filter (Sartorius AG, Geottingen, Germany) and stored at -20 ^{o}C until further analysis. Glucose, ethanol, glycerol, pyruvate, succinate and acetate were determined by high performance liquid chromatography (HPLC) analysis using an Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories, Hercules, CA). The column temperature was kept at 60 ^{o}C and the elution was performed using 5 mM H_2SO_4 with flow rate of 0.6 ml/min. Metabolites detection was performed by a RI-101differential refractometer detector (Shodex) and an UVD340U absorbance detector (Dionex) set at 210 nm.

Extracellular vanillin, vanillin β-D-glucoside (VG), protocatechuic acid (PAC), protocatechuic aldehyde (PAL) and vanillic acid were quantified by high performance liquid chromatography (HPLC) using Agilent 1100 series equipment with a Luna C18 column (Phenomenex). A gradient of methanol (+ 1% tetra-fluoroacetic acid) and water (+ 1% tetra-fluoroacetic acid) at a flow rate of 0.3 ml/min was used as mobile phase. The column was kept at 300 bar and 30 ^{o}C. Metabolite detection was performed using a UV diode-array detector set at 280 and 310 nm.

Acknowledgements

ARB acknowledge Ph.D. research grant from Fundação para a Ciência e Tecnologia (ref. SFRH / BD / 41230 / 2007). ARB and KRP acknowledge financial support from SYSINBIO - Systems Biology as a Driver for Industrial Biotechnology, Coordination and support action (call FP7-KBBE-2007-1). ARB is thankful to Manuel Quirós Asensio for helpful advice on experimental work. We are thankful to Kristian F. Nielsen and Hanne Jakobsen for help with analytical procedures.

References

1. Clark GS: **Vanillin**. *Perfumer & Flavorist* 1990, **15**:45,, 46, 50, 52-54.

2. Ander P, Hatakka A, Eriksson K-erik: **Vanillic Acid Metabolism by the White-Rot Fungus Sporotriehum pulverulentum**. *Enzyme* 1980, **202**:189-202.

3. Chatterjee T, De B, Bhattacharyya D: **Microbial conversion of isoeugenol to vanillin by Rhodococcus rhodochrous**. *Indian Journal of Chemistry* 1999, **38**:538-541.

4. Civolani C, Barghini P, Roncetti a R, Ruzzi M, Schiesser A: **Bioconversion of ferulic acid into vanillic acid by means of a vanillate-negative mutant of Pseudomonas fluorescens strain BF13.** *Applied and Environmental Microbiology* 2000, **66**:2311-7.

5. Hansen EH, Møller BL, Kock GR, et al.: **De novo biosynthesis of vanillin in fission yeast (Schizosaccharomyces pombe) and baker's yeast (Saccharomyces cerevisiae).** *Applied and Environmental Microbiology* 2009, **75**:2765-74.

6. Asadollahi M, Maury J, Patil KR, et al.: **Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering.** *Metabolic Engineering* 2009, **11**:328-34.

7. Bro C, Regenberg B, Förster J, Nielsen J: **In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production.** *Metabolic Engineering* 2006, **8**:102-11.

8. Nissen TL, Kielland-Brandt MC, Nielsen J, Villadsen J: **Optimization of ethanol production in Saccharomyces cerevisiae by metabolic engineering of the ammonium assimilation.** *Metabolic Engineering* 2000, **2**:69-77.

9. Ro D-K, Paradise EM, Ouellet M, et al.: **Production of the antimalarial drug precursor artemisinic acid in engineered yeast.** *Nature* 2006, **440**:940-3.

10. Raab AM, Gebhardt G, Bolotina N, Weuster-Botz D, Lang C: **Metabolic engineering of Saccharomyces cerevisiae for the biotechnological production of succinic acid.** *Metabolic Engineering* 2010:1-8.

11. Bailey JE: **Toward a science of metabolic engineering.** *Science* 1991, **252**:1668-75.

12. Nielsen J: **Metabolic engineering**. *Applied Microbiology and Biotechnology* 2001, **55**:263-283.

13. Ostergaard S, Olsson L, Nielsen J: **Metabolic engineering of Saccharomyces cerevisiae.** *Microbiology and Molecular Biology Reviews* 2000, **64**:34-50.

14. Patil KR, Akesson M, Nielsen J: **Use of genome-scale microbial models for metabolic engineering.** *Current Opinion in Biotechnology* 2004, **15**:64-9.

15. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnology and Bioengineering* 2003, **84**:647-57.

16. Patil KR, Rocha I, Förster J, Nielsen J: **Evolutionary programming as a platform for in silico metabolic engineering.** *BMC Bioinformatics* 2005, **6**:308.

17. Varma A, Boesch BW, Palsson BØ: **Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates.** *Applied and Environmental Microbiology* 1993, **59**:2465-73.

18. Famili I, Förster J, Nielsen J, Palsson BØ: **Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:13134-9.

19. Edwards JS, Ibarra RU, Palsson BØ: **In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.** *Nature Biotechnology* 2001, **19**:125-30.

20. Edwards JS, Palsson BO: **Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions.** *BMC Bioinformatics* 2000, **1**:1.

21. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13**:244-53.

22. Ibarra RU, Edwards JS, Palsson BO: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth**. *Nature* 2002, **420**:20-23.

23. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:15112-15117.

24. Ramachandra Rao S, Ravishankar G: **Vanilla flavour: production by conventional and biotechnological routes**. *Journal of the Science of Food and Agriculture* 2000, **80**:289-304.

25. Arana F: **Action of a B-Glucosidase in the curing of vanilla**. *Food Research* 1943, **8**:343-351.

26. Postma E, Verduyn C, Scheffers WA, Dijken JP Van: **Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae.** *Applied and Environmental Microbiology* 1989, **55**:468-77.

27. Rieger M, Käppeli O, Fiechter A: **The Role of Limited Respiration in the Incomplete Oxidation of Glucose by Saccharomyces cevevisiae**. *Journal of General Microbiology* 1983, **129**:653-661.

28. Urk H van, Postma E, Scheffers W a, Dijken JP van: **Glucose transport in crabtree-positive and crabtree-negative yeasts.** *Journal of General Microbiology* 1989, **135**:2399-406.

29. Bakker BM, Overkamp KM, Maris AJ van, et al.: **Stoichiometry and compartmentation of NADH metabolism in Saccharomyces cerevisiae.** *FEMS Microbiology Reviews* 2001, **25**:15-37.

30. **Saccharomyces Genome database** [http://www.yeastgenome.org/].

31. Magasanik B, Kaiser CA: **Nitrogen regulation in Saccharomyces cerevisiae.** *Gene* 2002, **290**:1-18.

32. Moreira dos Santos M, Raghevendran V, Kötter P, Olsson L, Nielsen J: **Manipulation of malic enzyme in Saccharomyces cerevisiae for increasing NADPH production capacity aerobically in different cellular compartments.** *Metabolic Engineering* 2004, **6**:352-63.

33. Pronk JT, Yde Steensma H, Dijken JP Van: **Pyruvate Metabolism inSaccharomyces cerevisiae**. *Yeast* 1996, **12**:1607-1633.

34. Hohmann S: **Characterization of PDC6, a third structural gene for pyruvate decarboxylase in Saccharomyces cerevisiae.** *Journal of Bacteriology* 1991, **173**:7963-9.

35. Flikweert MT, Zanden LVD, Janssen WMTM, et al.: **Pyruvate decarboxylase: An indispensable enzyme for growth ofSaccharomyces cerevisiae on glucose**. *Yeast* 1996, **140**:1723-257.

36. Grenson M, Dubois E, Piotrowska M, Drillien R, Aigle M: **Ammonia assimilation in Saccharomyces cerevisiae as mediated by the two glutamate dehydrogenases. Evidence for the gdhA locus being a structural gene for the NADP-dependent glutamate dehydrogenase.** *Molecular and General Genetics* 1974, **128**:73-85.

37. Cogoni C, Valenzuela L, González-Halphen D, et al.: **Saccharomyces cerevisiae has a single glutamate synthase gene coding for a plant-like high-molecular-weight polypeptide.** *Journal of Bacteriology* 1995, **177**:792-8.

38. Mitchell a P, Magasanik B: **Purification and properties of glutamine synthetase from Saccharomyces cerevisiae.** *The Journal of Biological Chemistry* 1983, **258**:119-24.

39. Miller SM, Magasanik B: **Role of NAD-linked glutamate dehydrogenase in nitrogen metabolism in Saccharomyces cerevisiae.** *Journal of Bacteriology* 1990, **172**:4927-35.

40. Moreira Dos Santos M, Thygesen G, Kotter P, Olsson L, Nielsen J: **Aerobic physiology of redox-engineered strains modified in the ammonium assimilation for increased NADPH availability**. *FEMS Yeast Research* 2003, **4**:59-68.

41. Kim P-J, Lee D-Y, Kim TY, et al.: **Metabolite essentiality elucidates robustness of Escherichia coli metabolism.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:13638-42.

42. Braus GH: **Aromatic amino acid biosynthesis in the yeast Saccharomyces cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway.** *Microbiological Reviews* 1991, **55**:349-70.

43. Luttik M a H, Vuralhan Z, Suir E, et al.: **Alleviation of feedback inhibition in Saccharomyces cerevisiae aromatic amino acid biosynthesis: quantification of metabolic impact.** *Metabolic Engineering* 2008, **10**:141-53.

44. Gertz EM: *Object-Oriented Software for Quadratic Programming*. ACM Transactions on the Mathematical Sofware 29; 2003:58-81.

45. Chung BKS, Lee D-Y: **Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network.** *BMC Systems Biology* 2009, **3**:117.

46. Reid RJD, Sunjevaric I, Keddache M, Rothstein R, Kedacche M: **Efficient PCR-based gene disruption in Saccharomyces strains using intergenic primers.** *Yeast* 2002, **19**:319-28.

47. Feng Q, Düring L, Mayolo AA de, et al.: **Rad52 and Rad59 exhibit both overlapping and distinct functions.** *DNA Repair* 2007, **6**:27-37.

48. Gietz D, St Jean A, Woods RA, Schiestl RH: **Improved method for high efficiency transformation of intact yeast cells.** *Nucleic Acids Research* 1992, **20**:1425.

49. Erdeniz N, Mortensen UH, Rothstein R: **Cloning-Free PCR-Based Allele Replacement Methods**. *Genome Research* 1997:1174-1183.

50. Boeke JD, Lacroute F, Fink GR: **Short communication A positive selection for mutants lacking orotidine-5 ' -phosphate decarboxylase activity in yeast : 5-fluoro-orotic acid resistance**. *Molecular and General Genetics* 1984:345-346.

51. Verduyn C, Postma E, Scheffers WA, Dijken JP Van: **Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation.** *Yeast* 1992, **8**:501-17.

Chapter 4: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 2 – Pathway and Regulatory Engineering

Part of this chapter was used for the manuscript:

Overexpression of O-methyltransferase leads to improved vanillin production in baker's yeast only when complemented with model-guided network engineering. **Ana Rita Brochado** and Kiran R. Patil. ***Submitted***.

Abstract

Vanillin is one of the most used flavoring agents worldwide. A yeast cell factory for vanillin production in its more soluble form, vanillin β-D-glucoside, was previously assembled and we recently reported an *in silico* based metabolic engineering strategy for improved vanillin production. Accumulation of several pathway intermediates was observed and promoting their conversion to the final product is the following step towards further strain improvement. The metabolic flux through the vanillin biosynthetic pathway is controlled by i) the availability of precursor and cofactors and ii) the enzyme activity of the pathway enzymes. In order to ensure a maximum outcome of the network engineering strategy previously implemented for boosting cofactors supply, limitation of the enzyme activity at the pathway level must not occur. The dominant effects between the local and global aspects are, however, hard to estimate *a priori*. The present study aims at deciphering if the metabolic flux towards vanillin is controlled by pathway enzyme availability or by global aspects, such as cofactor supply. Herein, we detect and optimize the expression of the rate limiting enzyme of the vanillin biosynthetic pathway. Additionally, attempting to further increase the ATP availability for vanillin biosynthesis, we suggest alleviating the yeast glucose repression phenomenon. Following the mentioned strategies, *hsOMT* and *HAP4* overexpression resulted in 60% increase in vanillin β-D-glucoside, with a final concentration of 400 mg/l in the fermentation broth.

Background

Vanillin, or 4-hydroxi-3-metoxibenzaldehyde, is the main compound responsible for vanilla aroma, one of the flavoring agents most widely used. Given its chemopreventive properties, vanillin can be used as antimicrobial and antioxidant agent. Several recent studies also suggest vanillin to be used for cancer treatment, as a metastasis suppressor [1–3]. Natural vanillin originates from the seed pod of the vanilla orchid, *Vanilla planifolia*. Approximately 40,000 manually pollinated flowers are required to produce 1 kg of vanillin; thereby naturally produced vanillin is not enough to fulfill the worldwide demand. Only about 0.25% of the annual vanillin market is obtained from natural sources, while the remaining demand is fulfilled by chemical synthesis from lignin or fossil hydrocarbons (in particular guaiacol) [1, 4]. Bioconversion of ferulic acid and eugenol by bacteria and fungi has been proposed as an environmental friendly process [5–9]. Alternatively, a *de novo* vanillin biosynthetic pathway from glucose was developed in yeast *Saccharomyces cerevisiae* [4]. The native metabolic precursor for *de novo* vanillin biosynthesis in yeast is 3-dehydroshikimate (3DHS), an intermediate of the shikimate pathway for aromatic amino acids biosynthesis (**Fig 3.1**). Three enzymes, a 3-dedhydroshikimate dehydratase (3DSD), an aryl-carboxylic acid reductase (ACAR) and an O-methyltransferase (hsOMT), catalyze the conversion from 3DHS to vanillin. In order to obtain a functional pathway, ACAR post-translational activation is performed by a heterologous phosphopantetheinyl transferase (PPTase), from *E. coli*, *C. glutamicum* or *N. farcinica* [4].

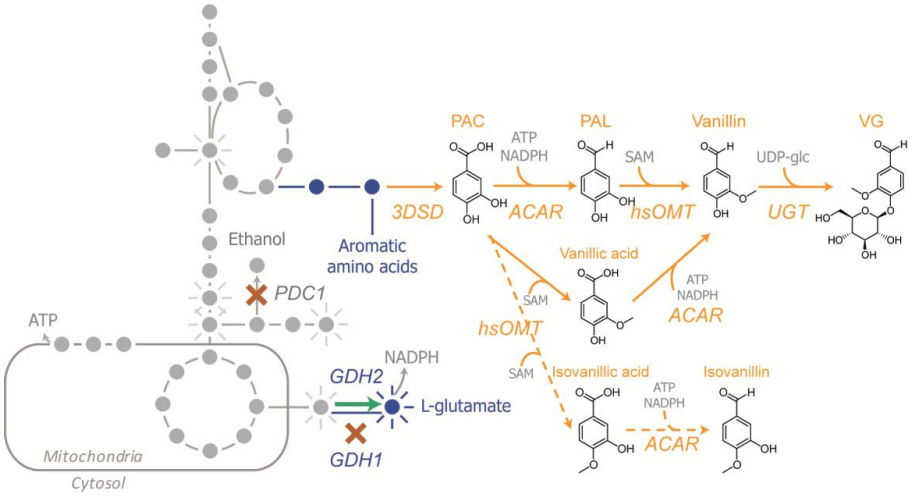


Figure 4.1: *De novo* vanillin β-D-glucoside biosynthesis in *S. cerevisiae*. A simplified metabolic map of the yeast central carbon (gray) and amino acid metabolism (blue) containing relevant reactions for the vanillin biosynthetic pathway (yellow) is shown. Metabolites are represented by filled circles in the case of central and amino acid

metabolism, and by their chemical structures in the case of vanillin biosynthetic pathway. Circles with multiple connecting lines represent highly connected metabolites. Reactions are represented by the connecting lines. Dashed yellow lines represent the synthesis of isovanillin, an undesired by-product. Italicized names refer to genes, while the other names refer to metabolites. The gene deletions (red crosses) and overexpressions (green arrows) correspond to a global engineering strategy carried out in previous work [10]. PAC, PAL and VG stand for protocatechuic acid, protocatechuic aldehyde and vanillin β-D-glucoside respectively. 3DSD, ACAR, hsOMT and UGT stand for 3-dehyroshikimate dehydratase, aryl-carboxylic acid reductase, O-methyltransferase and UDP-glycosyltransferase, respectively.

In a previous work [10], we achieved five-fold increase in vanillin production, in its less toxic and more soluble form, vanillin β-D-glucoside, based on a genome-scale modeling approach for designing metabolic engineering strategies [11–13]. Minimization of metabolic adjustment [14] was used as cellular biological objective function and OptGene [15] was used as modeling platform to identify genes within the yeast metabolic network [16], deletion of which would increase vanillin production. The strategy suggested by the modeling exercise was to, i) delete one of the pyruvate decarbolxilases (*PDC1*), ii) delete the most active glutamate dehydrogenase (*GDH1*) and iii) ensure sufficient nitrogen uptake by overexpressing the less used *GDH2* (**Fig. 4.1**). These modifications resulted in increased respiratory capacity of the cell, as well as shifted the cofactor usage of nitrogen uptake from NADPH to NADH. Both cofactors, ATP and NADPH, are required for the vanillin biosynthetic pathway to convert protocatechuic acid (PAC) in protocatechuic aldehyde (PAL) [4].

Nevertheless, apart from vanillin β-D-glucoside, several by-products were also accumulated in the fermentation broth, including intermediates of the vanillin biosynthetic pathway. At the most, vanillin β-D-glucoside accounts for approximately 25 % (mol/mol) of the total vanillin related compounds produced. Thus, decreasing the by-product formation will ultimately lead to increase the vanillin yield up to 4-fold, as well as decreased accumulation of potentially toxic compounds on the fermentation broth. The aim of the present study is to explore whether the metabolic flux through the vanillin biosynthetic pathway is regulated by enzyme availability at the pathway level, or by cofactors supply. To this end, we suggest i) detecting and optimizing the expression of rate limiting enzymes of the vanillin biosynthetic pathway and ii) alleviating the yeast glucose repression phenomenon in order to increase the ATP availability for vanillin production (through increased respiration). Towards the first goal, we tested two extra PPTases (from *Aspergillus nidulans* and *Nocardia sp.),* tackling the possibility that they would better activate ACAR in *S. cerevisiae*. Afterwards, we evaluated the effects of overexpressing *ACAR* and *hsOMT* on vanillin β-D-glucoside production. Finally, we assessed the effect of alleviating the glucose repression of respiration through *Hap4* overexpression.

Results and Discussion

Different PPTases for ACAR post-translational activation lead to altered vanillin β-D-glucoside production

Enzymes requiring activation by phosphopantetheinylation are not abundant in *S. cerevisiae*, therefore this microorganism displays a rather reduced repertoire of PPTases. It is also known that this class of enzymes can be highly organism specific, since PPTases accomplishing similar function across closely related organisms might not be able to activate both target enzymes [4, 17]. Hansen *et al.* [4] observed that only three, out of eleven tested bacterial PPTases, were able to activate ACAR in *S. cerevisiae*. Given that ACAR activity is highly influenced by the PPTase, we heterologously expressed two additional PPTases and studied their impact on vanillin β-D-glucoside production; i) *npgA* from *Aspergillus nidulans*, which was previously reported to be functional in *S. cerevisiae* by successfully activating a polyketide synthase [18] and a non-ribosomal peptide synthase [19] and ii) *npt* from *Nocardia sp.*, which was reported to be the PPTase activating ACAR in the native host [20]. Both genes were codon optimized and expressed in the reference strain VG0 (**Table 4.2**) from low-copy-number-replicating plasmids (CEN-ARS) under the yeast *TPI1* promoter (**Methods**). Furthermore, two of the PPTases previously used by Hansen *et al*. (*entD* from *E. coli* and PPTcg-1 from *C. glutamicum*) were also heterologously expressed for comparison.

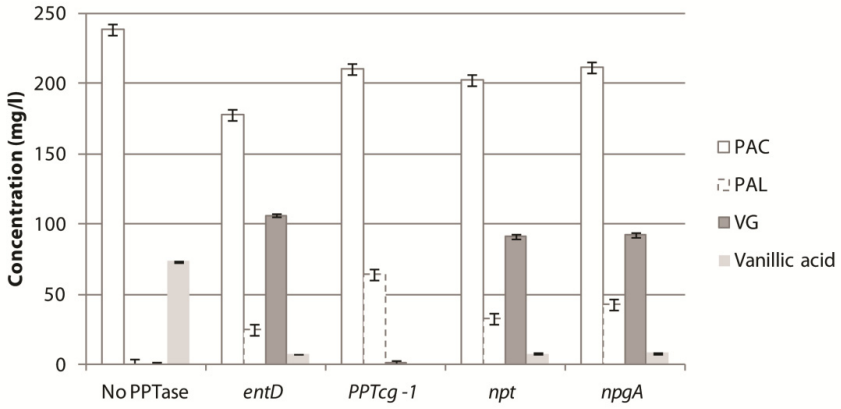


Figure 4.2: Impact of heterologous expression of different PPTases in ACAR activation in *S. cerevisiae*. Concentration of vanillin-related metabolites (vanillin β-D-glucoside (VG), vanillic acid, protocatechuic acid (PAC) and protocatechuic aldehyde (PAL)) in the fermentation broth (90 h in microtiter plates cultivation) for strain VG0 expressing four different PPTases. Results for the strain without PPTase are shown as a control.

ACAR catalyzes the conversion from PAC (protocatechuic acid) to PAL (protocatechuic aldehyde), as well as from vanillic acid to vanillin (**Fig. 4.1**). As expected, nor PAL neither vanillin are produced if there is no heterologous PPTase expression (negative control, **Fig. 4.2**). All PPTases seem to efficiently convert PAC to PAL, as PAL is accumulated in the extracellular broth. Nevertheless, the yeast strain carrying the gene *entD* (*E. coli)* showed the highest vanillin β-D-glucoside production (~100 mg/l) and, therefore, this was the selected PPTase for further work. The genes *npt* and *npgA* had very similar effects on vanillin β-D-glucoside production, and not so far from the best candidate, *entD*. If, at certain stage of further strain improvement, ACAR is found to be limiting step, these two genes might be taken into consideration once again.

hsOMT catalyses the rate-limiting step for *de novo* vanillin β-D-glucoside biosynthesis in *S. cerevisiae*

In our previous study we showed that the conversion of pathway intermediates to vanillin β-D-glucoside was incomplete even after a successful round of strain improvement by engineering the cell for increased cofactors and energy supply (strain VG4, **Table 4.2**) [10]. Typically, around 50% of the vanillin-related compounds accounts for PAC, suggesting that either ACAR or hsOMT (which catalyze the conversion from PAC into downstream pathway intermediates, **Fig. 4.1**) might be limiting the flux though the pathway. Consequently, to decipher whether one or both enzymes are limiting the metabolic flux towards vanillin, we overexpressed each gene by chromosomal integration of a second copy in the globally engineered strain VG4 (**Methods**, **Table 4.2, and Scheme 3.1**).

While overexpressing *ACAR* appears to have no significant effect of vanillin β-D-glucoside production (strain VG4-A1), the effect of overexpressing *hsOMT* is clear at the metabolite level (strain VG4-O1), with 30% increase of vanillin β-D-glucoside production (**Fig. 4.3**). The raise in vanillin β-D-glucoside production was accompanied by a decrease of accumulation of both intermediates, PAC and PAL, attesting the increased hsOMT activity in strain VG4-O1.

