UC San Diego
UC San Diego Electronic Theses and Dissertations

Title

A stoichiometric model of Escherichia coli 's macromolecular synthesis machinery and its integration with metabolism

Permalink
https://escholarship.org/uc/item/1z9752m7

Author

Thiele, Ines

Publication Date

2009

Peer reviewed|Thesis/dissertation

eScholarship.org
Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

A stoichiometric model of *Escherichia coli*'s macromolecular synthesis machinery and its integration with metabolism.

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics

by

Ines Thiele

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Steven P. Briggs, Co-Chair
Professor Alexander Hoffmann
Professor Milton H. Saier
Professor Julian I. Schroeder

2009

Copyright
Ines Thiele, 2009
All rights reserved.

The dissertation of Ines Thiele is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2009

DEDICATION

To Renate, Steffen, Dana, and Ronan.

TABLE OF CONTENTS

Signature Page . . . iii

Dedication . . . iv

Table of Contents . . . v

List of Figures . . . viii

List of Tables . . . x

Acknowledgements . . . xi

Vita and Publications . . . xiii

Abstract of the Dissertation . . . xv

Chapter 1 Introduction to molecular systems biology . . . 1
- 1.1 Constraint-based reconstruction and analysis . . . 2
- 1.2 Basic principles underlying the constraint-ba-sed reconstruction & analysis approach . . . 4
- 1.3 Reconstruction of metabolic networks in a nutshell . . . 6
- 1.4 Mathematical characterization of network capabilities . . . 8
- 1.5 Tools for analyzing network states. . . . 9
- 1.6 Preview of the dissertation . . . 11

Chapter 2 *Escherichia coli* . . . 13
- 2.1 Key properties . . . 13
- 2.2 *E.coli* by numbers . . . 16
- 2.3 Conclusion . . . 21

Chapter 3 A protocol for generating a high-quality genome-scale metabolic reconstruction . . . 22
- 3.1 Introduction . . . 23
- 3.2 General procedure . . . 29
- 3.3 Materials . . . 30
  - 3.3.1 Equipment . . . 30
  - 3.3.2 Equipment setup . . . 31
- 3.4 Procedure . . . 33
  - 3.4.1 Creating a draft reconstruction . . . 33
  - 3.4.2 Manual reconstruction refinement . . . 35
  - 3.4.3 Conversion from reconstruction to mathematical model . 59
  - 3.4.4 Network evaluation = 'Debugging mode' . . . 61
  - 3.4.5 Data assembly and Dissemination . . . 77
- 3.5 Timing . . . 78
- 3.6 Troubleshooting . . . 78

3.7 Anticipated Results 79

Chapter 4 State-of the art reconstructions of cellular networks 83
4.1 Available metabolic reconstructions 83
4.1.1 Metabolic reconstructions of *Escherichia coli* 84
4.2 Reconstruction jamborees 86
4.3 Reconstruction of other cellular networks 91
4.3.1 Reconstruction of signaling networks 91
4.3.2 Reconstruction of transcriptional regulatory networks . . 92
4.3.3 Reconstruction of transcription and translation 93
4.3.4 Reconstruction of integrated networks 93

Chapter 5 From Biology to computers - Representation of biological processes in computable format 94
5.1 Transcription: From DNA to RNA 95
5.2 Translation 100
5.3 Protein Maturation. 103
5.4 Metallo-ions incorporation. 104
5.5 Protein Folding. 108
5.6 mRNA degradation. 111
5.7 tRNA and rRNA processing. 112

Chapter 6 Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery 113
6.1 Introduction 114
6.2 Methods 118
6.3 Results and Discussion 123
6.3.1 Legacy data. 124
6.3.2 Reconstruction approach. 124
6.3.3 Unique properties of the 'E-matrix'. 124
6.3.4 'E-matrix' versus available databases. 126
6.3.5 Knowledge gaps. 127
6.3.6 Network topology. 128
6.3.7 Validation of the 'E-matrix' functionality - Ribosome production. 129
6.3.8 The effect of *in silico* rRNA operon deletions on ribosome production. 129
6.3.9 Integration of '-omics' data into 'E-matrix'. 132
6.3.10 Defining functional modules. 134
6.3.11 Integration with other cellular functions. 135
6.4 Conclusion 136

Chapter 7 Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional & translational machinery 138
7.1 Introduction 139
7.2 Material and Methods 142
7.3 Results 154

7.4 Conclusion 161

Chapter 8 An integrated model of macromolecular synthesis and metabolism of *Escherichia coli*. 163
8.1 Introduction 163
8.2 Materials and Methods 165
8.2.1 Constraint-based reconstruction and modeling appro-ach . 165
8.2.2 Reconstruction of the 'ME'-matrix 166
8.3 Results & Discussion 175
8.3.1 Properties of the reconstruction 175
8.3.2 Model Validation 180
8.3.3 Novel Application 185
8.4 Conclusion 196

Chapter 9 Conclusion: Towards whole-cell modeling 198
9.1 Constraint-based reconstruction and analysis approach as tool of choice 198
9.2 Challenges in reconstructing non-metabolic functions 199
9.3 Integration of metabolism and macromolecular synthesis - A stiff matrix 200
9.4 Applications of integrated models 201
9.5 What's next 202

Chapter 10 Glossary 205

Bibliography 209

LIST OF FIGURES

Figure 1.1: Bringing genomes to life 3
Figure 1.2: General workflow for reconstructing cellular networks 7

Figure 2.2: Dependency of cellular properties and growth rate in *E. coli*. 18

Figure 3.1: Overview of the procedure to iteratively reconstruct metabolic networks 24
Figure 3.2: Refinement of reconstruction content 28
Figure 3.3: Information used for draft reconstruction 33
Figure 3.4: Example of a draft reconstruction 34
Figure 3.5: Assessing the metabolic ”environment” or ”connectivity” of a metabolite 37
Figure 3.6: Examples of possible Gene-protein-reaction associations 41
Figure 3.7: Conversion of reconstruction into a condition-specific model 47
Figure 3.8: Example of biomass composition determination for *Pseudomonas putida* KT 2440. 48
Figure 3.9: Flow chart to calculate the fractional contribution of a precursor to the biomass reaction 50
Figure 3.10: Determination of the content of soluble pool 51
Figure 3.11: Calculation of biomass coefficient of ions 52
Figure 3.12: Growth-associated maintenance cost 54
Figure 3.13: Schematic representation of the assembly of the biomass reaction . . . 56
Figure 3.14: Layout of excel sheets as input for xls2model function. 59
Figure 3.15: Model in Matlab format and the COBRA Toolbox 60
Figure 3.16: Example of a stoichiometrically balanced cycle 64
Figure 3.17: Flow chart on debugging network reactions that cannot carry flux . . . 65
Figure 3.18: Gap analysis. 80
Figure 3.19: Network evaluation 81
Figure 3.20: Physiological properties that can be important for network evaluation 82

Figure 4.1: Growth of genome sequences and genome-scale metabolic reconstructions 84
Figure 4.2: List of information that needs to be associated with each metabolite and each reaction in the consensus reconstruction 88
Figure 4.3: Workflow of the *Salmonella* reconstruction jamboree 90

Figure 5.1: Examples of transcription units in *E. coli* 95

Figure 6.1: Overview of constraint-based reconstruction and analysis 116
Figure 6.2: Content of the ’E-matrix’ 117
Figure 6.3: Comparison of *in vivo* and *in silico* maximal number of ribosomes . . . 128
Figure 6.4: rRNA operon deletion study 130
Figure 6.5: Integration of ”-omics” data into ’E-matrix’ as reaction constraints . . 133
Figure 6.6: Schematic representation of *in silico* functional modules 135

Figure 7.1: Schematic illustration of alternate optimal solutions 140
Figure 7.2: Schematic representation of the mRNA and protein pools present in the *E*-matrix. 145

Figure 7.3: Schematic representation of the participation of tr/tr enzymes in network reactions 146
Figure 7.4: Illustration of reduction of flux span 149
Figure 7.5: Properties of AOS in E-matrix 155
Figure 7.6: Principal component analysis 160

Figure 8.1: Functional synergy between the metabolic network and the macromolecular synthesis network 164
Figure 8.2: Schematic depiction of the integration of the metabolic network and transcriptional/translational network 165
Figure 8.3: Stoichiometric coefficients in the ME-matrix 180
Figure 8.4: Sensitivity analysis 181
Figure 8.5: Comparison of *in vivo* growth phenotype predictions with *in silico* calculation 185
Figure 8.6: The growth of the ME-matrix and iAF1260 as a function of substrate uptake rate 186
Figure 8.7: Codon usage 190
Figure 8.8: Relative growth rates of biased strains 191
Figure 8.9: Growth rates biased strains 192
Figure 8.10: Genome parameters affecting *in silico* strain growth rates 194

LIST OF TABLES

Table 2.1: Macromolecular composition of an average *E. coli* cell 15
Table 2.2: Parameters related to the growth and macromolecular composition of bacterial cells 16
Table 2.3: Parameters related to the growth and macromolecular composition of bacterial cells 17
Table 2.4: Sigma factors in *E. coli* 21

Table 3.1: Data sources frequently used for metabolic reconstructions 26
Table 3.2: General error modes in metabolic networks 27
Table 3.3: List of experimental data used for reconstruction, modeling & network evaluation 36
Table 3.4: List of cellular compartments used in reconstructions 40
Table 3.5: Confidence score system 43
Table 3.6: List of spontaneous reactions 45
Table 3.7: Useful functions in the COBRA Toolbox for reconstruction 58
Table 3.8: List of tools for network refinement and expansion 72

Table 4.1: History of reconstruction of the *E. coli*'s metabolic network 85
Table 4.2: Parallel metabolic reconstructions 87

Table 5.1: List of reactions that synthesize iron-sulfur-cluster 106
Table 5.2: Template reactions for DnaK/J-GrpE dependent folding 109
Table 5.3: Template reactions for GroEL/ES dependent folding. 110

Table 6.1: List of rRNA transcription units & their basic characteristics 122
Table 6.2: Reactions per subsystems 125
Table 6.3: Overview of the 'E-matrix' content 126

Table 7.1: Overview of the '*E*-matrix' content 142
Table 7.2: List of *E. coli* cell specific parameters 144
Table 7.3: *in vivo* essential *E. coli* genes not expressed in all AOS 158

Table 8.1: Information used for synthesis reactions of *E. coli*'s genes 168
Table 8.2: Functional coverage of the ME-matrix 176
Table 8.3: Metallo-ions and prostethic groups included in the ME-matrix. 178
Table 8.4: Comparison of predicted and experimentally determined growth rates . . 182
Table 8.5: Improved essentiality 183
Table 8.6: Remaining false positives 184
Table 8.7: Codons recognition by tRNA in the ME-matrix - Part I 187
Table 8.8: Codons recognition by tRNA in the ME-matrix - Part II 188
Table 8.9: Network reactions with high reduced cost 196

ACKNOWLEDGEMENTS

I thank the National Institutes of Health for the research grants that have funded my graduate studies at UCSD.

I am very thankful to my advisor, Bernhard Palsson, for believing in me and my work. His scientific guidance, professional mentorship, motivation, and academic support had a great influence on my scientific development.

The graduate students and researchers in the Palsson lab deserve special thanks for their comraderie, knowledge, and support. In particular, my officemates, Monica, Vasiliy, and Karsten, which were always up for a joke and chat, deserve an extra thank-you. These conversations made some days just so much better. Neema became a "partner in crime" during these long discussions about science. I am very grateful for the support, patience and motivation of Nathan Price, who was certainly a key person for prolonging my stay in San Diego.

Most of all, I thank my parent, Renate and Steffen, my sister, Dana, and my fiance, Ronan, for all these phone calls and visits over the past years, which supported and strengthened me. Without you, this thesis would not have been possible.

The text of Chapter 1, in part or in full, is a reprint of the material as it appears in I. Thiele and B.Ø. Palsson, Bringing genomes to life: The use of genome-scale *in silico* models, Chapter 2, Introduction to Systems Biology, Humana Press (2007), and in I. Thiele and B.Ø. Palsson, Fundamentals of Constraint-Based Methods,Nature Biotechnology (submitted). I was the primary author of these publications and the co-author participated and directed the research, which forms the basis for Chapter 1.

The text of Chapter 3, in part or in full, is a reprint of the material as it appears in I. Thiele and B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction, submitted, 2009. I was the primary author of this publication and the co-author participated and directed the research which forms the basis for Chapter 3.

The text of Chapter 4, in part, is a reprint of the material as it appears in I. Thiele and B. Ø. Palsson, 2D genome annotation jamborees: A community effort in systems biology, submitted, 2009. I was the primary author of this publication and the co-author participated and directed the research which forms the basis for Chapter 4.

The text of Chapter 6, in full, is a reprint of the material as it appears in I. Thiele, N. Jamshidi, R.M.T. Fleming, B.Ø. Palsson, Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization, PLoS Comp. Biol., 2009. I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for Chapter 6.

The text of Chapter 7, in full, is a reprint of the material as it appears in I. Thiele, R.M.T. Fleming, A. Bordbar, J. Schellenberger, B.Ø. Palsson, Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional and translational machinery, *to be submitted.* I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for Chapter 7.

The text of Chapter 8, in full, is a reprint of the material as it appears in I. Thiele, R.M.T. Fleming, A. Bordbar, R. Que, B.Ø Palsson, An integrated model of macromolecular synthesis and metabolism of *Escherichia coli.*, *in preparation.* I was the primary author of this publication and the co-authors participated and directed the research, which forms the basis for Chapter 8.

VITA

2001	B. S., Technical Biology, University of Stuttgart, Germany
2004	M. S., Biotechnology, European School of Biotechnology, Strasbourg, France
2007	M. S., Bioinformatics, University of California, San Diego
2009	Ph. D., Bioinformatics, University of California, San Diego

PUBLICATIONS

Thiele, I., Fleming, R.M.T., Bordbar, A., Que, R., Palsson, B.. An integrated model of macromolecular synthesis and metabolism of *Escherichia coli.*, (In preparation).

Fleming, R.M.T., Thiele, I., Palsson, B.. Quantitative assignment of reaction directionality in constraint-based models of metabolism., (In preparation).

Thiele, I., Fleming, R.M.T., Bordbar, A., Palsson, B. ., Functional characterization of multiple equivalent states of *Escherichia coli*'s transcriptional and translational machinery., *Biophysical Journal*, (To be submitted).

Zhang, Y.*, Thiele, I.*, Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A.M., Wooley, J., Lesley, S.A., Wilson, I.A., Palsson, B. ., Osterman, A., Godzik, A., Structural genomics of the *Thermotoga maritima* sets the stage for the molecular level analysis of its metabolic network., *Science*, Submitted.

Thiele, I. and Palsson, B. ., A protocol for generating a high-quality genome-scale metabolic reconstruction., *Nature Protocols*, Submitted.

Thiele, I. and Palsson, B. ., 2D genome annotation jamborees: A community effort in systems biology., *Nature Biotechnology*, Submitted.

Fleming, R.M.T., Thiele, I., Provan, G., Palsson, B. ., H.P. Nasheuer. Integrated stoichiometric, thermodynamic and kinetic modeling of steady-state metabolism., *Biophysical Journal*, Accepted.

Feist, A.M., Herrgard, M., Thiele, I., Reed, J.L., Palsson, B. . Reconstruction of biochemical networks in microbial organisms., *Nature Reviews Microbiology*, 7(2) (2009).

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

Li, F., Thiele, I., Jamshidi, N., Palsson, B. . Functional assessment of the TLR receptor network., *PLoS Computational Biology*, 5(3): e1000312 (2009).

Nogales, J., Palsson, B. ., Thiele, I. A genome-scale metabolic reconstruction for *Pseudomonas putida* KT2440: *i*JN746 as cell factory., *BMC Systems Biology*, 2:79 (2008).

Feist A.M., Thiele, I., Palsson, B.., Genome-scale reconstruction, modeling, and simulation of *Escherichia coli*'s metabolic network, in Systems biology and biotechnology of Escherichia. coli, Eds: Lee, S.Y., *Springer*, to appear (2008).

Lewis, N.E., Jamshidi, N., Thiele, I., Palsson, B.., Metabolic Systems Biology: a constraint-based approach, Encyclopedia of Complexity and System Science, *Springer*, to appear (2008).

Thiele, I. and Palsson, B. ., Bringing genomes to life: The use of genome-scale *in silico* models, Chapter 2, Introduction to Systems Biology, *Humana Press* (2007).

Yeung, M., Thiele, I., Palsson, B. ., Estimation of the number of extreme pathways for metabolic networks, *BMC Bioinformatics*, 8:363 (2007).

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B. . Global reconstruction of the human metabolic network based on genomic and bibliomic data., *Proceedings of the National Academy of Sciences*, 104(6):1777-82 (2007).

Price, N.D., Thiele, I., and Palsson, B.., Candidate states of *Helicobacter pylori*'s genome-scale metabolic network upon application of loop law thermodynamic constraints., *Biophysical Journal*, 90(11): 3919-28 (2006).

Reed, J.L., Famili, I., Thiele, I., and Palsson, B. ., Towards multidimensional genome annotation., *Nature Reviews Genetics*, 7: 130-141 (2006).

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H-Y., Cohoon, M., de Crcy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., e, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Ostermann, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V., The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes., *Nucleic Acids Research*, 33(17):5691-5702 (2005).

Thiele, I.*, Vo, T.D.*, Price, N.D. and Palsson, B.., An expanded metabolic reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants., *Journal of Bacteriology*, 187(16): 5818-5830 (2005).

Thiele, I., Price, N.D., Vo, T.D. and Palsson, B.., Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet., *Journal of Biological Chemistry*, 280(12):11683-11695 (2005).

* These authors contributed equally to these works.

ABSTRACT OF THE DISSERTATION

A stoichiometric model of *Escherichia coli*'s macromolecular synthesis machinery and its integration with metabolism.

by

Ines Thiele
Doctor of Philosophy in Bioinformatics

University of California San Diego, 2009

Professor Bernhard Ø. Palsson, Chair,

Professor Steven P. Briggs, Co-Chair

Systems biology is a rapidly growing discipline. It is widely believed to have a broad transformative potential on both basic and applied studies in the life sciences. In particular, biochemical network reconstructions are playing a key role as they provide a framework for investigation of the mechanisms underlying the genotype-phenotype relationship. In this thesis, the procedure to reconstruct metabolic networks is illustrated and extended to other cellular processes. In particular, the constraint-based reconstruction and analysis approach was applied to reconstruct the transcriptional and translational (tr/tr) machinery of *Escherichia coli.* This reconstruction, denoted 'Expression-matrix' (E-matrix), represents stoichiometrically all known proteins and RNA species involved in the macromolecular synthesis machinery. It accounts for all biochemical transformations to produce active, functional proteins, tRNAs, and rRNAs known to be involved in macromolecular synthesis in *E. coli.* An initial study investigated basic properties of the E-matrix, including its capability to produce ribosomes, which was found to be in good agreement with experimental data from literature. Furthermore, quantitative gene expression data could be integrated with, and analyzed in the context of, the resulting constraint-based model. Adding mathematically derived constraints to couple certain reactions in the model allowed the quantitative representation of the size of steady state protein and RNA pools. Furthermore, the E-matrix was integrated with the genome-scale *E. coli* metabolic model and extended the transcriptional and translational reactions to encompass genes encoding

all the respective metabolic enzymes. The resulting Metabolite-Expression-matrix (ME-matrix), has exceeds the predictive capacity of the metabolic model and it can, for example, be used to predict the biomass yield since it represents the production of almost 2,000 proteins. *E. coli*'s ME-matrix is the first of its kind and represents a milestone in systems biology as demonstrates how to quantitatively integrate 'omics'-datasets into a network context, and thus, to study the mechanistic principles underlying the genotype-phenotype relationship. Possible applications are just beginning to become apparent and may include protein engineering, interpretation of adaptive evolution, and minimal genome design. An integration of the ME-matrix with remaining cellular processes, such as regulation, signaling, and replication, will be a next step to complete the first whole-cell model.

Chapter 1

Introduction to molecular systems biology

Computational modeling is an integral part of systems biology. In particular, metabolic reconstructions and constraint-based modeling have been proven to be very valuable in predicting phenotypic behavior of bacterial and eukaryotic cells. This thesis work is on the heart of systems biology with its effort towards whole-scale modeling of bacterial cells.

In general, one can distinguish at least two approaches to modeling in systems biology. *1.* In a so-called top-down approach, a multitude of high-throughput data (e.g., gene expression, proteomics, etc.) are generated under various environmental and genetic conditions and the interactions between the components are then interfered using statistical and computational methods. This top-down approach has been shown to be very useful for discovery of new biological functions. However, in most cases the interfered interactions between the cellular components are associated with uncertainty as they are not measured directly and thus rely on the quality of the employed tool. Many of these statistical and computational tools have a certain false-discovery rate associated which causes uncertainty in the constructed (interaction) network. The perhaps most comprehensive constructed network for an organism was done by Baliga and colleagues which elucidated the cellular network of *Halobacterium salinarum* using high-through data and comprehensive computational tools [31]. This model of *H. salinarum*'s genetic, regulatory and physiological properties has been shown to capture well known experimental observations and to predict behavior in previously undefined conditions (prospective predictive capabilities) [31].

2. The second approach is the bottom-up approach to reconstruct cellular networks based on genomic and bibliomic data [229, 205]. These bottom-up reconstructions are more defined in their scope (e.g., metabolism, signaling) than top-down networks but the links and interactions between the cellular components are better defined as they are obtained from more reliable data, such as genome annotation, biochemical and molecular studies. In particular, the bottom-up approach is very well defined for metabolic network and has been expanded to other cellular functions in more recent research efforts. The work presented in this thesis focuses on the bottom-up reconstruction method.