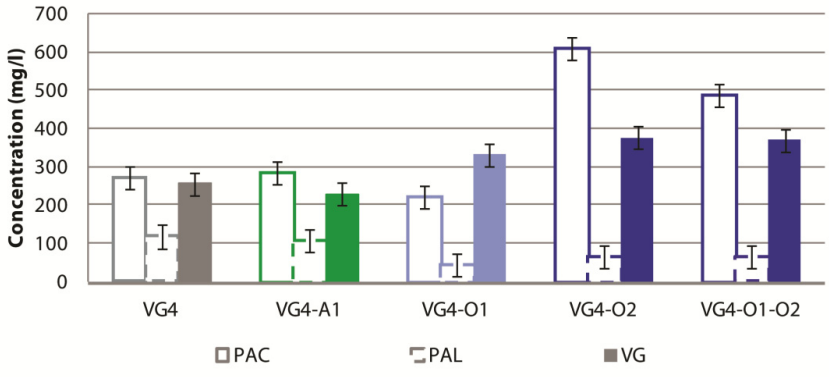


Figure 4.3: Effect of *ACAR* and *hsOMT* overexpression on vanillin β-D-glucoside production in *S. cerevisiae*. Concentration of vanillin-related metabolites on the fermentation broth (70 h cultivation in microtiter plates) for strains VG4, VG4-A1 (ACAR overexpressed), VG4-O1, VG4-O2 and VG4-O1-O2 (hsOMT overexpressed).

Exploring if stronger overexpression could further increase vanillin β-D-glucoside production, a third copy of *hsOMT* was integrated in strain VG4-O1, leading to strain VG4-O1-O2 (**Table 4.2**). Herein, the genomic integration site was selected based on the data from Bai *et al.* (2009) which systematically characterized genomic integration sites for heterologous gene expression in *S. cerevisiae* [21]. As the transcription levels vary between different chromosomal regions, selecting the appropriate integration site allows adjusting the expression level for the inserted gene [21–24]. We selected the integration site *YPRCΔ15* (#20) among the 24 sites characterized by Bai *et al.* (2009), because it showed the highest gene expression levels. Additionally, we chose *TEF1* promoter sequence (also a strong constitutive promoter) instead of *TPI1*, to avoid further increasing the competition for the same transcription factors; *TPI1* is already used for activation of all the genes from the vanillin biosynthetic pathway. We also constructed strain VG4-O2 (**Table 4.2**) by inserting a second copy of *hsOMT* in strain VG4 at *YPRCΔ15 locus*, so we could compare both overexpression strategies (VG4-O1 *vs* VG4-O2). Strain VG4-O2 produced around 13% more vanillin β-D-glucoside than strain VG4-O1 (**Fig. 4.3**), indicating that integration at *YPRCΔ15 locus* together with *TEF1* promoter is a better strategy for higher gene expression in strain VG4, compared to *Met15 locus* together with *TPI1* promoter, used for VG4-O1. This result also indicates that *hsOMT* expression level in strain VG4-O1 might still be limiting vanillin production. Simultaneous genomic integration of the two additional copies of *hsOMT* (strain VG4-O1-O2) resulted in the same titer as the one verified for strain VG4-O2. This led us to conclude that overexpression of *hsOMT* from *YPRCΔ15 locus* under *TEF1* promoter is sufficient for relieving the flux limitation through the vanillin pathway imposed by hsOMT.

Notably, all the strains having *hsOMT* overexpressed show a significant decrease in PAL accumulation, which is explained by the fact that hsOMT is catalyzing the conversion from PAL to vanillin. On the other hand, a rather counter-intuitive trend is verified for PAC, which increases when *hsOMT* is overexpressed from *YPRCΔ15 locus*. From previous experiments we observed that PAC accumulation takes place mostly when the cells are exponentially growing on glucose (as opposed to ethanol), as well as we often observed a positive correlation between biomass and PAC concentration (data not shown). Indeed, both strains overexpressing *hsOMT* from *YPRCΔ15 locus* have increased biomass formation, compared to VG4 (**Fig. 4.4A**).

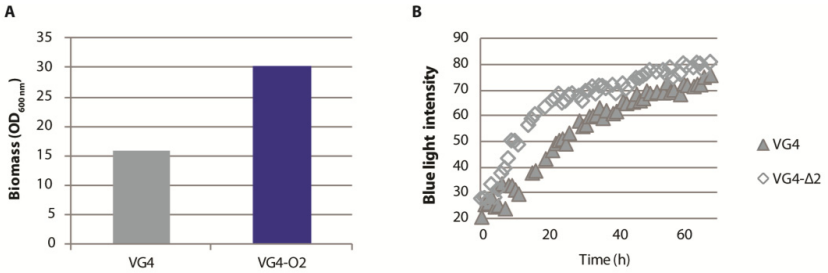


Figure 4.4: Effect of deletion of *YPRCΔ15 locus* on the growth of strain VG4. A) Final biomass concentration (measured by $OD_{600\ nm}$) of strains VG4 and VG4-O2 obtained after 70 h cultivation in microtiter plates. **B)** The growth of strains VG4 and VG4-Δ2 was followed over time using Enzyscreen (**Methods**). This experiment was carried out without expressing the PPTase, therefore no PAL, vanillin or vanillin β-D-glucoside are produced in these conditions.

Such strong effect on biomass yield is, most likely, the outcome of a combination of different factors. Nevertheless, we hypothesize that major contributions can come from having additional protein activity and consequent changes in the pool of by-products, as well as from the choice of overexpression *locus* itself. Although Bai *et al.* [21] mention that expressing a reporter gene form *YPRCΔ15 locus* did not significantly altered growth, we hypothesise that this might not be the case for the vanillin producing strains. Moreover, their study was carried out in a different strain background, which can be another source of variation [21]. We further explored the integration site related effects by constructing a control strain using the same strategy as used for VG4-O2, but the gene, promoter and terminator were omitted (VG4-Δ2, **Table 4.2**). Strains VG4 and VG4-Δ2 were grown prior to expression of the PPTase, so most of the vanillin related compounds were not produced. This way, we try to minimize the effects of cellular resources spent on the vanillin pathway, as well as effects coming from the accumulation of several different vanillin related compounds (only PAC and vanillin acid are produced). The control strain was found to have altered growth kinetics comparing to VG4, particularly evident while consuming glucose (**Fig. 4.4B**). Preliminary strain characterization suggests faster growth on glucose, as well as higher biomass yield on this substrate. More comprehensive conclusions would require better

strain characterization, namely to grow the control strain in vanillin producing conditions, so the effects of harboring the active pathway would also be taken into account. Nevertheless, these results do not contradict the fact that hsOMT is a rate-limiting enzyme for vanillin production, since in all *hsOMT* overexpressed strains, higher vanillin β-D-glucoside-to-PAL ratio is observed, independently of the overexpression strategy.

Another important aspect is that all the *hsOMT* overexpressed strains showed a significant increase in an unidentified compound, as detected in the chromatographic analysis. Production of this unknown compound was found to be strongly affected by *hsOMT* overexpression, so we speculate that the unidentified substance might be vanillic acid in its glucosylated form. Alternatively, hsOMT might lead to the formation of the by-product isovanillic acid, which would subsequently be reduced to isovanillin, in a reaction catalyzed by ACAR (**Fig. 4.1**). Accordingly, the unidentified compound can also be isovanillin, isovanillic acid or their glycosylated forms.

The results obtained so far indicate that the bottleneck of the pathway is no longer the hsOMT enzyme activity. In order to assess whether the control for the flux though the vanillin pathway moved from hsOMT to ACAR, we overexpressed *ACAR* in strain VG4-O2, again by genomic integration from *Met15 locus* (**Table 4.2**). We found that *ACAR* overexpression did not further increase vanillin β-D-glucoside.

Overexpression of hsOMT leads to improved vanillin production in baker's yeast only when complemented with model-guided network engineering

Overexpressing genes belonging to the pathway of interest for obtaining the desired product is a commonly used metabolic engineering strategy [25–27]. On the other hand, global metabolic network approaches, which target the metabolic network nodes distant from the pathway of interest, have also been shown to be effective. Ultimately the most efficient solution is, as we herein show, the complementary usage of both approaches [28, 29]. The two approaches are closely interrelated; global approaches often aim at boosting the pathway of interest through the supply of cofactors and precursors, while rather local approaches focus on relieving the flux through the pathway from enzyme limitations. Thus, whether a global strategy is pre-required for the success of a local strategy, or vice-versa, is difficult to estimate a priori.

Following the same strategy as before, we overexpressed hsOMT from *YPRCΔ15 locus* in strain VG0 (**Table 4.2**, **Methods**) and measured vanillin β-D-glucoside production.

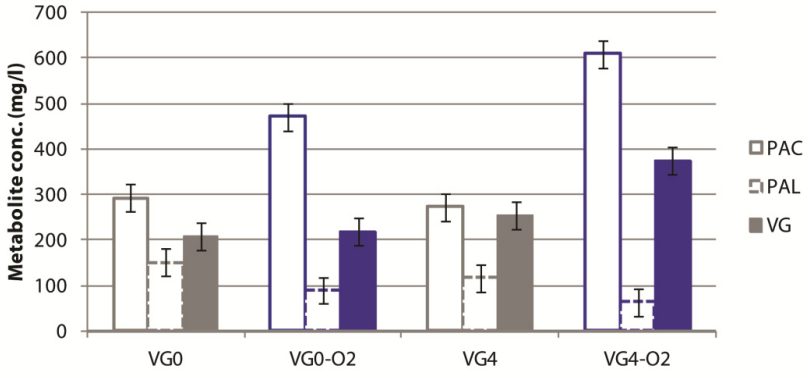


Figure 4.5: Overexpression of hsOMT leads to improved vanillin production only when complemented with model-guided network engineering. Concentration of vanillin-related metabolites on the fermentation broth of strains VG0 and VG4 before (gray) and after (blue) overexpression of *hsOMT*.

Figure 4.5 shows that overexpression of *hsOMT* in VG0 does not increase vanillin β-D-glucoside production, reinforcing that the cofactor and energy supply for vanillin biosynthesis is of primary importance, compared to the level of expression of the enzymes from vanillin biosynthesis pathway. Furthermore, the best cell factory is obtained when both strategies are combined, strain VG4-O2, resulting in a final titer of ~380 mg/l of vanillin β-D-glucoside. Also here, increased PAC accumulation for strain VG0-O2 is explained by increased biomass yield on glucose due to the integration *locus*.

Engineering glucose repression in *S. cerevisiae* leads to improved vanillin β-D-glucoside production

Tackling the problem of further increasing the availability of cofactor and energy supply for vanillin β-D-glucoside production, we propose to alleviate glucose repression in batch cultivation. When glucose concentration is above a given threshold, many genes involved in respiration, gluconeogenesis and utilization of non-glucose carbon sources are expressed at low levels or not at all, while alcoholic fermentation is highly active [30–32]. Comparing to alcoholic fermentation, respiration is much more efficient from the energetic perspective [32, 33]. This makes it an attractive solution for increasing vanillin production, which requires ATP. To this end, the activation of many enzymes is required and manipulating a regulatory circuit represents an elegant option. Glucose repression in yeast takes place mainly via two regulatory pathways: the induction pathway, ultimately responsible for derepressing hexose transporters, and the repression pathway, responsible for repressing most of the genes involved in respiration, among

other cellular processes [30, 34]. Among the repressed genes, there is the Hap2/3/4/5p regulatory complex, which activates transcription of genes coding for enzymes of the TCA cycle, as well as for the respiratory chain. It has been previously confirmed that overexpression of *HAP4,* coding for one of the subunits of the complex, activates several TCA cycle genes, as well as, respiratory genes in high glucose concentration (batch conditions) [33, 35]. Hereby, we hypothesize that *HAP4* overexpression will increase the respiration-to-fermentation ratio and thereby increase the ATP availability to produce vanillin β-D-glucoside.

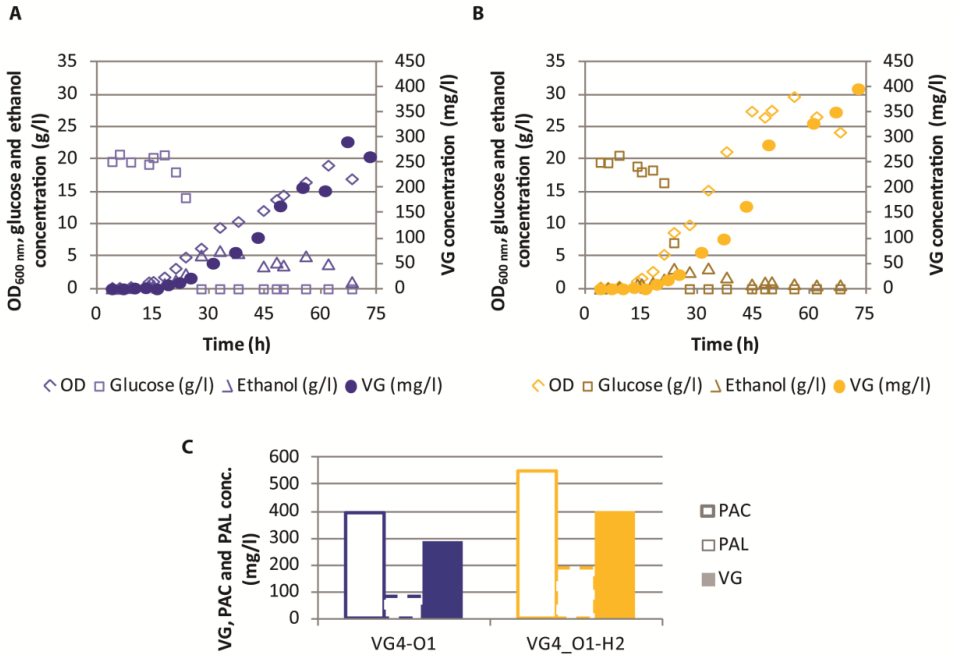


Figure 4.6: Overexpression of *HAP4* increases the production of vanillin β-D-glucoside. A) Fermentation profile of biomass ($OD_{600\ nm}$), glucose, ethanol and vanillin β-D-glucoside during aerobic batch cultivation in microtiter plates of strain VG4-O1. **B)** Fermentation profile of biomass ($OD_{600\ nm}$), glucose, ethanol and vanillin β-D-glucoside during aerobic batch cultivation in microtiter plates of strain VG4-O1-H2. **C)** Concentration of vanillin-related metabolites corresponding to the time-point with the highest concentration of vanillin β-D-glucoside, for strains VG4-O1 and VG4-O1-H2.

Overexpression of *HAP4* was accomplished by genomic integration of a second copy of the *HAP4* in *YPRCΔ15 locus,* under *TEF1* promoter, in strain VG4-O1 (VG4-O1-H2, **Table 4.2**). In agreement with previous studies [33, 36], overexpressing *HAP4* led to an increased biomass yield and decreased ethanol formation (**Fig. 4.6A&B**). The final concentration of vanillin β-D-glucoside on the fermentation broth of the strain VG4-O1-H2 was ~400 mg/l, approximately 30% higher than that of VG4-O1 (**Fig. 4.6C**). PAL concentration in the fermentation broth is also increased,

suggesting that alleviation of the glucose repression indeed contributed to better conversion from PAL to PAC, where ATP and NADPH are used as cofactors.

Overview of metabolic engineering strategies towards improving vanillin β-D-glucoside production

The improved biomass production observed for strain VG4-O1-H2 is not independent of the integration *locus*, as discussed before. A similar strategy for obtaining a control strain should be followed here in the future, in order to confirm the exact contribution of *HAP4* to the observed phenotype. Nevertheless, the production of vanillin related compounds obtained with strain VG4-O1-H2 can be also compared to that of VG4-O2, where integration site related effects on biomass are equal and hsOMT is overexpressed as well, although from a different *locus* (**Fig. 4.7**).

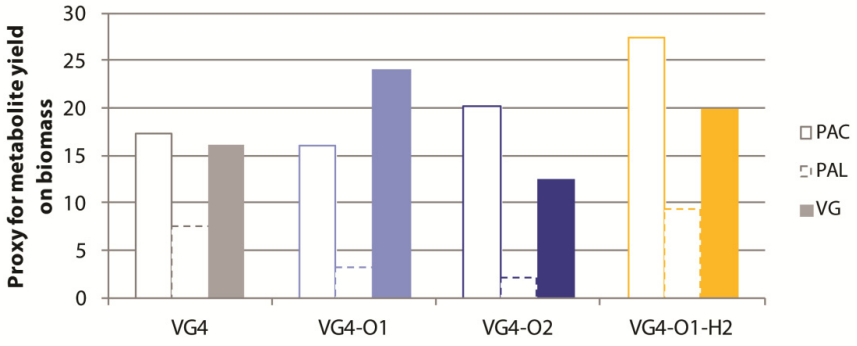


Figure 4.7: Overview of metabolic engineering strategies towards improving vanillin β-D-glucoside production. A proxy for metabolite yield on biomass, based on $OD_{600\ nm}$, is used to compare strain VG4, VG4-O1, VG4-O2 and VG4-O1-H2.

A general overview of all the metabolically engineered mutants shows that overexpression of *hsOMT* from the *Met15 locus* leads to the highest vanillin β-D-glucoside production from biomass (**Fig. 4.7**). However, biomass production is 2-fold higher when *hsOMT* or *HAP4* are overexpressed from *YPRCΔ15 locus*, thereby this strategy is, overall, more effective for converting glucose into vanillin β-D-glucoside. Independently of the integration site, *hsOMT* overexpression is highly beneficial for converting PAL into downstream compounds of the vanillin pathway. Additional *HAP4* overexpression further increases vanillin β-D-glucoside production, strain VG4-O1-H2 presents the highest vanillin β-D-glucoside titer obtained among all the mutants, ~400 mg/l. This strategy also results in a pronounced change in the profile of vanillin related compounds, with a 3-fold increase of PAL concentration. As previously discussed,

this suggests that, after *hsOMT* overexpression, cofactors like ATP probably became, once again, the bottleneck for converting PAC into PAL in strain VG4-O2. Also supporting this hypothesis is the fact that ACAR overexpression did not change the vanillin related compounds production profile (strain VG4-A1-O2). Further studies could include investigating the recurrent limitation of the hsOMT enzymatic step, focusing on both, boosting enzyme activity and cofactor supply (S-adenosylmethionine).

Conclusions and Future Perspectives

The present study aimed at improving a vanillin β-D-glucoside *S. cerevisiae* cell factory by boosting the conversion of vanillin pathway intermediates to the final product. Identification of *hsOMT*, which catalysis the conversion of PAL into vanillin, as the rate-limiting step of the vanillin biosynthetic pathway and its consequent overexpression led to 30% improvement of vanillin β-D-glucoside production. Additionally, we performed regulatory engineering, by overexpressing *HAP4,* with the goal of increasing the respiration-to-fermentation ratio, so that less ethanol would be produced and more ATP would be available for the vanillin biosynthetic pathway. This genetic modification yielded further increase in vanillin β-D-glucoside production. The final concentration in the fermentation broth was ~400 mg/l in batch cultivation, corresponding to ~20 mg_{VG}/g_{glc}, considerably higher than ~15 mg_{VG}/g_{glc}, previously reported for similar conditions [10]. Most importantly, higher accumulation of PAL in the fermentation broth was again observed, which suggests higher ATP availability for the vanillin biosynthetic pathway. Further strain characterization, such as chemostat cultivation or measurement of the respiratory quotient (RQ) in batch cultivation, could confirm higher respiratory capacity of this strain. Transcriptome analysis would be useful for carefully examining transcription levels of relevant genes, possibly affected by the overexpression of the transcription factor *HAP4*. ^{13}C isotope labeled fluxome analysis would also be of great interest for further characterizing the observed phenotype. Such data would also offer a chance of performing an integrated analysis with the yeast genome-scale model and find targets for the next metabolic engineering round. On a perspective of increasing enzyme specificity, protein engineering could be very valuable to reduce/eliminate by-product formation, namely vanillic acid and isovanillin related compounds [13, 37].

In conclusion, this study showed that accounting for the local pathway effects for metabolic engineering is of much higher value when it is combined with global approaches, which involve whole metabolic network, as well as regulatory effects. Although algorithms for including such

targets are already starting to appear, *e.g.* OptORF [38], this research field still in its infancy. Further efforts on characterizing regulatory networks are expected to provide valuable insights for integrating regulatory circuits with metabolic models [38, 39].

Methods

Strains and Plasmids

Common molecular biology protocols

Restriction enzymes and buffers were obtained from New England Biolabs and the conditions for restriction followed manufacture instructions. GFX^{TM} PCR DNA and Gel Band Purification Kit (GE Healthcare) was used for DNA purification. Ligation was performed with T4 DNA ligase (New England Biolabs) prior to plasmid transformation into chemo-competent *DH5α E. coli* cells. Ampicillin resistance was used as *E. coli* selection marker and plasmid extraction was performed using a GenElute HP Plamsid Miniprep Kit (Sigma-Aldrich). PCR amplification was performed using Phusion^{TM} Hot Start High-Fidelity DNA Polymerase (Finnzymes Oy, Espoo, Finland) to obtain all the DNA fragments, except for amplification of fragments for USER cloning. In such cases, PfuX7, a mutant Pfu DNA polymerase designed for advanced uracil-excision, was used instead [40]. All the primers used for this study were acquired from Sigma Aldrich and are listed in **Table ST4.1**.

USER cloning was performed as previously described [41] with minor modifications. The USER vectors (containing a USER cassette) were digested with *PacI*, followed by digestion with the appropriate nicking endonuclease, Nt.BbvCI. Purified digested vector (0.1 pmol) was mixed with 1 pmol purified PCR products (total concentration of the 3 PCR fragments, equimolar) and incubated with 1 U USER enzyme (New England Biolabs) for 20 min at 37°C, followed by 20 min at 25°C. 10-µl of the reaction mix was directly used to transform chemically competent *E. coli* cells.

S. cerevisiae heterologous DNA expression was achieved by using low-copy-number-replicating plasmids, CEN-ARS, for the PPTases genes and chromosomal integration for the remaining genes. The yeast transformation protocol was performed as described by Gietz *et al.* [42]. Chromosomal integration was achieved by gene targeting through homologous recombination of linear DNA fragments, which was either a digested vector, or bipartite PCR fragments. *MET15* or *URA3* from *Kluyveromyces lactis* [43] were used as selection markers. The *URA3* marker used

for chromosomal integration was flanked by direct repeats that allowed restoring of uracil auxotrophy by plating the cells in agar medium containing 5-Fluoroorotic acid (5-FOA) after each genetic manipulation [44].

PPTase-coding genes subcloning and expression in S. cerevisiae VG0 strain

Plasmids pJH589, pJH592, pARB118 and pARB117 (**Table 4.1**) were used for heterologous expression of the PPTase-coding genes *entD*, PPTcg-1, *npgA and npt* (respectively) in strain VG0 (**Table 4.2**). The backbone plasmid is identical for all the genes, including the yeast auxotrophic marker (*URA3*) and the promoter responsible for transcriptional activation of the genes (P_{TPI1}), in order to reduce variability.

Table 4.1: Plasmids used in this study.

Plasmid	Gene content	Plasmid type	Selection marker	Reference
pJH665	*UGT72E2 (Arabidopsis thaliana)*	*CEN*-ARS (*S. cerevisiae*)	*URA3*	Hansen *et al.* [4]
pRS426	*n.a.*	Episomal (*S. Cerevisiae*)	*URA3*	Christianson *et al.* [45]
pWJ1042	Recyclable *URA3 (Kluyveromyces lactis)*	*CEN*-ARS (*S. Cerevisiae*)	*URA3*	Reid *et al.* [46]
pJH589	*endD* (*E. coli*)	*CEN*-ARS (*S. cerevisiae*)	*URA3*	Hansen *et al.* [4]
pJH592	PPTcg-1 (*C. Glutamicum*)	*CEN*-ARS (*S. Cerevisiae*)	*URA3*	Hansen *et al.* [4]
pESC-npgA-pcbAB	*npgA* (*Aspergillus nidulans*)	*CEN*-ARS (*S. Cerevisiae*)	*URA3*	Siewers *et al.* [19]
pJ204-npt	*npt* (*Nocardia sp.), synthetic codon optimized*	*n.a.*	*n.a.*	DNA-2-Go™
pJH674	*ACAR* (*Nocardia sp.), synthetic codon optimized*	Integration *(S. cerevisiae)*	*HphMX*	Hansen *et al.* [4]
pJH543	*hsOMT* (*Homo sapiens), synthetic codon optimized*	Integration *(S. Cerevisiae)*	*NatMX*	Hansen *et al.* [4]
pARB118	*npgA* (*Aspergillus nidulans*)	*CEN*-ARS (*S. Cerevisiae*)	*URA3*	This study
pARB117	*npt* (*Nocardia sp.), synthetic codon optimized*	*CEN*-ARS (*S. Cerevisiae*)	*URA3*	This study
pARB014	*endD* (*E. Coli*)	Integration (*S. Cerevisiae*)	*URA3*	This study
pARB010	*n.a.*	Integration *(S. Cerevisiae)*	*URA3*	This study
pARB030	*n.a.*	Integration *(S. Cerevisiae)*	*MET15*	This study
pARB032	*ACAR* (*Nocardia sp.), synthetic codon optimized*	Integration *(S. Cerevisiae)*	*MET15*	This study
pARB033	*hsOMT* (*Homo sapiens), synthetic codon optimized*	Integration *(S. Cerevisiae)*	*MET15*	This study
pARBU2000	*PacI*/Nt.BbvCI USER cassette [41]	Episomal (*S. Cerevisiae*)	*n.a.*	This study
pARBU2101	600 bp fragments flanking the integration site *YPRCΔ15* (*S. cerevisiae*) [21]	Episomal (*S. Cerevisiae*)	Recyclable *URA3*[a]	This study
pARBU2131	*hsOMT* (*Homo sapiens), synthetic codon optimized*	Episomal (*S. cerevisiae*)	Recyclable *URA3* [a]	This study
pARBU2191	*HAP4 (S. cerevisiae)*	Episomal (*S. Cerevisiae*)	Recyclable *URA3* [a]	This study

[a] *URA3* from *Kluyveromyces lactis* flanked by direct repeats for restoration of uracil auxotrophy [43].

Table 4.1 contains all the plasmids used and obtain in this study. Plasmid pARB118, containing the gene *npgA* from *Aspergillus nidulans,* was obtained from pJH665 by replacing the gene *UGT72E2* by *npgA,* using the restriction endonucleases *XbaI* and *KpnI*. The vector was treated with Antarctic Phosphatase (New England Biolabs) after restriction with *XbaI* in order to create a blunt end, and only then restricted with *KpnI*. The gene *npgA* was obtained by PCR amplification with primers O1_f and O2_r, using pESC-npgA-pcbAB as template. Plasmid ARB117, containing the gene *npt* from *Nocardia* sp. (codon optimized) was obtained from plasmid pJH665 by replacement of the gene *UGT72E2. XbaI* and *KpnI* were used to restrict pJ204-npt (DNA-2-Go™) in order to obtain the DNA fragment containing the gene, which was cloned in the restricted and purified vector pJH665 using the same enzymes.

ACAR*,* hsOMT *and* HAP4 *subcloning and overexpression in vanillin β-D-glucoside producing* S. cerevisiae *strains*

Plasmid pARB014 was obtained from pJH589 by restriction with *FseI* and recircularization of the purified vector without CEN-ARS. Plasmid pARB010 was obtained from pARB014 by restriction with *NotI* and recircularization of the purified vector without the P_{TPI1}–*entD*–T_{TPI1} fragment. pARB030 was obtained by replacing the *URA3* marker for *S. cerevisiae* in pARB010 by the marker *MET15*, using *AscI*. *MET15* (incl. native promoter, 450 bp, and terminator, 215 pb) was obtained by PCR amplification from CEN.PK 113-7D genomic DNA using primers O5_f and O6_r. pARB032 and pARB033 were obtained by inserting a fragment containing P_{TPI1}–*ACAR*–T_{TPI1} and P_{TPI1}–*hsOMT*–T_{TPI1}, respectively, on purified pARB030 after restriction with *Not1*. The fragments containing P_{TPI1}–*ACAR*–T_{TPI1} and P_{TPI1}–*hsOMT*–T_{TPI1} were obtained by *NotI* restriction of pJH674 and pJH543 [4], respectively.

pARBU2000 contains a *PacI*/*Nt.BbvCI* USER cassette [41] and it was the basic USER vector constructed and used in this study. It was obtained by restricting pRS426 with *XbaI* followed by recircularization in order to remove the *URA3* original *S. cerevisiae* marker. The intermediate vector was restricted with *NotI* in order to insert the fragment containing the *PacI*/*Nt.BbvCI* USER cassette. The *PacI*/*Nt.BbvCI* USER cassette (restricted with *NotI*) was kindly provided by Bjarne G. Hansen. pARBU2101 was obtained by USER cloning the *URA3* recyclable marker (*K. Lactis*) between the flanking regions of *S. cerevisiae* integration site #20 from Bai *et al.* (*YPRCΔ15*) [21]. The *URA3* recyclable marker was obtained by PRC amplification using primers UO1_f and UO2_r, and pWJ1042 [46] as template. The *YPRCΔ15* flanking regions were obtained by PRC amplification using primers UO3_f, UO4_r, UO5_f and UO6_r, and CEN.PK 113-7D genomic DNA as template. Note that primer UO5_f was designed such that the USER cassette is

recovered while the fragments are correctly assembled on the plasmid. pARBU2131 and pARBU2191 were obtained by USER cloning the promoter region of *TEF1* (400 bp) and the terminator region of *CYC1* (200 bp) flanking the genes *hsOMT* and *HAP4*, respectively, on pARBU2101. P_{TEF1}, T_{CYC1} and *HAP4* DNA fragments were obtained by PRC amplification using primers UO7_f and UO8_r, UO9_f and UO10_r and UO11_f and UO12_r, respectively while using CEN.PK 113-7D genomic DNA as template. *hsOMT* fragment was PCR amplified from pARB033 using primers UO13_f and UO14_r. pARBU2131 and pARBU2191 can be used for expressing *hsOMT* and *Hap4* from plasmid (multi-copy), as well as PCR template for obtaining the fragments for chromosomal integration (bipartite method) of the genes. In this work, we followed the second approach; the primers used to obtain the PCR fragments for chromosomal integration of *hsOMT* and *HAP4* were UO3_f, O10_r and O9_f, UO6_r.

Strains VG0 and VG4 by Brochado *et al.* [10] were used as background for the yeast strains obtained during this study. **Table 4.2** lists all the strains, as well as their relevant genotype.

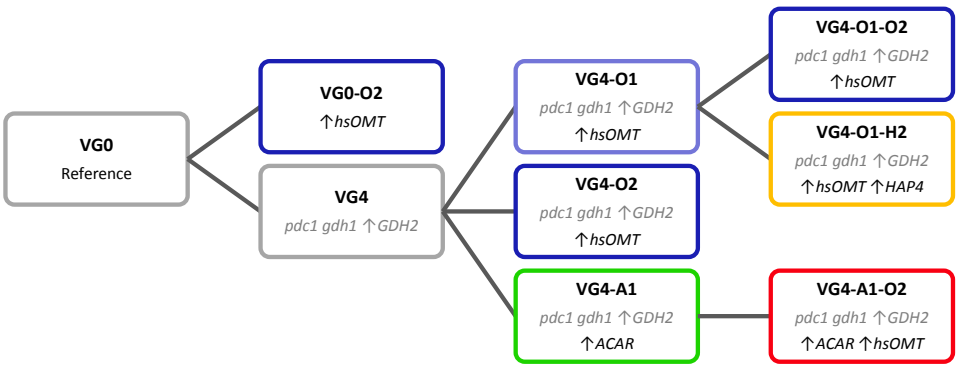
Table 4.2: Yeast strains obtained in this study.

Strain[a]	Relevant genotype	Vanillin pathway overexpressed genes
VG0	*met15*Δ0::*MET15*	
VG4	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0::*MET15*	
VG0-O2	*met15*Δ0::*MET15 yprcΔ15::hsOMT*	*hsOMT*
VG4-O1	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0::P_{TPI1}-*hsOMT[MET15]*	*hsOMT*
VG4-O2	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0::*MET15 yprcΔ15*::P_{TEF1}-*hsOMT*	*hsOMT*
VG4A	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0:: P_{TPI1}-*ACAR[MET15]*	*ACAR*
VG4-A1-O2	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0:: P_{TPI1}-*ACAR[MET15] yprcΔ15*:: P_{TEF1}-*hsOMT*	*ACAR & hsOMT*
VG4-O1-O2	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0:: P_{TPI1}-hs*OMT[MET15] yprcΔ15*:: P_{TEF1}-*hsOMT*	*hsOMT*
VG4-O1-H2	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0:: P_{TPI1}-hs*OMT[MET15] yprcΔ15*:: P_{TEF1}-*HAP4*	*hsOMT & HAP4*
VG4-*Δ2*	*pdc1 gdh1 GDH2*::P_{PGK1}-*GDH2 met15*Δ0::*MET15 yprcΔ15*	
VG0-*Δ2*	*met15*Δ0::*MET15 yprcΔ15*	

[a] All strains are derivatives from strain VG0 from Brochado *et al.* [10]. Only the relevant phenotype, not common to all the strains, is show in the table. Methionine prototrophy was restored in all the strains by gnomic integration of gene *MET15* in its original *locus*.

Scheme 3.1 contains an overview of all the strains obtained in this study. *hsOMT* overexpressed strains are represented in blue, *ACAR* overexpression is represented in green. Combined overexpression of *hsOMT* and *ACAR* is represented in red, while the combined overexpression of *hsOMT* and *HAP4* is represented in yellow. Strains not containing additional overexpression of any of the genes form the vanillin biosynthetic pathway are represented in gray.

VG0 is the reference strain used for this study (gray, **Scheme 3.1**). VG4 was previously engineered for improved vanillin production and it lacks the genes *PDC1* and *GDH1*, while *GDH2* is overexpressed. *hsOMT* overexpressed strain VG4-O1 was obtained from VG4 by chromosomal integration of the liberalized vector ARB033 (*Nar1*). Strains VG0-O2, VG4-O2 and VG4-O1-O2 (dark-blue) where obtained from strains VG0, VG4 and VG4-O1, respectively, by chromosomal integration of a bipartite substrate containing the recyclable *URA3* marker, as well as hsOMT under P_{TEF1}. Therefore, these strains are also hsOMT overexpressed, but the genomic integration site (*YPRCΔ15*) and promoter differs from those of VG4-O1. The uracil auxotrophy of these strains was re-established after isolation in 5-FoA [43]. Note that this is a fundamental step, since vanillin β-D-glucoside production requires the expression of a PPTase, which occurs from a low-copy-number plasmid with *URA3* as selective marker. ACAR overexpressed strain VG4A (green, **Scheme 1**) was obtained from strain VG4 by chromosomal integration of liberalized vector ARB032 (*Nar1*). ACAR and hsOMT overexpressed strain VG4-A1-O2 (red, **Scheme 1)** was obtained from VG4A by chromosomal integration of a bipartite substrate (recyclable *URA3* marker and hsOMT under P_{TEF1}) similarly to VG4-O2. *hsOMT* and *HAP4* overexpressed strain VG4-O1-H2 (yellow, **Scheme 1**) was obtained from VG4-O1 by chromosomal integration of a bipartite substrate containing the recyclable *URA3* marker and *HAP4* under P_{TEF1}.



Scheme 3.1: Overview of the yeast strains obtained during this study. VG0 and VG4 (gray) are the reference and previously metabolically engineered strains, respectively. The relevant genotype of the strains is represented bellow the name (gray, in the cases of genetic modifications carried out in previous work). Letters *A, H* and *O* represent the overexpressed genes *ACAR*, *HAP4* and *hsOMT,* respectively. Numbers *1* and *2* following *A, H* and *O* correspond to integration sites *MET15* and *YPRCΔ15*, respectively. VG4-O1 (light-blue) and VG4-O1-O2 (dark-blue) have *hsOMT* overexpressed from the *MET15 locus*. VG0-O2, VG4-O2, VG4-O1-O2 (dark-blue) and VG4-A1-O2 (red) have *hsOMT* overexpressed from the *YPRCΔ15 locus*. VG4-A1 (green) and VG4-A1-O2 (red) have ACAR overexpressed from the *MET15 locus*. VG4-O1-H2 (yellow) is has *Hap4* overexpressed from the *YPRCΔ15 locus*.

Media composition

Lysogeny broth (LB) medium was used to grow *Escherichia coli* cells for cloning purposes. The medium composition is as follows: 10 g/L of tryptone, 5 g/L of yeast extract and 10 g/L of NaCl. The medium pH was adjusted to 7.5 with NaOH/HCl 2N, before autoclavation and sterile filtered ampicillin was added to a final concentration of 100 mg/L, when needed, after autoclavation. For solid medium, 20 g/L agar was added prior to autoclavation.

YPD medium was used to grow *S. cerevisiae* cells for cloning purposes. The medium composition is as follows: 10 g/L of yeast extract, 20 g/L of peptone and 20 g/L of glucose (glucose was autoclaved separately). For solid medium, 20 g/L agar was added prior to autoclavation.

Synthetic complete (SC) dropout media, SC URA- and SC Met-, were used as selective media to grow *S. cerevisiae* for cloning purposes. The SC medium composition is as follows: 6.7 g/l of yeast nitrogen base without amino acids, 40 mg/l of adenine sulphate, 20 mg/l of L-arginine, 100 mg/l of L-aspartic acid, 100 mg/l of L-glutamic acid, 20 mg/l of L-histidine, 60 mg/l of L-leucine, 30 mg/l of L-lysine, 20 mg/l of L-methionine, 50 mg/l of L-phenylalanine, 375 mg/l of L-serine, 200 mg/l of L-threonine, 40 mg/l of L-tryptophan, 30 mg/l of L-tyrosine, 150 mg/l of L-valine, 20 mg/l of L-uracil and 20 g/l glucose. SC URA- and SC Met- media lack uracil and methionine, respectively. Glucose was autoclaved separately. The pH was adjusted to 5.8 prior to autoclavation. For solid medium, 20 g/L agar was added prior to autoclavation.

A defined minimal medium as described by Verduyn *et al.* (1992) [11] with 20 g/l glucose as sole carbon source was used for *S. cerevisiae* cultivation in batch cultivation in 24-well microtiter plates [47]. The medium composition is as follows: 7.5 g/l $(NH_4)_2SO_4$, 14.4 g/l KH_2PO_4, 0.5 g/l Mg_2SO_4, 2.0 ml/l trace metal solution, 1.0 ml/l vitamins solution, 0.05 ml/l antifoam 204 (Sigma-Aldrich A-8311) and 100 mg/l L-methionine when needed. Trace metal solution contained 3 g/L $FeSO_4.7H_2O$, 4.5 g/L $ZnSO_4.7H_2O$, 4.5 g/L $CaCl_2.6H_2O$, 0.84 g/L $MnCl_2.2H_2O$, 0.3 g/L $CoCl_2.6H_2O$, 0.3 g/L $CuSO_4.5H_2O$, 0.4 g/L $NaMoO_4.2H_2O$, 1 g/L H_3BO_3, 0.1 g/L KI and 15 g/L $Na_2EDTA.2H_2O$. Vitamins solution included 50 mg/l d-biotin, 200 mg/l *para*-amino benzoic acid, 1.0 g/l nicotinic acid, 1.0 g/l Ca-pantothenate, 1.0 g/l pyridoxine HCL, 1.0 g/l thiamine HCl and 25 mg/l m-inositol. The pH was adjusted to 6 by addition of 2N NaOH prior to autoclavation, the glucose was autoclaved separately and methionine and vitamins solutions were sterile filtered (0.2 µm pore-size Ministart®-Plus Sartorius AG, Geottingen, Germany) and added after autoclavation.

Microtiter plates batch cultivation

S. cerevisiae cell cultivation was carried out in white, flat-bottom, 24-well microtiter plates (Cornig Inc., NY) covered by a sterile sandwich-lid with a black rubber inner layer (Enzyscreen, The Netherlands) to prevent contamination. The cultures were grown in minimal medium, initial $OD_{600\ nm}$ of 0.1 in a total volume of 900 µl per well, using Growth Porfiler 1152 (Enzyscreen, The Netherlands), shaking frequency of 200 rpm, with controlled temperature at 30 ^{o}C.

Biomass determination

Samples were maintained at 4 ^{o}C post sampling and the biomass concentration was monitored by optical density at 600 nm ($OD_{600\ nm}$). $OD_{600\ nm}$ was measured throughout all the fermentation in a Shimadzu UV mini 1240 spectrophotometer (Shimadzu Europe GmbH, Duidberg, Germany). The samples were diluted with distilled water in order to obtain measurements in the linear range of 0 to 0.6 $OD_{600\ nm}$

Glucose and external metabolites analysis

The fermentation samples were filtered using a 0.45 µm pore-size syringe-filter (Sartorius AG, Geottingen, Germany) and stored at -20 ^{o}C until further analysis. Glucose, ethanol, glycerol, pyruvate, succinate and acetate were determined by high performance liquid chromatography (HPLC) analysis using an Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories, Hercules, CA). The column temperature was kept at 60 ^{o}C and the elution was performed using 5 mM H_2SO_4 with flow rate of 0.6 ml/min. Metabolites detection was performed by a RI-101differential refractometer detector (Shodex) and an UVD340U absorbance detector (Dionex) set at 210 nm.

Extracellular vanillin, vanillin β-D-glucoside (VG), protocatechuic acid (PAC), protocatechuic aldehyde (PAL) and vanillic acid were quantified by high performance liquid chromatography (HPLC) using Agilent 1100 series equipment with a Luna C18 column (Phenomenex). A gradient of methanol (+ 1% tetra-fluoroacetic acid) and water (+ 1% tetra-fluoroacetic acid) at a flow rate of 0.3 ml/min was used as mobile phase. The column was kept at 300 bar and 30 ^{o}C. Metabolite detection was performed using a UV diode-array detector set at 280 and 310 nm.

Acknowledgments

ARB thanks Ana Machás, Małgorzata E. Bejtka, Emauela Monteiro and Joanna Guzicka for helpful contribution for establishing the molecular biology tools used for constructing all the mutants. ARB thanks Bjarne G. Hansen for all the valuable discussions regarding USER cloning. ARB thanks Louise Mølgaard by all valuable discussions regarding general cloning information, as well as for kindly sharing some important plasmids.

References

1. Walton N: **Vanillin**. *Phytochemistry* 2003, **63**:505-515.

2. Lirdprapamongkol K, Kramb J-P, Suthiphongchai T, et al.: **Vanillin suppresses metastatic potential of human cancer cells through PI3K inhibition and decreases angiogenesis in vivo.** *Journal of Agricultural and Food Chemistry* 2009, **57**:3055-63.

3. Liang J-an, Wu S-lu, Lo H-yi, Hsiang C-yun, Ho T-yun: **Vanillin Inhibits Matrix Metalloproteinase-9 Expression through Down-Regulation of Nuclear Factor- kB Signaling Pathway in Human Hepatocellular Carcinoma Cells**. *Molecular Pharmacology* 2009, **75**:151-157.

4. Hansen EH, Møller BL, Kock GR, et al.: **De novo biosynthesis of vanillin in fission yeast (Schizosaccharomyces pombe) and baker's yeast (Saccharomyces cerevisiae).** *Applied and Environmental Microbiology* 2009, **75**:2765-74.

5. Ander P, Hatakka A, Eriksson K-erik: **Vanillic Acid Metabolism by the White-Rot Fungus Sporotriehum pulverulentum**. *Enzyme* 1980, **202**:189-202.

6. Chatterjee T, De B, Bhattacharyya D: **Microbial conversion of isoeugenol to vanillin by Rhodococcus rhodochrous**. *Indian Journal of Chemistry* 1999, **38**:538-541.

7. Civolani C, Barghini P, Roncetti a R, Ruzzi M, Schiesser A: **Bioconversion of ferulic acid into vanillic acid by means of a vanillate-negative mutant of Pseudomonas fluorescens strain BF13.** *Applied and Environmental Microbiology* 2000, **66**:2311-7.

8. Gioia D Di, Luziatelli F, Negroni A, et al.: **Metabolic engineering of Pseudomonas fluorescens for the production of vanillin from ferulic acid.** *Journal of Biotechnology* 2010, **156**:309-16.

9. Srivastava S, Luqman S, Khan F, Chanotiya CS, Darokar MP: **Metabolic pathway reconstruction of eugenol to vanillin bioconversion in Aspergillus niger**. *Bioinformation* 2010, **4**:320-325.

10. Brochado AR, Matos C, Møller BL, et al.: **Improved vanillin production in baker's yeast through in silico design.** *Microbial Cell Factories* 2010, **9**:84.

11. Nielsen J: **Metabolic engineering**. *Applied Microbiology and Biotechnology* 2001, **55**:263-283.

12. Nielsen J, Jewett MC: **Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae.** *FEMS yeast research* 2008, **8**:122-31.

13. Lee JW, Kim TY, Jang Y-S, Choi S, Lee SY: **Systems metabolic engineering for chemicals and materials.** *Trends in Biotechnology* 2011, **29**:370-8.

14. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:15112-15117.

15. Patil KR, Rocha I, Förster J, Nielsen J: **Evolutionary programming as a platform for in silico metabolic engineering.** *BMC Bioinformatics* 2005, **6**:308.

16. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13**:244-53.

17. Guo S, Bhattacharjee JK: **Posttranslational activation, site-directed mutation and phylogenetic analyses of the lysine biosynthesis enzymes alpha-aminoadipate reductase Lys1p (AAR) and the phosphopantetheinyl transferase Lys7p (PPTase) from Schizosaccharomyces pombe.** *Yeast* 2004, **21**:1279-88.

18. Wattanachaisaereekul S, Lantz AE, Nielsen ML, Nielsen J: **Production of the polyketide 6-MSA in yeast engineered for increased malonyl-CoA supply.** *Metabolic Engineering* 2008, **10**:246-54.

19. Siewers V, Chen X, Huang L, Zhang J, Nielsen J: **Heterologous production of non-ribosomal peptide LLD-ACV in Saccharomyces cerevisiae.** *Metabolic Engineering* 2009, **11**:391-7.

20. Venkitasubramanian P, Daniels L, Rosazza JPN: **Reduction of carboxylic acids by Nocardia aldehyde oxidoreductase requires a phosphopantetheinylated enzyme.** *The Journal of Biological Chemistry* 2007, **282**:478-85.

21. Bai D, Siewers V, Huang L, Nielsen J: **Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae**. *Yeast* 2009, **26**:545-551.

22. Yamane S, Yamaoka M, Yamamoto M, et al.: **Region specificity of chromosome III on gene expression in the yeast Saccharomyces cerevisiae.** *The Journal of General and Applied Microbiology* 1998, **44**:275-281.

23. Thompson A, Gasson MJ: **Location Effects of a Reporter Gene on Expression Levels and on Native Protein Synthesis in Lactococcus lactis and Saccharomyces cerevisiae**. *Applied and Environmental Microbiology* 2001, **67**:3434-3439.

24. Yuan G-C, Liu Y-J, Dion MF, et al.: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science* 2005, **309**:626-30.

25. Atsumi S, Hanai T, Liao JC: **Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels.** *Nature* 2008, **451**:86-9.

26. Radakovits R, Eduafo PM, Posewitz MC: **Genetic engineering of fatty acid chain length in Phaeodactylum tricornutum.** *Metabolic Engineering* 2011, **13**:89-95.

27. Ro D-K, Paradise EM, Ouellet M, et al.: **Production of the antimalarial drug precursor artemisinic acid in engineered yeast.** *Nature* 2006, **440**:940-3.

28. Yim H, Haselbeck R, Niu W, et al.: **Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol.** *Nature Chemical Biology* 2011, **7**:444-445.

29. Farmer WR, Liao JC: **Improving lycopene production in Escherichia coli by engineering metabolic control.** *Nature Biotechnology* 2000, **18**:533-7.

30. Gancedo JM: **Yeast carbon catabolite repression.** *Microbiology and Molecular Biology Reviews* 1998, **62**:334-61.

31. Carlson M: **Glucose repression in yeast**. *Current Opinion in Microbiology* 1999, **2**:202-207.

32. Fraenkel DG: *Yeast Intermediary Metabolism*. Cold Spring Harbor Laboratory Press; 2011.

33. Maris a J van, Bakker BM, Brandt M, et al.: **Modulating the distribution of fluxes among respiration and fermentation by overexpression of HAP4 in Saccharomyces cerevisiae.** *FEMS Yeast Research* 2001, **1**:139-49.