1.1 Constraint-based reconstruction and analysis

The constraint-based reconstruction and analysis (COBRA) approach is one possible modeling approach that uses stoichiometric information about biochemical transformations taking place in a target organism to construct the model. While a metabolic reconstruction is unique to the target organism one can derive many different condition-specific models from a single reconstruction. The conversion of a metabolic reconstruction of an organism into models requires the imposition of physicochemical and environmental constraints to define systems boundaries [204, 219, 229]. The conversion also includes the transformation of the reaction list into a computable, mathematical matrix format. In this so-called S matrix, where S stands for stoichiometric, the rows correspond to the network metabolites and the columns to the network reactions (Figure 1.1). This conversion can be done automatically (e.g., using the Matlab-based COBRA Toolbox [21]). Once in this format, numerous mathematical tools can be used to interrogate the metabolic network properties *in silico*.

Many of the published mathematical tools have been reviewed [219, 70] and encoded in Matlab format [21]. One of the most frequently used mathematical COBRA tool is flux balance analysis (FBA). FBA is a formalism in which a reconstructed network is framed as a linear programming (LP) optimization problem and a specific objective function (e.g., growth, by-product secretion) is maximized or minimized [219]. A large subset of the COBRA tools relies on LP. While LP-based tools are very helpful in studying reconstructed

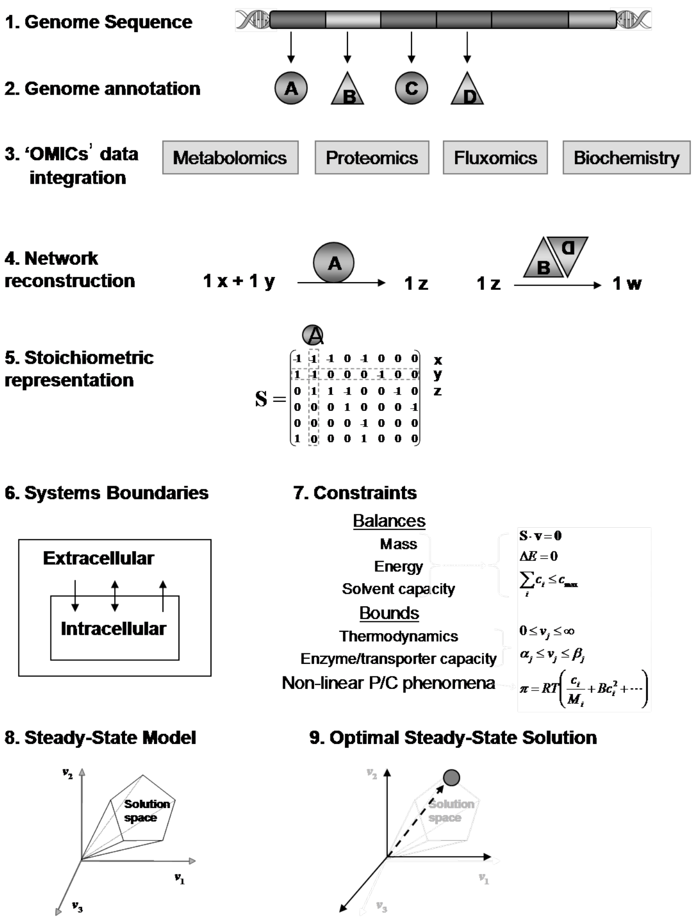

1. Genome Sequence
2. Genome annotation
A
B
C
D
3. 'OMICs' data integration
Metabolomics
Proteomics
Fluxomics
Biochemistry
4. Network reconstruction
1 x + 1 y
A
1 z
1 z
B
D
1 w
5. Stoichiometric representation
A
S =
-1 -1 -1 0 -1 0 0 0
1 -1 0 0 0 -1 0 0
0 1 1 -1 0 0 -1 0
0 0 0 1 0 0 0 -1
0 0 0 0 -1 0 0 0
1 0 0 0 1 0 0 0
x
y
z
6. Systems Boundaries
Extracellular
Intracellular
7. Constraints
Balances
Mass
Energy
Solvent capacity
Bounds
Thermodynamics
Enzyme/transporter capacity
Non-linear P/C phenomena
$\mathbf{S}\cdot\mathbf{v}=0$
$\Delta E=0$
$\sum_i c_i \le c_{max}$
$0\le v_j\le\infty$
$\alpha_j\le v_j\le\beta_j$
$\pi=RT\left(\frac{c_i}{M_i}+Bc_i^2+\cdots\right)$
8. Steady-State Model
v_2
Solution space
v_1
v_3
9. Optimal Steady-State Solution
v_2
Solution space
v_1
v_3

Figure 1.1: **Bringing genomes to life.** Starting from the genome sequence an initial component list of the network is obtained. Using additional data such as biochemical and other 'omics' data the initial component list is refined as well as information about the links between the network components. Once the network links, or reactions, are formulated, the stoichiometric matrix can be constructed using the stoichiometric coefficients that link the network components. The definition of the system boundaries transforms from a network reconstruction into a model of a biological system. Every network reaction is elementary balanced and may obey further constraints (e.g., enzyme capacity). These constraints allow the identification of candidate network solution which lay within the set of constraints. Different mathematical tools can be employed to study these allowable steady-state network states under various aspects such as optimal growth, by-product secretion and others.

metabolic networks, some questions may better be addressed without having to choose an objective function. Those methods are called unbiased methods, in contrast to biased LP-based methods, because they identify all feasible flux distributions under the given set of environmental constraints rather than only the optimal distributions. The COBRA approach [219, 229] has been successfully used to build and analyze genome-scale *in silico* reconstructions for representatives of many organisms.

The numerous mathematical tools have been used for

1. Identification and filling of knowledge gaps, e.g., missing gene annotations [231]
2. Prediction of the outcome of adaptive evolution [118, 85, 128]
3. Design of engineered production strains [210]
4. Understanding of topological features of metabolic networks [228, 6, 15, 280].

A recent review illustrates the variety of questions that have been addressed to *Escherichia coli*'s metabolic network using different biased and unbiased COBRA methods [80].

1.2 Basic principles underlying the constraint-ba-sed reconstruction & analysis approach

The basic principles which underly the constraint-based reconstruction and analysis approach COBRA approach are listed in the following section inform of axioms. These principle lie out why this approach works, why it is so powerful in assisting the understanding of biological systems and why it is valid to extend the COBRA approach to cellular functions besides metabolism.

Axiom #1: All cellular functions are based on chemistry. A simple but consequential statement, as it implies the fundamental events in a cell can be described by chemical equations. These equations in turn come with chemical information and physico-chemical principles.

Axiom #2: Annotated genome sequences along with experimental data enable the reconstruction of genome-scale metabolic networks. The reconstruction process is a grand scale systematic assembly of information in a QC/QA-ed setting that leads to a Biochemically, Genetically and Genomically (BiGG) structured knowledgebase. The reconstruction process has been reviewed elsewhere [79, 229], and a growing number of reconstructions are available.

Axiom #3: Cells function in a context-specific manner. When a cell is placed in a particular environment, it expresses a subset of its genes in response to environmental cues. The abundance of cellular components can be profiled using transcriptomic, proteomic, and metabolomic methods. Such high-throughput data can be mapped onto a network reconstruction to tailor it to the particular condition being considered.

Axiom #4: Cells operate under a series of constraints. Factors constraining cellular functions fall into four principal categories [204]: physico-chemical (i.e., see axiom #5), topological (crowding effects), environmental (axiom #3) and regulatory (basically self-imposed constraints, or restraints). These constraints cannot be violated allowing the estimation of all functional (i.e., physiological) states, which a genome-scale reconstruction can achieve. Mathematically, such statements are translated into fundamental subspaces associated with the stoichiometric matrix (S), whose properties can be characterized [206].

Axiom #5: Mass (and energy) is conserved. This statement is one of the basic physical laws. Since all proper chemical equations can be described by stoichiometric coefficients, and since a set of chemical equations can be described by S this means that all steady states (normally close to the homeostatic states of interest) of a network can be described by a simple linear equation, $S \cdot v = 0$ [206], where v is a vector of fluxes through chemical reactions. Thus, the computation of the functional states of a network is enabled based on the known underlying chemistry.

Axiom #6: Cells evolve under a selection pressure in a given environment. This statement has implicit optimality principles built into it. Consequently, if we know the selection pressure, we can state a so-called objective function and determine optimal states given a network reconstruction and governing constraints. The most familiar case of such computations is flux-balance analysis (FBA). A broad spectrum of methods has been developed under this umbrella [206, 219, 70], collectively called COBRA methods.

These axiomatic statements lead to simple mathematics associated with BiGG knowledge-bases. Such formal representation allows for queries of the knowledge-base and the formulation of genome-scale models (GEMs). The result is a myriad of uses as mentioned above.

So why does COBRA work so well? It appears that we can now enumerate cellular components, describe their interactions chemically, formulate a mathematical description of the totality of such interactions, identify the constraints that the resulting network operates under, and apply optimality principles to evaluate likely physiological functions in a given environment. This train sequence of events seems to provide a consistent framework on which a mechanistic basis for the microbial metabolic genotype-phenotype relationship can be formulated. The underlying process is based on an emerging paradigm to relate the genotype to the phenotype through reconstruction and modeling:

Genome sequence and high-throughput data → networks reconstruction though the formulation of BiGG → conversion the knowledge in BiGG to a mathematical representation → computation of phenotypes and other applications using GEM.

1.3 Reconstruction of metabolic networks in a nutshell

The genome annotation, or 1D annotation, provides the most comprehensive list of components in a biological network. In metabolic network reconstructions, the genome annotation is used to identify all potential gene products involved in metabolism of an organism (Figure 1.1 and 1.2. The links in metabolic networks are the reactions carried out by metabolic gene products (Figure 1.1). In order to assign cellular components with the metabolic reactions, different information is required and provided by various sources, including organism-specific and non-organism-specific databases and literature. Since some of the information sources are more reliable than others, a confidence scoring system may be used to distinguish them. Once the network reactions are defined, the metabolic network can assembled in a step-wise fashion by starting with central metabolism, which contains the fueling reactions for the cell, and moving on to the biosynthesis of individual macromolecular building blocks (e.g., amino acids, nucleotides, and lipids). The step-wise assembly of the network facilitates the identification of missing steps within the pathway that were not defined by the 1D annotation. Once well defined metabolic pathways are

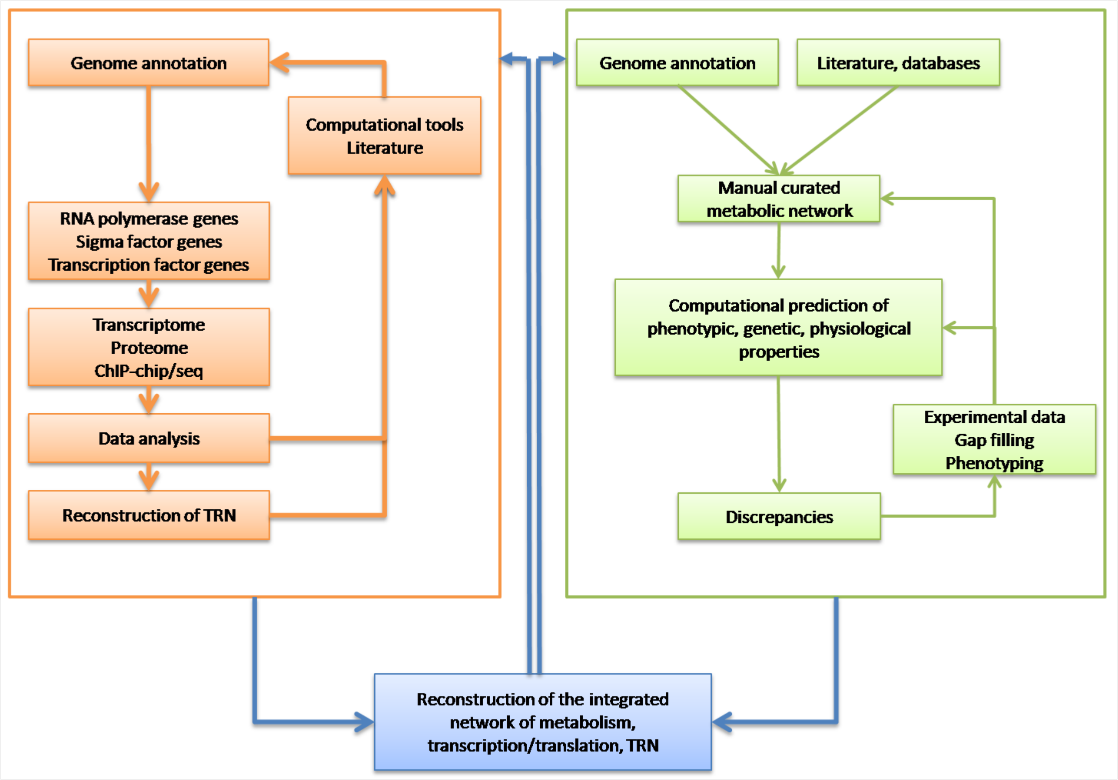

Genome annotation
Computational tools
Literature
RNA polymerase genes
Sigma factor genes
Transcription factor genes
Transcriptome
Proteome
ChIP-chip/seq
Data analysis
Reconstruction of TRN
Genome annotation
Literature, databases
Manual curated metabolic network
Computational prediction of phenotypic, genetic, physiological properties
Experimental data
Gap filling
Phenotyping
Discrepancies
Reconstruction of the integrated network of metabolism, transcription/translation, TRN

Figure 1.2: **General workflow for reconstructing cellular networks. Left:** Transcriptional regulatory networks. **Right:** Metabolic networks.

assembled, reactions can be added that do not fit into these pathways but are supported by the 1D annotation or biochemical studies. Such enzymes might be involved in the utilization of other carbon sources or connect different pathways.

Even genomes of well-studied organisms harbor genes of unknown functions (e.g., 20% for *E. coli*). Subsequently, metabolic networks constructed solely based on genomic evidence often contain many network gaps, so called blocked reactions. Physiological data may help to determine whether a pathway is functional in the organism or not and thus may provide evidence for the missing reactions. This procedure is called gap filling and it is a crucial step in network reconstruction.

The gap filling process is followed by a detailed network evaluation. Here, the network is examined to see if it can generate the precursor metabolites, such as biomass components, and metabolites the organism is known to produce or degrade. Furthermore, the comparison of the network behavior with various experimental observations, such as secretion products and gene essentiality, will ensure similar properties and capabilities of the *in silico* metabolic network and the biological system. This sequential, iterative process of network evaluation is labor intensive, but it will ensure high accuracy and quality by network adjustments, refinements, and expansions.

1.4 Mathematical characterization of network capabilities

The stoichiometric matrix, denoted as S, is formed by the stoichiometric coefficients of the reactions that comprise a reaction network (see Figure 1.1). This matrix is organized such that every column corresponds to a reaction, and every row corresponds to a compound. Mathematically, the stoichiometric matrix, S, transforms the flux vector $v = (v_1, v_2, , v_n)$, which contains the reaction rates, into a vector that contains the time derivatives of the concentrations $x = (x_1, x_2, , x_m)$:

$$\frac{dx}{dt} = S \cdot v \equiv 0 \tag{1.1}$$

At steady-state, there is no accumulation or depletion of metabolites in a metabolic network, so the rate of production of each metabolite in the network must equal its rate of consumption. The stoichiometric matrix thus contains chemical and network information. Bounds that further constrain the values of individual variables can be identified, such as fluxes, concentrations and kinetic constants.

As mentioned in Section 1.2 cellular functions are limited by different types of constraints, which can be grouped in four general categories: fundamental physico-chemical, spatial or topological, condition-dependent environmental, and regulatory or self-imposed constraints. While the first two categories of constraints are assumed to be independent from the environment, the latter two types of constraints may vary in the simulation.

1.5 Tools for analyzing network states.

The analysis of organism's phenotypic functions on a genome-scale using constraint-based modeling has developed rapidly in recent years. The plethora of steady-state flux analysis methods can be broadly classified into the following categories: i) finding best or optimal states in the allowable range; ii) investigating flux dependencies; iii) studying all allowable states; iv) altering possible phenotypes as a consequence of genetic variations; and v) defining and imposing further constraints. In this section, we will discuss some of the numerous methods that have been developed. A more comprehensive list of methods can be found in [219].

Once a GEM is obtained it can be used for simulations and thus analyze and predict cellular functions.One can also take the point of view that COBRA methods are query tool that are used to interrogate the BiGG knowledge-base about functions it allows. Such computations are performed using a variety of constraint-based methods. Some computations are exploratory, while others predict phenotypic functions under the assumption that cellular functions are optimal (Axiom #6, Section 1.2). The challenge with the latter of course is, we do not always know what is being optimized. Here we give four examples of the uses of COBRA methods. Details beyond the scope of this primer are found in Price *et al.* [219] and Durot *et al.* [70].

Integration of high-throughput data. In addition to the application of environmental constraints, high-throughput data, including gene expression and proteomic data, can be employed to define the subset of active gene products (and thus, metabolic reactions) in a given condition. Subsequently, reaction constraints for highly expressed genes can be set to be, e.g., maximal, while reactions associated with genes that are not or weakly expressed are constraint such that the reaction flux cannot be maximally. These additional constraints lead to reduction of the set of candidate steady-states.

Flux-balance analysis (FBA). A majority of COBRA methods rely on FBA [219] that is based on linear programming (LP) to investigate the consequences of genetic and environmental perturbations on allowable network states. The corresponding LP problem is formulated as
Maximize $z = c^T \cdot v \equiv < c_i \cdot v_i >$
Subject to $S \cdot v = 0$
$v_{i,min} \leq v_i \leq v_{i,max}$ $v_i \in \Re$, for all i network reactions
where z is the objective function representing a linear combination of metabolic fluxes v_i. The vector c indicates which network reaction(s) of v contribute to z and their coefficient(s) (c_i). $v_{i,min}$ and $v_{i,max}$ represent the lower and upper bound on each reaction, respectively. By modifying the constraints and objective functions, a large range of biological and biotechnological questions can be addressed [219]. We note that the objective function is set by the user and represents a bias.

Identification of essential and synthetically lethal genes. A powerful COBRA application is the systematic identification of essential genes. A computational single gene deletion study is performed by deleting one gene at the time and then tracing through the GPRs the corresponding catalyzed metabolic reaction(s). The removal of a reaction i from the network is realized by setting its lower bound and upper bound to zero (i.e., $v_{i,min} = v_{i,max} = 0$) and optimizing for an objective function, often the biomass reaction, $v_{biomass}$. If the maximal value for biomass production is zero, then the gene is predicted to be lethal. Deleting two genes simultaneously will allow the determination of synthetic lethal genes [103].

Flux variability analysis (FVA). Another important COBRA method is FVA, which allows the determination of the network flexibility in a given condition. A particular subset of FVA, investigates the set of alternate optimal solutions, by setting a cellular objective (e.g., the biomass reaction) to its maximum ($v_{biomass,min} = v_{biomass,max} = max$). Subsequently, every network reaction is chosen as objective function, minimized and maximized, resulting in the reaction's flux span.Narrow or no, non-zero flux span ($v_{i,min} = v_{i,max} \neq 0$) indicates essential reactions to achieve the optimal conditions, while reactions with no, zero flux span ($v_{i,min} = v_{i,max} = 0$) do not contribute to the objective at all. Such assessment gets at the robustness and redundancy characteristics of a network.

1.6 Preview of the dissertation

The bottom-up reconstruction approach is well established for metabolism. In this thesis, the same COBRA approach used for metabolic reconstruction was employed and expanded to describe the cellular processes taking place during transcription and translation (tr/tr). In particular, a comprehensive reconstruction was developed for the synthesis and function of the macromolecular synthesis machinery in *Escherichia coli.* Subsequently, an method was developed to integrated this network with a metabolic reconstruction of *E. coli* leading to the first integrated stoichiometric, genome-scale representation of multiple cellular functions. This integrated model of macromolecular synthesis and metabolism is a first significant step towards whole-cell modeling of organisms. Moreover, this thesis illustrates a new range of questions that can be addressed with this integrated model which would not be possible with any of the models individually.

The chapters in this dissertation thus deal with the following topics:

- **Chapter 1**: This chapter describes the principle reconstruction approaches in system biology with emphasis on bottom-up reconstruction and constraint-based modeling.
- **Chapter 2**: This chapter represents a primer of capabilities of an *E. coli* cell.
- **Chapter 3**: A standard operating procedure is described here, based on which high-quality, genome-scale metabolic network have been reconstruction and which provides a framework for reconstructions of the tr/tr machinery.
- **Chapter 4**: Illustration of the content covered in some of the available metabolic reconstructions and currently available techniques to reconstruct other cellular functions.
- **Chapter 5**: The complex processes underlying transcription and translation in bacteria are presented and computational formulations are proposed. These formulations build the foundation for the tr/tr reconstruction.
- **Chapter 6**: The tr/tr reconstruction of *E. coli* is described and some applications are presented.
- **Chapter 7**: A limitation of the FBA-based modeling is that cellular pools of macromolecules are not explicitly represented. Here, an approach is presented to incor-

porate macromolecular pools into the FBA framework and is applied to the tr/tr reconstruction.

- **Chapter 8**: The tr/tr reconstruction is integrated with the metabolic reconstruction and the properties of the "merged" network are investigated.
- **Chapter 9**: Conclusion and outlook.

The text of this chapter, in part or in full, is a reprint of the material as it appears in I. Thiele and B.Ø. Palsson, Bringing genomes to life: The use of genome-scale *in silico* models, Chapter 2, Introduction to Systems Biology, Humana Press (2007), and in I. Thiele and B.Ø. Palsson, Fundamentals of Constraint-Based Methods, Nature Biotechnology (Submitted). I was the primary author of these publications and the co-author participated and directed the research, which forms the basis for this chapter.