34. Westergaard SL, Oliveira AP, Bro C, Olsson L: **A Systems Biology Approach to Study Glucose Repression in the Yeast Saccharomyces cerevisiae**. *Biotechnology* 2007, **96**:134-145.

35. Blom J, Mattos MJTD, Grivell LA: **Redirection of the Respiro-Fermentative Flux Distribution in Saccharomyces cerevisiae by Overexpression of the Transcription Factor Hap4p**. *Applied and Environmental Microbiology* 2000, **66**:1970-3.

36. Dueñas-Sánchez R, Codón AC, Rincón AM, Benítez T: **Increased biomass production of industrial bakers' yeasts by overexpression of Hap4 gene.** *International Journal of Food Microbiology* 2010, **143**:150-60.

37. Leonard E, Kumaran P, Thayer K, Xiao W-hai, Mo JD: **Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:13654-13659.

38. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.** *BMC Systems Biology* 2010, **4**:53.

39. Osterlund T, Nookaew I, Nielsen J: **Fifteen years of large scale metabolic modeling of yeast: Developments and impacts.** *Biotechnology Advances* 2011.

40. Nørholm MHH: **A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering.** *BMC Biotechnology* 2010, **10**:21.

41. Nour-Eldin HH, Hansen BG, Nørholm MHH, Jensen JK, Halkier B a: **Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments.** *Nucleic Acids Research* 2006, **34**:e122.

42. Gietz D, St Jean A, Woods RA, Schiestl RH: **Improved method for high efficiency transformation of intact yeast cells.** *Nucleic Acids Research* 1992, **20**:1425.

43. Erdeniz N, Mortensen UH, Rothstein R: **Cloning-Free PCR-Based Allele Replacement Methods**. *Genome Research* 1997:1174-1183.

44. Boeke JD, Lacroute F, Fink GR: **Short communication A positive selection for mutants lacking orotidine-5 ' -phosphate decarboxylase activity in yeast : 5-fluoro-orotic acid resistance**. *Molecular and General Genetics* 1984:345-346.

45. Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P: **Multifunctional yeast high-copy-number shuttle vectors.** *Gene* 1992, **110**:119-22.

46. Reid RJD, Sunjevaric I, Keddache M, Rothstein R, Kedacche M: **Efficient PCR-based gene disruption in Saccharomyces strains using intergenic primers.** *Yeast* 2002, **19**:319-28.

47. Verduyn C, Postma E, Scheffers WA, Dijken JP Van: **Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation.** *Yeast* 1992, **8**:501-17.

Chapter 5: Minimization of Metabolite Balance Explains Genetic Interactions within the Yeast Metabolic Network

This chapter is based on the manuscript:

"Minimization of Metabolite Balance explains complex genetic interactions within the yeast metabolic network". **Ana Rita Brochado**, Sergej Andrejev, Costas D. Maranas and Kiran R. Patil. ***Sumbitted***.

Abstract

Genome-scale metabolic networks provide a comprehensive structural framework for integrative data analysis as well as for metabolic modeling. The solution space for the metabolic flux state of the cell is typically very large and optimization-based approaches are often necessary for identifying the active metabolic state under specific environmental conditions. The objective function to be used is problem dependent and one of the most relevant parameters for successful modeling. Although linear combination of selected fluxes is widely used for formulating metabolic objective functions, we found that the mathematical implementation of these optimization problems is context dependent owing to the non-uniqueness of the stoichiometric coefficients scaling. We hereby propose a new method, Minimization of Metabolite Balance (MiMBl), which allows consistent formulation of the desired biological principles, thereby providing robust predictions for gene essentiality and genetic interactions. We demonstrate how MiMBl can be used together with existing algorithms to provide mechanistic insights into the yeast genetic interactions network.

Introduction

The fundamental role of metabolism within a living cell has become a focal point of study in many disciplines, such as cell biology, physiology, medicine and synthetic biology. The assembly of all reactions and metabolites into a genome-scale metabolic network provides a comprehensive structural framework for integrative data analysis [1, 2], as well as for quantitative modeling of cellular metabolism [3–6]. As the solution space for the metabolic flux state of the cell is typically very large, constraint based optimization approaches are often applied for simulating metabolic fluxes. In essence, these approaches search for an optimal flux distribution that maximizes or minimizes an appropriate biological objective function while satisfying the mass balance and metabolite exchange constraints. Among these, Flux Balance Analysis [7] is a widely used simulation tool that utilizes a linear programming formulation for maximization of growth (synthesis of biomass constituents) as biological objective function. FBA has been applied with various degrees of success, albeit mostly for "wild-type" or unperturbed metabolic networks [8, 9]. In addition to FBA, different frequently used objective functions were reviewed and tested against experimental data by Schuetz *et al*. [10], including minimization of overall intracellular flux, maximization of biomass or ATP yield, among others. In case of genetically or environmentally perturbed networks, Minimization of Metabolic Adjustment algorithm - MoMA [11] - has been reported to better represent the biological observations [11–14]. The hypothesis underlying MoMA is that the fluxes in a perturbed cell (*e.g.* a mutant) will be redistributed so as to be as similar as possible to the wild-type.

The biological principles exemplified by simulation tools for both wild-type and perturbed networks are undeniably fascinating, which is confirmed by their numerous applications – including prediction of genetic interactions [2, 15, 16], metabolic engineering [13, 14, 17] and microbial community modeling [18]. Several of the commonly used objective functions rely on the use of linear combination of fluxes, e.g., MoMA or maximization of ATP production (**Table 5.1**). We found that the mathematical formulation of this class of problems (i.e. where linear combination of fluxes is part of the objective function) is sensitive to the representation of the reaction stoichiometry with results strongly dependent on the adopted scaling of the stoichiometric coefficients. We find that this non-uniqueness in prediction may confound the interpretation of the intended biological objective function. The stoichiometric representation of any reaction is subjective, often scaled to have coefficient of 1 for one of the reactants/products. Importantly, as the modeling problem involves hundreds of reactions, there is no unique possible stoichiometric representation of the metabolic network.

We motivate the need for rethinking the problem formulation for metabolic modeling by illustrating how the current methods lead to incoherent biological predictions when alternatively representing the reaction stoichiometry. Tackling a proper problem formulation, we propose a new methodology for metabolic modeling – Minimization of Metabolite Balance (MiMBl), which accounts for reaction stoichiometry in the objective function by mapping the flux space into the metabolite space. As intended, MiMBl shows robust predictions independently of the stoichiometry representation. We demonstrate the biological relevance of the new formulation with increased power for predicting genetic interactions in the metabolic network of *S. cerevisiae*. In a recent study reporting a large genetic interactions dataset covering the *S. cerevisiae* metabolic network [2], FBA was found to have limited capability for predicting the experimentally observed interactions, partially due to the lack of regulatory information. Within this study we successfully challenged MiMBl to accomplish the task of extending the range of genetic interactions that can be predicted. By combining the results from MiMBl and FBA we obtain novel insights into the operating mechanisms underlying genetic interactions within metabolic networks.

Results and Discussion

Minimization of sum of fluxes

Minimization of the sum of intracellular flux is a routinely used objective function for estimating intracellular fluxes [10, 19, 20]. By using the *i*FF708 *S. cerevisiae* genome-scale metabolic model [21], we illustrate how the use of this objective function leads to inconsistent predictions when using different, but equivalent, reaction stoichiometry. *i*FF708 was a model of choice for this illustration as it is one of the highly curated yeast models that is most suitable for predicting intracellular flux distributions. On the other hand, for studying large scale genetic interactions in yeast we will use the more recent and comprehensive model *i*AZ900 (also manually curated) which allows us to analyse a larger part of the dataset. Linear scaling of all stoichiometric coefficients of a given reaction (*e.g.* multiplication by a scalar θ, **Methods**) preserves the stoichiometry and ought not to affect the simulation outcome for a correct problem formulation. However, in this case, scaling of a single reaction (RPI1) results in diverting the carbon flow from glycolysis to pentose phosphate pathway, which is one of the most important metabolic branch points (**Figs. 4.1 and S1**). This deviation was verified not to be consequence of

alternative optima of the same mathematical solution (**Fig. S4.1**), thus representing divergent biological solutions.

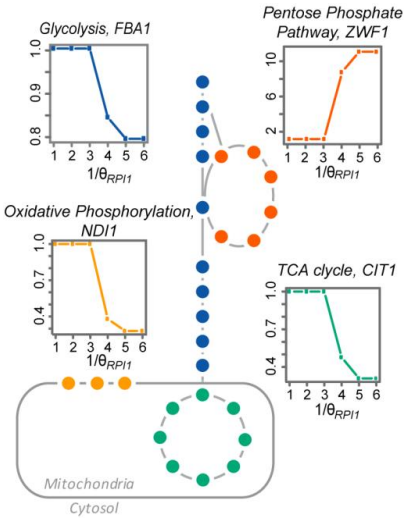


Figure 5.1: Minimization of overall intracellular flux leads to divergent predictions for flux distribution when using alternative stoichiometry representations. Shown are predicted fluxes through key pathways within the S. cerevisiae central carbon metabolism, using alternative stoichiometric representation of reaction RPI1 (θRPI1, Methods). θ_{RPI1} is represented on the x-axis, while fold-change of fluxes relatively to θ=1 is represented on the y-axis. A representative reaction from each of the pathways was selected to illustrate the flux re-arrangement; FBA1 for glycolysis, ZWF1 for pentose phosphate pathway, CIT1 for tricarboxilic acid cycle and NID1 for oxidative phosphorylation. Note that θ=1 is an arbitrary reference, as the stoichiometric representation of any reaction is subjective, often scaled to have coefficient of 1 for one of the reactants/products.

In order to understand the nature of the problem leading to the susceptibility of the solution towards alternative representation of the stoichiometric matrix, we used a toy-model (**Fig. 5.2a**). As a case study, *minimization of metabolic adjustment* was chosen as a biological principle and formulated as minimization of Manhattan Distance (most commonly used formulation of MoMA, termed lMoMA [22]). In the wild-type toy-model, flux goes from *A* to *D* via *R5*. The goal is to predict flux distribution in the mutant lacking *R5*. The biological principle of *minimization of metabolic adjustment* dictates rewiring of the flux through *R6*. However, lMoMA found contradictory optimal solutions, i.e. solutions that re-route the flux via *R2-R3-R4* or *R6*, depending on the stoichiometric representation of *R6* (**Fig. 5.2b**). Insight into the cause of this behavior can be gained by analyzing the optimal objective function values, *i.e.* distances, as function of θ_{R6} (**Fig. 5.2d**). Decreasing θ_{R6} (i.e. increasing $1/\theta_{R6}$) implies higher numerical value of the flux through *R6*, hence higher contribution of *R6* to the distance. Consequently, after a

certain value of θ_{R6}, the activation of the longer *R2-R3-R4* pathway more than compensates the use of *R6*. The two solutions are not alternative optima, as the objective function value (i.e. distance) neither remains constant nor linearly scales with θ_{R6}. Such non-linear dependency of the objective function value on the scalar θ_{R6} violates the requirement of a correct problem formulation. The analytical proof is presented in the Methods section and shows that the optimality condition for the linear programming problem after scaling is not guaranteed to be satisfied. Notably, widely used FBA-like problems (max/minimization of a single flux as objective function) are perfectly robust concerning scaling of the stoichiometric coefficients. As a single flux is used objective function, the relative values of all the remaining fluxes (which depend on the stoichiometry representation) does not influence the optimal solution to be found (for theoretical proof, see **Supplementary notes**).

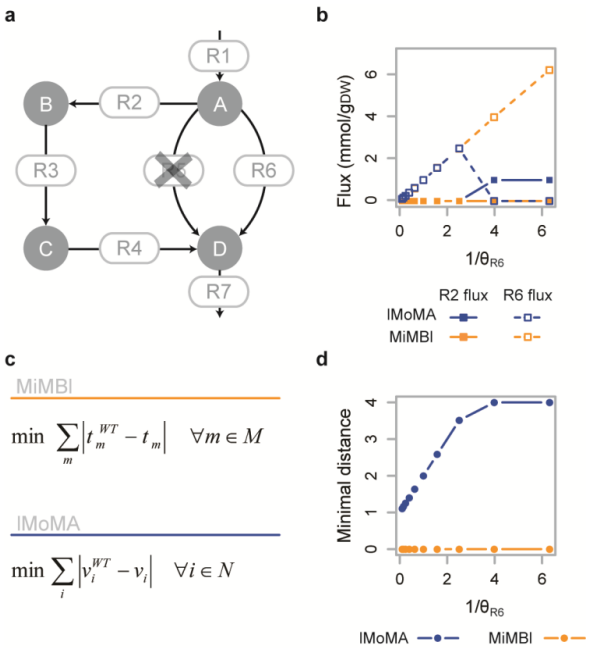


Figure 5.2: A toy-model illustrating how, and why, alternative stoichiometry representations influence simulation of minimization of metabolic adjustment by using lMoMA or MiMBl. a) Toy-model: *R1* to *R7* and *A* to *D* represent reactions and metabolites, respectively. In the wild-type, or reference, flux goes from *A* to *D* via *R5*. *R6* and *R2-R3-R4* are two alternative pathways for flux re-distribution after deletion of *R5*. **b)** Flux through reactions *R2* (full symbols) and *R6* (open symbols) obtained after simulation with lMoMA (blue) and MiMBl (yellow) using alternative representations of reaction *R6* (given by different θ_{R6}, Methods). c) Formulation of objective functions for lMoMA and MiMBl (Methods). d) Optimal objective function value (distance) obtained for *minimization of metabolic adjustment* using lMoMA (blue) and MiMBl (yellow) as function of θ_{R6}.

The mathematical caveat illustrated above means that the contribution of the desired biological objective function towards the obtained solution is inseparable from that of the artifacts of

stoichiometry representation. Importantly, in large metabolic networks the effects of stoichiometric representation of reactions are cumulative. As shown in the following, this problem can be solved by proper normalization of the objective function variables with respect to stoichiometric representation of the reactions. To achieve such normalization, we devised two approaches, normalized lMoMA (normlMoMA) and Minimization of Metabolite Balance (MiMBl). In normlMoMA, each variable in the objective function is normalized by its value in the wild-type flux distribution. Albeit being simple, this normalization method has three major drawbacks: i) many reactions often have null fluxes in the wild-type, thus posing a problem for normalization (**Methods, Supplementary notes** and **Fig. S4.2**); ii) it requires a reference flux distribution to obtain the normalization factors, making it inappropriate to formulate objective functions such as minimization of overall intracellular flux; and iii) the influence of each flux on the metabolic adjustment would be exclusively due to its fold change, not taking into account that reactions carrying higher fluxes could have a stronger impact on the predicted flux distribution, as implied in the original concept of *minimization of metabolic adjustment*.

Minimization of Metabolites Balance - MiMBl

To obtain a biological meaningful and mathematically robust normalization, we propose Minimization of Metabolite Balance (MiMBl) as a new method for metabolic modeling. The objective function in MiMBl is formulated as a linear combination of metabolite turnovers (t_M). The turnover of a metabolite is the sum of all fluxes producing (or consuming) it, multiplied by the corresponding stoichiometric coefficients (**Methods**). The objective function for *minimization of metabolic adjustment* is reformulated to include metabolite turnovers instead of fluxes **(Fig. 5.2c)**. Because the stoichiometric coefficients are taken into account while calculating t_M, MiMBl is robust to the linear scaling of the stoichiometric matrix, analytical proof of which is presented in the **Methods** section. In case of the toy-model (**Fig. 5.2a, d**), this robustness is illustrated by the invariant nature of the objective function as well as the flux distribution. Note that the flux through R6 linearly scales with θ_{R6}, while the turnover of all metabolites is conserved. The normalization implied in MiMBl formulation is suitable for addressing a variety of biological questions involving different objective functions, such as minimization of overall flux (by using a null vector for wild-type flux distribution) or maximization of ATP yield (by maximizing the ATP turnover for a given substrate uptake rate), among others (**Table 5.1**).

Table 5.1: Formulation of different biological objective functions using MiMBl.

Biological objective function	Previous objective function	Description	MiMBl objective function	Description
Minimization of metabolic adjustment	$\min \sum_{i \in N} \left\| v_i^{WT} - v_i \right\|$ [11]	Minimization of Manhattan distance between the vectors containing the reference and mutant flux distributions	$\min \sum_{m \in M} \left\| t_m^{WT} - t_m \right\|$	Minimization of Manhattan distance between the vectors containing the reference and mutant intracellular metabolites turnover
Minimization of overall intracellular flux	$\min \sum_{i \in N} \left\| v_i \right\|$ [19]	Minimization of the sum of all intracellular fluxes	$\min \sum_{m \in M} \left\| t_m^{WT} - t_m \right\|$ $t^{WT} = 0$	Minimization of the sum of intracellular metabolites turnover
Maximization of ATP yield	$\max \frac{\sum_{v_{ATP} \in N} v_{ATP}}{v_{\text{glucose}}}$ [23]	Maximization of the sum of all reactions producing ATP	$\max \frac{\left\| t_{ATP}^{WT} - t_{ATP} \right\|}{v_{\text{glucose}}}$ $t^{WT} = 0$	Maximization of ATP turnover
Minimization of ATP yield	$\min \frac{\sum_{v_{ATP} \in N} v_{ATP}}{v_{\text{glucose}}}$ [24]	Minimization of the sum of all reactions producing ATP	$\min \frac{\left\| t_{ATP}^{WT} - t_{ATP} \right\|}{v_{\text{glucose}}}$ $t^{WT} = 0$	Minimization of ATP turnover
Minimization of redox potential	$\min \frac{\sum_{v_{NADH} \in N} v_{NADH}}{v_{\text{glucose}}}$ [23]	Minimization of the sum of all reactions producing NADH	$\min \frac{\left\| t_{NADH}^{WT} - t_{NADH} \right\|}{v_{\text{glucose}}}$ $t^{WT} = 0$	Minimization of NADH turnover
Maximization of biomass*	$\max v_{Growth}$ (Varma & Palsson, 1993, Price *et al*, 2004)	Maximization of biomass flux	$\max t_{biomass}$	Maximization of biomass turnover

***Note:** Typically, biomass production within metabolic models is formulated as one equation accounting for all the biomass constitutes. As there is usually only one equation producing biomass, the flux of such reaction equals the biomass turnover. Therefore, for simulating such biological objective, the original formulation – FBA – and MiMBl are completely equivalent.

While mapping the flux space into the metabolite space for the objective function formulation, as we do for MiMBl, it is possible that, for a few cases, alternative flux distributions are found around a given metabolite. We therefore introduce a second optimization step that reinforces the proximity to the reference flux distribution. This is achieved by using a normlMoMA routine where the optimal objective function value found in the first MiMBl optimization step is used as an additional constraint (**Methods**). Nevertheless, highly connected metabolites ensure a degree of network connectivity, which is sufficient for decreasing the number of situations where alternative flux distributions around the same metabolite are picked by MiMBl. Indeed, we did

not find any case in the simulations performed for this study where growth prediction was altered in the second optimization step. An example case where the second optimization step will be more relevant is simulations involving export of metabolites, where the choice of a particular transporter (as in the reference flux distribution) among several alternative options is desired.

In order to estimate to what extent the lack of normalization of stoichiometric coefficients within the objective function influences the biological interpretation of simulation results, we used lMoMA for simulating gene knockouts in the *S. cerevisiae* genome-scale metabolic model *i*FF708 [21]. In case of simulations for triple gene knockouts, more than 200,000 triplets were found such that their predicted phenotype switched from lethal to non-lethal (or vice-versa) for two synonymous matrix representations (**Table ST5.1**). From a biotechnological perspective, predictions from genome-scale modeling have direct influence on the choice of gene targets selected for metabolic engineering. By using lMoMA, we identified metabolic engineering strategies (by simulating all possible combinations of knockout of (up to) three genes, **Methods**) for production of two different compounds in yeast: succinate – a native product, and vanillin-glucoside – a heterologous product. Not only a significant fraction of mutants had divergent predictions for product yield when using two alternative stoichiometric matrices, but also several highly ranked strategies in one case were low priority targets in the other (**Fig. 5.3a, b, c)**. Moreover, we also observed that the number of predicted synthetic lethal pairs differed by more than two-fold when using alternative stoichiometric matrix representations (**Table ST5.2**). These inconsistencies have immediate implications on the consequent biological interpretation, as well as on the experimental design, and can be successfully overcome by using MiMBl (**Fig. 5.3d, Table ST5.3**).

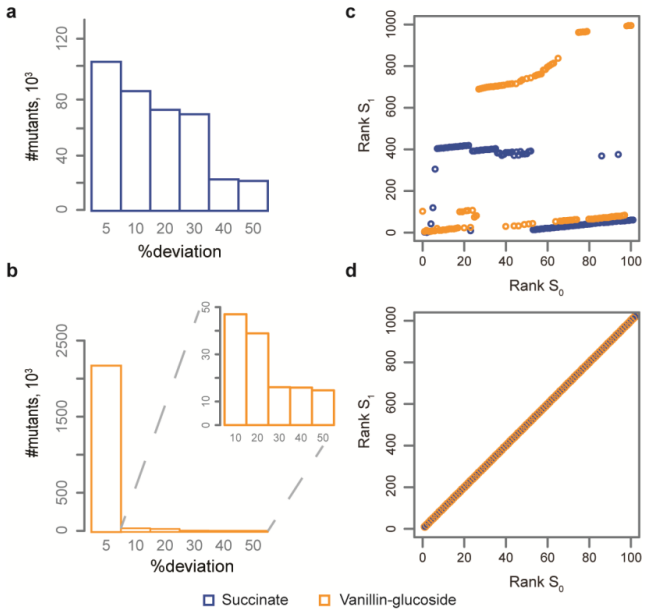


Figure 5.3: **Stoichiometry representation impacts the design of metabolic engineering strategies depending on the nature of the objective function formulation.** Shown is the comparison of predicted succinate and vanillin-glucoside yields for triple gene knockout mutants obtained with two alternative stoichiometric matrices (S_0 and S_1, **Methods**). Number of mutants diverging in their lMoMA-predicted **a)** succinate and **b)** vanillin-glucoside yields for the two alternative representations of stoichiometry. The x-axis represents the percentage of deviation of product formation by the mutants relative to S_0. **c)** Comparison of ranks of lMoMA-predicted metabolic engineering strategies for improving succinate and vanillin-glucoside production, obtained by using S_0 and S_1. **d)** Comparison of ranks of MiMBl-predicted metabolic engineering strategies for improving succinate and vanillin-glucoside production, obtained by using S_0 and S_1.

Predicting genetic interactions using MiMBl

To what extent MiMBl contributes for increasing biological understandings gained from the application of optimization-based metabolic modeling? We used one of the most recent and comprehensive *S. cerevisiae* models, *i*AZ900 [26], to run simulations for single and double gene knockouts and challenged MiMBl to predict the epistasis scores of all significantly interacting non-essential gene pairs reported by Szappanos *et al.* [2]. Genetic interaction networks are valuable resources towards deciphering the complex genotype-phenotype relationship. A genetic interaction between two genes occurs when the phenotype displayed by a double deletion mutant is different than the one expected based on the phenotype of the single mutants. Accordingly, two genes can display positive, negative or non-epistatic interaction. In order to capture most of the biological variability contained in the experimental dataset, we propose the use of two different objective functions, *maximization of growth* (FBA) and *minimization of metabolic adjustment* (MiMBl). FBA is ought to cover situations where

maximization of growth is the cellular objective, while MiMBl will account for regulatory effects inherent to the wild type flux distribution. FBA is a conservative method for finding genetic interactions compared to MiMBl, since the parameter used to define and measure genetic interactions is also the objective of optimization, *i. e*., growth. Within the metabolic network, often there are several optimal solutions theoretically satisfying maximum biomass formation. In case of a single/double gene deletion mutant where an alternative optimal pathway exists, FBA will always find such an alternative solution, even though it may not be biologically plausible due to regulatory constraints, and, thereby may miss potential genetic interactions. On the other hand, MiMBl will help in capturing more refined regulatory effects where the loss of growth is a side effect of minimizing the flux rerouting relative to the wild type.

The subset of experimental genetic interactions involving non-essential genes from the yeast metabolic model contains 17419 interactions (939 positive, 1806 negative and 14674 non-epistatic interactions) connecting 520 genes (**Methods**). Around 84 % of these interactions are non epistatic. Importantly, both MiMBl and FBA are able to capture > 97 % of non-epistatic pairs, meaning that the metabolic model and both simulation algorithms are accurately predicting the majority of the interactions within the dataset. Nevertheless, the remainder 16 % of the interactions, which we subsequently explore, holds the most interesting information concerning functional gene relationships.

In order to assess the performance of the different algorithms, we carried out a sensitivity *versus* precision analysis. Precision was calculated as the fraction of experimentally validated interactions among all predicted interactions, while the sensitivity represents the fraction of the experimentally validated interactions captured by the analysis (**Methods**). A computational epistasis cutoff (ε_{cutoff}, **Methods**) was used to call a particular interaction to be epistatic (either positive ($\varepsilon > \varepsilon_{cutoff}$) or negative ($\varepsilon < -\varepsilon_{cutoff}$)) or non-epistatic ($|\varepsilon| < \varepsilon_{cutoff}$). The performance of both algorithms (MiMBl, FBA and lMoMA) is summarized as ROC (partial receiver operating characteristic) curves for both, positive and negative epistasis (**Fig. 5.4 a, b**). The sensitivity and precision of the FBA predictions obtained in this study are within the same range as previously reported by Szappanos *et al.* [2]. MiMBl shows less precision than FBA in case of both positive (~20 %, ~30 % respectively) and negative interactions (~50 %, ~60 % respectively), but its sensitivity is considerably higher in both cases (~9 %, ~4 % for positive, **Fig. 5.4a;** ~5 %, ~3 % for negative, **Fig. 5.4b**), which reflects the conservative nature of FBA in predicting genetic interactions. Notably, for the entire range of genetic interaction cutoffs, MiMBl sensitivity is twice as high as lMoMA and similar trend is observed for precision (**Fig. 5.4 a, b**). As previously

reported by Szappanos and co-workers, lMoMA does not improve FBA predictions. Such behavior further emphasizes that a proper mathematical formulation of the biological principle (objective function) has a major impact on the ability to interpret *in vivo* observations.

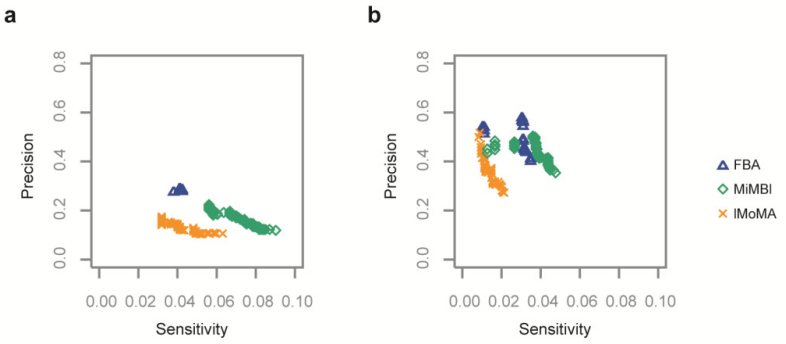


Figure 5.4: **Comparison of the performance of MiMBl, lMoMA and FBA for predicting genetic interactions in *S. cerevisiae*. a, b)** The accuracy of prediction of the algorithms was assessed by comparing the sensitivity and precision of their predictions (partial receiver operating characteristic ROC curves) for positive (a) and negative (b) interactions. Sensitivity reflects the fraction of experimentally validated interactions captured by the algorithm while precision is experimentally validated interactions among all predicted interactions.

We chose a strict interaction cutoff ($|\varepsilon_{cutoff}|$= 0.013) for further analysis of the predicted interactions (**Fig. S4.6**). For this cutoff, the correctly predicted genetic interactions map contains 142 interactions (73 positive and 69 negative) connecting 86 genes (**Fig. 5.5a**). MiMBl not only captures all interactions, except one, predicted by FBA, but also contributes with 48 additional interactions (~34 % of all accurate predictions). MiMBl predictions thus span almost all of those from FBA (**Fig. 5.5b**), which we attribute to the fact that many metabolites within the metabolic model are directly contributing to biomass formation. Consequently, if the turnover of most metabolites is kept constant upon gene deletions, the biomass turnover (growth) will also remain constant. On the other hand, FBA is not able to capture many genetic interactions found by MiMBl (**Fig. 5.5b**). These will involve mutants where the loss of fitness upon gene deletion is caused by the change from an *in vivo* well-tuned pathway to an alternative pathway containing different metabolites or enzymes. For many of such cases, there are alternative pathways that sustain the same growth as the reference and FBA is ought to find such solutions, regardless of the magnitude of the turnover adjustment that is required for the cell. Because of this feature, MiMBl is capable of capturing a part of the regulatory constraints on the operation of cellular metabolism, which lMoMA failed to convey (**Fig. 5.5b**).

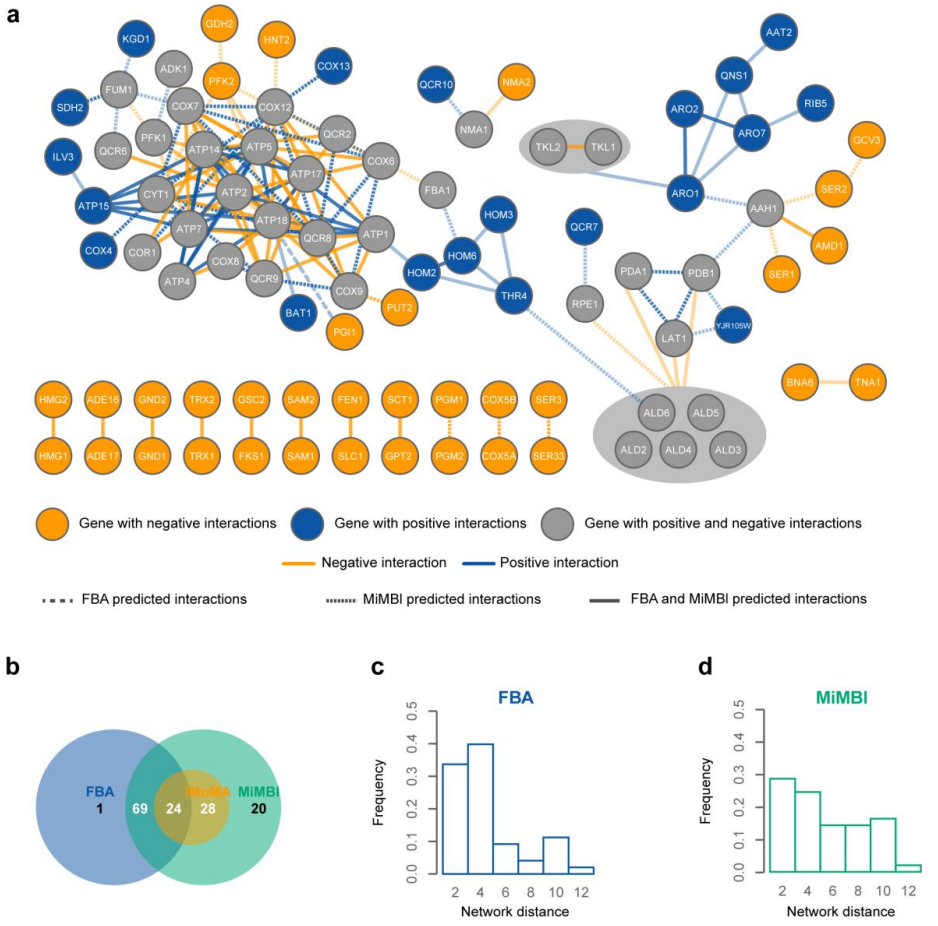


Figure 5.5: Understanding genetic interactions by using MiMBl. a) *S. cerevisiae* genetic interaction network accurately predicted using MiMBl and FBA. Genetic interactions (edges) predicted by both algorithms together and independently are represented in the map (FBA – dashed line, MiMBl – dotted line, both – full line). Positive and negative interactions are distinguished by colour (yellow and blue, respectively) and the opacity of the edges illustrates the network distance between the two genes (nodes) involved in the interaction. Opacity increases as the distance between the genes decreases. Gray nodes represent genes which display both, positive and negative interactions. A gray circle is used to enclose isoenzymes when one of them was found to interact with other genes in the metabolic network. **b)** Venn diagram showing the overlap of accurately predicted interactions by FBA, MiMBl and lMoMA. **c, d)** Distribution of the network distances between genetic interacting genes captured by FBA (c) and MiMBl (d). As MiMBl accounts for the majority of FBA predicted interactions, exclusive MiMBl interactions are represented in order to illustrate the ability of MiMBl to capture genetic interaction of further distant genes with the network.

The regulatory constraints imposed by MiMBl assume even stronger relevance in the case of positive interactions, where MiMBl exclusively accounts for almost 50% of the all the successfully predicted interactions (**Fig. 5.5a**). In fact, FBA's ability of predicting positive interaction is limited, as the maximum predicted biomass formation of a double deletion mutant would never be higher than the highest predicted for the two single mutants. If a single deletion

mutant has the maximum predicted fitness of 1 (meaning that the fitness of the mutant is the same as that of the wild-type), positive interactions involving the deleted gene will be impossible to predict. As FBA is bound to find the optimal solution that provides the highest growth, single mutants with maximum fitness are much more often predicted than the ones found by MiMBl, where minimal adjustment of the metabolic network is preferred over maintaining maximum growth. Indeed, MiMBl predicts decreased single mutant fitness for twice more gene knockouts than FBA (~38.4 *vs* 18.1 %). Consequently, MiMBl also displayed an improved capacity to predict both positive and negative epistasis involving the same gene. More than 80 % of the genes display this feature *in vivo*. Interestingly, 30 % of the genes involved in MiMBl predicted epistasis interact both positively and negatively, while FBA predicts that only 14 % of the genes do so (**Fig. 5.5a**).

MiMBl predicts genetic interactions between metabolically distant genes

MiMBl was also able to capture interactions between more distant genes within the metabolic network (a measure for pleiotropy). To measure this feature, we calculated the network distance between each pair of genes predicted to interact (**Methods**). MiMBl captured interactions between genes that are significantly more distant as opposed to FBA (~40% more distant for negative epistasis, p-value = 0.022; ~10% more distant for both positive and negative epistasis, p-value = 0.089; **Fig. 5.5c, d**). Inspection of the flux rerouting and metabolite turnover changes revealed MiMBl's ability to finely adjust (or maintain homeostasis) the use of highly connected metabolites (e.g. cofactors).

Genetic interactions and isoenzymes

When it comes to double gene deletion events, isoenzymes are a rather interesting class to study, since the deletion of two enzymes often have a drastic effect on the cellular phenotype. Indeed, in the experimental dataset, ~36.3 % of the isoenzyme gene pairs displays negative interaction. *In silico*, almost 20 % of the negative interactions contained in the interaction map hereby predicted (**Fig. 5.5a**) are between isoenzymes. On the other hand, only by excluding isoenzymes, potential interactions occurring between one gene within a group of isoenzymes and other genes within the metabolic network can be captured. Furthermore, if such interactions are ought to be explained by metabolic means, this will only be the case if the deletion of one isoenzyme is not compensated by other isoenzyme. *In vivo*, most of the single deletion of genes coding for isoenzymes contained in the dataset (> 85%) are reported to have maximum fitness, thereby indicating that if these genes will display genetic interactions they will

be due to other mechanisms than flux rerouting, *e.g.* regulatory function. The remaining 15 % of the genes coding for isoenzymes, which potentially lack *in vivo* compensation of enzyme activity, participate in around 10 % of the total positive and negative interactions. In order to explore whether such interactions could be associated to a metabolic fluxes re-distribution, we performed additional simulations with total suppression of the reactions carried out by these isoenzymes. Half of these reactions were found to be *in silico* essentials for growth, therefore no interactions involving these genes could be predicted. A more detailed approach would be needed to study such interactions, for example by simulating decreased enzyme activity. Nevertheless, four additional negative and one positive interactions were accurately predicted (**Fig. 5.5a**).

Using MiMBl and FBA for understanding genetic interactions

Use of MiMBl not only allowed us to expand the range of genetic interactions predicted by FBA, but also the combined use of these two complementary algorithms enabled finding of relevant interactions where only one or both of the simulation principles apply. For example, the interaction between *PGK2* and *GDH2,* exclusively captured by MiMBl, is due to balancing of NADH and glutamate, two of the most connected metabolites in the network. As there are alternative pathways for fulfilling NADH and glutamate requirement (despite implying higher metabolic adjustments), FBA could not capture this interaction. A similar effect is observed for the negative interaction between isoenzymes *SER3* and *SER33*. In the absence of both genes, FBA predicts the needed supply of serine to be totally fulfilled by rerouting the metabolic fluxes via the glyoxylate shunt and threonine biosynthesis. On the other hand, MiMBl predicts that the supply of serine will be shared between the two alternative pathways, but the rescue cannot be complete, because the corresponding metabolic adjustment costs overweighs the growth. This prediction is in very good agreement with the experimental verification that the double mutant growth is impaired and can be restored by adding glycine to the medium, which is the intermediate for serine synthesis via glyoxylate or threonine [27].

We demonstrated that the use of optimization-based algorithms that are stoichiometry representation independent is fundamental for biological applications. These include metabolic engineering, gene essentiality and genetic interaction predictions. We anticipate that the biological relevance of consistent problem formulation will be even more apparent when simulating cells with complex biological objective functions such as human tissues. A proper problem formulation is therefore of primary importance. To this end, we developed a new, consistent method for metabolic modeling – MiMBl – which we thus used to perform a

systematic study of the yeast genetic interaction network. Our results show that the combined use of different objective functions is of primary importance in order to achieve a more complete understanding of the operating principles behind complex biological phenomena such as genetic interactions. We were able to almost double the accurately predicted genetic interactions by using MiMBl, emphasizing the impact of flux regulation within the metabolic network.

Methods

Yeast genome-scale metabolic models

The *Saccharomyces cerevisiae* genome-scale stoichiometric model *i*FF708 [21] was used to study the impact of scaling the stoichiometric matrix on the simulation results when using sum of fluxes in the objective function. The metabolic network was pre-processed so as to remove reactions that carry no flux under the simulated conditions (*i.e.* blocked reactions). Isogenes – genes coding for isoenzymes – were retained. Genes/reactions deemed to be essential by FBA were not considered for knockout simulations within this study, since the prediction of their knockout does not alter when using other algorithms or alternative stoichiometric matrices. Flux Balance Analysis [7] was used to simulate wild-type flux distribution when required, together with the constraints based on experimental data.

For analyzing the predictions of the intracellular flux distribution, the *i*FF708 model was constrained with physiological data for a wild-type strain grown in batch cultivation under aerobic conditions [28] and simulations were performed using minimization of overall intracellular flux. In the case studies for determining the impact of alternative stoichiometry representation on the model predictions for growth, genetic interactions and succinate production, the network was constrained in agreement with the experimental conditions as described by Szappanos *et al.* [2]. In the vanillin case study, the vanillin-glucoside heterologous pathway according to Brochado *et al.* [14] was introduced in the metabolic network and physiological data from the same study was used to obtain the necessary constraints. The target selection for metabolic engineering was done by simulating the deletion of all possible combinations (up to 3 genes) of (FBA-) non-essential genes. Following the simulations, the metabolic engineering strategies were ranked based on yield on carbon source (glucose) of the product of interest (vanillin or succinate).

The *S. cerevisiae* model *i*AZ900 [26] was used to accomplish the genetic interactions study. Genes coding for blocked reactions as well as false essentials predicted from single gene deletion simulations were removed from the analysis, thereby reducing the errors due to misprediction of single-mutant fitness. The metabolic network was constrained in agreement with the experimental conditions as described by Szappanos *et al.* [2].

All algorithms developed and used in this study were formulated as Linear Programming (LP) optimization problems and implemented in C++ using GLPK (http://www.gnu.org/s/glpk/). All simulations were performed on IBM BladeCenters running CentOS Linux (64bit). Detailed description of the LP formulations used for solving FBA, minimization of overall intracellular flux and minimization of metabolic adjustment (lMoMA) are provided in **Supplementary methods**.

Normalized lMoMA

Normalized lMoMA was formulated as follows:

$$\min \sum_i \frac{1}{\left|v_i^{WT}\right|} \left|v_i^{WT} - v_i\right| \qquad \forall i \in N : v_i^{WT} \neq 0$$
$$s.t. \quad S \cdot v = 0$$
$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \in N$$

Where *N* is the set of all reactions, *M* is the set of all intracellular metabolites, *S* is the stoichiometric matrix and v_i is a flux for reaction *i*. *WT* stands for wild-type (or reference), v_i^{lb} and v_i^{up} are the lower and upper bounds for the flux of reaction *i*.

Minimization of metabolites balance – MiMBl

Metabolite turnover is defined as the sum of all fluxes producing (or consuming) it, multiplied by the stoichiometric coefficients:

$$t_m = \sum_{i \in N_m} \alpha_{m,i} v_i \qquad \forall m \in M, N_m \subset N$$

N_m is the subset of N producing or consuming metabolite *m* and $\alpha_{m,i}$ is the stoichiometric coefficient of metabolite *m* in reaction *i*. Note that $\alpha_{m,i}$ is always a positive number in the definition above, irrespective of *m* being a substrate or a product.

MiMBl was formulated as two sequential linear programming problems, as follows:

1st optimization:

$$\min \sum_{m \in M} \left| t_m^{WT} - t_m \right|$$

$$s.t.$$

$$S \cdot v = 0$$

$$t_m = \sum_i \alpha_{m,i} v_i \qquad \forall i \in N_m, N_m \subset N : S_{m,i} \neq 0$$

$$\alpha_{m,i} \equiv \left| S_{m,i} \right| \qquad \forall i \in N, m \in M$$

$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \in N$$

$$v_i \geq 0 \qquad \forall i \in N$$

2nd optimization:

$$\min \sum_i \frac{1}{\left| v_i^{WT} \right|} \left| v_i^{WT} - v_i \right| \qquad \forall i \in N : v_i^{WT} \neq 0$$

$$s.t \qquad S \cdot v = 0$$

$$v_i^{lb} \leq v_i \leq v_i^{ub} \qquad \forall i \in N$$

$$\sum_{m \in M} \left| t_m^{WT} - t_m \right| = \min \sum_{m \in M} \left| t_m^{WT} - t_m \right|$$

Alternative stoichiometry representations

Alternative stoichiometry representations were obtained by multiplying a given reaction (or a set of reactions) by a scalar θ (or a set of scalars). For any reaction r: $\alpha_{Ar} Y + \alpha_{Br} X \rightarrow \alpha_{Cr} Z$, an equivalent representation is given by:

$$\theta \alpha_{Y,r} Y + \theta \alpha_{X,r} X \rightarrow \theta \alpha_{Z,r} Z \qquad \forall \theta > 0$$

where *Y, X* and *Z* represent the metabolites participating in the reaction *r* and α_{Yr}, α_{Xr}, α_{Zr} represent the corresponding stoichiometric coefficients. Note that when the stoichiometry of reaction *r* is scaled by θ, the corresponding flux value will be scaled by 1/θ for the same optimal solution. For illustrating the impact of linear scaling of the reactions stoichiometry on the internal flux distribution, the reaction *RPI1* of *i*FF708 model was divided by the scalar θ. For illustrating the impact of using alternative stoichiometry representations on the design of metabolic engineering strategies, two synonymous stoichiometric matrices were used: i) the original matrix from the yeast model (S_0) and ii) an equivalent matrix (S_1) where the stoichiometric coefficients of the reactions *SERxtO, PDC6, FUR1, GAP1_21, PNP1_1,* and *CYSxtO* were divided by 100, 100, 0.1, 0.01, 100 and 0.1, respectively. A third equivalent matrix (S_2) was

generated by dividing the coefficients of a single reaction (*PGK1*) by 0.1. The results of the comparison between S_0 and S_2 are presented in **Fig. S4.5**.

Impact of scaling stoichiometry on the optimal solution – Analytical evidence

The impact of scaling the constraints of a given linear programming problem depends on whether such changes guarantee the optimality conditions after the scaling. Consider the problem:

$$\begin{aligned} &\min \sum_{i \in N} c_i v_i \\ &s.t. \quad S \cdot v = b \\ &\qquad v_i \geq 0 \end{aligned}$$

Where c_i is the cost coefficient of variable v_i in the objective function. Here, a linear combination of non-normalized fluxes is used in the objective function, similarly to e.g. minimization of intracellular flux and lMoMA. Assuming that B is an optimal basis matrix for the problem, the following optimization condition is satisfied:

$$\begin{aligned} &\overline{c_j} = c_j - c_B' B^{-1} S_j \\ &\overline{c_j} \geq 0, \quad \forall j \in N \end{aligned}$$

where j is the index of variable v in matrix S, $\overline{c_j}$ is the reduced cost of the variable v_j, c_j is the objective function coefficient of v_j, c_B is the vector containing the objective coefficients of basic variables and S_j is the j^{th} column of matrix S [29]. Linear scaling the problem by the matrix Θ will result in the following reduced cost for each variable:

$$\overline{c_{j\,\Theta}} = c_j - c_B' \left(B\Theta_B\right)^{-1} \theta_{jj} S_j$$

Where Θ is a nxn positive diagonal matrix (**scaling matrix**) and θ_{jj} is the scaling factor for the j^{th} column of matrix S. In the cases of entries $\theta_{jj} \neq 1$ the corresponding columns of S are accordingly scaled. Analogously, Θ_B is the scaling matrix corresponding to the basic variables.

Unless all entries of Θ are identical,

$$\exists\, \theta \in \Re^+ : c_j - c_B' \left(B\Theta_B\right)^{-1} \theta_{jj} S_j \leq 0 \qquad \forall j \in N$$

Therefore the **optimality condition is not guaranteed**.

Corollary 1: When all (diagonal) entries of Θ are identical (uniform scaling matrix), and therefore equal to θ_{jj}, the optimality condition is simplified to

$$\overline{c_{j\,\Theta}} = c_j - c_B{}'\Theta_B{}^{-1}B^{-1}\theta_{jj}S_j$$
$$where\ \Theta_B{}^{-1} = \frac{1}{\theta_{jj}}I$$
$$\overline{c_{j\,\Theta}} = c_j - c_B{}'B^{-1}S_j = \overline{c_j}$$
$$\because \overline{c_j} \geq 0 \Rightarrow \overline{c_{j\,\Theta}} \geq 0$$

The optimality condition can thus be guaranteed only when the matrix S is uniformly scaled. Note that due to the nature of the biological problem, the genuine representation of S might not be known, thereby Θ cannot be guaranteed to be a uniform scaling matrix. More importantly, for metabolic modeling purposes (where flux units and ranges are problem dependent), it is nevertheless undesirable that the solution is sensitive to non-uniform scaling and thus context dependent.

Now consider the following problem:

$$\min \sum_{m\in M} \gamma_m t_m$$
$$s.t. \quad S\cdot v = b$$
$$t_m \geq 0$$

Where γ_m is cost coefficient of variable t_m in the objective function. The new problem biologically corresponds to the previous one, after mapping the flux space into metabolite space. We term it as a MiMBl-like problem formulation.

$$\text{As } t_m = \sum_{i\in N_m} \alpha_{m,i} v_i \qquad \forall m \in M, N_m \subset N$$

Recall that $\alpha_{m,i}$ is the stoichiometric coefficient of metabolite m in reaction i. The objective function can be re-written as function of v_i:

$$\min \sum_{m\in M}\left(\gamma_m\left(\sum_{i\in N_m}\alpha_{m,i}v_i\right)\right) \qquad \forall m \in M, N_m \subset N$$
$$=$$
$$\min \sum_{i\in N}\left(\sum_{m\in M}\left(\gamma_m\alpha_{m,i}\right)\cdot v_i\right)$$

Therefore, the objective function coefficients of each v_i is a function of the stoichiometric coefficients $\alpha_{m,i}$: $c_i = \sum_{m \in M} \gamma_m \alpha_{m,i}$.