Chapter 2

Escherichia coli

2.1 Key properties

Escherichia coli was one of the first model systems for molecular biology, which was discovered in 1885 by Theodor Escherich [77]. *E. coli* is a facultative anaerobe that colonizes the lower gut of animals but it can also survive in pure water. Therefore, its life cycle is composed of a series of 'shocks': the cold and nutrient-deprived shock in water, the acid shock on its trajectory to the lower gut, the heat shock in the gut, etc. *E. coli* is a gram-negative bacterium (Figure 2.1) and does not sporulate. It has a single circular chromosome with 4,401 open reading frames (ORFs), which enable it to survive and growth in these different environmental conditions [237]. Due to its short doubling time and easy genetic manipulation *E. coli* is commonly used as model organism in many biological research areas. Subsequently, many genetic, molecular, and biochemical properties are well studied and understood. The large body of scientific literature about its biological functions and capabilities is abundant and available, making *E. coli* a good candidate for 'bottom-up' network reconstructions. Furthermore, *in silico* derived hypotheses can be readily tested with experiments.

Metabolic capabilities of *E. coli* As mentioned *E. coli* is a facultative anaerobe found in the lower gut of animals but also in water. It can grow on glucose as sole carbon and energy source in aerobic and anoxic condition. Glucose is transported into the cell by the phophotransferase system and is then catabolized to pyruvate via glycolysis. Under aerobic conditions, pyruvate is used to generate NADH in the Krebs cycle, which is then reoxidized by the respiratory chain. During fermentation, the respiratory chains, which

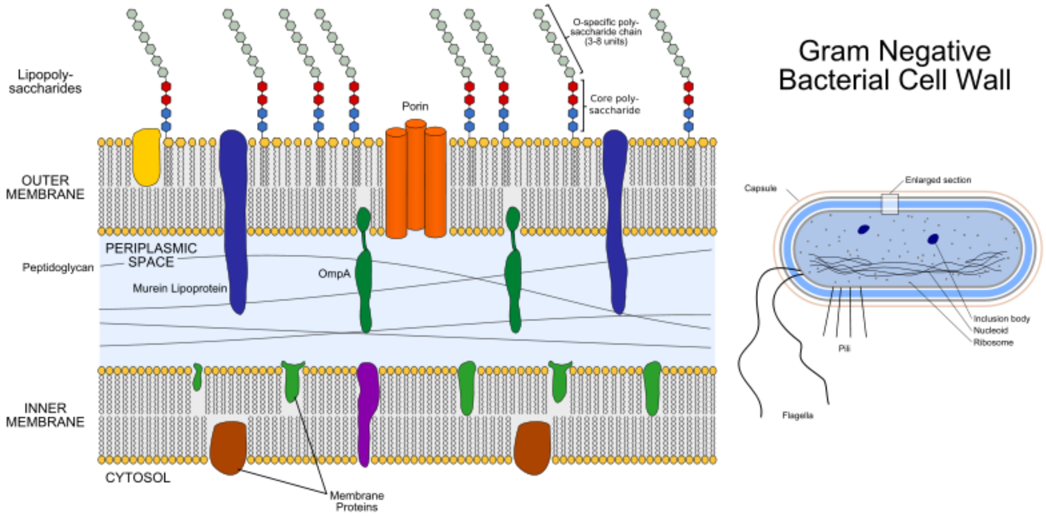

O-specific poly-saccharide chain (3-8 units)
Gram Negative Bacterial Cell Wall
Lipopoly-saccharides
Core poly-saccharide
Porin
OUTER MEMBRANE
Capsule
Enlarged section
PERIPLASMIC SPACE
Peptidoglycan
OmpA
Murein Lipoprotein
Inclusion body
Nucleoid
Ribosome
Pili
INNER MEMBRANE
Flagella
CYTOSOL
Membrane Proteins

Figure 2.1: Diagram of a gram-negative cell wall. Taken from http://en.wikipedia.org/wiki/File:Gram_negative_cell_wall.svg

is linked to oxygen, and other terminal, alternative electron accepters are not functional. Pyruvate is converted nor those linked to alternative electron acceptors are functional. The major fermentation products are acetate, ethanol, lactate, and formate. *E. coli* can grow on a wide variety of nutrients, including simple and complex carbohydrates. *E. coli* can produces vitamin K_{12}, which can be used by the host (e.g., human).

RNA and Protein synthesis in *E. coli.* Macromolecules are the major constituents of cells. In *E. coli*, proteins and RNA account for 55 and 20.5% of the total dry weight, respectively, (Table 2.1) [185]. The synthesis of proteins (translation) and RNA (transcription) involves numerous cellular factors that concatenate the relevant precursors, amino acids or nucleotides, under consumption of energy. In fact, these processes consume a major part of the energy produced by the cell. The transcription of RNA from a DNA template is done by the DNA-dependent RNA polymerase (RNAP) by binding to the promoter region of an open reading frame (Figure 2.1). After binding the DNA-RNAP-complex forms a so-called open complex upon which the RNAP leave the transcription initiation site and moves along the DNA strand to synthesize a nascent RNA strand until the RNAP reaches a termination site. This process is aided by numerous accessory proteins (transcription factors), which ensure accuracy and fidelity of the process. Different sets of transcription factors are necessary for rRNA, tRNA and mRNA transcription. In bacteria, transcription and translation are not spatially or temporally separated with the

Table 2.1: Macromolecular composition of an average *E. coli* cell at a doubling time of T_D = 40 minutes (grown at 37°C in glucose minimal medium). Adapted from [185].

Macromolecuel	% of total dry weight	Weight per cell ($10^{15}\times$ weight, grams)
Protein	55	155
RNA	20.5	155
- 23S rRNA		31.0
- 16S rRNA		16.0
- 5S rRNA		1.0
- transfer		8.6
- messenger		2.4
DNA	3.1	9.0
Lipid	9.1	26.0
Murein	2.5	7.0
Glycogen	2.5	7.0
Total Macromolecules	96.1	273.0
Soluble pool	2.9	8.0
- building blocks		7.0
- metabolites, vitamins		1.0
Inorganic ions	1.0	3.0
Total dry weight	100.0	284.0
Total dry weight/cell		2.8×10^{-13}g
Water (at 70% of cell)		6.7×10^{-13}g
Total weight of one cell		9.5×10^{-13}g

consequence that ribosomes bind to the growing mRNA molecule and translate the template into proteins. The ribosome consists of two subunits (50S and 30S) which together are build of more than 50 different proteins and three rRNA (5S, 16S and 23S rRNA). The amino acids are delivered to the ribosome via their corresponding tRNA. In addition to the ribosomes numerous accessory factors are necessary to enable the translation process.

Regulation in *E. coli.* The transcription rate of most genes is highly regulated by activators and repressors. These transcription factors (TF) bind to specific sites on the DNA upstream from the RNAP binding site. In the case of transcriptional activators, this corporative binding helps the formation of the open DNA-RNAP complex and thus successful transcription. Transcriptional repressors often block the RNAP binding site and thus significantly reduce the transcription rate. Subsequently, the locations and orientations of these binding sites, as well as the affinity of the TFs to particular variants of the site, determine the expression levels of a gene in response to changes in the active

TF concentrations inside the cell. It has been demonstrated that the known organization of promoter regions in bacteria allows the implementation of a wide class of regulatory logic functions within a single promoter [37], so that even a single node in the regulatory network can be relatively complex.

2.2 *E.coli* by numbers

Table 2.2: **Parameters related to the growth and macromolecular composition of bacterial cells.** Adapted from [183]. nt = nucleotides, aa = amino acids, pol = RNA polymerase, rib = ribosome.

Parameter	Symbol	Value
Doubling time	τ	24 - 100 min
Deoxyribonucleotide residues per genome	kbp/genome	4,700
Ribonucleotide residues per rRNA precursor	nt/prib	6,400
Ribonucleotide residues per 70S ribosome	nt/rib	4,566
Amino acid residues per 70S ribosome	aa/rib	7,336
Ribonucleotide residues per tRNA	nt/tRNA	80
Amino acid residues per RNA polymerase	aa/pol	3,407
Fraction of total RNA that is stable RNA	f_s	0.98
Fraction of stable RNA that is tRNA	f_t	0.14
Fraction of total Protein that is r-protein	α_r	0.09-0.22
Fraction of total Protein that is RNA polymerase	α_p	0.009-0.01
Peptide chain elongation rate	c_p	12-22 aa/sec
Stable RNA chain elongation rate	c_s	85 nt/sec
mRNA chain elongation rate	c_m	40-55 nt/sec

The presented thesis work, and much of the work on *E. coli* systems biology published in the last 20 years, was enabled by work from a group of people that seeked to understand the biology of bacterial cells in great detail, and more importantly, quantified cellular components in different growth states. This key work was done using mainly *E. coli* and *Salmonella typhimurium.*

In this section, I will summarize and list the numbers related to cellular growth which were discovered in the last 60 years, or so, and which are essential for reconstructing and modeling cellular networks of *E. coli.* Although these numbers may differ between organism, they are often used and adapted for other organisms (e.g., for the biomass reaction in metabolic networks), as the associated, detailed studies on *E. coli* have not yet been carried out for other organism in the same extend.

Table 2.3: **Parameters related to the growth and macromolecular composition of bacterial cells.** Adapted from [183]. nt = nucleotides, aa = amino acids, rib = ribosome. RNAP = RNA polymerase. [a] The number of mRNA molecules was calculated based on the total number of nucleotides per cell minus the number of nucleotides present in tRNA and ribosomes per cell.

μ (1/h)	2.5	2	1.5	1	0.6
doubling time	24	30	40	60	100
time chromosome replication (min)	42	43	45	50	67
time btw replication & division (min)	23	24	25	27	30
10^3mRNA per cell	7.7	4.6	3.4	1.4	0.9
mRNA elongation rate ($\frac{nt}{sec}$)	55	52	50	45	39
stable RNA elongation rate(sec)	85	85	85	85	85
stable RNA initiation rate (min)	58	39	23	10	4
mass/cell ($\frac{\mu g_{DW}}{10^9}$ cells)	865	641	433	258	148
10^3 RNAP per cell	11.4	8	5	2.8	1.5
peptide chain elongation rate ($\frac{aa}{sec}$)	21	20	18	16	12
10^6 nt per cell	390	244	143	73	37
10^3 rib per cell	72	45.1	26.3	13.5	6.8
10^6 nt per rib per cell	328.75	205.93	120.09	61.64	31.05
10^3 tRNA per cell	669	419	244	125	63
10^6 nt per tRNAs per cell	53.52	33.52	19.52	10	5.04
mRNA per cell[a]	7728	4553.4	3394.2	1359	911.2

The results of numerous studies were made accessible in two books, which have been basis for the present thesis, as well as for generations of reconstructions of *E. coli*'s metabolism [184, 183, 185]. For instantce, the work done by Schaechter *et al.* in 1958 showed that the physiological state of a cell, i.e., cell size and composition, is dependent on the growth rate regardless how the growth rate is achieved [246]. This implies that cells grown in two media which enable the same growth rate, would have the same cellular composition and size. A key result from this work was that there is a relationship between growth rate and cell composition. Rigorous measurements of the different cellular parameters allowed the determination of equations relating the cell composition in exponential cultures to basic cell cycle parameters (Table 2.2 and 2.3) [183].

As mentioned earlier, this thesis heavily relies on these cellular component numbers. However, the cellular parameters were not determined in the aforementioned studies for all required growth rates. Therefore, we extrapolated measured and calculated parameters listed in Table 2.3 and obtained the equations of the fitted curves(Figure 2.2).

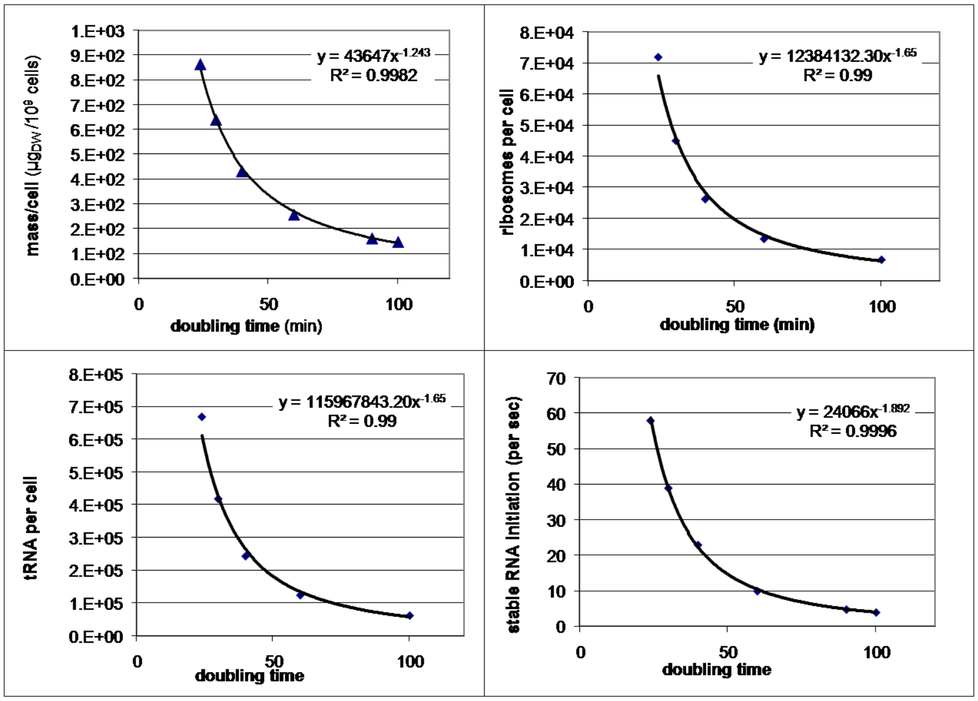


Figure 2.2: **Dependency of cellular properties and growth rate in *E. coli.*** Experimentally measured and calculated values are shown in blue. Linear regression curves, their corresponding equations, and Pearson correlation are shown in black.

Overall, we used the following, fitted equations in the following chapters as basis for calculating cellular components and parameters:

- Time between replication and division (min): $0.12.727 \cdot T_D^{0.1852}$, $R^2 = 0.9974$
- Time chromosome replication (min): $35.52e^{0.0062\ cdotT_D}$, $R^2 = 0.9889$
- RNA polymerase per cell $\frac{molecules}{cell}$: $1,012,059.04 \cdot T_d^{-1.43}$, $R^2 = 1$
- mRNA elongation rate $(\frac{nt}{sec})$: $-11.056 \cdot ln(T_D) + 90.14$, $R^2 = 0.9963$
- tRNA molecules per cell $\frac{molecules}{cell}$: $115,967,843.20 \cdot T_D^{-1.65}$, $R^2 = 0.9974$
- Ribosomes per cell $\frac{molecules}{cell}$: $12,384,132.30 \cdot T_D^{-1.65}$, $R^2 = 0.999$
- Stable RNA initiation rate $(\frac{1}{min})$: $24,065 \cdot T_D^{-1.8924}$, $R^2 = 0.9996$
- Mass per cell $(\frac{\mu g_{DW}}{10^9 cells})$: $43,652 \cdot T_D^{-1.2428}$, $R^2 = 0.9982$
- Nucleotides per cell $\frac{molecules}{cell}$:$66,521,577,620.36 \cdot T_D^{-1.64}$, $R^2 = 0.99$

- Peptide chain elongation rate ($\frac{aa}{sec}$): $74.203 \cdot T_D^{-0.387}$, $R^2 = 0.9734$
- mRNA molecules per cell $\frac{molecules}{cell}$: $877,333 \cdot T_D^{-1.5217}$, $R^2 = 0.9742$

The elongation rate of stable RNA synthesis (transcription) has been found to be independent of the growth rate (85 nucleotides per second) [183]. In contrast, the rate of transcription initiation of ribosomal RNA can range from four initiations per minute to 61 initiations per minute depending on the growth rate (Figure 2.2) [183]. Furthermore, in cases of deletion of one or more of the seven ribosomal RNA operons, the rRNA synthesis rate was found to be increased [49] to compensate for the loss. Therefore, the number of ribosomal RNA operons is not limiting on the rRNA synthesis. The mRNA elongation rate varies with the growth rate and can be as fast as 55 nucleotides per second (Table 2.3). At the beginning of translation, on nascent mRNA, there are approx. one ribosome per 120 nucleotides. As the chain growth the number of ribosomes on the mRNA increases and can be as close as one ribosome per 54 nucleotides [183].

The number of **ribosomes** per cell has been found to be correlated with the growth rate [183] (Figure 2.2). There are approx. nine tRNA per ribosome in exponentially growing *E. coli* cells and almost no variation in this ratio has been found at growth rate higher than 0.5 $\frac{1}{h}$ [183].

There are three translation initiation factors (IF): IF1, IF2 bound to GTP, and IF3. Translation initation is thought to be the rate limiting step in protein synthesis [222]. There are about two to three molecules of each IF per ribosome in the cell. There are three natural forms of IF2: IF2α, IF2β_1, and IF2β_2, which are caused by differential transcription termination on the corresponding operon (metY-yhbC-nusA-infB-rbfA-truB-rpsO-pnp) [222].

The peptide elongation chain rate is approx. 20 amino acids per second (Table 2.3) and each tRNA is required to cycle through the ribosome on average twice per second. There are about six elongation factor Tu (EF-TU) per ribosome at high growth rates [222]. EF-Tu is therefore one of the most abundant proteins in the cell ($\tilde{1}0\%$) [222]. This elongation factor is required for GTP-dependent deposition of aminoacyl-tRNA into the A-site of the ribosome. Elongation factor G (EF-G) is complexes with GTP and is necessary for the translocation of the mRNA/tRNA complex in the ribosome. There is about one EF-G per ribosome in the cell [222]. There are about two molecules of EF-Ts per ten ribosomes [222].

There are three release factors (RF1, RF2, RF3) and ones ribosomal release factor (RRF). The ratio between RF1 and RF2 is approx. 1:5, which is growth rate independent, while their abundance is increasing with the growth rate ($\tau = 0.3h^{-1}$: RF1 = 1,200 molecules per cell and RF2 = 5,900 molecules per cell; $\tau = 2.4h^{-1}$: RF1 = 4,900 molecules per cell and RF2 = 24,900 molecules per cell) [222]. The concentration of RF3 is about 60 fold less than the concentration of RF1 and RF2 [222]. RRF is present as 0.75 molecules per ribosome, whereas about 30% of RRF is bound to the ribosome [222].

There are about one **aminoacyl-tRNA synthetase** per ten ribosomes. This means that every aminoacyl-tRNA synthetase needs to aminoacylate about ten molecules of its cognate tRNA per second to sustain protein synthesis [183]. There is one aminoacyl-tRNA synthetase for each amino acid expect for lysine for which two lysyl-tRNA synthetases exist. At a doubling time of $T_D = 49$ minutes, there are about 1,300 to 2,600 aminoacyl-tRNA synthetase in an *E. coli* cell but the concentration of different charged tRNA isoacceptor families varies a lot: e.g., $T_D = 60$ minutes, there are approx. 700 tRNA^{Gln} but 8,000 tRNA^{Val} [222]. The turn-over rate of aminoacyl-tRNA synthetases is between 2 to 8 per second: e.g., glutamyl-tRNA synthetase charges two tRNA per second while threonyl-tRNA synthetase charges about 48 tRNA per second [222]. The aminoacyl-tRNA synthetase increases with growth rate. The charging level of tRNA is between 70 an 90% and in average there are two to three tRNA molecules bound to each translating ribosome [222].

The **RNA polymerase** (RNAP) of *E. coli* consists of four distinct subunits: α, β, β' and a sigma factor. The α subunit has be found to be present in *E. coli* in excess, while the availability of β and β' determine the amount of core-enzyme [183]. The transcription of β and β' is two-fold controlled (with an upstream promoter and an antitermination site). It has been found that there are 1,500 RNAP molecules per cell at a doubling time of $\tau =$ 100 minutes and 11,400 molecules at $\tau = 24$ minutes (Table 2.3).

***E. coli* has seven sigma subunits** that are thought to bind to the RNA polymerase holoenzyme prior to binding at the promoter site on the DNA (Table 2.4). Sigma

70 was found to have the highest affinity for the RNA polymerase in *E. coli*, whereas sigma 38 (σ^S) showed the lowest affinity [168].

2.3 Conclusion

This chapter illustrated that many cellular properties of *E.coli* are known and have been quantified, which is crucial for mathematical modeling. Furthermore, these properties apply to other cells and organisms. In more recent years, high-throughput technologies enabled the detection and measurement of thousands of *E. coli* components although relative and absolute quantification has not been done yet to the level of accuracy necessary for modeling, it is expected that coming years will provide those data.

Table 2.4: **Sigma factors in *E. coli*.** Data taken from [168].