Similarly to the previous problem, the following optimality condition is satisfied, so v is an optimal solution.

$$\overline{c_j} = c_j - c_B' B^{-1} S_j$$
$$\overline{c_j} \geq 0 \qquad \forall j \in N$$

Scaling the optimality condition will result in:

$$\overline{c_{j\Theta}} = c_j \theta_{jj} - c_B' \Theta_B \left(B \Theta_B\right)^{-1} \theta_{jj} S_j$$
$$= c_j \theta_{jj} - c_B' \Theta_B \Theta_B{}^{-1} B^{-1} \theta_{jj} S_j$$
$$= \theta_{jj} \cdot \left(c_j - c_B' B^{-1} S_j\right)$$
$$\left(c_j - c_B' B^{-1} S_j\right) \geq 0 \; and \; \theta_{jj} > 0$$
$$\therefore \overline{c_{j\Theta}} \geq 0$$

Unlike the previous situation (sum of non-normalized fluxes in the objective function), using a MiMBl-like problem formulation guarantees that the optimality condition is always satisfied, independently of the stoichiometry representation.

Genetic interactions – epistasis score

The fitness (*f*) of each single and double mutant was calculated by normalization of the mutant growth to the wild-type growth. A variety of phenotypic traits can be used to quantify epistasis, growth being the most commonly used, due to its accurate experimental quantification in an efficient high-throughput manner. The epistatic interaction (ε) of each double gene combination (A and B) was calculated based on the following metrics:

$$\varepsilon = f_{AB} - f_A \cdot f_B$$

To address the accuracy of the different algorithms (FBA, lMoMA and MiMBl) in predicting genetic interactions, we did a *precision versus sensitivity* analysis given the range of computational epistasis score cutoffs $|\varepsilon_{cutoff}|<0.05$. *Precision* is the fraction of experimentally validated interactions among all predicted interactions, while the *sensitivity* represents the fraction of the experimentally validated interactions captured by the analysis.

$$Precision = \frac{True\ positives}{All\ predicted\ positives}$$

$$Sensitivity = \frac{True\ positives}{All\ experimental\ positives}$$

All significant genetic interactions among the non-essential genes from Szappanos *et al.* [2] dataset involving genes contained in *i*AZ900 model were included in the present analysis. The experimental data was filtered by using a confidence threshold of $|\varepsilon|<0.08$ and $P<0.05$ [2, 30].

R 2.12.1 was used to perform statistics calculations and to generate the plots. The Venn diagram from **Fig. 5.4b** was generated by using the R package "VennDiagram" by Chen & Boutros (2011). Cytoscape 2.8.2 was used to generate the genetic interactions map from **Fig. 5.4a**.

Metabolic/network distance

The metabolic connectivity graph obtained from the metabolic model *i*AZ900 was used to calculate the metabolic/network distance between two genes. We define network distance between the two as the number of reactions belonging to the shortest path between the two genes on the connectivity graph. A pair of directly connected metabolic genes was considered as being separated by distance of 2. This way, genes coding for the same reaction (e.g. isoenzymes and complexes) have a distance of 1. As highly connected metabolites, as cofactors (ATP, NADH, NADPH, FADH2, pyrophosphate and orthophosphate) are not likely to connect genes with related metabolic functions, this subset of metabolites was excluded from the connectivity graph [32]. Despite being highly connected, mitochondrial protons were kept part of the connectivity graph to ensure the integrity of the respiratory chain. Graph Template Library (GTL) was used to implement the algorithm for network distance calculations.

Acknowledgements

ARB acknowledge Ph.D. research grant from Fundação para a Ciência e Tecnologia (ref. SFRH / BD 579 / 41230 / 2007). Authors are grateful to Aleksej Zelezniak, Arnau Montagud, Martin Beck and Wolfgang Huber for insightful comments and discussions.

References

1. Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR: **Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes.** *PLoS Computational Biology* 2010, **6**:e1000729.

2. Szappanos B, Kovács K, Szamecz B, et al.: **An integrated approach to characterize genetic interaction networks in yeast metabolism.** *Nature Genetics* 2011, **43**:656-62.

3. Kümmel A, Panke S, Heinemann M: **Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data.** *Molecular Systems Biology* 2006, **2**:2006.0034.

4. Oberhardt MA, Palsson BØ, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Molecular Systems Biology* 2009, **5**:320.

5. Jerby L, Shlomi T, Ruppin E: **Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.** *Molecular Systems Biology* 2010, **6**.

6. Osterlund T, Nookaew I, Nielsen J: **Fifteen years of large scale metabolic modeling of yeast: Developments and impacts.** *Biotechnology Advances* 2011.

7. Varma A, Palsson BØ: **Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors**. *Journal of Theoretical Biology* 1993, **165**:477-502.

8. Ibarra RU, Edwards JS, Palsson BO: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth**. *Nature* 2002, **420**:20-23.

9. Famili I, Förster J, Nielsen J, Palsson BØ: **Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:13134-9.

10. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.** *Molecular Systems Biology* 2007, **3**:119.

11. Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:15112-15117.

12. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in Saccharomyces cerevisiae**. *Genome Research* 2005, **15**:1421-30.

13. Asadollahi M, Maury J, Patil KR, et al.: **Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering.** *Metabolic Engineering* 2009, **11**:328-34.

14. Brochado AR, Matos C, Møller BL, et al.: **Improved vanillin production in baker's yeast through in silico design.** *Microbial Cell Factories* 2010, **9**:84.

15. Suthers PF, Zomorrodi A, Maranas CD: **Genome-scale gene/reaction essentiality and synthetic lethality analysis.** *Molecular Systems Biology* 2009, **5**:301.

16. Snitkin ES, Segrè D: **Epistatic Interaction Maps Relative to Multiple Metabolic Phenotypes**. *PLoS Genetics* 2011, **7**.

17. Alper H, Jin Y-S, Moxley JF, Stephanopoulos G: **Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli.** *Metabolic Engineering* 2005, **7**:155-64.

18. Wintermute EH, Silver P: **Emergent cooperation in microbial metabolism.** *Molecular Systems Biology* 2010, **6**:407.

19. Blank LM, Kuepfer L, Sauer U: **Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast.** *Genome Biology* 2005, **6**:R49.

20. Lewis NE, Hixson KK, Conrad TM, et al.: **Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models.** *Molecular Systems Biology* 2010, **6**.

21. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13**:244-53.

22. Becker S a, Feist AM, Mo ML, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nature Protocols* 2007, **2**:727-38.

23. Knorr AL, Jain R, Srivastava R: **Bayesian-based selection of metabolic objective functions.** *Bioinformatics* 2007, **23**:351-7.

24. Gulik WM van, Heijnen JJ: **A metabolic network stoichiometry analysis of microbial growth and product formation.** *Biotechnology and Bioengineering* 1995, **48**:681-98.

25. Price ND, Reed JL, Palsson BØ: **Genome-scale models of microbial cells: evaluating the consequences of constraints.** *Nature Reviews Microbiology* 2004, **2**:886-97.

26. Zomorrodi AR, Maranas CD: **Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data.** *BMC Systems Biology* 2010, **4**:178.

27. Otero JM: **Industrial Systems Biology and Metabolic Engineering of Saccharomyces cerevisiae**. *Technology* 2009.

28. Hoek P Van, Dijken JP Van, Pronk JT: **Effect of Specific Growth Rate on Fermentative Capacity of Baker's Yeast**. *Applied and Environmental Microbiology* 1998, **64**:4226-4233.

29. Bertsimas D, Tsitsiklis JN: *Introduction to Linear Optimization*. First Edit. Belmont, Massachusetts: Athena Scientific; 1997.

30. Costanzo M, Baryshnikova A, Bellay J, et al.: **The genetic landscape of a cell**. *Science* 2010, **327**:425-31.

31. Chen H, Boutros PC: **VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.** *BMC Bioinformatics* 2011, **12**:35.

32. Kharchenko P, Church GM, Vitkup D: **Expression dynamics of a cellular metabolic network.** *Molecular Systems Biology* 2005, **1**.

Chapter 6: Conclusions and Future Perspectives

Metabolic engineering for vanillin production and analysis of genetic interactions in *S. cerevisiae* were used to contextualise the application of genome-scale metabolic models for prediction and systematic analysis of phenotypes resulting from perturbed genotypes. The results obtained within this thesis provide experimental evidence for the current capabilities and limitations for predicting metabolic phenotypes based on the genotype by using metabolic models. While prediction of the metabolic behaviour of single gene deletion mutants can be successfully achieved, as we showed by experimentally demonstrating the application of the yeast genome-scale stoichiometric model to metabolic engineering, predicting metabolic phenotypes of multiple gene deletions or overexpression mutants is still a challenge. A newly developed computational tool herein presented is expected to contribute for enhancing the predictive power of metabolic phenotypes by means of metabolic modeling, particularly relevant for multiple gene deletion mutants.

The yeast genome-scale model was successfully used to guide the design of a metabolic engineering strategy for increasing heterologous production of vanillin in *S. cerevisiae* (**Chapter 3**). The biological objective function used to perform metabolic simulations proved to be a relevant parameter, as FBA (**F**lux **B**alance **A**nalysis) predicted that no vanillin improvement would take place even after deleting six metabolic reactions. By using MoMA (**M**inimization of **M**etabolic **A**djustment), and thereby assuming the mutant's metabolic proximity towards the reference strain instead of its optimal growth, several strategies comprising the deletion of one reaction were suggested. A thorough analysis of the predicted gene targets was performed accounting for different modes of yeast physiology and final target selection for experimental work was done based on prioritizing strategies which were predicted to compel the smallest metabolic adjustment. Implementation of the metabolic engineering strategies following the *in silico* guided design successfully resulted in three mutants with up to 2-fold increased vanillin production. The results from physiological characterization and flux variability analysis of the predicted flux distribution of the mutants suggested that increased vanillin production was associated with improved cofactors supply for the vanillin biosynthetic pathway.

Despite the success of taking a global approach for improving vanillin production by using the yeast genome-scale metabolic model, lower vanillin yield than expected was obtained, as accumulation of several intermediate compounds of the vanillin biosynthetic pathway was observed. In order to achieve maximum potential of the global approach for metabolic

engineering, the activity of the pathway enzymes must not impose a limitation. Indeed, In **Chapter 4** it was shown that the O-methyltransferase for vanillin biosynthesis was limiting the conversion of pathway intermediates to the final product, vanillin glucoside. Moreover, the limitation was found to prevail exclusively on the globally engineered mutant, supporting the fact that local pathway engineering is of higher value when it is combined with global approaches, which take the whole metabolic network into account.

Model incompleteness and lack of regulatory information are often pointed as potential factors for limiting the success of application of the genome-scale models. Although the usage of MoMA as biological objective function indirectly accounts for regulation imbedded in the wild type flux distribution, direct inclusion of regulatory genes on metabolic networks is still limited due to lack of knowledge of regulation at the genome-scale level. Additionally, regulatory genes modulate the expression of several genes, rendering the mathematical description of the relationship between the expression level of the genes and the corresponding flux changes an intricate task. **Chapter 4** present experimental examples of overexpression of a transcription factor for metabolic engineering, using vanillin and succinate production in *S. cerevisiae* as case studies. In the case of production of vanillin, alleviation of the yeast glucose repression by overexpression of *HAP4* lead to increased respiratory capacity, decreased ethanol formation and higher vanillin production, most likely due to higher ATP availability. In the case of succinate, the same strategy raised production by 15%, but it was also found to have a strong negative effect on biomass formation. Better characterization of the improved strains, particularly by obtaining the transcriptome and fluxome, could further explain the observed phenotypes. Concerning succinate production in particular, clarifying the reasons for the observed decreased fitness, as well as explaining the origin of the produced succinate, either cytosolic or mitochondrial, would reveal relevant information for future strain improvement.

The results obtained in **Chapter 4** confirm the potential of using regulatory genes as targets for metabolic engineering. In a recent study, Sauer and co-workers (2010) performed systematic flux analysis of 119 yeast transcription factor deletion mutants, under 43 different experimental conditions. Their results demonstrate that, despite the altered expression profile of numerous genes, the metabolic network operation mode is very robust towards transcription factor deletion. However, the results herein presented, also in agreement with previous work on overexpression of transcription factors, *e.g*. van Maris *et al.* (2001) and Sopko *et al.* (2006), suggest that transcription factor overexpression, instead of deletion, may in fact strongly affect growth rate, as well as the distribution of metabolic fluxes and modulate the activity of central

pathways in yeast. Efforts towards the systematic clarification of the relationship between changes in gene expression and metabolic fluxes upon overexpression of transcription factors, could largely contribute for understanding regulatory mechanisms.

At last, **Chapter 5** describes a new computational tool focusing on the formulation of objective functions for optimization-based algorithms using genome-scale models. MiMBl – **Mi**nimization of **M**etabolites **B**a**l**ance – presents an alternative methodology to formulate new and currently used biological principles in a stoichiometry representation independent manner. Several biological principles are currently formulated by using linear combination of fluxes, *e.g.* MoMA and minimization of intracellular flux, which was herein shown to be susceptible to alternative representations of stoichiometry. This is an undesirable feature, as alternative representation of the stoichiometric coefficients can be used to describe the same biochemical reaction, thus simulation algorithms should be insensitive towards such artifacts. MiMBl was used to perform a systematic study of the yeast genetic interaction network while assuming minimization of metabolic adjustment as biological principle and showed considerably higher prediction accuracy than lMoMA. Despite the apparently modest accuracy of prediction of genetic interactions using metabolic modeling, MiMBl accurately captured the largest fraction of genetic interactions comparing to other simulation algorithms, namely FBA. Nevertheless, the results herein obtained support the observations previously reported by Constanzo *et al.* (2010), where correlation between genetic interaction degree and different cellular features, such as single mutant fitness, physical interactions and gene multi-functionality was observed, suggesting that the wide map of genetic interactions is explained by diverse biological traits. Analogously, as FBA and MiMBl describe different biological principles, their combined use captures genetic interactions potentially explained by different biological traits. Indeed, by performing such analysis, we observed that FBA mostly captures negative generic interactions, as a direct consequence of its objective function, while MiMBl is more suitable for describing both positive and negative interactions, due to the prevalence of minimizing metabolic adjustment relatively to the reference over maximizing growth. Furthermore, MiMBl accurately captures interactions between genes which are more distant in the metabolic network than FBA. Ultimately, differently predicted genetic interactions between both approaches provide mechanistic insights about modes of operation of metabolic networks.

The use of genome-scale metabolic models is currently increasing; the scope of their applications is expanding, as well as the number of organisms with their metabolic network reconstructed. The work herein developed evidenced the effectiveness of such models for

investigating genotype to phenotype relationship, particularly contributing to the analysis of perturbed metabolic networks. MiMBl versatility for implementation of objective functions can be explored in the future for formulation of many different biological principles potentially governing the metabolic phenotype. To this end, availability of large-scale datasets, such as internal flux distribution of the single and double gene deletion libraries, as well as of overexpression libraries, in combination with transcriptome, proteome and metabolome, would largely increase the potential of uncovering biological principles underlying metabolism. Furthermore, developing new computational tools which can describe complex relationships between these cellular components would enhance systematic investigation of how changes in genotype influence the phenotype. Consequent improvement of the predictive power of metabolic models would contribute for increasing their usefulness in supporting fundamental and applied research.

Supplementary Information

Chapter 2: Systems Biology, Genome Scale Metabolic Modeling and Metabolic Engineering of *S. cerevisiae* – Acts and Facts

Table ST2.1: Database and other online resources useful for research in *S. cerevisiae* and other yeasts.

Table ST2.2: Comprehensive list of added-value small molecules produced in *S. cerevisiae*.

Chapter 3: Metabolic Engineering for Vanillin Production in S. cerevisiae. Part 1 – Metabolic Modeling for Strain Improvement

Table ST3.1: List of primers used in this study.

Chapter 4: Metabolic Engineering for Vanillin Production in S. cerevisiae. Part 2 – Pathway and Regulatory Engineering

Table ST4.2: List of oligonucleotides used in this study.

Chapter 5: Minimization of Metabolite Balance Explains Genetic Interactions within the Yeast Metabolic Network

Supplementary Methods

Supplementary Method 5.1: Flux Balance Analysis.

Supplementary Method 5.2: Minimization of overall intracellular flux.

Supplementary Method 5.3: Minimization of metabolic adjustment - lMoMA.

Supplementary Figures

Figure S5.1: Profiles obtained for the objective function value (minimization of overall intracellular flux) using alternative stoichiometry representations.

Figure S5.2: A toy-model illustrating how, and why, alternative stoichiometry representations influence simulation of minimization of metabolic adjustment by using normalized lMoMA – normlMoMA.

Figure S5.3: Stoichiometry representation (S_0 and S_1) impacts the design of metabolic engineering strategies for improving succinate production in *S. cerevisiae* depending on the nature of the objective function formulation.

Figure S5.4: Stoichiometry representation (S_0 and S_1) impacts the design of metabolic engineering strategies for improving vanillin-glucoside production in *S. cerevisiae* depending on the nature of the objective function formulation.

Figure S5.5: Stoichiometry representation (S_0 and S_2) impacts the design of metabolic engineering strategies for improving succinate and vanillin-glucoside yields in *S. cerevisiae* depending on the nature of the objective function formulation.

Figure S5.6: Comparing the performance of MiMBl, FBA and lMoMA for predicting yeast genetic interactions using i*AZ900* metabolic model.

Supplementary Tables

Table ST5.1: Number of lMoMA-predicted lethal gene/reaction knockouts in *S. cerevisiae* that differ between alternative representations of stoichiometry (S_1 and S_2), relative to S_0.

Table ST5.2: lMoMA-predicted epistatic interactions within *S. cerevisiae* genome-scale metabolic model.

Table ST5.3: MiMBl-predicted epistatic interactions within *S. cerevisiae* genome-scale metabolic model.

Supplementary Notes

Supplementary Note 5.1: Toy-Model.

Supplementary Note 5.2: normlMoMA.

Supplementary Note 5.3: Impact of scaling stoichiometry on finding the optimal solution for metabolic flux distributions using FBA-like objective functions – Analytical evidence.

Chapter 2: Systems Biology, Genome Scale Metabolic Modeling and Metabolic Engineering of S. cerevisiae – Acts and Facts

Table ST2.1: Database and other online resources useful for research in *S. cerevisiae* and other yeasts. Adapted from [33].

Database	URL	Comments
SGD (Saccharomyces Genome Database)	http://www.yeastgenome.org/	The most widely used genome database. It contains a wide range of data and community resources. It employs Gene Ontology (GO) for functional annotation.
CYGD (Comprehensive Yeast Genome Database)	http://mips.helmholtz-muenchen.de/genre/proj/yeast	Particularly useful for protein sequence and structure. It employs FunCat for functional annotation, often complementary to GO.
EUROSCARF (European Saccharomyces cerevisiae Archive for Functional Analysis)	http://web.uni-frankfurt.de/fb15/mikro/euroscarf/index.html	Index to the *S. cerevisiae* strains and plasmids repository. It contains the complete deletion collections and several other important resources.
PROPHECY (Profiling of phenotypic characteristics in yeast)	http://prophecy.lundberg.gu.se/	Quantitative phenotypic information for the complete library of *S. cerevisiae* deletion mutants.
BioGRID (Biological General Repository for Interaction Datasets)	http://thebiogrid.org/	A curated database of genetic and protein-protein interactions for a number of organisms, *including S. cerevisiae*.
YEASTRACT (Yeast Search for Transcriptional Regulators and Consensus Tracking)	http://www.yeastract.com/	A curated repository for regulatory associations between transcription factors and target genes in *S. cerevisiae*.
SGRP (Saccharomyces Genome Resequencing Project)	http://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html	Contains the genome sequence of *S. cerevisiae* and *S. paradoxus*, as well as of several wild and industrial strains.
WoLF PSORT (Protein subcellular localization prediction)	http://wolfpsort.org/	An online tool for prediction of sucellular localization of proteins in *S. cerevisiae* and other fungi.
YGOB (Yeast Gene Order Browser)	http://wolfe.gen.tcd.ie/ygob/	An easy-to-use online viewer that enables synthetic comparisons between 16 yeast species, as well as the reconstruction of their ancestral genome
NetPhosYeast	http://www.cbs.dtu.dk/services/NetPhosYeast/	A web based tool for prediction of serine and threonine phosphorylation sites in yeast proteins.
YTPdb (Yeast Transport Proteins database)	http://homes.esat.kuleuven.be/~sbrohee/ytpdb/	A manually annotated set of S. cerevisiae proteins either demonstrated or predicted to be transporters.
YEASTNET (A consensus reconstruction for yeast metabolism)	http://www.comp-sys-bio.org/yeastnet/	A consensus reconstruction of the *S. cerevisiae* metabolic network.

Table ST2.2: Comprehensive list of added-value small molecules produced in *S. cerevisiae*. Cortesy from Erica Valentini and Tomasz Boruta.

Native Compounds Produced by *S. carevisiae**		
Compounds	**Applications**	**Reference**
2-methyl-1-butanol	Biofuel.	Zhang, K., Sawaya, M. R., Eisenberg, D. S., & Liao, J. C. (2008). Expanding metabolism for biosynthesis of nonnatural alcohols. Proceedings of the National Academy of Sciences of the United States of America, 105(52), 20653-8. doi:10.1073/pnas.0807157106
2-phenylethanol	Rose-like fragrance.	Stark, D., Münch, T., Sonnleitner, B., Marison, I. W., & von Stockar, U. (2002). Extractive bioconversion of 2-phenylethanol from L-phenylalanine by Saccharomyces cerevisiae. *Biotechnology progress*, *18*(3), 514-23. doi:10.1021/bp020006n
3-methyl-1-butanol	Biofuel.	Atsumi, S., Hanai, T., Liao, J. C., Non fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 2008, 451, 86-89.
D-Ribose	Food ingredient.	Toivari, M. H., Ruohonen, L., Miasnikov, A. N., Richard, P., & Penttilä, M. (2007). Metabolic engineering of Saccharomyces cerevisiae for conversion of D-glucose to xylitol and other five-carbon sugars and sugar alcohols. *Applied and environmental microbiology*, *73*(17), 5471-6. doi:10.1128/AEM.02707-06
Ergosterol	Provitamin D2.	Veen, M., Stahl, U. and Lang, C. (2003) Combined overexpression of genes of the ergosterol biosynthetic pathway leads to accumulation of sterols in Saccharomyces cerevisiae. FEMS Yeast Research 4(1), 87-95.
Ethanol	Biofuel and fundamental ingredient of alcoholic beverages.	Bro, C., Regenberg, B., Förster, J., & Nielsen, J. (2006). In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. *Metabolic Engineering*, *8*(2), 102-11. doi:10.1016/j.ymben.2005.09.007
Fumarate	Starting material for the production of chemicals.	Xu, G., Liu, L., & Chen, J. (2012). Reconstruction of cytosolic fumaric acid biosynthetic pathways in Saccharomyces cerevisiae. *Microbial cell factories*, *11*(1), 24. doi:10.1186/1475-2859-11-24
Glutathione	Antioxidant, used in the treatment of liver disease.	Yoshida, H., Hara, K. Y., Kiriyama, K., Nakayama, H., Okazaki, F., Matsuda, F., Ogino, C., et al. (2011). Enzymatic glutathione production using metabolically engineered Saccharomyces cerevisiae as a whole-cell biocatalyst. *Applied Microbiology and Biotechnology*, *91*(4), 1001-6. doi:10.1007/s00253-011-3196-4
Glycerol	Feedstock for the production of chemicals.	Overkamp, K.M., Bakker, B.M., Kötter, P., van Tuijl, A., de Vries, S., van Dijken, J. P. and, & Pronk, J. T. (2002). Metabolic engineering of glycerol production in Saccharomyces cerevisiae. *Applied and environmental microbiology*, *68*(6), 2814-2821. doi:10.1128/AEM.68.6.2814
Isobutanol	Biofuel .	Chen, X., Nielsen, K. F., Borodina, I., Kielland-Brandt, M. C., & Karhumaa, K. (2011). Increased isobutanol production in Saccharomyces cerevisiae by overexpression of genes in valine metabolism. *Biotechnology for biofuels*, *4*(1), 21. BioMed Central Ltd. doi:10.1186/1754-6834-4-21

Lanosterol	Emulsifier.	Veen, M., Stahl, U. and Lang, C. (2003) Combined overexpression of genes of the ergosterol biosynthetic pathway leads to accumulation of sterols in Saccharomyces cerevisiae. FEMS Yeast Research 4(1), 87-95.
L-Glycerol 3-phosphate	Starting material for enzymatic synthesis of carbohydrates.	Nguyen, H. T. T., Dieterich, A., Athenstaedt, K., Truong, N. H., Stahl, U. and Nevoigt, E. (2004) Engineering of Saccharomyces cereviasiae for the production of L-glycerol 3-phosphate. Metabolic Engineering 6, 155-163.
Malate	Used in food industry as a taste enhancer and acidifying agent, a building block for a variety of chemicals	Zelle, R. M., de Hulster, E., van Winden, W. A., de Waard, P., Dijkema, C., Winkler, A. A., Geertman, J.-M. A., van Dijken, J., Pronk, J. T. and van Maris, A. J. A. (2008) Malic acid production by Saccharomyces cerevisiae : engineering of pyruvate carboxylation, oxaloacetate reduction and malate export. Applied and Environmental Microbiology 74(9), 2766-2777.
Pyruvate	Used as a dietary supplement, shown to enhance treatment of dermatological conditions, such as acne or pigment disorders.	van Maris, A. J., Geertman, J. M., Vermeulen, A., Groothuizen, M. K., Winkler, A. A., Piper, M. D., van Dijken, J. P. and Pronk, J. T. (2004) Directed evolution of pyruvate decarboxylase-negative Saccharomyces cerevisiae, yielding a C2-independent, glucose-tolerant, and pyruvate-hyperproducing yeast. Applied and Environmental Microbiology 70, 159-166.
S-adenosyl-L-methionine	Chemotherapeutic agent.	Shobayashi, M., Mukai, N., Iwashita, K., Hiraga, Y., & Iefuji, H. (2006). A new method for isolation of S-adenosylmethionine (SAM)-accumulating yeast. *Applied microbiology and biotechnology*, *69*(6), 704-10. doi:10.1007/s00253-005-0009-7
Succinate	A building block for a variety of chemicals. Chelator.	Raab, A. M., Gebhardt, G., Bolotina, N., Weuster-Botz, D., & Lang, C. (2010). Metabolic engineering of Saccharomyces cerevisiae for the biotechnological production of succinic acid. *Metabolic Engineering*, *12*(6), 518-25. Elsevier. doi:10.1016/j.ymben.2010.08.005
Trehalose	Stabilizer for vaccines, enzymes, pharmaceuticals.	Schiraldi, C., Di Lernia, I., & De Rosa, M. (2002). Trehalose production: exploiting novel approaches. *Trends in biotechnology*, *20*(10), 420-5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12220904
Xylitol	Used as a sweetener, suitable for diabetics due to its insulin-independent metabolism.	Toivari, M. H., Ruohonen, L., Miasnikov, A. N., Richard, P., & Penttilä, M. (2007). Metabolic engineering of Saccharomyces cerevisiae for conversion of D-glucose to xylitol and other five-carbon sugars and sugar alcohols. *Applied and environmental microbiology*, *73*(17), 5471-6. doi:10.1128/AEM.02707-06
Zymosterol	Precursor of cholesterol lowering substances.	Heiderpriem, R. W., Livant, P. D., Parish, E. J., Barbuch, R. J., Broaddus, M. G., Bard, M. (1992) A simple method for the isolation of zymosterol from a sterol mutant of Saccharomyces cerevisiae. J. Steroid Biochem. Mol. Biol. 43, 741-743.

Heterologous Compounds Produced by *S. cerevisiae*

Compounds	Applications	Reference
6-MSA	Member of polyketide family with farmaceutical importance.	Kealey, J. T., Liu, L., Santi, D. V., Betlach, M. C., & Barr, P. J. (1998). Production of a polyketide natural product in nonpolyketide-producing prokaryotic and eukaryotic hosts. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(2), 505-9.
8-Epi-cedrol	Sesquiterpenoid alcohol of potential pharmaceutical interest.	Jackson, B. E., Hart-Wells, E. a, & Matsuda, S. P. T. (2003). Metabolic engineering to produce sesquiterpenes in yeast. *Organic letters*, *5*(10), 1629-32. doi:10.1021/ol034231x
Alpha-santalene	A precursor of alpha-santalol, a component of sandalwood oil, applied for the production of cosmetics and perfumes.	Scalcinati, G.,et al. (2012) Dynamic control of gene expression in Saccharomyces cerevisiae engineered for the production of plant sesquitepene alpha-santalene in a fed-batch mode. Metab.Eng. ,doi:10.1016/j.ymben.2012.01.007
Amorpha-4,11-diene	Precursor for antimalarial drug artemisinin.	Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature, 440(7086), 940-3.
Apigenin	Anxiolytic effects.	Leonard, E., Yan, Y., Lim, K. H. and Koffas, M. A. G. (2005) Investigation of two distinct flavone synthases for plant-specific flavone biosynthesis in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(12), 8241-8248.
Artemisinic acid	Precursor of antimalarial drug artemisinin.	Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, *440*(7086), 940-3.
Ascorbic acid	Vitamin C.	Branduardi, P., Fossati, T., Sauer, M., Pagani, R., Mattanovich, D., & Porro, D. (2007). Biosynthesis of vitamin C by yeast leads to increased stress resistance. *PloS one*, *2*(10), e1092. doi:10.1371/journal.pone.0001092
Astaxanthin	Red carotenoid present in fish, crustaceans, birds. Used for food colouring and as a food supplement.	Ken Ukibe, Keisuke Hashida, Nobuyuki Yoshida, and Hiroshi Takagi. Metabolic Engineering of Saccharomyces cerevisiae for Astaxanthin Production and Oxidative Stress Tolerance. APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 2009, p. 7205–7211
Beta-amyrin	Precursor of triterpene saponins of pharmaceutical interest for the development of cholesterol-lowering supplements.	Madsen et al. (2011) Linking Genotype and Phenotype of Saccharomyces cerevisiae Strains Reveals Metabolic Engineering Targets and Leads to Triterpene Hyper-Producers.March 2011 | Volume 6 | Issue 3 | e14763

Bisabolene	A precursor of bisabolane, a fuel compound	Peralta-Yahya PP, Ouellet M, Chan R, Mukhopadhyay A, Keasling JD, Lee TS: Identification and microbial production of a terpene-based advanced biofuel. Nat Commun 2011, 2:.
Beta-Carotene	Precursor of vitamin A, antioxidant, anti-cancer properties, immune system stimulator.	Verwaal, R., Wang, J., Meijnen, J.-P., Visser, H., Sandmann, G., Berg, J. a van den, et al. (2007). High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. Applied and environmental microbiology, 73(13), 4342-50.
n-butanol	Biofuel.	Steen, E. J., Chan, R., Prasad, N., Myers, S., Petzold, C.J., Redding, A., Ouellet, M. and Keasling, J.D. (2008) Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol. Microbial Cell Factories 7: 36.
Casbene	Antifungal diterpene.	Kirby J, Nishimoto M, Park JG et al. (2010) Cloning of casbene and neocembrene synthases from Euphorbiaceaeplants and expression in Saccharomyces cerevisiae. Phytochemistry 71: 1466–1473.
Chrysin	Anti-inflammatory effect.	Leonard, E., Yan, Y., Lim, K. H. and Koffas, M. A. G. (2005) Investigation of two distinct flavone synthases for plant-specific flavone biosynthesis in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(12), 8241-8248.
Cubebol	Flavouring agent, taste cooling and refreshing.	Møller, K., Nielsen, K. F., Asadollahi, M. A., Schalk, M., Clark, A., & Nielsen, J. (2008). Production of Plant Sesquiterpenes in Saccharomyces cerevisiae : Effect of ERG9 Repression on Sesquiterpene Biosynthesis. *Biotechnology*, *99*(3), 666-677
Eriodictyol	Bitter-masking agent.	Yan, Y., Kohli, A. and Koffas, M. A. G. (2005) Biosynthesis of natural flavanones in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(9), 5610-5613.
Ethylene	Used in chemical industry for the production of polyethylene, ethylene oxide and other important compounds. Applied in agriculture as a plant hormone.	Pirkov, I., Albers, E., Norbeck, J., & Larsson, C. Ã. (2008). Ethylene production by metabolic engineering of the yeast Saccharomyces cerevisiae. *Biotechnology*, *10*, 276-280. doi:10.1016/j.ymben.2008.06.006
Farnesol	Anti- fungal activity and as a precursor in the chemical synthesis of perfume Ambrox and polyprenols. Biofuel for diesel or jet fuel.	Song L: Recovery of E,E-farnesol from cultures of yeast erg9 mutants: extraction with polymeric beads and purification by normal-phase chromatography. Biotechnol Prog 2009, 25:1111-1114.
Geraniol	Used in flavours (Citronella).	Oswald, M., Fischer, M., Dirninger, N., & Karst, F. (2007). Monoterpenoid biosynthesis in Saccharomyces cerevisiae. FEMS yeast research, 7(3), 413-21.

Homoeriodictyol	Taste-modifier with large potential in food applications and pharmaceuticals.	Yan, Y., Kohli, A. and Koffas, M. A. G. (2005) Biosynthesis of natural flavanones in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(9), 5610-5613.
Hydrocortisone	Steroid, a starting material for a variety of steroid drugs.	Szczebara, F. M., Chandelier, C., Villeret, C., Masurel, A., Bourot, S., Duport, C., Blanchard, S., Groisillier, A., Testet, E., Costaglioli, P., Cauet, G., Degryse, E., Balbuena, D., Winter, J., Achstetter, T., Spagnoli, R., Pompon, D. and Dumas, B. (2003) Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nat Biotechnol 21: 143-149.
Linalol	Aroma for hygiene products and cleaning agents.	Oswald, M., Fischer, M., Dirninger, N., & Karst, F. (2007). Monoterpenoid biosynthesis in Saccharomyces cerevisiae. FEMS yeast research, 7(3), 413-21.
L-lactic acid	Substrate production of PLA (poly lactic acid), renewable alternative for petroleum-based plastics.	Ishida, N., Saitoh, S., Ohnishi, T., Tokuhiro, K., Nagamori, E., Kitamoto, K., et al. (2006). Metabolic engineering of Saccharomyces cerevisiae for efficient production of pure L-(+)-lactic acid. *Applied Biochemistry and Biotechnology*, *131*(1-3), 795-807.
Luteolin	Anti-ischemic effects.	Leonard, E., Yan, Y., Lim, K. H. and Koffas, M. A. G. (2005) Investigation of two distinct flavone synthases for plant-specific flavone biosynthesis in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(12), 8241-8248.
Lycopene	Antioxidant, anti-cancer properties.	Verwaal, R., Wang, J., Meijnen, J.-P., Visser, H., Sandmann, G., Berg, J. a van den, et al. (2007). High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. Applied and environmental microbiology, 73(13), 4342-50.
Mannitol	Sweetener, hyperosmolar agent.	Costenoble, R., Adler, L., Niklasson, C., Liden, G. (2003) Engineering of the metabolism of Saccharomyces cerevisiae for anaerobic production of mannitol. FEMS Yeast Research 3, 17-25
Naringenin	Antioxidant, free radical scavenger, anti-inflammatory, carbohydrate metabolism promoter.	Yan, Y., Kohli, A. and Koffas, M. A. G. (2005) Biosynthesis of natural flavanones in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(9), 5610-5613.
Nicotianamine	Antihypertensive agent. Chelates metal cations and participates in metal nutrient transport.	Wada, Y., Kobayashi, T., Takahashi, M., Nakanishu, H., Mori, S. and Nishizawa, N. K. (2006). Metabolic engineering of Saccharomyces cerevisiae producing nicotianamine: potential for industrial biosynthesis of a novel antihypertensive substrate. *Bioscience, Biotechnology, and Biochemistry*, *70*(6), 1408-1415. doi:10.1271/bbb.50660
Patchoulol	Patchouli scent, also used in the synthesis of the chemotherapy drug Taxol.	Møller, K., Nielsen, K. F., Asadollahi, M. A., Schalk, M., Clark, A., & Nielsen, J. (2008). Production of Plant Sesquiterpenes in Saccharomyces cerevisiae : Effect of ERG9 Repression on Sesquiterpene Biosynthesis. *Biotechnology*, *99*(3), 666-677
P-coumaric acid	Starting material for flavors, fragrances, pharmaceuticals,	Vannelli, T., Wei Qi, W., Sweigard, J., Gatenby, A. a, & Sariaslani, F. S. (2007). Production of p-hydroxycinnamic acid from glucose in Saccharomyces cerevisiae and Escherichia coli by expression of heterologous genes from

	biocosmetics, health and nutrition products.	plants and fungi. Metabolic engineering, 9(2), 142-51.
PH3B (poly[(R)-3-hydroxybutyrate])	Biodegradable polyester material.	Carlson, R. and Srienc, F. (2006) Effects of recombinant precursor pathway variations on poly[(R)-3-hydroxybutyrate] synthesis in Saccharomyces cerevisiae. J Biotechnol 124: 561-573.
Pinocembrin	Antioxidant.	Yan, Y., Kohli, A. and Koffas, M. A. G. (2005) Biosynthesis of natural flavanones in Saccharomyces cerevisiae. Applied and Environmental Microbiology 71(9), 5610-5613.
Polyunsaturated fatty acids	Dietary supplements.	Tavares, S., Grotkjær, T., Obsen, T., Haslam, P., Napier, J. A., Gunnarsson, N., Tavares, S., et al. (2011). Metabolic Engineering of Saccharomyces cerevisiae for Production of Eicosapentaenoic Acid , Using a Novel Δ 5-Desaturase from Paramecium tetraurelia Metabolic Engineering of Saccharomyces cerevisiae for Production of Eicosapentaenoic Acid , Using a Novel ⌬ 5-Desaturase from Paramecium tetraurelia ⍰. Society. doi:10.1128/AEM.01935-10
Propane-1,2-diol	Production of cosmetics, detergents, polyester resins, pharmaceuticals, used as antifreeze and stabilizing agent.	Lee, W., & Dasilva, N. A. (2006). Application of sequential integration for metabolic engineering of 1,2-propanediol production in yeast. *Metabolic engineering*, *8*(1), 58-65.
Propane-1,3-diol	Feedstock in polypropylene terephthalate manufacturing. Applied in production of organic chemicals and polymers.	Cameron, D.C., Altaras, N.E., Hoffman, M.L. and Shaw, A.J. (1998) Metabolic engineering of propanediol pathways. Biotechnol Progr 14: 116-125.
Resveratrol	Phenolic compound of pharmacological importance also present in red wine. Antioxidant, cholesterol-lowering substance, anti-inflammatory agent. Believed to have preventive activity against cancer.	Becker, J., Armstrong, G., Vandermerwe, M., Lambrechts, M., Vivier, M., & Pretorius, I. (2003). Metabolic engineering of for the synthesis of the wine-related antioxidant resveratrol. *FEMS Yeast Research*, *4*(1), 79-85. doi: 10.1016/S1567-1356(03)00157-0.
Taxa-4(20),11(12)-	Precursor of anticancer drug Taxol.	Dejong, J. M., Liu, Y., Bollon, A. P., Long, R. M., Jennewein, S., Williams, D., et al. (2006). Genetic engineering of taxol biosynthetic genes in Saccharomyces cerevisiae. *Biotechnology and bioengineering*, *93*(2), 212-24.

Dern-5alpha-acetoxy-10beta-ol		
Tranilast	Antihistaminic used to treat skin disease associated with excessive fibrosis, regulates expression of cytokines.	Eudes A, Baidoo E, Yang F, Burd H, Hadi M, Collins F, Keasling J & Loque´ D (2011) Production of tranilast [N-(3',4'-dimethoxycinnamoyl)-anthranilic acid] and its analogs in yeast Saccharomyces cerevisiae. Appl Microbiol Biotechnol 89: 989–1000.
Valencene	Aroma component of citrus fruit.	Møller, K., Nielsen, K. F., Asadollahi, M. A., Schalk, M., Clark, A., & Nielsen, J. (2008). Production of Plant Sesquiterpenes in Saccharomyces cerevisiae : Effect of ERG9 Repression on Sesquiterpene Biosynthesis. *Biotechnology*, *99*(3), 666-677.
Vanillin	Flavouring agent.	Brochado, A. R., Matos, C., Møller, B. L., Hansen, J., Mortensen, U. H., & Patil, K. R. (2010). Improved vanillin production in baker's yeast through in silico design. *Microbial cell factories*, *9*(1), 84. doi:10.1186/1475-2859-9-84

Heterologous Compounds for Future Production by *S. cerevisiae*

Compounds	**Applications**	**Reference**
4-methylthiobutyl-desulfoglucosinolate	Effective cancer prevention agent.	Halkier, B. A., & Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. Annual review of plant biology, 57, 303-33. doi:10.1146/annurev.arplant.57.032905.105228
Artemisinin	Antimalarial drug.	Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature, 440(7086), 940-3.
Isopropanol	Biofuel.	Hanai, T., Atsumi, S., Liao, J. C., Engineered synthetic pathway for isopropanol production in Escherichia coli. Appl. Environ. Microbiol. 2007, 73, 7814-7818
Limonene	Used in food manufacturing and some medicines.	Chang, M. C. Y., & Keasling, J. D. (2006). Production of isoprenoid pharmaceuticals by engineered microbes. Nature chemical biology, 2(12), 674-81.
Menthol	Commercial applications are: flavours, fragrances, cleaning products, anticancer agent, antimicrobial agent.	Chang, M. C. Y., & Keasling, J. D. (2006). Production of isoprenoid pharmaceuticals by engineered microbes. Nature Chemical Biology, 2(12), 674-81. doi: 10.1038/nchembio836.
Penicillin G	An antibiotic compound	Gidijala, L., Kiel, J. a K. W., Douma, R. D., Seifar, R. M., van Gulik, W. M., Bovenberg, R. a L., Veenhuis, M., et al. (2009). An engineered yeast efficiently secreting penicillin. PloS one, 4(12), e8317. doi:10.1371/journal.pone.0008317
Propanol	Biofuel.	Atsumi, S., Hanai, T., Liao, J. C., Non fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 2008, 451, 86-89.

***Note:** A metabolite is considered to be native if it is present in the consensus yeast metabolic reconstruction, Herrgård, Swainston *et al.* (2008) "A consensus yeast metabolic reconstruction obtained from a community approach to systems biology" Nature Biotechnol. 26, 1155-1160.

Chapter 3: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 1 – Metabolic Modeling for Strain Improvement

Table ST3.3: List of primers used in this study.

Primer name	Primer sequence	Purpose
His3_Fw	attgctatgacccgggCATAACACAGTCCTTTCCCGC	Amplify *HIS3* marker from pWJ1213
His3_Rev	attgctatgtcccgggTCACACCGCATAGATCCGTC	Amplify *HIS3* marker from pWJ1213
MarkSeq_Fw	CTGTGTCTATGAAAGTCGACGCG	Sequencing marker in plasmid ARB021
MarkSeq_Rev	GGGGCTGGCTTAACTATGC	Sequencing marker in plasmid ARB021
dKL5'	gtcagcggccgcatccctgcTTCGGCTTCATGGCAATTCCCG	Amplify N-terminal URA3 from plasmid pWJ1042
Kl URA3 3'-int	GAGCAATGAACCCAATAACGAAATC	Amplify N-terminal URA3 from plasmid pWJ1042
Kl URA3 5'-int	CTTGACGTTCGTTCGACTGATGAGC	Amplify C-terminal URA3 from plasmid pWJ1042
cKL3'	cacggcgcgcctagcagcGGTAACGCCAGGGTTTTCCCAGTCAC	Amplify C-terminal URA3 from plasmid pWJ1042
GDH1(UP)_Fw	GTCATCATTTCAAATATATG	Amplify *GDH1* upstream fragment from VG0 genomic DNA
GDH1(UP)_Rev	gcagggatgcggccgctgacATAGTCTAAAAGAAAGAAAA	Amplify *GDH1* upstream fragment from VG0 genomic DNA
GDH1(DW)_Fw	ccgctgctaggcgcgccgtgTTCTTTTTCTTTTTGGTCTC	Amplify *GDH1* downstream fragment from VG0 genomic DNA
GDH1(DW)_Rev	AAAGTATACGTAATCTAAGT	Amplify *GDH1* downstream fragment from VG0 genomic DNA
GDH1_Ver_FW	TTGCAAGTTAAAGCGGTC	Analytical PCR for verifying *GDH1* deletion from VG2 genomic DNA
GDH1_Ver_REV	GCCCATGCATTTTCAGT	Analytical PCR for verifying *GDH1* deletion from VG2 genomic DNA
PDC1(UP)_Fw	TCGTTTAAGAGAAATTCTCC	Amplify *PDC1* upstream fragment from VG0 genomic DNA
PDC1(UP)_Rev	gcagggatgcggccgctgacGCGATTTAATCTCTAATTAT	Amplify *PDC1* upstream fragment from VG0 genomic DNA
PDC1(DW)_Fw	ccgctgctaggcgcgccgtgTTTGATTGATTTGACTGTGT	Amplify *PDC1* downstream fragment from VG0 genomic DNA
PDC1(DW)_Rev	GTGATGGCACATTTTTGCAT	Amplify *PDC1* downstream fragment from VG0 genomic DNA
PDC1_Ver_FW	AGCAATGGCTTGCTTAATAG	Analytical PCR for verifying *PDC1* deletion from VG2 genomic DNA
PDC1_Ver_REV	ATTTGCAAAATGCATAACCT	Analytical PCR for verifying *PDC1* deletion from VG2 genomic DNA
GDH2(UP)_Fw	AGCAATGTCATACTGGCC	Amplify *GDH2* upstream fragment from VG3 genomic DNA
GDH2(UP)_Rev	gcagggatgcggccgctgacTTGAGATCGTGACAATCAC	Amplify *GDH2* upstream fragment from VG3 genomic DNA
PGK1_GDH2(Dw)_Fw	ccgctgctaggcgcgccgtgTCTAACTGATCTATCCAAAACTG	Amplify *PGK1_GDH2* (DW) fragment from pPGK1_GDH2
PGK1_GDH2 (DW)_Rev	GAATCATCCATTTCAATCC	Amplify *PGK1_GDH2* (DW) fragment from pPGK1_GDH2
PGK1verifb	GTCACACAACAAGGTCCTA	Analytical PCR for verifying GDH2 overexpression from VG4 genomic DNA
Gdh2verifb	GGTTTTCTACAATCTCCAAAAGAG	Analytical PCR for verifying GDH2 overexpression from VG4 genomic DNA

[a]Small case letters indicate non homologous region of the primer, rather they code for restriction sites for cloning purposes or tails for fusion PCR.

[b]Primers by TL Nissen et al. (2000)

Chapter 4: Metabolic Engineering for Vanillin Production in *S. cerevisiae*. Part 2 – Pathway and Regulatory Engineering

Table ST4.4: List of oligonucleotides used in this study.