Sigma factor	Abundance (molecules per cell)
σ^{70}	700
σ^{54} (σ^N)	110
σ^{38} (σ^S)	< 1
σ^{32} (σ^H)	< 10
σ^{28} (σ^F)	370
σ^{24} (σ^E)	< 10
σ^{18} (σ^{FecI})	< 1

Chapter 3

A protocol for generating a high-quality genome-scale metabolic reconstruction

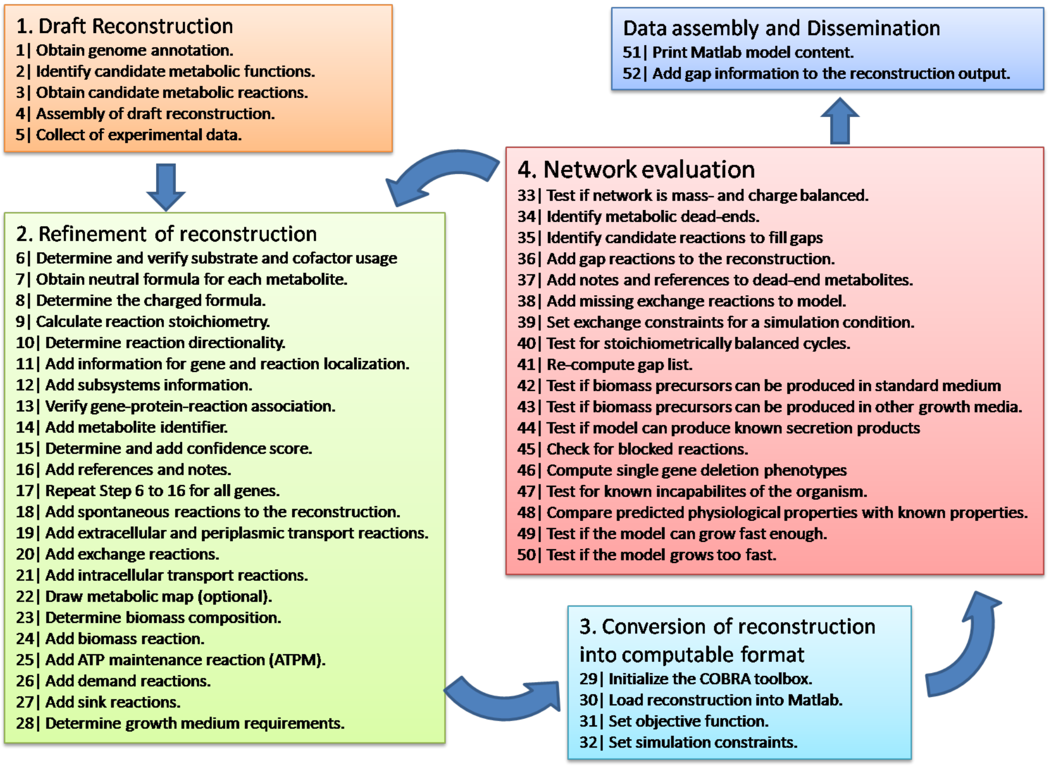
Network reconstructions have become a common denominator in systems biology. Bot-tom-up metabolic network reconstructions have developed over the past ten years. Reconstructions represent structured knowledge-bases that abstract pertinent information on the biochemical transformations taking place within specific target organisms. The conversion of a reconstruction into a mathematical format facilitates myriad computational biological studies including evaluation of network content, hypothesis testing and generation, analysis of phenotypical characteristics, and metabolic engineering. To date, metabolic reconstructions for more than 30 organisms have been published and this number is expected to increase rapidly. However, these reconstructions differ in quality and coverage, which may minimize their predictive potential and use as a knowledge-base. Here, we present a comprehensive protocol describing each reconstruction step necessary to construct a high-quality genome-scale metabolic reconstruction. The protocol also discusses common trials and tribulations that can occur at different steps of the reconstruction process. Therefore, this protocol provides a helpful manual for all stages of the reconstruction process.

3.1 Introduction

Metabolic network reconstructions have become an indispensable tool for studying the systems biology of metabolism. The number of organisms for which metabolic reconstructions have been created is increasing at a pace similar to whole genome sequencing [279]. However, the quality of metabolic reconstructions differs considerably, which is partially caused by varying amounts of available data for the target organisms, but also partially by a missing standard operating procedure that describes the reconstruction process in detail. This protocol details a procedure by which a quality-controlled quality-assured (QC/QA) reconstruction can be built to ensure high quality and comparability between reconstructions. In particular, the protocol points out the data which are necessary for the reconstruction process and that should accompany the reconstructions. Moreover, standard tests are presented, which are necessary to verify the functionality and applicability of reconstruction-derived metabolic models. Finally, this protocol presents strategies to debug non- or malfunctioning models. While the reconstruction process has been reviewed conceptually by numerous groups [70, 79, 193, 229] and a good general overview of the necessary data and steps is available, no detailed description of the reconstruction, debugging, and iterative validation process has been published. This protocol seeks to make this process explicit and generally available.

The presented protocol describes the procedure necessary to reconstruct metabolic networks indented to be used for computational modeling, including the constraint-based reconstruction and analysis (COBRA) approach [218]. These network reconstructions, and *in silico* models, are created in a bottom-up fashion based on genomic and bibliomic data, and thus represent a biochemical, genetic, and genomic (BiGG) knowledge-base for the target organism [229]. These BiGG reconstructions can be readily converted into mathematical models and their properties can be determined. For example, they can be used to simulate maximal growth of a cell in a given environmental condition using flux balance analysis (FBA) [249, 289]. In contrast, the generation of networks derived from top-down approaches (high-throughput data based interference of component interactions) is not discussed here as they do not result in functional, mathematical models.

The metabolic reconstruction process described herein is usually very labor- and time intensive, spanning from six months for well-studied, medium genome sized bacteria, to two years (and six people) for the metabolic reconstruction of human metabolism [66].



1. Draft Reconstruction
1| Obtain genome annotation.
2| Identify candidate metabolic functions.
3| Obtain candidate metabolic reactions.
4| Assembly of draft reconstruction.
5| Collect of experimental data.

2. Refinement of reconstruction
6| Determine and verify substrate and cofactor usage
7| Obtain neutral formula for each metabolite.
8| Determine the charged formula.
9| Calculate reaction stoichiometry.
10| Determine reaction directionality.
11| Add information for gene and reaction localization.
12| Add subsystems information.
13| Verify gene-protein-reaction association.
14| Add metabolite identifier.
15| Determine and add confidence score.
16| Add references and notes.
17| Repeat Step 6 to 16 for all genes.
18| Add spontaneous reactions to the reconstruction.
19| Add extracellular and periplasmic transport reactions.
20| Add exchange reactions.
21| Add intracellular transport reactions.
22| Draw metabolic map (optional).
23| Determine biomass composition.
24| Add biomass reaction.
25| Add ATP maintenance reaction (ATPM).
26| Add demand reactions.
27| Add sink reactions.
28| Determine growth medium requirements.

3. Conversion of reconstruction into computable format
29| Initialize the COBRA toolbox.
30| Load reconstruction into Matlab.
31| Set objective function.
32| Set simulation constraints.

4. Network evaluation
33| Test if network is mass- and charge balanced.
34| Identify metabolic dead-ends.
35| Identify candidate reactions to fill gaps
36| Add gap reactions to the reconstruction.
37| Add notes and references to dead-end metabolites.
38| Add missing exchange reactions to model.
39| Set exchange constraints for a simulation condition.
40| Test for stoichiometrically balanced cycles.
41| Re-compute gap list.
42| Test if biomass precursors can be produced in standard medium
43| Test if biomass precursors can be produced in other growth media.
44| Test if model can produce known secretion products
45| Check for blocked reactions.
46| Compute single gene deletion phenotypes
47| Test for known incapabilites of the organism.
48| Compare predicted physiological properties with known properties.
49| Test if the model can grow fast enough.
50| Test if the model grows too fast.

Data assembly and Dissemination
51| Print Matlab model content.
52| Add gap information to the reconstruction output.

Figure 3.1: *Overview of the procedure to iteratively reconstruct metabolic networks.* In particular, the steps, or stages, 2 to 4 are continuously iterated until model prediction is similar to the phenotypic characteristics of the target organism and/or all experimental data for comparison are exhausted.

Often, the reconstruction process is iterative, as demonstrated by the metabolic network of *Escherichia coli*, whose reconstruction has been expanded and refined over the last 19 years [80]. As the number of reconstructed organisms increases, the need to find automated, or at least semi-automated, ways to reconstruct metabolic networks straight from the genome annotation is growing. Despite growing experience and knowledge, to date, we are still not able to completely automatically reconstruct high-quality metabolic networks that can be used as predictive models. Recent reviews highlight current problems with genome annotations and databases, which make automated reconstructions challenging and thus, require manual evaluation [79, 96, 229]. Organism-specific features such as substrate and cofactor utilization of enzymes, intracellular pH, and reaction directionality remain problematic, but some organism-specific databases and approaches exist, which may be used for further automation. We describe here the manual reconstruction process in detail.

A limited number of software tools and packages are available (freely and commercially), which aim to assist and facilitate the reconstruction process (Table 3.1). The presented protocol can, in principle, be combined with those reconstruction tools. For generality, we present the entire procedure using a spreadsheet, namely Excel workbook (Microsoft Inc), and a numeric computation and visualization software package, namely Matlab (Mathwork Inc). Free spreadsheets (e.g., Open office and Google Docs) could be used instead of the listed spreadsheet. Alternatively, MySQL databases may be used as they are very helpful to structure and track data. Matlab was also used to encode the COBRA Toolbox, which is a suite of COBRA functions commonly used for simulation [21]. This Toolbox was extended to facilitate the reconstruction, debugging, and manual curation process described herein.

The protocol describes in detail the process to generate metabolic reconstructions applicable for representatives of all domains of life. The process of reconstructing prokaryotic and eukaryotic metabolic networks is, in principle, identical, although eukaryote reconstructions are more challenging due to size of genomes, coverage of knowledge, and the multitude of cellular compartments. Specific properties and pitfalls are highlighted in the protocol.

The described reconstruction and debugging process requires organism specific information. The minimum information includes the genome sequence, from which key

Table 3.1: **Data sources frequently used for metabolic reconstructions.**

Name	**Link**	**Comments**
	Genome Databases	
Comprehensive Microbial Resource (CMR)	http://cmr.jcvi.org	
Genomes Online Database (GOLD)	www.genomesonline.org/	
TIGR	www.tigr.org/db.shtml	
NCBI Entrez Gene	www.ncbi.nlm.nih.gov/sites/entrez	
SEED database [203]	theseed.uchicago.edu/FIG/index.cgi	Comparative genomics tool
	Biochemical Databases	
KEGG [130]	www.genome.jp/kegg/	
BRENDA [18]	www.brenda-enzymes.info/	
Transport DB [234]	www.membranetransport.org/	
PubChem [295]	http://pubchem.ncbi.nlm.nih.gov/	
pKa DB	www.acdlabs.com/products/ phys_chem_lab/pka/	Commercial software
	Organism-specific databases	
Ecocyc [135]	ecocyc.org/	*E. coli* database
Gene Cards	www.genecards.org/	Human gene database
	Protein Localization databases	
PSORT [91]	www.psort.org/psortb/	Support vector machine (SVM) based
PA-SUB [167]	www.cs.ualberta.ca/b̃ioinfo/PA/Sub/	SVM based
	COBRA simulation environments	
CellNetAnalyzer [145, 146]	http://www.mpi-magdeburg.mpg.de	Matlab is required
COBRA Toolbox [21]	http://systemsbiology.ucsd.edu/	Matlab is required
MetaFluxNet [154, 158]	http://mbel.kaist.ac.kr/lab/mfn/	Stand alone package

metabolic functions can be obtained, and physiological data, such as growth conditions, which allow the comparison of model prediction to refine the network's content. In general, the more information about physiology, biochemistry, and genetics that is available for the target organism, the better the predictive capacity of the models. This property becomes obvious considering that the network evaluation and validation process relies on comparing predicted phenotypes (e.g., growth rate) with experimental observations. Additional cellular objectives may be considered to compare with experimental data but they are not discussed here in detail [39, 66, 94, 245, 251].

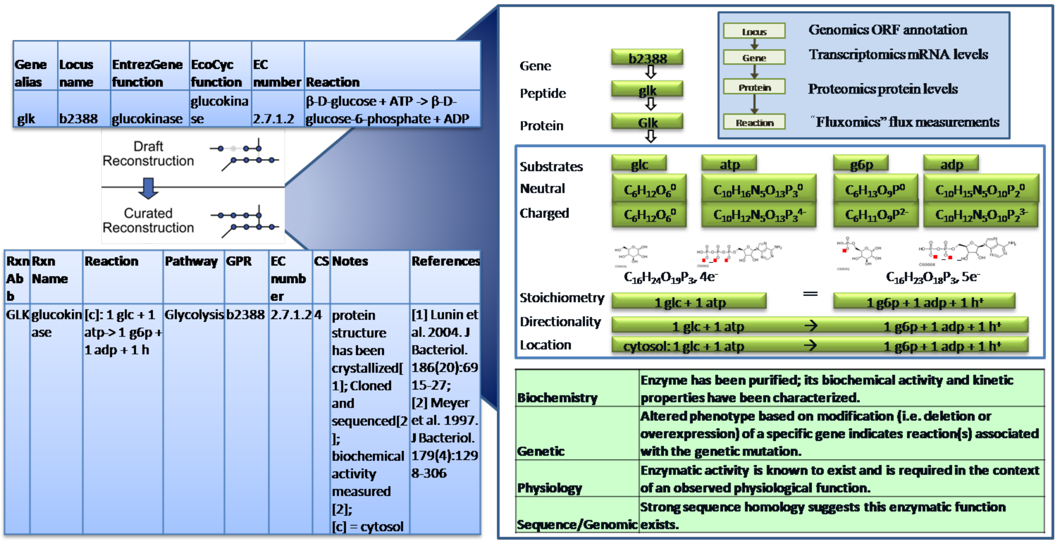
Table 3.2: **General error modes in metabolic networks.**

Error mode	**Action**
Wrong reaction constraints	Check reaction constraints if they are applied correctly.
Missing transport reactions	Add transport reactions.
Missing exchange reactions	Add exchange reactions.
Cofactor cannot be consumed or produced.	Follow Figure 3.17
Shuttling of compounds across compartment.	Adjust reversibility of transport reactions.

Although this protocol presents the reconstruction process in terms of metabolic networks, the same approach can, and has been, applied for reconstructing signaling [161, 209] and transcription/ translation networks [277]. Regulatory networks have not been constructed in a fully stoichiometric manner yet, although a pseudo-stoichiometric approach has been proposed [95]. The reconstruction process for these networks is not as well established as for metabolic networks, and is thus still subject to active research.

Lastly, myriad data sources are used during the reconstruction process rendering metabo-lic network reconstructions as knowledge-bases, which summarize and structure the available BiGG knowledge about the target organism. A list of frequently used organism-unspecific and some of the organism-specific resources are listed in Table 3.1. Note that the quality and wealth of organism-specific information will directly affect the quality and coverage of the metabolic reconstruction. Great resources are organism-specific books that have been published for a growing number of organisms. In cases where organism-specific information is scarce, data from phylogenetic neighbors may be of great help. It is important to ensure that, in cases where the reconstruction relies extensively on relative information, the overall behavior of the model matches the target organism. This assurance can be achieved by carefully comparing the predictions with experimental and physiological data, such as growth conditions, secretion products, and knock-out phenotypes.

The resulting knowledge-bases can be queried, used for mapping experimental data (e.g., gene expression, proteomic, fluxomic, and metabolomic data), and converted into a mathematical format to investigate metabolic capabilities and generate new biological hypotheses. The multitude of possible applications of BiGG knowledge-bases distinguishes them from other, automated efforts. By introducing standards in content and format with this protocol it will soon be possible to compare metabolic reconstructions between different organisms, which will further enhance our understanding of the evolutionary processes and may provide a complementary approach to comparative genomics.



Gene alias	Locus name	EntrezGene function	EcoCyc function	EC number	Reaction
glk	b2388	glucokinase	glucokinase	2.7.1.2	β-D-glucose + ATP -> β-D-glucose-6-phosphate + ADP

Draft Reconstruction
Curated Reconstruction

Rxn Abb	Rxn Name	Reaction	Pathway	GPR	EC number	CS	Notes	References
GLK	glucokinase	[c]: 1 glc + 1 atp-> 1 g6p + 1 adp + 1 h	Glycolysis	b2388	2.7.1.2	4	protein structure has been crystallized[1]; Cloned and sequenced[2]; biochemical activity measured [2]; [c] = cytosol	[1] Lunin et al. 2004. J Bacteriol. 186(20):6915-27; [2] Meyer et al. 1997. J Bacteriol. 179(4):1298-306

Gene
Peptide
Protein
b2388
glk
Glk
Locus
Gene
Protein
Reaction
Genomics ORF annotation
Transcriptomics mRNA levels
Proteomics protein levels
"Fluxomics" flux measurements
Substrates
Neutral
Charged
glc
atp
g6p
adp
$C_6H_{12}O_6^0$
$C_{10}H_{16}N_5O_{13}P_3^0$
$C_6H_{13}O_9P^0$
$C_{10}H_{15}N_5O_{10}P_2^0$
$C_6H_{12}O_6^0$
$C_{10}H_{12}N_5O_{13}P_3^{4-}$
$C_6H_{11}O_9P^{2-}$
$C_{10}H_{12}N_5O_{10}P_2^{3-}$
$C_{16}H_{24}O_{19}P_3$, $4e^-$
$C_{16}H_{23}O_{18}P_3$, $5e^-$
Stoichiometry
1 glc + 1 atp
=
1 g6p + 1 adp + 1 h^+
Directionality
1 glc + 1 atp → 1 g6p + 1 adp + 1 h^+
Location
cytosol: 1 glc + 1 atp → 1 g6p + 1 adp + 1 h^+

Biochemistry	Enzyme has been purified; its biochemical activity and kinetic properties have been characterized.
Genetic	Altered phenotype based on modification (i.e. deletion or overexpression) of a specific gene indicates reaction(s) associated with the genetic mutation.
Physiology	Enzymatic activity is known to exist and is required in the context of an observed physiological function.
Sequence/Genomic	Strong sequence homology suggests this enzymatic function exists.

Figure 3.2: **Refinement of reconstruction content.** The draft reconstruction is converted into a curated reconstruction by re-evaluation of the content. In particular, the metabolic reactions, obtained from biochemical databases or the literature, need to be tested for mass- and charge balancing. Many resources omit protons and water. Furthermore, adjusting metabolites to a particular pH may change their charged formulae and thus may require correction of the network reaction. For instance, the reaction catalyzed by the glucokinase which was obtained from KEGG [295] is not mass- and charge-balanced when charged metabolite formula at pH 7.2 is considered. The right hand side (RHS) is missing an H^+ and the charge is unbalanced. Adding a proton to the RHS balances both sides of the equation in terms of protons and electrons/charge. Abbreviations: glc - D-glucose, g6p - D-glucose-6-phosphate, atp - adenosine-triphosphate, adp - adenosine-diphosphate, H+ - proton. CS - confidence level.

3.2 General procedure

The metabolic network reconstruction process described herein consists of four major stages or steps followed by its prospective use in Step 5 (Figure 3.1).

- The first step is the generation of a draft reconstruction based on the genome annotation of the target organism and biochemical databases. This draft reconstruction, or automated reconstruction, is thus a collection of genome encoded metabolic functions, some of which may be falsely included while other ones are missing (e.g., due to missing or incomplete annotations). Software tools such as Pathway tools [135] or metaSHARK [215] can be used for the generation of the draft reconstruction but they do not replace the manual curation.

- The second step of the reconstruction process concentrates on the curation and refinement of the content. We highlight in this protocol parts that need special attention. In particular, the metabolic functions and reactions collected in the draft reconstruction are individually evaluated against organism-specific literature (and expert opinion). This manual evaluation is important since 1) not all annotations have a high confidence score (e.g., low e-value), and 2) biochemical databases are mostly organism-unspecific, listing enzymes activities found in various organisms, not all of which may be present in the target organism (Figure 3.2). Including organism-unspecific reactions can affect the predictive behavior of the models. Furthermore, information about biomass composition, maintenance parameters, and growth conditions are collected in this step, which will provide a basis for the simulations in Steps 3 and 4.

- In the third step, the reconstruction is converted into a mathematical format and condition-specific models are defined. This step can be mostly automated. Moreover, systems boundaries are defined, converting the general reconstruction into a condition-specific model. Note that the initial model may differ in scope and boundaries to the final model, which is obtained after multiple iterations of validation and refinement, and which is used to simulate phenotypic behavior in a prospective manner. Figure 3.7 illustrates the conversion of a reconstruction into mathematical format.

- The fourth step in the reconstruction process consists of network verification, evaluation, and validation. Common error modes in metabolic reconstructions are listed in Table 3.2. The metabolic model created in the third step is tested, among other thing, for its ability to synthesize biomass precursors (such as amino acids, nucleotides triphosphates, and lipids). This evaluation generally leads to the identification of missing metabolic functions in the reconstruction, so called network gaps, which are added by repeating Steps 2 and 3. This illustrates how the reconstruction process is an iterative procedure. An important issue is when to stop the iterative process and call a reconstruction ”finished”. This decision is normally based on the definition of the scope and purpose of the reconstruction.

Once the necessary content and capability is reached, one can start to use the reconstruction in a prospective manner, which represents a fifth step in the reconstruction process that is not address here.

The literature search as well as physiological data may suggest the presence of metabolic functions for which no responsible gene can be identified in the genome. The presented protocol does not describe how more refined annotation can be obtained. The interested reader should refer to available work and reviews [13, 202, 203, 263].

3.3 Materials

3.3.1 Equipment

- A standard personal computer that can run Matlab Version 6.0 or above of Matlab (Mathwork Inc.), a numerical computation and visualization software (http://mathworks.com)
- The COBRA Toolbox (version 1.4 or above) is provided at http:// systemsbiology.ucsd.edu/downloads/COBRAToolbox
- The SBML Toolbox for Matlab which allows reading models in SBML format (http://sbml.org/Software/SBMLToolbox)
- A linear programming (LP) solver. Multiple solvers are currently supported by the COBRA Toolbox:
    - glpk (freeware): http://www.gnu.org/software/glpk/

  - LINDO (LINDO Systems Inc.) Matlab API (commercial): http://www.lindo.com
  - CPLEX (ILOG Inc.) through the Tomlab (Tomlab Optimization Inc.) optimization environment (commercial, but best LP solver available) http://tomopt.com/
  - Mosek (MOSEK ApS) (commercial): http://www.mosek.com
- Extreme pathway software package, X3, provided at http://systemsbiology.ucsd.edu/downloads/Extreme_Pathway_Analysis
- Excel (Microsoft Inc., http://office.microsoft.com/en-us/excel/default.aspx)

3.3.2 Equipment setup

COBRA Toolbox. The COBRA Toolbox [21] consists of files which should be placed in a local folder on the user's computer. After opening Matlab, a path should be set to the local folder, containing the COBRA Toolbox (Matlab ' File ' Set Path ' Add with Subfolder, choose the corresponding folder and save). All working files (SBML and xls files) should also be stored in the local folder, in order to allow access to the reconstruction and models. A full documentation of the COBRA Toolbox can be found in the "doc" subfolder within the main Toolbox folder, which has all help files as html files. Furthermore, help for Matlab and COBRA Toolbox functions can be accessed via Matlab's "help" facility by typing "help function_name" on Matlab command line.

SBML file. Comprehensive documentation on SBML, the file format, and model setup, can be found at the official SBML website (http://sbml.org/documents/, level 2 version 1). The SBML file describing the model has to include at least the following information: stoichiometry of each reaction, upper/lower bounds of each reaction, and objective function coefficients for each reaction. Additionally, gene-reaction associations can be added to the "Notes" section.

Spreadsheet. The first two reconstruction steps are illustrated in this protocol using spreadsheets. It is important that the order of the columns in the spreadsheet match the example given in Figure 3.14 and in the supplemental files.

Variables. The imported model from the spreadsheets is contained in a model structure (see Figure 3.15 for details on this structure). All functions in the COBRA Toolbox access the information stored in the model structure. The values computed by the COBRA Toolbox are fluxes, which can be best understood as reaction rates. The units for fluxes used throughout this protocol are $\frac{mmol}{g_{DW} \cdot h}$, where g_{DW} is the dry weight of the cell in grams.

Installation. The Matlab software, SBML Toolbox, and one or more of the suggested LP solvers should be installed following the instructions of the software providers. Note that the SBML Toolbox and the LP solver also need to be accessible in the Matlab path (see above). Sample installation instructions for the lp_solve LP solver on Windows can be found in Becker *et al.* [21]. The COBRA Toolbox is initiated by typing in the Matlab command window:

```
>> changeCobraSolver(solverName);
```

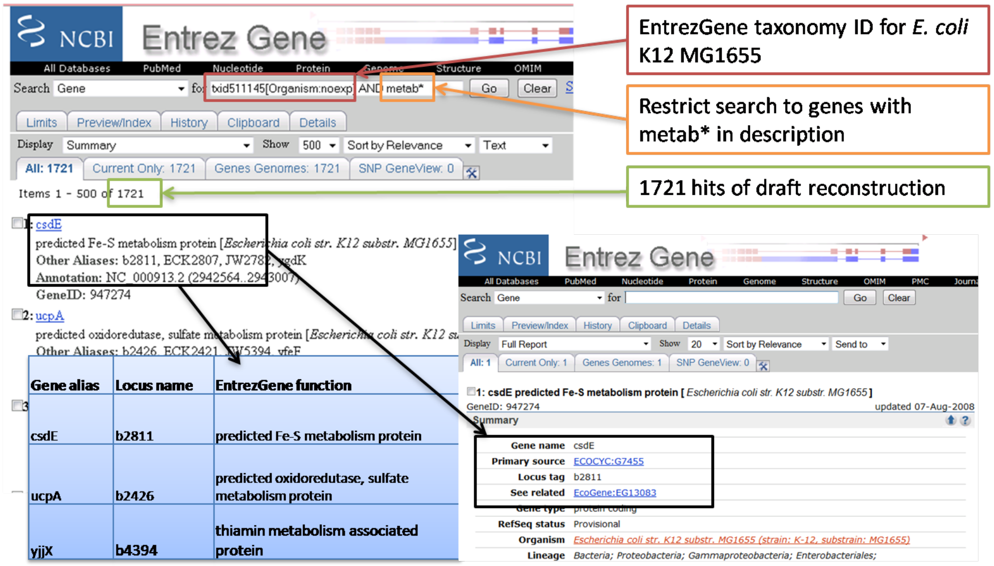
where 'solverName' is, e.g., 'lp_solve'

```
>> initCobraToolbox;
```

CRITICAL STEP The SBML Toolbox and the LP solver should be tested for functionality following the software provider's instructions before attempting to use the COBRA Toolbox.

X3 is the software package used to determine stoichiometrically unbalanced cycles, or Type III pathways. X3.exe needs to be placed and extracted in a local folder. The help can be accessed by opening the DOS command line, changing to the local folder, and typing X3 -h. The extreme pathway tool will be called from Matlab. It can be downloaded from http://systemsbiology. ucsd.edu/downloads /Extreme_Pathway_Analysis.

We will illustrate many steps of the protocol using KEGG [130] because it is freely accessible and very helpful for the illustrated pathway-by-pathway reconstruction process. However, one has to keep in mind two properties of KEGG [130]: 1. It is NOT organism-specific data; hence, not all reactions associated with an enzyme may be catalyzed by the enzyme of the target organism, and 2. KEGG [130] may NOT update the genome annotation of the target organism on a regular basis, hence the information may be outdated and need a "second opinion" from another more recent resource. 3. Not all reactions in the KEGG [130] database are mass- and charge-balanced; they omit protons and water


NCBI
Entrez Gene
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
OMIM
PMC
Journa
Search
Gene
for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display
Show
Sort by Relevance
Text
Send to
All: 1721
Current Only: 1721
Genes Genomes: 1721
SNP GeneView: 0
All: 1
Current Only: 1
Genes Genomes: 1
SNP GeneView: 0
Summary

Search Gene for txid511145[Organism:noexp] AND metab*

Display Summary Show 500

Items 1 - 500 of 1721

☐1: csdE
predicted Fe-S metabolism protein [*Escherichia coli str. K12 substr. MG1655*]
Other Aliases: b2811, ECK2807, JW2782, ygdK
Annotation: NC_000913.2 (2942564..2943007)
GeneID: 947274

☐2: ucpA
predicted oxidoredutase, sulfate metabolism protein [*Escherichia coli str. K12 su*
Other Aliases: b2426, ECK2421, JW5394, yfeF

Gene alias	Locus name	EntrezGene function
csdE	b2811	predicted Fe-S metabolism protein
ucpA	b2426	predicted oxidoredutase, sulfate metabolism protein
yjjX	b4394	thiamin metabolism associated protein

Display Full Report Show 20

☐1: csdE predicted Fe-S metabolism protein [*Escherichia coli str. K12 substr. MG1655*]
GeneID: 947274 updated 07-Aug-2008

Gene name	csdE
Primary source	ECOCYC:G7455
Locus tag	b2811
See related	EcoGene:EG13083
Gene type	protein coding
RefSeq status	Provisional
Organism	*Escherichia coli str. K12 substr. MG1655 (strain: K-12, substrain: MG1655)*
Lineage	*Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;*

EntrezGene taxonomy ID for *E. coli* K12 MG1655

Restrict search to genes with metab* in description

1721 hits of draft reconstruction

Figure 3.3:

Collecting information for draft reconstruction using, e.g., EntrezGene as annotation source.

molecules in some cases.

3.4 Procedure

The order of the steps in the different stages is a recommendation. This means that the order of steps can be altered within each stage, and with some limitations between the stages, as long as they are completed. The quality of the reconstruction is generally ensured by performing the steps although some logical order may apply.

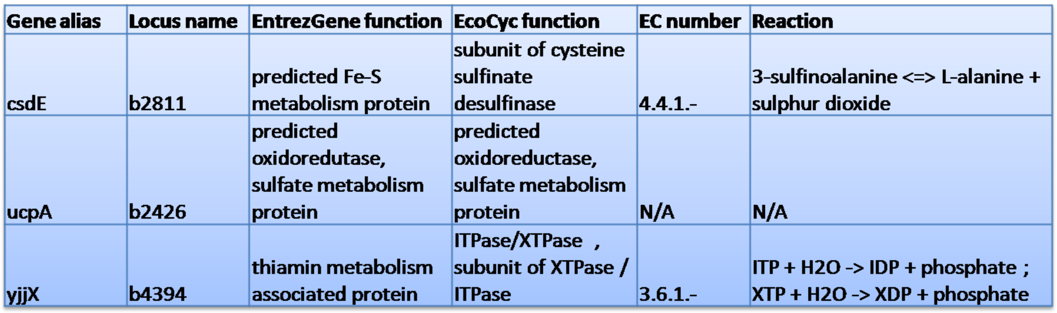
3.4.1 Creating a draft reconstruction

Note that the creation of a draft reconstruction and the manual reconstruction refinement (next stage) may be combined for bacterial reconstructions with main emphasis on the reconstruction refinement.

Step 1| Obtain genome annotation. The genome annotation can be obtained from various sources, including sequencing centers (e.g., TIGR) and the National Center for Biotechnology Information (NCBI) depository. The following information should be retrieved for each gene: genome position, coding region, strand, locus name, alias, gene function (i.e., current annotation), protein classification (e.g., Enzyme Commission (E.C.) number [181]). The genomic information is important to unambiguously define the gene in respect to the organism's genome as well as to allow data mapping (e.g., gene expression) in subsequent studies.

CRITICAL STEP Since the draft reconstruction, and, to some extent, the curated reconstruction, relies mainly on the genome annotation, it is important to download the most recent version available to ensure that updates and corrections since the genome's original publication are accounted for. In some cases, the genome-sequencing group created organism specific database (e.g., for *Helicobacter pylori* [30] and *E. coli* [134]). Table 3.1 lists some of the commonly used databases for annotations.

CRITICAL STEP In eukaryotic organisms, information regarding alternate transcripts must also be collected, since different splice forms may have distinct functionality or cellular localization.



Gene alias	Locus name	EntrezGene function	EcoCyc function	EC number	Reaction
csdE	b2811	predicted Fe-S metabolism protein	subunit of cysteine sulfinate desulfinase	4.4.1.-	3-sulfinoalanine <=> L-alanine + sulphur dioxide
ucpA	b2426	predicted oxidoredutase, sulfate metabolism protein	predicted oxidoreductase, sulfate metabolism protein	N/A	N/A
yjjX	b4394	thiamin metabolism associated protein	ITPase/XTPase , subunit of XTPase / ITPase	3.6.1.-	ITP + H2O -> IDP + phosphate ; XTP + H2O -> XDP + phosphate

Figure 3.4: Example of a draft reconstruction.

Step 2| Identify candidate metabolic functions. This step is straight-forward once the genome annotation has been obtained. Different approaches may be applied to collect candidate metabolic functions including searching for E.C. numbers (complete and partial) [181] and for metabolic terms (e.g., dehydrogenase, kinase, etc.) (Figure 3.3). If gene ontology (GO) [11] or cluster of orthologous groups of proteins (COG) [274] information was obtained with the genome annotation, they can be used as well to find metabolic enzymes.

CRITICAL STEP It is important to understand that this step aims to obtain a list of candidates, which in no way will be complete or comprehensive. Many false-positives may be present in the list. For example, proteins involved in DNA methylation or rRNA modification also have E.C. numbers, but their functions are normally not considered in metabolic reconstructions. Another example involves kinases that may be involved in signal transfer reactions or annotated as 'histidine kinase-like' and thus, no specific function can be derived from this annotation. A more targeted query for metabolic annotations could be designed to reduce the number of false-positives but it does not replace manual curation.

Step 3| Obtain candidate metabolic reactions for these functions (e.g., from KEGG [130]). Comprehensive reaction databases such as KEGG [130], Brenda [18], and publically available reconstructions can be used as a resource to combine the gene functions with metabolic reactions.

Step 4| Assembly of draft reconstruction. All candidate metabolic genes and their potential reactions are collected in a spreadsheet. This spreadsheet will serve as a starting point for the manual curation process (see Figure 3.4 for an example).

Step 5| Collection of experimental data. The manual curation process relies heavily on experimental, organism-specific information. All possible information needs to be retrieved. The following steps will include reviewing scientific literature during which the information listed in Table 3.3 should be collected. Alternatively, additional experimental data can be generated by growing and measuring various metabolic capabilities and properties of the target organism.

3.4.2 Manual reconstruction refinement

In this part, the entire draft reconstruction will be re-evaluated and refined. For each gene and reaction entry, two questions will be asked: 1) Should this entry be here? 2) Is there an entry missing to connect the entry with the remainder of the network?

Step 6| Determine and verify substrate and cofactor usage. Substrate and cofactor specificity of enzymes differs between organisms. Organism-unspecific databases, such as KEGG [130] and Brenda [18], list all possible transformations of an enzyme that have been identified in any organism. As a rule of thumb, one can assume that enzymes, which have only one reaction associated in KEGG [130], for example, do not require organism refinement. However, enzymes that are associated with multiple reactions, with

Table 3.3: **List of experimental data that can be used for the different stages of reconstruction, modeling and network evaluation.** L - Literature, DB - Databases, G - Genome, GE - Growth experiments, Ph - Phenotyping, PS - Protein structures, CG - Comparative genomics, Mu - Microarray data, P - Proteomic data, EM - Exometabolmic data, M - Metabolomic data, F - Flucomic data, ko - Single gene deletion, B - Biochemical essays

Data type	**Purpose**	**L**	**DB**	**G**	**GE**	**Ph**	**PS**	**CG**	**Mu**	**P**	**EM**	**M**	**F**	**ko**	**B**
Gene function	Refinement	X	X	X				X							X
Protein function	Refinement	X	X	X			X	X							X
Reaction mechanism	Refinement	X	X	X											X
Growth media	Transport, simulations	X			X	X					X	X			
Carbon sources	Transport, simulations	X			X	X					X	X			
Gene/protein presence/ absence	Condition-specific/ cell-type models		X						X	X				X	
Reaction constraints	Simulations				X	X			X	X	X		X		
Network evaluation	Debugging				X	X		X			X	X	X	X	
Gap filling	Debugging	X	X	X		X			X	X					X
Cofactor/ substrate specificity	Refinement	X	(X)				X								X
Directionality	Refinement	X	(X)												X

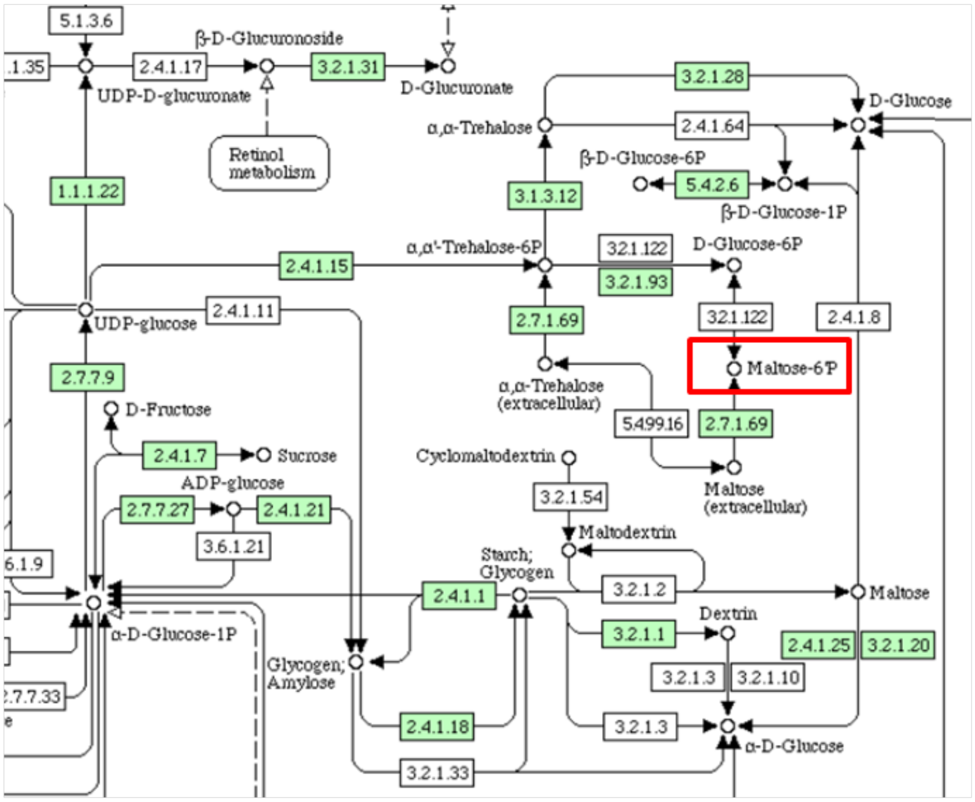

5.1.3.6
β-D-Glucuronoside
.1.35
2.4.1.17
3.2.1.31
UDP-D-glucuronate
D-Glucuronate
3.2.1.28
D-Glucose
α,α-Trehalose
2.4.1.64
Retinol metabolism
β-D-Glucose-6P
5.4.2.6
1.1.1.22
3.1.3.12
β-D-Glucose-1P
α,α'-Trehalose-6P
3.2.1.122
D-Glucose-6P
2.4.1.15
3.2.1.93
2.4.1.11
UDP-glucose
3.2.1.122
2.4.1.8
2.7.1.69
Maltose-6P
2.7.7.9
α,α-Trehalose (extracellular)
D-Fructose
5.4.99.16
2.7.1.69
2.4.1.7
Sucrose
Cyclomaltodextrin
ADP-glucose
Maltose (extracellular)
2.7.7.27
2.4.1.21
3.2.1.54
Maltodextrin
3.6.1.21
6.1.9
Starch; Glycogen
3.2.1.2
Maltose
2.4.1.1
Dextrin
α-D-Glucose-1P
3.2.1.1
2.4.1.25
3.2.1.20
Glycogen; Amylose
3.2.1.3
3.2.1.10
.7.7.33
2.4.1.18
3.2.1.3
α-D-Glucose
3.2.1.33

Figure 3.5: **Assessing the metabolic "environment" or "connectivity" of a metabolite.** Here, Maltose-6-phosphate is highlighted on the KEGG [130] map for "Starch and Sucrose Metabolism". All annotated *E. coli* genes (MG1655) in KEGG [130] were highlighted in green. Enzymes that are currently not annotated or not found are shown in white boxes. Maltose-6-phosphate is a dead-end metabolite in *E. coli*'s metabolic reconstruction [78]. The enzyme 3.2.1.122 is currently not annotated (in KEGG [130]). There are only two enzymes in the KEGG database [130] that seem to produce/consume Maltose-6-phosphate: 2.7.1.69 and 3.2.1.122. In contrast, D-Glucose-6-Phosphate is highly connected in the *E. coli* reconstruction [78].

varying substrates and/or cofactors, require manual refinement. Information about substrate and cofactor utilization can be obtained from organism-specific biochemical studies and/or organism-specific databases.

CRITICAL STEP This step can be very time consuming and laborious as it may be difficult to find the necessary information. This step is often associated with an intensive literature search. It is important to pay great attention to this step as false inclusion of substrates or cofactors can greatly change the *in silico* behavior (i.e., predictive potential) of the reconstruction.

CRITICAL STEP If no organism-specific information can be found in the literature, and no data are available for phylogenetically similar organisms, all reactions associated with the enzyme may be added to the reconstruction, but should be marked with the lowest confidence score. In the case of problems during subsequent simulations, these low confidence reactions can be easily identified. Alternatively, one can choose to only include the main reaction(s) associated with the pathway that is currently considered. The remaining reactions may be noted somewhere so that they can be readily retrieved if necessary.

CAUTION It is important to note if no evidence for all/most reactions associated with an enzyme could be found because the metabolites may be dead-end metabolites. Connecting low confidence dead-end metabolites with the remaining network in the gap analysis (**Steps 34 to 37**) may change the predictive potential of the reconstruction. Therefore, it is important for the gap analysis to have this information readily available.

CRITICAL STEP In some cases, it is possible to exclude certain reactions to be entered in the reconstruction. For example, exotic sounding substrates or products may be excluded as they are not very likely to ever be connected to the network. Furthermore, reactions containing generic terms, such as protein, DNA, electron acceptor, etc., should not be included as they are not specific enough and normally serve in databases as space holders until more knowledge and biochemical evidence is available.

CRITICAL STEP This step is an ideal point in the manual reconstruction process to identify missing functions in the draft reconstruction. Using KEGG [130] maps, for example, one can analyze the metabolic ”environment” of the reaction(s) under inspections. If the genome annotation of the target organism is present in KEGG [130], one can highlight the genes on the map. This gives an estimate of the ”connectivity” of the reaction with its metabolic surrounding (Figure 3.5). Missing reactions/functions may become evidence for which experimental/ annotation evidence should be collected (see **Steps 34 to 37** for gap analysis).

Step 7| Obtain a neutral formula for each metabolite in the reaction. The neutral formula can be readily obtained from various resources, including KEGG [130], Brenda [18], and PubChem [295]. While PubChem [295] is more comprehensive, KEGG [130] is certainly the most accessible resource, especially when KEGG [130] is used for obtaining the reaction.

CRITICAL STEP No database is perfect, therefore, one should always double-check the entries with one's expectation. In KEGG [130], for example, the neutral formula does not always agree with the molecule structure drawn in the images. Other resources such as biochemical textbooks or PubChem [295] should be used to ensure that the correct neutral formula is used. Furthermore, obtaining the molecular structure of the metabolites is important for verifying the neutral formula and deriving the charged formula.

Step 8| Determine the charged formula for each metabolite in the reaction. Retrieve the molecular structure for each metabolite, if you have not already done so in **Step 7**. The protonation state, and thus the charged formula, depends on the pH of interest. Often metabolic networks are reconstructed assuming an intracellular pH of 7.2. However, the intracellular pH of bacterial cells may vary depending on environmental conditions and bacteria. Also, the pH of organelles may be different, e.g., peroxisome. The protonated formula is calculated based on the pK_a value of the functional groups. Software packages such as Pipeline Pilot and pK_a DB can predict pK_a values for a given compound (Table 3.1). See Figure 3.2 for an example calculation of charged formula.