Name	Sequence	Purpose
O1_f	ATGGTGCAAGACACATCAAGC	*npgA* (*Aspergillus nidulans*) amplification from pESC-npgA-pcbAB
O2_r	attgctatgaGGTACCTTAGGATAGGCAATTACACACCC	
O3_f	GGAGTTTAGTGAACTTGCAA	*npgA* (*Aspergillus nidulans*) verification in pARB118
O4_r	GATGACTACGGAAAGCTTT	
O5_f	attgctatgaGGCGCGCCGAGTTTGGCCGAGTGGTT	*MET15* (*S. cerevisiae*) amplification from CEN.PK 113 7D gDNA
O6_r	attgctatgaGGCGCGCCGAAACCTCCATCATCCTCTT	
O7_f	CTGTGTCTATGAAAGTCGACGCG	*MET15* (*S. cerevisiae*) verification in pARB030
O8_r	GGGGCTGGCTTAACTATGC	
UO1_f	AGTCGCCCCCUTTCGGCTTCATGGCAATTCCCG	Recyclable URA3 marker amplification from pWJ1042
UO2_r	ACTCACTGCUGGTAACGCCAGGGTTTTCCCAGTCAC	
UO3_f	GGGTTTAAU ATAAAGCAGCCGCTACCAAA	*YPRCΔ15* - UP fragment amplification from CEN.PK 113 7D gDNA
UO4_r	AGGGGGCGACUAATGGAAGGTCGGGATGAG	
UO5_f	AGCAGTGAGUGCTGAGGGTTTAATTAAGTCCTCAGCCCTTTC GTCATGGACACTTCT	*YPRCΔ15* - DW fragment amplification from CEN.PK 113 7D gDNA
UO6_r	GGACTTAAUTACCAACGGACTTACCTTCAG	
UO7_f	GGGTTTAAUCACACACCATAGCTTCAAA	P_{TEF1} amplification from CEN.PK 113 7D gDNA
UO8_r	AGGGGGCGACUTTTGTAATTAAAACTTAGATTAG	
UO9_f	AGCAGTGAGUCATGTAATTAGTTATGTCACGC	T_{CYC1} amplification from CEN.PK 113 7D gDNA
UO10_r	GGACTTAAUGCAAATTAAAGCCTTCGA	
UO11_f	AGTCGCCCCCUATGACCGCAAAGACTTTTCT	*HAP4* amplification from CEN.PK 113 7D gDNA
UO12_r	ACTCACTGCUCAACATGCCTATTTCAAAATAC	
UO13_f	AGTCGCCCCCUATGGGTGACACTAAGGAGC	*hsOMT* amplification from pARB033
UO14_r	GGACTTAAUCGCGTCGACTTTCATAGA	
O9_f	CTTGACGTTCGTTCGACTGATGAGC	Internal *URA3* primers for amplifing bipartite fragments for yeast transformation
O10_r	GAGCAATGAACCCAATAACGAAATC	
O11_f	ATCTGCTGCGCTACCACTG	Amplification of *YPRCΔ15 locus* from CEN.PK 113 7D gDNA (verification)
O12_r	GGCTTGTGGTCACCTGTCAT	

Note 1: Lower case letters indicate "jungle" to ensure DNA restriction by the endonucleases.

Note 2: UO_# distinguishes the USER oligonucleotides.

Note 3: Gray highlighting designates restriction sites or USER tails, in case of USER cloning primers.

Note 4: Underlined text designate *PacI/Nt.BbvCI* User cassette.

Chapter 5: Minimization of Metabolite Balance Explains Genetic Interactions within the Yeast Metabolic Network

Supplementary methods

All the simulation algorithms were formulated as Linear Programming (LP) optimization problems and implemented in C++ using the GLPK (http://www.gnu.org/s/glpk/).

Supplementary Method 5.1: Flux Balance Analysis [2] was formulated as follows:

$$\begin{aligned} &\max v_{Growth} \\ &s.t. \quad S \cdot v = 0 \\ &\qquad v_i^{lb} \le v_i \le v_i^{ub} \qquad \forall i \in N \end{aligned}$$

Supplementary Method 5.2: Minimization of overall intracellular flux [3] was formulated as follows:

$$\begin{aligned} &\min \sum_i |v_i| \qquad \forall i \in N \\ &s.t. \quad S \cdot v = 0 \\ &\qquad v_i^{lb} \le v_i \le v_i^{ub} \qquad \forall i \in N \end{aligned}$$

Supplementary Method 5.3: Minimization of metabolic adjustment (lMoMA, [4]) was formulated as follows:

$$\begin{aligned} &\min \sum_i |v_i^{WT} - v_i| \qquad \forall i \in N \\ &s.t. \quad S \cdot v = 0 \\ &\qquad v_i^{lb} \le v_i \le v_i^{ub} \qquad \forall i \in N \end{aligned}$$

Supplementary figures

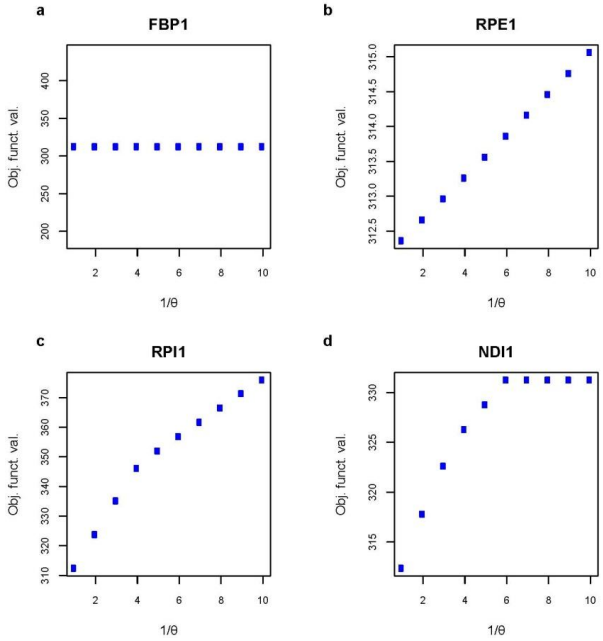


Figure S5.1: Profiles obtained for the objective function value (minimization of overall intracellular flux) using alternative stoichiometry representations of *S. cerevisiae* genome-scale model [5]. This analysis is complementary to and based on the same simulation constraints as used for **Fig. 5.1** in the main text. As the contribution of each flux to the objective function changes based on the corresponding stoichiometry representation, different situations could be described, leading either to the same (a, b) or distinct (c, d) optimal solutions. To illustrate these different situations, four reactions within the model were linearly scaled one at a time by multiplying by a scalar θ as described in **Methods**. **a)** Linear scaling of the reaction *FBP1*. As *FBP1* carries no flux under the simulated conditions, the scaling of this reaction does not affect the objective function value. **b)** Linear scaling of the reaction *RPE1*. For the range of θ tested, the objective function value perfectly correlated with the scaling factor of the reaction *RPE1*, which indicates that all obtained solutions are in fact the same optimal solution (or alternative optimal solutions, depending on the model complexity). This profile means that there is no pathway alternative to *RPE1* that can become part of the optimal solution. **c)** Linear scaling of the reaction *RPI1*. For the range of tested θ, at least two slopes are observed when correlating the objective function value with 1/θ, indicating that at least two different optimal solutions were found for the same problem. **d)** Linear scaling of the reaction *NDI1*. Similarly to that of *RPI1*,

scaling of *NDI1* leads to different optimal solutions. However, in this case, the objective function value stabilizes after a given θ, which means that this flux no longer influences the optimization. Such profile suggests that the optimal solution found after the given value of θ does no longer involve *NID1*, but an alternative pathway, which became preferred for minimizing the objective function.

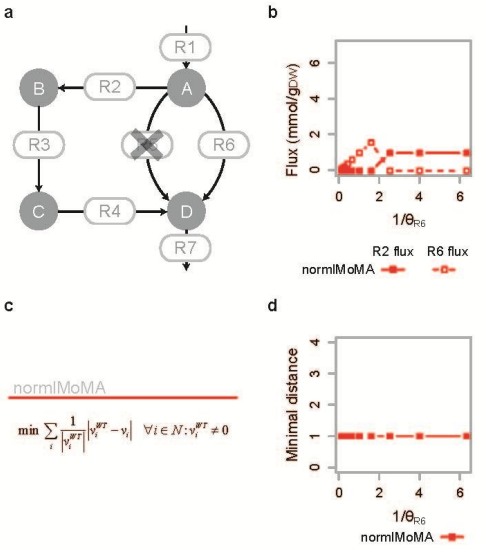


Figure S5.2: A toy-model illustrating how, and why, alternative stoichiometry representations influence simulation of minimization of metabolic adjustment by using normalized lMoMA – **normlMoMA**. **a)** Toy-model: *R1* to *R7* and *A* to *D* represent reactions and metabolites, respectively. In the wild-type, or reference, flux goes from *A* to *D* via *R5*. *R6* and *R2*-*R3*-*R4* are two alternative pathways for flux re-distribution after deletion of *R5*. **b)** Flux through reactions *R2* (full symbols) and *R6* (open symbols) obtained after simulation with normlMoMA by using alternative representations of reaction *R6* (given by different θ_{R6}, **Methods**). **c)** Formulation of normlMoMA objective function (**Methods**). **d)** Optimal objective function value (distance) obtained for minimization of metabolic adjustment as function of θ_{R6}.

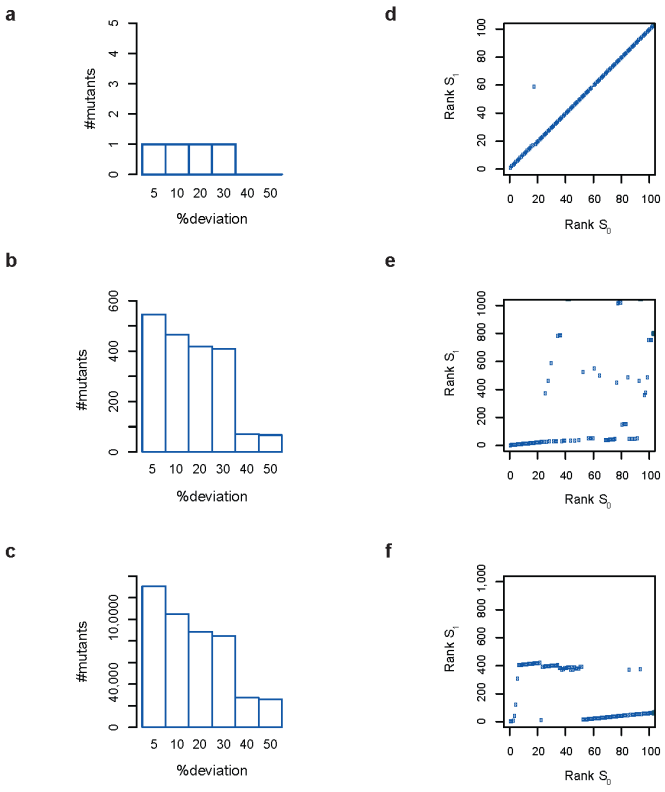


Figure S5.3: Stoichiometry representation impacts the design of metabolic engineering strategies for improving succinate production in *S. cerevisiae* depending on the nature of the objective function formulation. Shown is the comparison of predicted succinate yield for **a)** single, **b)** double and **c)** triple gene knockout mutants obtained with two alternative stoichiometric matrices (S_0 and S_1, **Methods**). The number of mutants diverging in their lMoMA-predicted succinate yield for the two alternative representations of stoichiometry is represented on the y-axis, while the percentage of deviation of product formation by the mutants relative to S_0 is represented on the x-axis. **d-f)** Comparison of ranks of lMoMA-predicted metabolic engineering strategies for improving succinate production obtained by using S_0 and S_1 for d) single, e) double and f) triple gene knockout mutants.

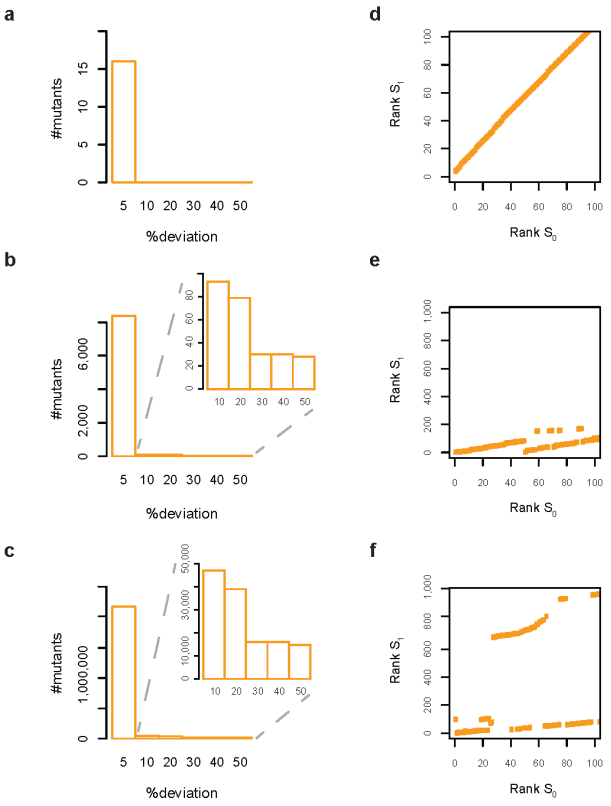


Figure S5.4: Stoichiometry representation impacts the design of metabolic engineering strategies for improving vanillin-glucoside production in *S. cerevisiae* depending on the nature of the objective function formulation. Shown is the comparison of predicted vanillin-glucoside yield for **a)** single, **b)** double and **c)** triple gene knockout mutants obtained with two alternative stoichiometric matrices (S_0 and S_1, **Methods**). The number of mutants diverging in their lMoMA-predicted vanillin-glucoside yield for the two alternative representations of stoichiometry is represented on the y-axis, while the percentage of deviation of product formation by the mutants relative to S_0 is represented on the x-axis. **d-f)** Comparison of ranks of lMoMA-predicted metabolic engineering strategies for improving vanillin-glucoside production obtained by using S_0 and S_1 for d) single, e) double and f) triple gene knockout mutants.

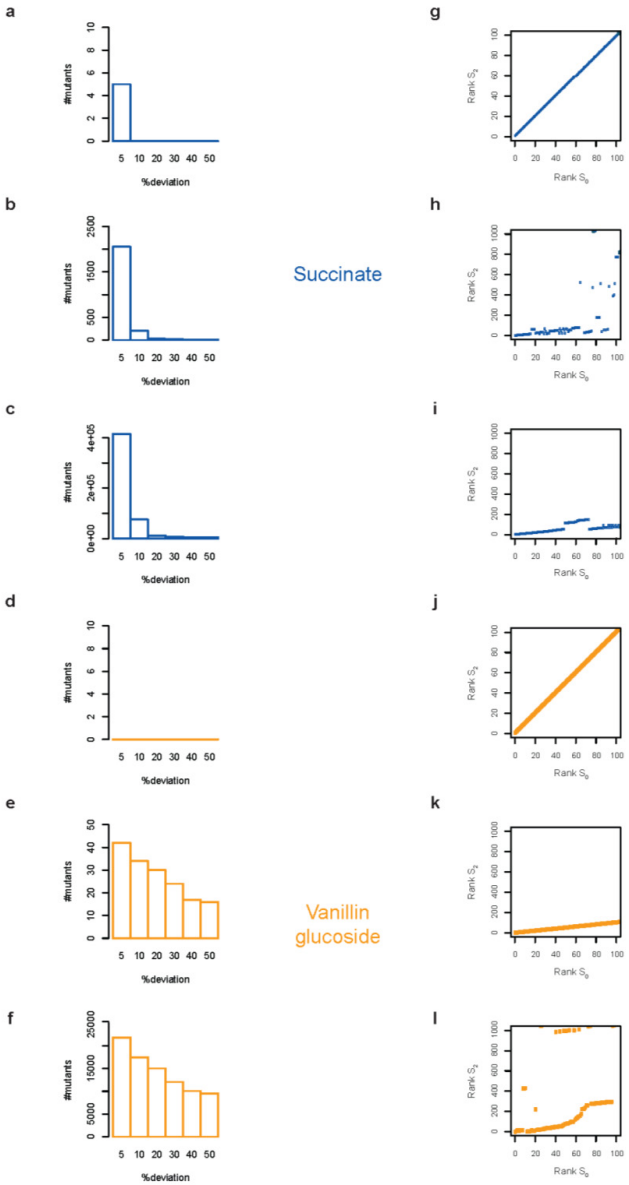


Figure S5.5: Stoichiometry representation impacts the design of metabolic engineering strategies for improving succinate and vanillin-glucoside yields in *S. cerevisiae* depending on the nature of the objective function formulation. **a-f)** Number of mutants diverging in their lMoMA-predicted a-c) succinate and d-f) vanillin-glucoside yields for two alternative representations of stoichiometry, S_0 and S_2 (**Methods**). Results for a,d) single, b,e) double and c,f) triple gene knockout mutants are presented. **g-l)** Comparison of ranks of lMoMA-predicted metabolic

engineering strategies for improving g-i) succinate and j-l) vanillin-glucoside production obtained by using S_0 and S_2. Results for g,j) single, h,k) double and i,l) triple gene knockout mutants are presented.

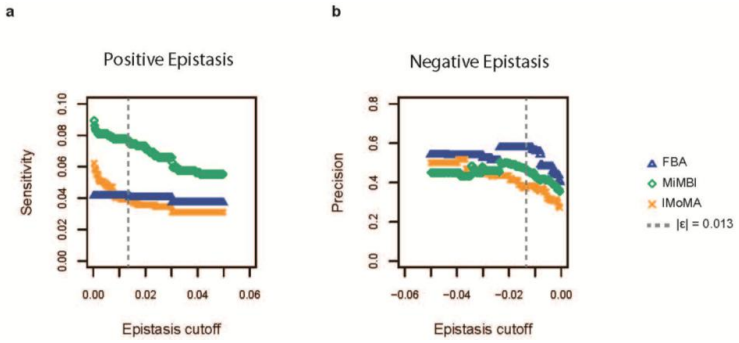


Figure S5.6: Comparing the performance of MiMBI, FBA and lMoMA for predicting yeast genetic interactions using i*AZ900* metabolic model.

Supplementary tables

Table ST5.1: Number of lMoMA-predicted lethal gene/reaction knockouts in *S. cerevisiae* that differ between alternative representation of stoichiometry (S_1 and S_2), relative to S_0 (**Methods**). The yeast genome scale model [5] was constrained as per Szappanos *et al*, 2011, and FBA was used to generate the reference flux distributions. Single, double and triple gene/reaction deletions were simulated.

Number of Deletions	S_2		S_1		Total number of combinations	
	Reactions	Genes	Reactions	Genes	Reactions	Genes
1	0	0	4	3	387	424
2	0	13	1526	1166	74691	89676
3	189	4565	289664	225548	9585345	12614424

Table ST5.2: lMoMA-predicted epistatic interactions within *S. cerevisiae* genome-scale metabolic model [5]. Simulations were performed using three alternative representations of stoichiometry, S_0, S_1 and S_2 (**Methods**). The yeast genome-scale metabolic model was constrained as in Szappanos *et al.* 2011.

	S_0	S_2	S_1
Positive interactions	2219	2154	2087
Negative interactions	840	781	742
Synthetic lethals	197	217	97
Total number of interactions			89676

Table ST5.3: MiMBl-predicted epistatic interactions within *S. cerevisiae* genome-scale metabolic model [5]. Simulations were performed using two alternative representations of stoichiometry, S_0 and S_1 (**Methods**). The yeast genome-scale metabolic model was constrained as in Szappanos *et al.* 2011.

	S_0	S_1
Positive interactions	2019	2019
Negative interactions	811	811
Synthetic lethals	198	198
Total number of interactions		89676

Supplementary notes

Supplementary Note 5.1: Toy-model

Reaction ID	*Reaction*	*[lower bound, upper bound]*
R1:	Axt -> A	[1,1]
R2:	A -> B	[0,100]
R3:	B -> C	[0,100]
R4:	C -> D	[0,100]
R5:	A -> D	[0,100]
R6:	A -> D	[0,100]
R7:	D -> Dxt	[0,100]

Where *R1* to *R7* represent reactions, *A, B C* and *D* represent internal metabolites, Axt and Dxt represent external metabolites. Constraints are represented as [*lower bound, upper bound*].

Reference flux distribution

R1	1
R2	0
R3	0
R4	0
R5	1
R6	0
R7	1

Reference intracellular metabolite turnovers

A	1
B	0
C	0
D	1

Supplementary Note 5.2: normlMoMA

One possible approach to normalize the objective function variables with respect to the stoichiometric representation, can be achieved by dividing each variable by its value on the wild-type flux distribution – normalized lMoMA (normlMoMA). Albeit being simple, this normalization method has several drawbacks, the first one being the fact that many reactions have null fluxes on the wild-type flux distribution. The simplest solution to this problem would be to disregard the contribution of null fluxes to the objective function, biologically meaning that the onset of new enzymes would be costless. This normalization solves the inconsistency of the initial problem, since the calculated distance is constant independently of the matrix representation. However, as *R2-R3-R4* and *R6* (toy-model, **Fig. S5.2**) have null fluxes on the wild-type flux distribution, and therefore they do not take part of the objective function, they represent alternative optima to the problem. Moreover, exclusion of the wild-type null fluxes from the objective function alters the biological principle underlying MoMA and increases the number of alternative optimal solutions.

Supplementary Note 5.3: Impact of scaling stoichiometry on finding the optimal solution for metabolic flux distributions using FBA-like objective functions – Analytical evidence

Consider the problem

$$\begin{aligned} &\min\ c_a v_a \\ &s.t.\quad S \cdot v = b \\ &\qquad v_i \geq 0, \forall i \in N \end{aligned}$$

Where a is the index of variable v_a on the matrix S. Note that this is a particular case of having a sum of non-normalized variables on the objective function, where all the entries of the

objective function coefficients vector c are zero, except for the a^{th} entry $(=c_a)$. As described in **Methods**, after scaling matrix S with Θ, the optimality condition becomes:

$$c_j - c_B'\left(B\Theta_B\right)^{-1}\theta_{jj}S_j \geq 0$$

where j is the index of variable v in matrix S, $\overline{c_j}$ is the reduced cost of the variable v_j, c_j is the objective function coefficient of v_j, c_B is the vector containing the objective coefficients of basic variables, S_j is the j^{th} column of matrix S, Θ is a nxn positive diagonal matrix (**scaling matrix**) and θ_{jj} is the scaling factor for the j^{th} column of matrix S [6].

Three different cases should be considered and in all of them we will show that the **optimality condition is satisfied**.

a) $a \notin B \; and \; a \neq j$

$\therefore c_B = [0] \; and \; c_j = 0$

$\therefore \overline{c_{j\Theta}} = 0$

The optimality condition is satisfied.

b) $a \notin B \; and \; a = j$

$\therefore c_B = [0] \Rightarrow \overline{c_j} = c_j \; and \; c_j \geq 0 \; (optimality)$

$\therefore \overline{c_{j\Theta}} = c_j \Rightarrow \overline{c_{j\Theta}} \geq 0$

The optimality condition is satisfied.

c) $a \in B$

$if \; a = j, then \; j \in B \Rightarrow \overline{c_j} = \overline{c_{j\Theta}} = 0$

$if \; a \neq j \Rightarrow c_j = 0$

$\therefore -c_B'\left(B\Theta_B\right)^{-1}\theta_{jj}S_j \geq 0$

As $c_B'B^{-1}S_j \leq 0$ and all entries of the matrix Θ are positive, the optimality condition is always satisfied.

References

1. Szappanos B, Kovács K, Szamecz B, et al.: **An integrated approach to characterize genetic interaction networks in yeast metabolism.** *Nature Genetics* 2011, **43**:656-62.

2. Varma A, Palsson BØ: **Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors**. *Journal of Theoretical Biology* 1993, **165**:477-502.

3. Blank LM, Kuepfer L, Sauer U: **Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast.** *Genome Biology* 2005, **6**:R49.

4. Becker S a, Feist AM, Mo ML, et al.: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nature Protocols* 2007, **2**:727-38.

5. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13**:244-53.

6. Bertsimas D, Tsitsiklis JN: *Introduction to Linear Optimization*. First Edit. Belmont, Massachusetts: Athena Scientific; 1997.

CMB is an Engineering Center of Excellence funded by the Danish Research Agency. It is a collaboration between an acknowledged research manager, his/her institute and university, and the Research Agency. An Engineering Center of Excellence is a research institute of first-class quality with tradition for cooperation with industry.

Center for Microbial Biotechnology
Department of Systems Biology
Technical University of Denmark
Building 223
DK-2800 Kgs. Lyngby
Denmark

Phone: +45 4525 2525
Fax: +45 4588 4148

www.cmb.dtu.dk

ISBN-nr: 978-87-91494-13-0