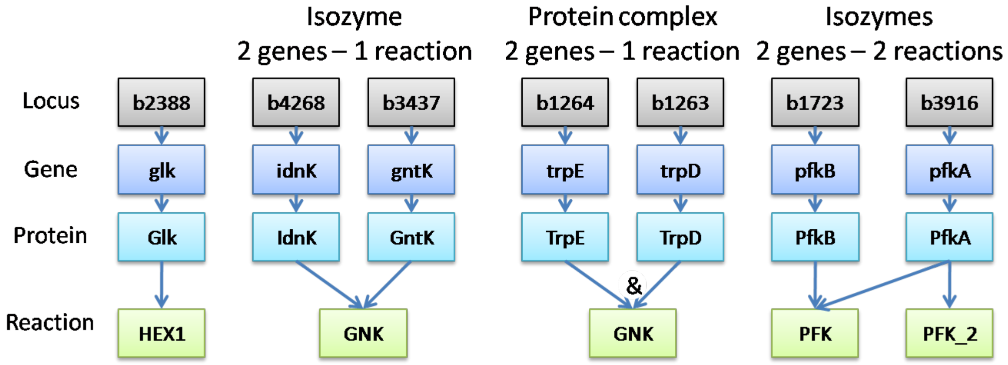
Step 9| Calculate reaction stoichiometry. Once the charged formula is obtained for each metabolite, the reaction stoichiometry can be determined. Protons and water may need to be added to the reaction in this step as some databases and many biochemical textbooks omit these molecules. Therefore, every element and the charge need to balance on both sides of the reaction. This step is easy for many central metabolic reactions but may become challenging for more complex reactions.

CRITICAL STEP Unbalanced reactions may lead to synthesis of protons or energy (ATP) out of nothing (see Figure 3.19 for examples).

Step 10| Determine reaction directionality. Biochemical data for the target organism are very important for this step but may not be available. New approaches are available that allow the estimation of the standard Gibbs free energy of formation ($\Delta_f G^{'\circ}$)

Table 3.4: **List of cellular compartments used in reconstructions** (may not be complete). For example, membrane compartments were not considered as reconstructions generally do not account for these, and detailed measurements are difficult to do. A - Achaea. B - Bacteria. EP - Eukaryotic pathogens. F - Fungi. PE - Photosynthetic eukarya. Y - Bakers yeast. H - Human. ° Symbol may vary between reconstructions. * Lysosome has been defined as a compartment but has not been used yet in any reconstruction. [a] *Leishmania major* [43]. [b] *Aspergillus nidulans* [59]. [c] *Chlamydomonas reinhardtii* [172].

Compartment	**Symbol**°	**A**	**B**	**EP**[a]	**F**[b]	**PE**[c]	**Y**	**H**
Extracellular space	[e]	X	X		X	X	X	X
Periplasm	[p]		X					
Cytoplasm	[c]	X	X	X	X	X	X	X
Nucleus	[n]			X			X	X
Mitochondrion	[m]			X	X		X	X
Chloroplast	[h]					X		
Lysosome*	[l]							
Vacuole	[v]						X	X
Golgi apparatus	[g]						X	X
Endoplasmatic reticulum	[r]			X			X	X
Peroxisome	[x]						X	X
Flagellum	[f]			X				
Glyoxysome	[o]				X			
Glycosome	[y]			X				
Acidocalcisome	[a]			X				


Isozyme
2 genes – 1 reaction
Protein complex
2 genes – 1 reaction
Isozymes
2 genes – 2 reactions
Locus
Gene
Protein
Reaction
b2388
b4268
b3437
b1264
b1263
b1723
b3916
glk
idnK
gntK
trpE
trpD
pfkB
pfkA
Glk
IdnK
GntK
TrpE
TrpD
PfkB
PfkA
&
HEX1
GNK
GNK
PFK
PFK_2

Reaction Abbreviation	Reaction Name	E. C. number	GPR
HEX1	hexokinase (D-glucose:ATP)	2.7.1.1	(b2388)
GNK	gluconokinase	2.7.1.12	(b3437) or (b4268)
ANS	anthranilate synthase	4.1.3.27	(b1264) and (b1263)
PFK	phosphofructokinase	2.7.1.11	(b1723) or (b3916)
PFK_2	phosphofructokinase	2.7.1.11	b3916

Figure 3.6: Examples of possible Gene-protein-reaction (GPR) associations and their representation in the reconstruction in Boolean format, taken from the *E. coli* metabolic reconstruction [78].

and of reaction ($\Delta_r G'^{\circ}$) in a biochemical system [83, 124]. If such data is not available, the following rule of thumb may be applied: 1) all reactions involving transfer of phosphate from ATP to an accepter molecule should be irreversible (with the exception of the ATP synthetase, which is known to occur in reverse); 2) reactions involving quinones are generally irreversible. While all KEGG [130] reactions are reversible, the KEGG [130] maps show the directionality of the reactions (which can also be downloaded as flat file from KEGG [130] ftp server (http://www.genome.jp/kegg/download/ftp.html). We normally assume the reaction directionality given in textbooks or KEGG maps, if no other information is available.

CRITICAL STEP Assigning the wrong direction to a reaction may have significant impact on the model's performance. In general, one should leave a reaction reversible if no information is available and the aforementioned rules of thumb do not apply. However, models with too many reversible reactions (too loose constraints) may have so called futile cycle, which overcome the proton gradient by freely exchanging metabolites and protons across compartments. Therefore, assigning the correct reversibility to transport reactions is especially important (see also **Steps 18 and 19**).

Step 11| Add information for gene and reaction localization. This information may be difficult to obtain. The compartments that have been considered in various metabolic reconstructions are listed in Table 3.4. Algorithms such as PSORT [91] and PASUB [167] can be used to predict the cellular localization of proteins based on nucleotide or amino acid sequences. A recently published protocol describes the use of online tools to predict the subcellular location of eukaryotic and prokaryotic proteins [75]. High-throughput experimental approaches are available to locate individual proteins, including immunofluorescence [240] and GFP tagging of individual proteins [116].

CRITICAL STEP In the absence of appropriate data, proteins should be assumed to reside in the cytosol. Incorrect assignment of the location of a reaction can lead to additional gaps in the metabolic network and misrepresentation of the network properties, especially if intracellular transport reactions need to be added for which no evidence is available either (see **Step 21** for details).

Step 12| Add subsystems information to reaction. This will be of great help for the debugging and network evaluation work. The subsystem assignment can be done either based on biochemical textbooks or KEGG [130] maps. Note that a reaction or an enzyme can appear in multiple KEGG [130] maps; therefore, the subsystem should reflect its primary function.

Step 13| Verify gene-protein-reaction (GPR) association. The genome annotation provides information about the GPR association, i.e., it indicates which gene has what function (Figure 3.6). The verification and refinement necessary in this step includes determining: i) if the functional protein is a heteromeric enzyme complex; ii) if the enzyme (complex) can carry out more than one reaction and iii) if more than one protein can carry out the same functions (i.e., isozymes exist). For the first case (i), the genome annotation often has refined information, e.g.,: protein X, catalytic subunit - which indicates that there is at least one more subunit needed for the function of the protein complex. Furthermore, KEGG [130] may list subunits in some cases. Often, a more comprehensive database and/or literature search is required. The protein complex composition may differ between organisms. The second case can also be identified from biochemical databases or literature. Note that multitasking of enzymes may differ between organisms.

CRITICAL STEP Mistakes or mis-assignments in the GPR associations will change

Table 3.5: **Confidence score system that is currently employed for metabolic reconstructions.**

Evidence type	**Confidence score**	**Examples**
Biochemical data	4	Protein purification and biochemical assays, experimentally solved protein structures. Comparative gene-expression studies can be also used as evidence (such as Chhabra *et al.* [46]).
Genetic data	3	Knockout characterization, Knock-in characterization, overexpression.
Physiological data	2	Physiological evidence for existence of reaction. E.g., known secretion products imply indirectly the existence of transporter as well as metabolic reactions. However, no enzyme/gene has been directly identified in target organism; thus, the actual reaction mechanism may differ from the reaction(s) included in the reconstruction.
Sequence data	2	Genome annotation, SEED annotation [203].
Modeling data	1	No supporting evidence available but reaction is required for modeling, e.g., growth or known by-product secretion. These reactions represent hypothesis and the actual reaction mechanism may be different from the included reactions.
Not evaluated	0	

results of *in silico* gene deletion studies. However, discrepancies between *in silico* and *in vivo* results can be used to refine knowledge and reconstructions (see **Step 46**)

Step 14| Add metabolite identifier. Metabolite identifiers are mecessary to enable the use of reconstructions for high-throughput data mapping (e.g., metabo-lomic or fluxomic data) and for comparison of the network content with other metabolic reconstructions. Therefore, metabolites and reactions need to be recognizable by other scientists and by software tools. Each metabolite should be associated with at least one of the following identifiers: ChEBI [36], Kegg [130], and PubChem [296]. In many cases, having one of the identifiers is sufficient to automatically obtain the other two identifiers. Furthermore, database-independent representations of metabolites such as SMILES [294] and InCHI strings [48, 298] are also helpful when associated with each metabolite. These representations represent the exact chemical structure of compounds. Additionally, collecting Molfiles (MDL file format, http://www.symyx.com/), which hold information about the atoms, bonds, connectivity and coordinates of a molecule, will be very useful, e.g., if you are using online software for pK_a determination (see **Step 10** for details). The supplemental material contains this information for the central metabolic reconstruction of *E. coli.*

Step 15| Determine and add confidence score. The confidence score represents a fast way of assessing the amount of information available for a metabolic function, pathway, or the entire reconstruction. Every network reaction is associated with a confidence score reflecting the information and evidence currently available. The confidence score ranges from 0 to 4, where 0 is the lowest and 4 is the highest evidence score (Table 3.5). Note that multiple information types result in a cumulative confidence score. For example, a confidence score of 4 may represent physiological and sequence evidence.

Step 16| Add references and notes based on experimental information. In **Steps 6 to 13** many organism-specific, experimental data is collected that needs to be associated with the reconstruction in the form of references and notes. This allows other users of the reconstruction to easily retrace the evidence and supporting material for reaction and gene inclusion.

Step 17| Repeat Steps 6 to 16 for all genes identified in the draft reconstruction. Also repeat these steps for metabolic functions that were identified from bibliomic sources

Table 3.6: **List of spontaneous reactions present in *E. coli*'s metabolic reconstruction** [78]. Note that this list is not complete. H^+ - proton. H_2O - water. GTP guanosine triphosphates. XTP Xanthosine triphosphates.

Abbreviation	**Name**	**Reaction**	**Pathway**
AOBUTDs	L-2-amino-3-oxobutanoate decarboxylation (spontaneous)	[L-2-Amino-3-oxobutanoate + H^+ →Aminoacetone + CO2	Threonine and Lysine Metabolism
DHPTDCs	4,5-dihydroxy-2,3-pentanedione cyclization (spontaneous)	4,5-dihydroxy-2,3-pentane-dione → H_2O + 4-hydroxy-5-methyl-3(2H)-furanone	Methionine Metabolism
FALGTHLs	formaldehyde glutathione ligase (spontaneous)	formaldehyde + Reduced glutathione ↔ hydroxymethylglutathione	Cofactor and Prosthetic Group Biosynthesis
G5SADs	L-glutamate 5-semialdehyde dehydratase (spontaneous)	L-glutamate 5-semialdehyde → 1-pyrroline-5-carboxylate + H^+ + H_2O	Arginine and Proline Metabolism
GTPHs	GTP amine hydrolysis (spontaneous)	GTP + H^+ + H_2O → Ammonium + XTP	Nucleotide Salvage Pathway
METOX1s	methionine oxidation (spontaneous)	hydrogen peroxide + L-methionine → H_2O + L-methionine sulfoxide	Methionine Metabolism

during the reconstruction process and whose genes could not determined.

Step 18| Add spontaneous reactions to the reconstruction. An excerpt of typical spontaneous reactions included in metabolic reconstructions is listed in Table 3.6. Note that only those spontaneous reactions that have at least one metabolite connecting them to the rest of the reconstruction should be added. This is to avoid too many dead-end metabolites caused by spontaneous reactions. In more recent reconstructions, spontaneous reactions have been associated with an artificial gene (s0001) and protein (S0001). By doing so, reaction and gene essentiality studies are easier to analyze. Furthermore, this artificial GPR makes it easy to distinguish between spontaneous and orphan reactions, i.e., reactions without known gene.

Step 19| Add extracellular and periplasmic transport reactions to the reconstruction. This addition is done based on experimental data. The rule here is that for every metabolite that is known to be taken up from the medium or that is known to be secreted into the medium, a transport reaction should exist (from extracellular space to periplasm and from periplasm to cytoplasm).

CRITICAL STEP Note that the transport from extracellular space to periplasm is mostly enabled by porins, which are unspecific and facilitate diffusion, while transport across the inner membrane is done by metabolite-specific transport systems, such as ABC transport, antiport, symport, etc. Some of these transport systems are already annotated in the genome.

CRITICAL STEP Include transport reactions for metabolites that can diffuse through the membranes. Small, hydrophilic compounds can diffuse through the outer membrane [125].

Step 20| Add exchange reactions to the reconstruction. Exchange reactions need to be added for all extracellular metabolites. The exchange reactions represent the systems boundaries (Figure 3.7).

Step 21| Add intracellular transport reactions to the reconstruction. (For multi-compartment reconstructions only). Intracellular transport reactions need to be added for all metabolites that are supposed to ”move” between compartments. Inner cellular transport systems are not very well studied and many of these are not annotated in the genome. Finding experimental data is often not easy. A general approach should be

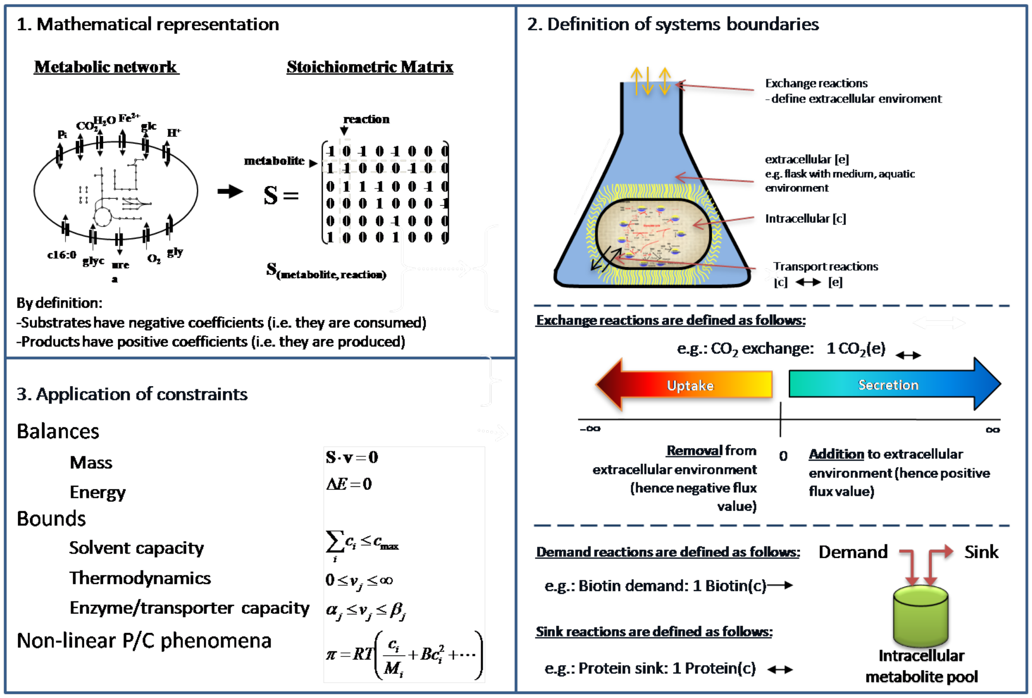
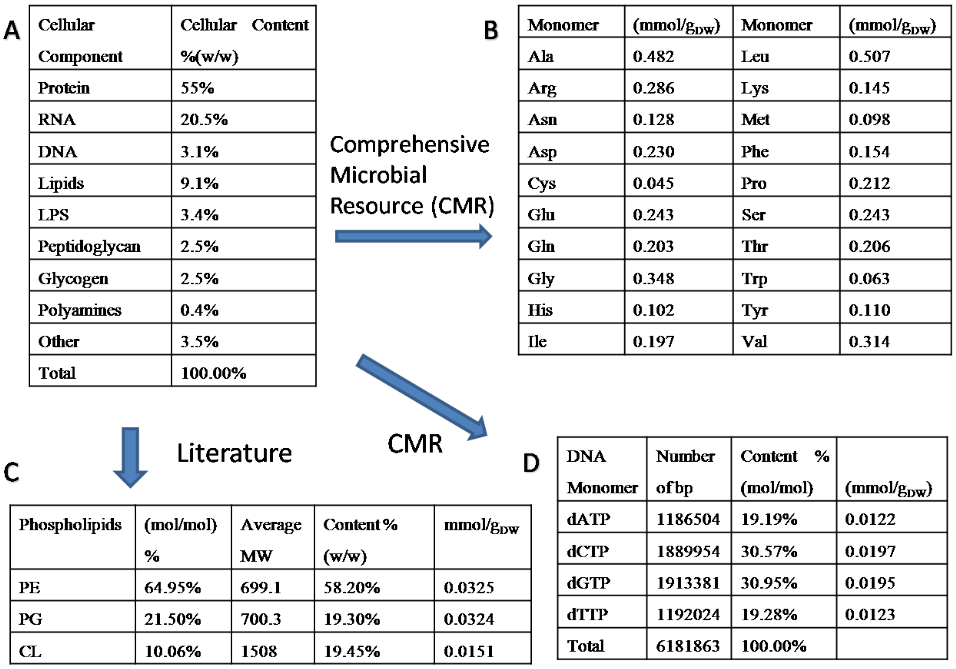

1. Mathematical representation
Metabolic network
Stoichiometric Matrix
P_i CO_2 H_2O Fe^{3+} glc H^+
c16:0 glyc urea O_2 gly
reaction
metabolite
S =
$S_{(metabolite, reaction)}$
By definition:
-Substrates have negative coefficients (i.e. they are consumed)
-Products have positive coefficients (i.e. they are produced)
3. Application of constraints
Balances
Mass $S \cdot v = 0$
Energy $\Delta E = 0$
Bounds
Solvent capacity $\sum_i c_i \le c_{max}$
Thermodynamics $0 \le v_j \le \infty$
Enzyme/transporter capacity $\alpha_j \le v_j \le \beta_j$
Non-linear P/C phenomena $\pi = RT\left(\frac{c_i}{M_i} + Bc_i^2 + \cdots\right)$
2. Definition of systems boundaries
Exchange reactions
- define extracellular enviroment
extracellular [e]
e.g. flask with medium, aquatic environment
Intracellular [c]
Transport reactions
[c] ↔ [e]
Exchange reactions are defined as follows:
e.g.: CO_2 exchange: 1 CO_2(e) ↔
Uptake
Secretion
-∞
0
∞
Removal from extracellular environment (hence negative flux value)
Addition to extracellular environment (hence positive flux value)
Demand reactions are defined as follows:
e.g.: Biotin demand: 1 Biotin(c)→
Demand
Sink
Sink reactions are defined as follows:
e.g.: Protein sink: 1 Protein(c) ↔
Intracellular metabolite pool

Figure 3.7: **Conversion of reconstruction into a condition-specific model.** This conversion requires three main steps. 1. The first step involves the mathematical representation by a stoichiometric matrix, S, of the network reaction list. The columns of S correspond to the network reactions, while the rows represent the network metabolites. The substrates in a reaction are defined to have a negative coefficient, while products have a positive value. The metabolites participating in a reaction have non-zero entry in the S matrix. 2. Now that the reconstruction is in a computer-readable format, the systems boundaries need to be defined. In particular, this means that for all metabolites that can be consumed or secreted by the target cell a so-called exchange reaction needs to be added to the reconstruction. The exchange reactions can be employed in later simulation to define for example environmental conditions (e.g., carbon source). 3. As a last step, constraints will be added to the reconstruction, thus rendering it to a condition-specific model. Mass conservation is a basic physical law. All steady-states can be thus described by $S \cdot v = 0$ where v is a vector of reaction fluxes. Adding further constraints such as thermodynamics (reaction directionality), enzyme capacity or regulation (i.e., presence or absence of an enzyme) to the model will lead to a smaller, more confined set of feasible steady-states flux solutions.


A
B
C
D
Comprehensive Microbial Resource (CMR)
CMR
Literature

A

Cellular Component	Cellular Content %(w/w)
Protein	55%
RNA	20.5%
DNA	3.1%
Lipids	9.1%
LPS	3.4%
Peptidoglycan	2.5%
Glycogen	2.5%
Polyamines	0.4%
Other	3.5%
Total	100.00%

B

Monomer	(mmol/g_{DW})	Monomer	(mmol/g_{DW})
Ala	0.482	Leu	0.507
Arg	0.286	Lys	0.145
Asn	0.128	Met	0.098
Asp	0.230	Phe	0.154
Cys	0.045	Pro	0.212
Glu	0.243	Ser	0.243
Gln	0.203	Thr	0.206
Gly	0.348	Trp	0.063
His	0.102	Tyr	0.110
Ile	0.197	Val	0.314

C

Phospholipids	(mol/mol) %	Average MW	Content % (w/w)	mmol/g_{DW}
PE	64.95%	699.1	58.20%	0.0325
PG	21.50%	700.3	19.30%	0.0324
CL	10.06%	1508	19.45%	0.0151

D

DNA Monomer	Number of bp	Content % (mol/mol)	(mmol/g_{DW})
dATP	1186504	19.19%	0.0122
dCTP	1889954	30.57%	0.0197
dGTP	1913381	30.95%	0.0195
dTTP	1192024	19.28%	0.0123
Total	6181863	100.00%	

Figure 3.8: **Example of biomass composition determination for *Pseudomonas putida* KT 2440. A.** Chemical composition of *E. coli* adopted from [185] and utilized as a template for *P. putida* KT2440, since no extensive information is available. **B.** Protein composition in *P. putida* broken down by monomer contribution in $\frac{mmol}{g_{DW}}$. **C.** Phospholipid contributions to the biomass function where PE is phosphatidylethanolamine, PG is phosphatidylglycerol, and CL is cardiolipin. dNTP composition of the entire *P. putida* chromosomal genome. CMR - Comprehensive Microbial Resource (See Table 3.1 for the link).

to minimize the number of intracellular transport reactions to the ones that really need to be there. If too many transport reactions are added in a reconstruction, they can cause cycles, futile cycles, or Type III pathways. This is a common problem in reconstructions with multiple compartments.

CRITICAL STEP For the directionality of intracellular transport reactions, one should consider the nature of the pathway in the compartment. For instance, if the pathway is biosynthetic, it is very likely that i) the precursor(s) is only imported, ii) the product(s) of the pathway is only exported from the compartment, and iii) intermediates are not transported at all. Another problem is the mechanism of transport. Many transport reactions are in symport or antiport with either protons, cations, or other metabolites. However, not much information is available for intracellular transporters, but the mechanism used in the model may affect the predictive potential.

CRITICAL STEP If a corresponding reaction from extracellular/periplasmic space to cytoplasm is known (and is not an ABC transport reaction); one can adopt this mechanism for the intracellular transport. Otherwise assume (facilitated) diffusion reaction as mechanism. Make sure that those reactions receive a low confidence score (1 for modeling purpose) to enable easy identification if necessary.

Step 22| Draw metabolic map. (optional) If appropriate drawing software is available, the creation of organism-specific maps is very useful for gap analysis, network evaluation, and data mapping.

Step 23| Determine biomass composition. The biomass reaction accounts for all known biomass constituents and their fractional contributions to the overall cellular biomass. If detailed information is not available for the target organism, a biomass reaction from a reconstructed relative is often adapted. Ideally, the detailed biomass composition of the target organism is experimentally determined [24, 107, 120]. Note that the unit of the biomass reaction is $\frac{1}{h}$ since all biomass precursor fractions are converted to $\frac{mmol}{g_{DW}}$. Therefore, the biomass reaction sums the mole fraction of each precursor necessary to produce 1 g dry weight of cells. To formulate the biomass reaction, the following information has to be assembled:

1. **Determine the chemical composition of the cell.** Figure 3.8A shows the chemical composition of an *E. coli* cell. This information may be available in the literature.

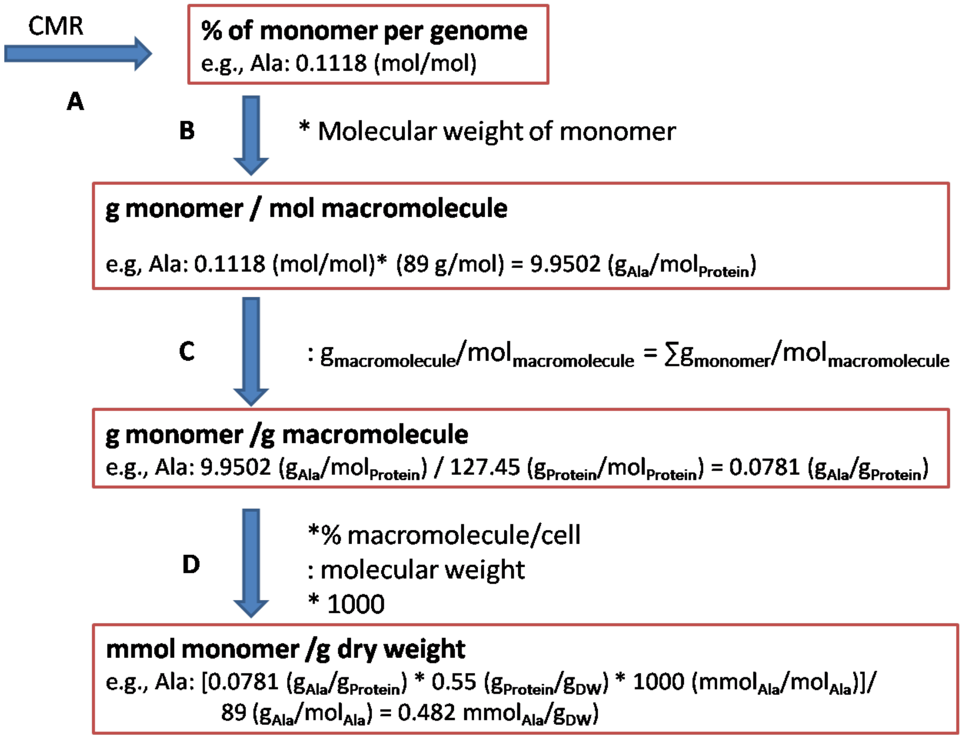
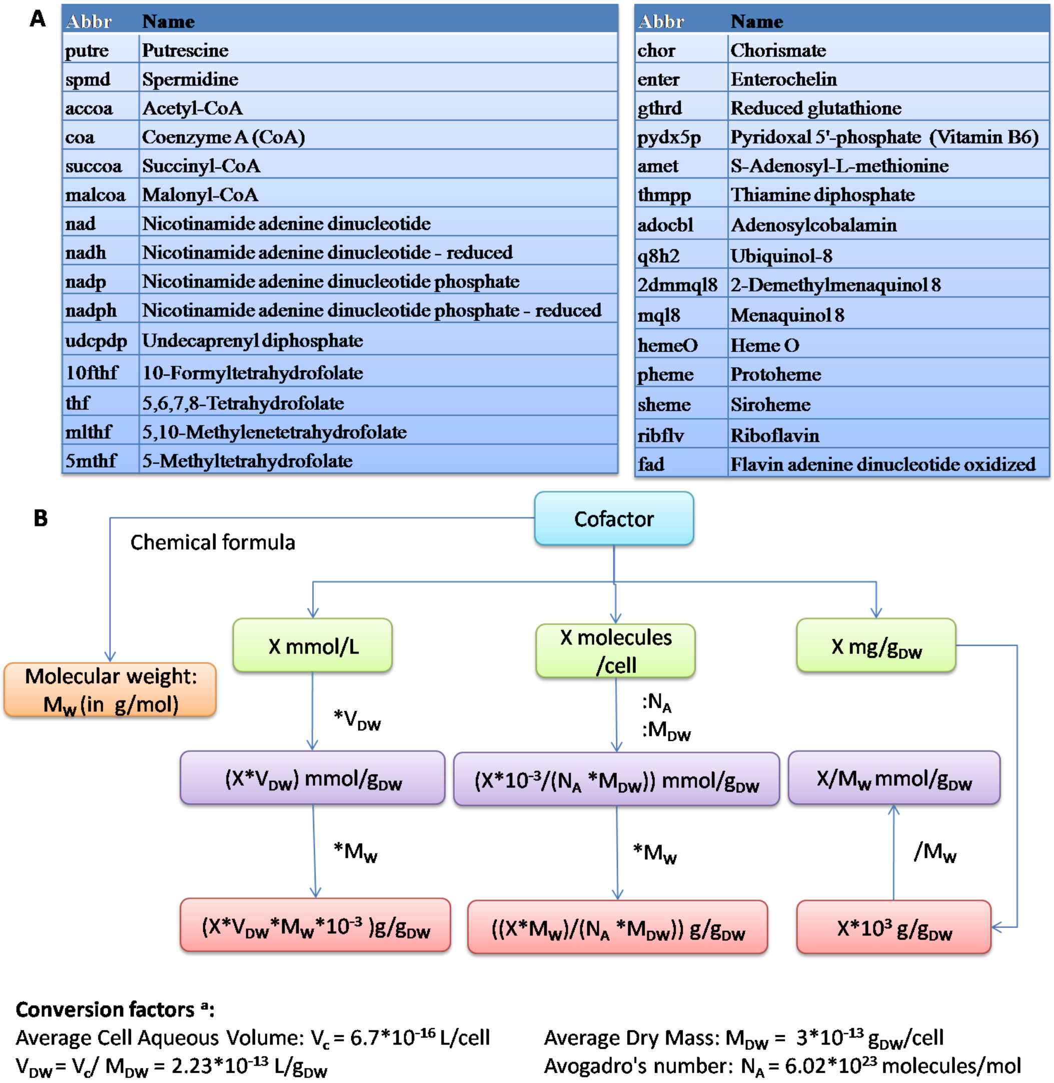

CMR
A
% of monomer per genome
e.g., Ala: 0.1118 (mol/mol)
B
* Molecular weight of monomer
g monomer / mol macromolecule
e.g, Ala: 0.1118 (mol/mol)* (89 g/mol) = 9.9502 ($g_{Ala}/mol_{Protein}$)
C
: $g_{macromolecule}/mol_{macromolecule} = \sum g_{monomer}/mol_{macromolecule}$
g monomer /g macromolecule
e.g., Ala: 9.9502 ($g_{Ala}/mol_{Protein}$) / 127.45 ($g_{Protein}/mol_{Protein}$) = 0.0781 ($g_{Ala}/g_{Protein}$)
D
*% macromolecule/cell
: molecular weight
* 1000
mmol monomer /g dry weight
e.g., Ala: [0.0781 ($g_{Ala}/g_{Protein}$) * 0.55 ($g_{Protein}/g_{DW}$) * 1000 ($mmol_{Ala}/mol_{Ala}$)]/
89 (g_{Ala}/mol_{Ala}) = 0.482 $mmol_{Ala}/g_{DW}$)

Figure 3.9: **Flow chart to calculate the fractional contribution of a precursor to the biomass reaction based on genome information obtained from the Comprehensive Microbial Resource** (CMR, see Table 3.1 for link). This approach can be used for amino acids, nucleotide triphosphates (ATP, GTP, CTP, UTP), and deoxy-nucleotide triphosphates (dATP, dGTP, dCTP, dTTP). The steps are illustrated for L-alanine, **A.** From CMR the fractional contribution of alanine to the proteome is obtained (see Figure 3.8B). **B.** To convert the molar percentage into weight of alanine per mole protein, the molar percentage is multiplied by the molecular weight of alanine. Once the weight of amino acid per mole protein is obtained for all amino acids, they are summed to obtain the weight of protein per mole protein. **C.** The weight of alanine per mole protein is converted into weight alanine per weight protein by multiplying with the sum of all amino acid's weight. **D.** Finally, the weight of alanine is multiplied by the cellular content of protein (see Figure 3.8A) and divided by its molecular weight to obtain the mole alanine per cell dry weight. Multiplying this molar contribution by a factor of 1000 will result in a final unit of mmol alanine per gram dry weight.


A
B
Chemical formula
Cofactor
Molecular weight: M_W (in g/mol)
X mmol/L
X molecules /cell
X mg/g_{DW}
$*V_{DW}$
$:N_A$
$:M_{DW}$
$(X*V_{DW})$ mmol/g_{DW}
$(X*10^{-3}/(N_A *M_{DW}))$ mmol/g_{DW}
X/M_W mmol/g_{DW}
$*M_W$
$*M_W$
$/M_W$
$(X*V_{DW}*M_W*10^{-3})$g/g_{DW}
$((X*M_W)/(N_A *M_{DW}))$ g/g_{DW}
$X*10^3$ g/g_{DW}

A

Abbr	Name
putre	Putrescine
spmd	Spermidine
accoa	Acetyl-CoA
coa	Coenzyme A (CoA)
succoa	Succinyl-CoA
malcoa	Malonyl-CoA
nad	Nicotinamide adenine dinucleotide
nadh	Nicotinamide adenine dinucleotide - reduced
nadp	Nicotinamide adenine dinucleotide phosphate
nadph	Nicotinamide adenine dinucleotide phosphate - reduced
udcpdp	Undecaprenyl diphosphate
10fthf	10-Formyltetrahydrofolate
thf	5,6,7,8-Tetrahydrofolate
mlthf	5,10-Methylenetetrahydrofolate
5mthf	5-Methyltetrahydrofolate

Abbr	Name
chor	Chorismate
enter	Enterochelin
gthrd	Reduced glutathione
pydx5p	Pyridoxal 5'-phosphate (Vitamin B6)
amet	S-Adenosyl-L-methionine
thmpp	Thiamine diphosphate
adocbl	Adenosylcobalamin
q8h2	Ubiquinol-8
2dmmql8	2-Demethylmenaquinol 8
mql8	Menaquinol 8
hemeO	Heme O
pheme	Protoheme
sheme	Siroheme
ribflv	Riboflavin
fad	Flavin adenine dinucleotide oxidized

Conversion factors [a]**:**

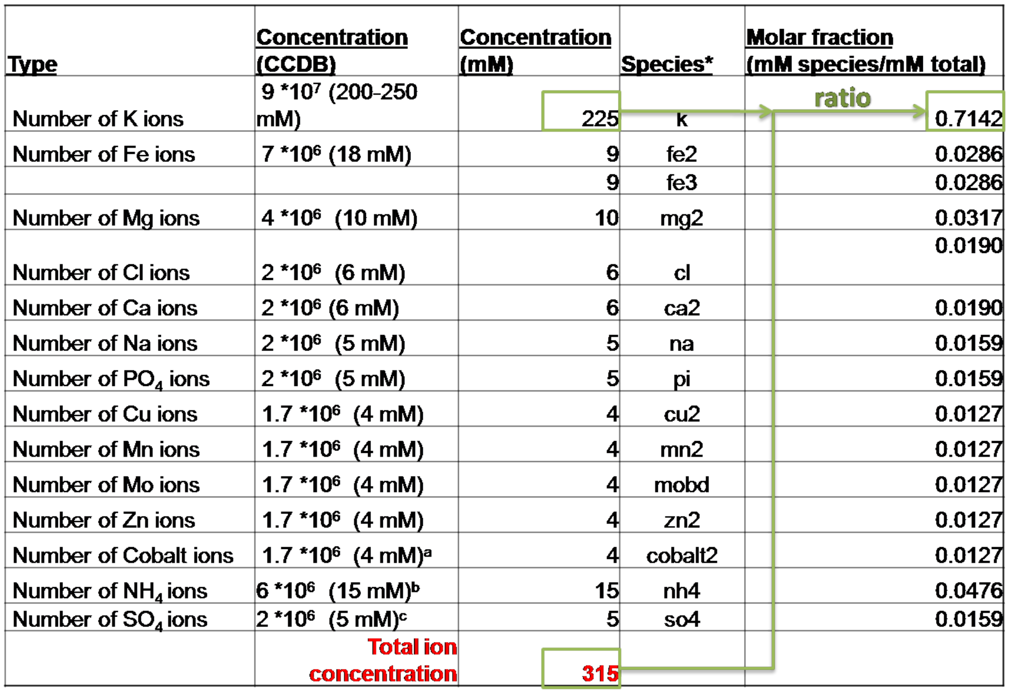
Average Cell Aqueous Volume: $V_c = 6.7*10^{-16}$ L/cell

$V_{DW} = V_c / M_{DW} = 2.23*10^{-13}$ L/g_{DW}

Average Dry Mass: $M_{DW} = 3*10^{-13}$ g_{DW}/cell

Avogadro's number: $N_A = 6.02*10^{23}$ molecules/mol

Figure 3.10: **Determination of the content of soluble pool.** **A.** The soluble pool can contain numerous polyamines, vitamins and cofactors. This list represents the metabolites considered by the *E. coli* reconstruction [78]. **B.** Depending on the available information from literature, measurements or database entries the conversion into $\frac{mmol}{g_{DW}}$ and $\frac{g}{g_{DW}}$ is shown. The value in the purple box corresponds to the stoichiometric coefficient in the biomass reactions for the precursor. [a] Information was obtained from Cybercell Database (CCDB, see Table 3.1 for link).



Type	Concentration (CCDB)	Concentration (mM)	Species*	Molar fraction (mM species/mM total)
Number of K ions	9 *10^7 (200-250 mM)	225	k	0.7142
Number of Fe ions	7 *10^6 (18 mM)	9	fe2	0.0286
		9	fe3	0.0286
Number of Mg ions	4 *10^6 (10 mM)	10	mg2	0.0317
Number of Cl ions	2 *10^6 (6 mM)	6	cl	0.0190
Number of Ca ions	2 *10^6 (6 mM)	6	ca2	0.0190
Number of Na ions	2 *10^6 (5 mM)	5	na	0.0159
Number of PO_4 ions	2 *10^6 (5 mM)	5	pi	0.0159
Number of Cu ions	1.7 *10^6 (4 mM)	4	cu2	0.0127
Number of Mn ions	1.7 *10^6 (4 mM)	4	mn2	0.0127
Number of Mo ions	1.7 *10^6 (4 mM)	4	mobd	0.0127
Number of Zn ions	1.7 *10^6 (4 mM)	4	zn2	0.0127
Number of Cobalt ions	1.7 *10^6 (4 mM)[a]	4	cobalt2	0.0127
Number of NH_4 ions	6 *10^6 (15 mM)[b]	15	nh4	0.0476
Number of SO_4 ions	2 *10^6 (5 mM)[c]	5	so4	0.0159
	Total ion concentration	315		

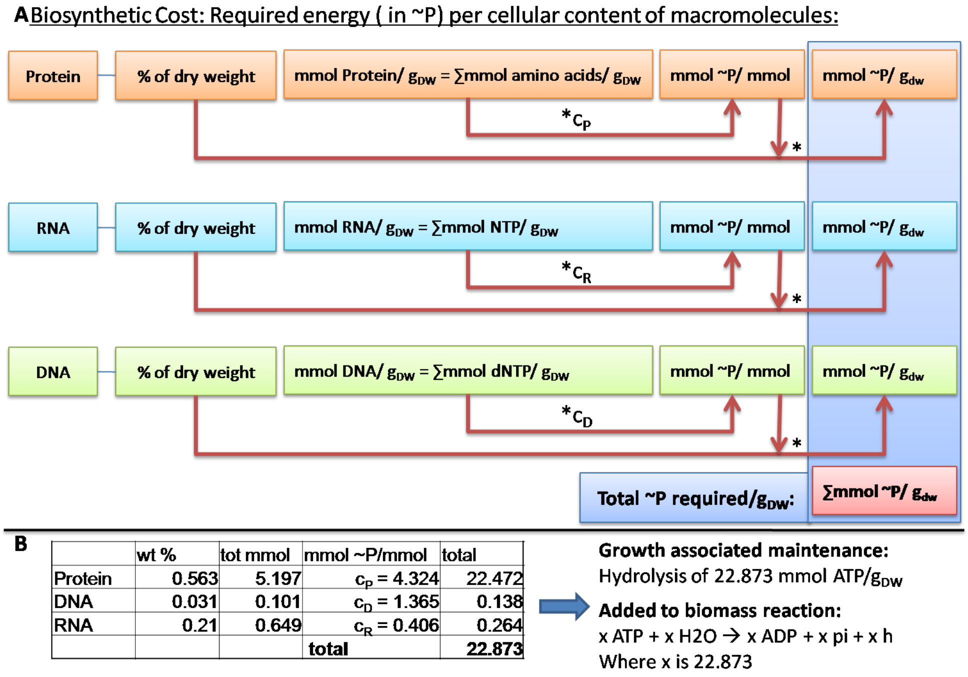
Figure 3.11: **Calculation of biomass coefficient of ions, many of which are necessary for structure and/or catalytic activity of enzymes.** * Ion abbreviation in *E. coli* reconstruction. [a] Assumed to be the same as most other metals. *b* Assumed to be second most abundant cation (based on Neidhardt *et al.* [185]). [c] Assumed to be the same as phosphate (PO_4). CyberCell Database (CCDB, see Table 3.1 for the link).

2. **Determine the amino acid content.** This step assumes that there are no direct measurements available. Therefore, organism-specific information can be gathered from Comprehensive Microbial Resource (CMR), for example, (Table 3.1). The amino acid content can be determined by selecting the Genome Tools tab, followed by Analysis Tools, and finally Codon Usage. Using the molar percentage and molecular weight of each amino acid, one can calculate the weight per mol protein. Summing the individual amino acid values gives a total molecular weight of the protein content. Subsequently, one can calculate the weight percent per amino acid. The calculated weight percent is then multiplied by the cellular content percentage of the macromolecule and divided by the molecular weight of the individual monomer (Figure 3.8B, Figure 3.9).

3. **Determine the nucleotide content.** Next, the nucleotide composition of the cell can be determined using a similar approach (Figure 3.9). From the aforementioned Genome Tools tab, Summary Information was selected, followed by DNA Molecule Info. The number of each dNTP (i.e., dATP, dCTP, dGTP, and dTTP) present in the genome is listed on the summary page, and the resulting composition, based on the same calculations that were previously performed for each amino acid (Figure 3.8D). In order to determine the RNA composition of the cell, the codon usage that was accessed in the amino acid step can be utilized. Remember that RNA incorporates uracil instead of thymine, therefore, the codon usage needs to be read with every T replaced by a U. Tabulating the frequency of each RNA monomer and following the calculations outlined above results in the determination of the biomass coefficients for RNA contribution. (Note that ribosomal ATP contribution will be combined with energy and maintenance requirements of the cell in the final biomass function to give the proper coefficient for ATP).

4. **Determine the lipid content.** The lipid composition of the cell is slightly more complicated to calculate because it includes contributions from both fatty acids and phospholipids. First, the average molecular weight of a fatty acid in the cell needs to be determined by incorporating the average fatty acid composition of the cell (requires experimental data from literature). Using the average molecular weight of each fatty acid and summing the weight contributions of each, the average molecular weight for a fatty acid chain can be determined. This weight can then be used to calculate the average molecular weight of various lipids within the cell. Such computation is

performed by summing the molecular weight of the core structure of the molecule and the molecular weight of the fatty acids attached to the core structure based on the average molecular weight of one fatty acid that was determined above. The molar percentages of the three major phospholipids, phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and cardiolipin (CL), may be found in the literature. Thus, the phospholipid contributions to the biomass function can be then determined (Figure 3.8C).


A
Biosynthetic Cost: Required energy (in ~P) per cellular content of macromolecules:
Protein
% of dry weight
mmol Protein/ g_{DW} = Σmmol amino acids/ g_{DW}
mmol ~P/ mmol
mmol ~P/ g_{dw}
*c_P
*
RNA
% of dry weight
mmol RNA/ g_{DW} = Σmmol NTP/ g_{DW}
mmol ~P/ mmol
mmol ~P/ g_{dw}
*c_R
*
DNA
% of dry weight
mmol DNA/ g_{DW} = Σmmol dNTP/ g_{DW}
mmol ~P/ mmol
mmol ~P/ g_{dw}
*c_D
*
Total ~P required/g_{DW}:
Σmmol ~P/ g_{dw}
B

	wt %	tot mmol	mmol ~P/mmol	total
Protein	0.563	5.197	c_P = 4.324	22.472
DNA	0.031	0.101	c_D = 1.365	0.138
RNA	0.21	0.649	c_R = 0.406	0.264
			total	22.873

Growth associated maintenance:
Hydrolysis of 22.873 mmol ATP/g_{DW}
Added to biomass reaction:
x ATP + x H2O → x ADP + x pi + x h
Where x is 22.873

Figure 3.12: **Growth-associated maintenance cost. A.** Calculation of growth-associated maintenance cost. **B.** Sample calculation for *E. coli.* Adapted from Feist *et al.* [78]. The energy necessary for the synthesis of the macromolecules from the building blocks were obtained from Table 4 - 6 of Chapter 3 in Neidhardt *et al.* [185]. The coefficient c_P, c_D, c_R were calculating the total energy necessary for the macromolecules divided by the total number of building blocks (See also Neidhardt *et al.* [185] and Feist *et al.* [78] for more details).

5. **Determine the content of the soluble pool** (polyamines and vitamins and cofactors). The soluble pool contains, for example, spermidine, coenzyme A, and folic acid. Figure 3.10 lists the composition of the soluble pool in the *E. coli* metabolic network and how their fractional distributions to the biomass reaction were calculated.

6. **Determine the ion content.** The calculation of the molar fraction of the ions

is illustrated in Figure 3.11. It assumes that concentration data are available or can be estimated for each ion. Information about the ion content can be obtained from different resources, including primary literature and databases. In the case of the *E. coli* metabolic reconstruction, the ion concentration was obtained from the CyberCell Database. The reported concentration (c_i) for each ion species i, needs to be converted into mM. All ion species should be added (total ion concentration, c_{total}). The molar fraction (f_i) of each ion species i is then calculated by dividing c_i with c_{total}:
$f_i = \frac{c_i}{c_{total}}$ where $c_{total} = \sum c_i$.

7. **Growth associated maintenance.** The energy required for macromolecular synthesis, e.g., proteins, must be also accounted for in the biomass reaction. Therefore, the total amount (mmol) of macromolecule (Protein, DNA, and RNA) is determined using the information compiled above. Neidhardt *et al.* [185] lists the amount of phosphate bonds necessary to synthesize a macromolecule which is then multiplied with the total amount of phosphate bonds necessary (Figure 3.12). These phosphate bonds are accounted for by adding ATP hydrolysis to the biomass reaction (x ATP + x $H_2O \rightarrow$ x ADP + x P_i + x H^+, where x is the number of required phosphate bonds). Additionally, the *E. coli* biomass reaction accounted for unknown growth-associated maintenance cost based on experimental data, based on chemostat growth data (see Feist *et al.* [78] for details).

CAUTION The composition of the biomass reaction plays an important role for *in silico* gene deletion experiments. If a biomass precursor is not accounted for in the biomass reactions, the synthesis reactions may not be required for growth (i.e., they are non-essential). Consequently, the associated genes may not be essential either. This means that the presence or absence of metabolites in the biomass reaction may affect the *in silico* essentiality of reactions and their associated gene(s). In contrast, the fractional contribution of each precursor plays a minor role for gene and reaction essentiality studies. When one wishes to predict the optimal growth rate accurately, the fractional distribution of each compound may play an important role.

Step 24| Add biomass reaction to the reconstruction. In **Step 23**, all biomass precursors and their fractional contribution to the overall cell composition was determined and calculated in $\frac{mmol}{g_{DW}}$ (Figure 3.13). In this step, all precursors are assembled in one single reaction - the biomass reaction - which is then added to the reaction list of the reconstruction.

CAUTION Note that some metabolites might be produced. For instance, in the *E. coli* biomass reaction, proton (H^+), orthophosphate (P_i) and some other metabolites are produced [78]. These metabolites originate mainly from the growth associated ATP hydrolysis (see above).

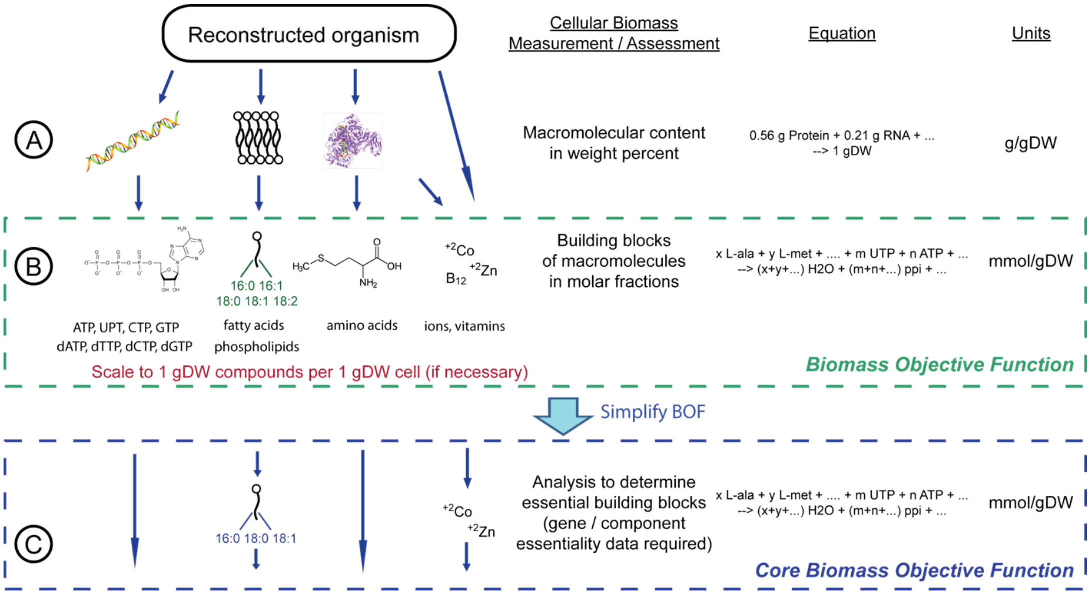

Reconstructed organism
Cellular Biomass Measurement / Assessment
Equation
Units
A
Macromolecular content in weight percent
0.56 g Protein + 0.21 g RNA + ... --> 1 gDW
g/gDW
B
ATP, UPT, CTP, GTP dATP, dTTP, dCTP, dGTP
16:0 16:1 18:0 18:1 18:2
fatty acids phospholipids
amino acids
+2Co +2Zn B12
ions, vitamins
Building blocks of macromolecules in molar fractions
x L-ala + y L-met + ... + m UTP + n ATP + ... --> (x+y+...) H2O + (m+n+...) ppi + ...
mmol/gDW
Scale to 1 gDW compounds per 1 gDW cell (if necessary)
Biomass Objective Function
Simplify BOF
C
16:0 18:0 18:1
+2Co +2Zn
Analysis to determine essential building blocks (gene / component essentiality data required)
x L-ala + y L-met + ... + m UTP + n ATP + ... --> (x+y+...) H2O + (m+n+...) ppi + ...
mmol/gDW
Core Biomass Objective Function

Figure 3.13: **Schematic representation of the assembly of the biomass reaction based on the information retrieved in Step 23**. **A.** the weight percent of the macromolecules is obtained from experiments or literature. The macromolecules are DNA, RNA, cell wall, proteins, ions and cofactors. **B.** Once all the biomass precursors and their fractional contribution to the dry weight is determined, they are assembled into one reaction. The fractions may be scaled such that the biomass compounds add up to 1 g_{DW} per 1 g_{DW} cell. **C.** If desired, the biomass objective function can be simplified to account only for core building blocks, while condition - specific building blocks (e.g., antigenes) are removed. Adapted from [79].

Step 25| Add non-growth associated ATP maintenance reaction (ATPM). More recent reconstructions include an ATP hydrolysis reaction ($1\ ATP + 1\ H_2O \rightarrow 1\ ADP + 1\ P_i + 1\ H^+$), which represents non-growth associated ATP requirements of the cell to maintain, for example, Turgor pressure [78]. The value for the reaction rate can be estimated from growth experiments. Based on such measurements, the reaction flux rate was constrained to 8.39 $\frac{mmol}{g_{DW} \cdot h}$ in the *E. coli* metabolic model [78].

CAUTION An unconstrained ATPM reaction can change the model prediction in some cases. For example, if the computed growth rate of the model is too high, check the flux value through the ATPM in the optimal solution (see also **Step 50**)

Step 26| Add demand reactions to the reconstruction. Demand reactions are unbalanced network reactions that allow the accumulation of a compound, which otherwise is not allowed in steady-state models due to the mass-balancing requirement (i.e., in steady state the sum of influx equals the sum of efflux for each metabolite). Most of the demand reactions will be added in the gap filling process (**Steps 34 to 37**). At this stage, demand functions should only be added for compounds that are known to be produce by the organism, e.g., certain cofactors, lipopolysaccharide, and antigens, but i) for which information is available about their fractional distribution to the biomass or ii) which may be only produced in some environmental conditions.

Step 27| Add sink reactions to the reconstruction. Sink reactions are similar to demand reactions but are defined to be reversible and thus provide the network with metabolites (see Figure 3.7 for examples). These sink reactions are of great use for compounds that are produced by non-metabolic cellular processes but need to be metabolized. **CAUTION** Adding too many sink reactions may enable the model to grow without any resources in the medium. Therefore, sink reactions have to be added with care.

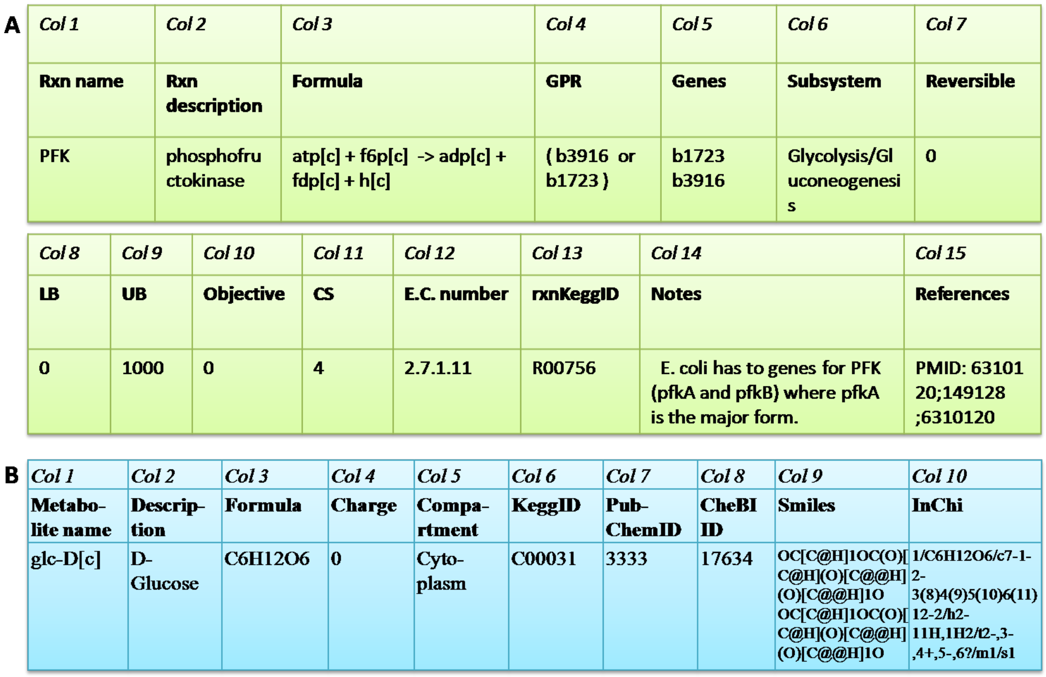
Step 28| Determine growth medium requirements. Information about growth-enabling media is of great help in the following two sections. Thus, if possible, they should be collected prior to the conversion and debugging stage. The following information should be collected: 1) Which metabolites are present? 2) Are there any auxotrophies? 3) Define the composition of a base medium, e.g., water, protons, ions, etc. 4) Obtain information about rich medium composition. This data will be crucial for simulations and network evaluations. If uptake or secretion rates are available, then they should be documented as well. While this step is easy for the experimentalist, researchers which cannot grow the target organism have to identify the growth requirements from literature. In some cases, research studies describe minimal, defined, or rich medium compositions. In other cases, the culturing conditions reported in some experimental study must be sufficient.

Table 3.7: Useful functions in the COBRA Toolbox for reconstruction. $^{s}*$ For details on syntax, please refer to the COBRA Toolbox and [21].

Action	**COBRA Toolbox command***	**Comments**
Add reaction	model = addReaction(model,'newRxn1','A + 2 B −− > C')	Adds reactions such as: A + 2 B → C
Add demand reaction	[model,rxnNames] = addDemandReaction(model, metaboliteNameList)	Adds reactions such as: A →
Add sink reaction	[model] = AddSinkReactions(model, metabolites, bounds)	Adds reactions such as: A ⇔
Remove reaction	modelOut = removeRxns(model, rxnRemoveList)	
Write model in file	writeCbModelRecon(model, format, fileName)	File format can be sbml, plain text, xls
Open sbml model	model = readCbModel(fileName)	
Open reconstruction from xls	model = xls2model(RxnFileName, MetFileName)	
Change gene association	model = changeGeneAssociation(model, rxnName, grRule, geneNameList, systNameList)	
Change reaction bounds	model = changeRxnBounds(model, rxnNameList, value, boundType)	
Change objective function	model = changeObjective(model, rxnNameList, objectiveCoeff)	
Print constraint reactions	PrintConstraints(model, MinInf, MaxInf)	Prints all reactions with upper bounds (lower) less (greater) than MaxInf (MinInf)
Load reconstruction into Matlab	model = xls2model(RxnFileName, MetFileName);	

3.4.3 Conversion from reconstruction to mathematical model

Up to here we collected the information in a spreadsheet. The next steps involve converting the reaction list (table) into a mathematical format (matrix). We will use Matlab for that.



A

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*
Rxn name	**Rxn description**	**Formula**	**GPR**	**Genes**	**Subsystem**	**Reversible**
PFK	phosphofructokinase	atp[c] + f6p[c] -> adp[c] + fdp[c] + h[c]	(b3916 or b1723)	b1723 b3916	Glycolysis/Gluconeogenesis	0

Col 8	*Col 9*	*Col 10*	*Col 11*	*Col 12*	*Col 13*	*Col 14*	*Col 15*
LB	**UB**	**Objective**	**CS**	**E.C. number**	**rxnKeggID**	**Notes**	**References**
0	1000	0	4	2.7.1.11	R00756	E. coli has to genes for PFK (pfkA and pfkB) where pfkA is the major form.	PMID: 6310120;149128;6310120

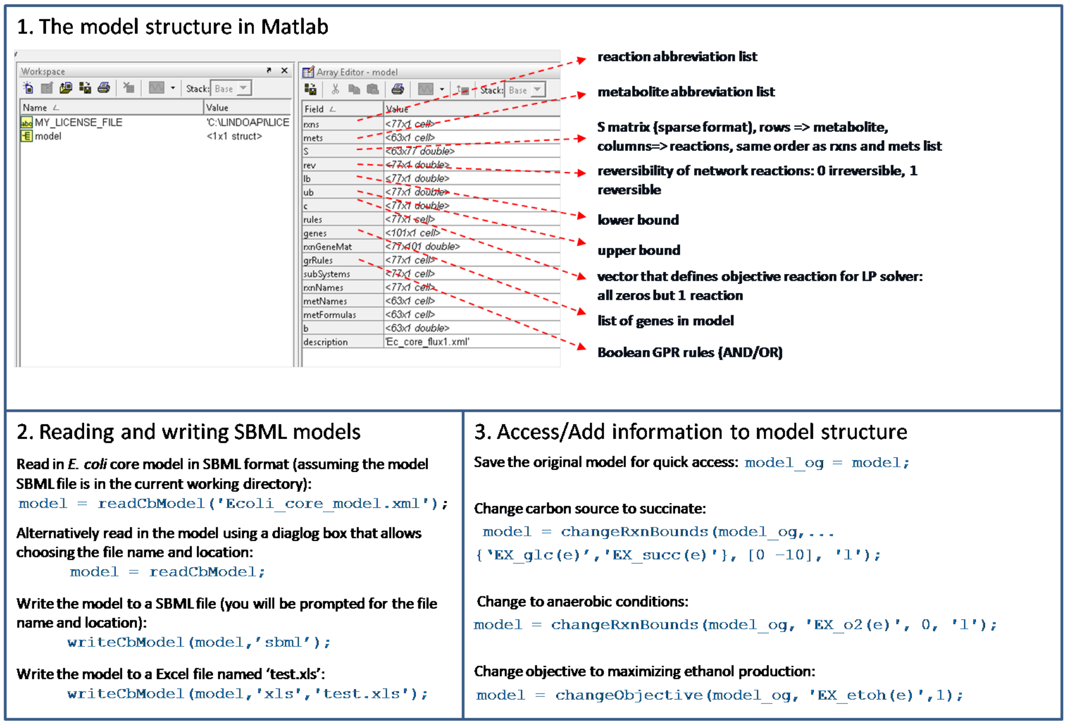
B

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*	*Col 8*	*Col 9*	*Col 10*
Metabolite name	**Description**	**Formula**	**Charge**	**Compartment**	**KeggID**	**PubChemID**	**CheBI ID**	**Smiles**	**InChi**
glc-D[c]	D-Glucose	C6H12O6	0	Cytoplasm	C00031	3333	17634	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1

Figure 3.14: **Layout of excel sheets as input for xls2model function. A.** Reaction file. **B.** Metabolite file. At least one identifier from Column 6 to 8 should be included. Column 9 and 10 are optional.

Step 29| Initialize the COBRA Toolbox. Install Matlab, the required Toolboxes (SBML Toolbox and COBRA Toolbox), and an LP solver [21]. Start Matlab as described in the installation instructions. Within Matlab, move to the directory where the COBRA Toolbox was installed. Initiate the COBRA Toolbox by entering the command initCobraToolbox in the Matlab command line. Note that the default LP solver can be changed by editing the initCobraToolbox script or at any time during a Matlab session by using the changeCobraSolver function included in the Toolbox. A list of frequently used COBRA Toolbox functions is given in Table 3.7. See also the Nature protocol on the COBRA Toolbox for details on initializing and using the Toolbox [21]. Furthermore, the supplemental material contains a Matlab primer, which aims to facilitate the use of Matlab for novices.

TROUBLESHOOTING



1. The model structure in Matlab

Workspace

Name	Value
MY_LICENSE_FILE	'C:\LINDOAPI\LICE
model	<1x1 struct>

Array Editor - model

Field	Value
rxns	<77x1 cell>
mets	<63x1 cell>
S	<63x77 double>
rev	<77x1 double>
lb	<77x1 double>
ub	<77x1 double>
c	<77x1 double>
rules	<77x1 cell>
genes	<101x1 cell>
rxnGeneMat	<77x101 double>
grRules	<77x1 cell>
subSystems	<77x1 cell>
rxnNames	<77x1 cell>
metNames	<63x1 cell>
metFormulas	<63x1 cell>
b	<63x1 double>
description	'Ec_core_flux1.xml'

reaction abbreviation list

metabolite abbreviation list

S matrix (sparse format), rows => metabolite, columns=> reactions, same order as rxns and mets list

reversibility of network reactions: 0 irreversible, 1 reversible

lower bound

upper bound

vector that defines objective reaction for LP solver: all zeros but 1 reaction

list of genes in model

Boolean GPR rules (AND/OR)

2. Reading and writing SBML models

Read in *E. coli* core model in SBML format (assuming the model SBML file is in the current working directory):

```
model = readCbModel('Ecoli_core_model.xml');
```

Alternatively read in the model using a diaglog box that allows choosing the file name and location:

```
        model = readCbModel;
```

Write the model to a SBML file (you will be prompted for the file name and location):

```
        writeCbModel(model,'sbml');
```

Write the model to a Excel file named 'test.xls':

```
        writeCbModel(model,'xls','test.xls');
```

3. Access/Add information to model structure

Save the original model for quick access: `model_og = model;`

Change carbon source to succinate:

```
 model = changeRxnBounds(model_og,...
{'EX_glc(e)','EX_succ(e)'}, [0 -10], 'l');
```

Change to anaerobic conditions:

```
model = changeRxnBounds(model_og, 'EX_o2(e)', 0, 'l');
```

Change objective to maximizing ethanol production:

```
model = changeObjective(model_og, 'EX_etoh(e)',1);
```

Figure 3.15: **Model in Matlab format and the COBRA Toolbox.**

Step 30| Load reconstruction into Matlab. Save the reaction list in an excel sheet with the same order of columns as shown in Figure 3.14 ('RxnFileName'). A second file containing metabolite information needs to be saved as well ('MetFileName'). See supplemental material for an example of the input files. The following COBRA Toolbox function should be used to read the reconstruction into Matlab:

>> model = xls2model(RxnFileName,MetFileName);

The loaded metabolic model is stored in a structure named 'model' in Matlab. This structure contains all the information about the reconstruction in the different fields of the structure. Figure 3.15 provides a description of the individual fields and their content.

TROUBLESHOOTING

Step 31| Set objective function. The following COBRA Toolbox function is used to set the objective function of the model:

>> model = changeObjective(model, rxnNameList, objectiveCoeff)

The reaction(s) that should be set as the objective function is given by 'rxnNameList'.

They will receive a corresponding coefficient ’objectiveCoeff’. This means that a single reaction or a linear combination of multiple reactions can be chosen as objective function. **CAUTION** The COBRA Toolbox is set up in a way that the coefficient(s) has to be a positive number. When minimizing, the input option to the COBRA toolbox function optimizeCBmodel.m can be set to ’min’. The default option of the ’optimizeCBmodel’ function is maximizing (’max’) (see Table 3.7).

Step 32| Set simulation constraints. Use the following function to set the constraints of the model:

```
>> model = changeRxnBounds(model,rxnNameList,value,boundType)
```

The list of reactions for which the bounds should be changed is given by ’rxnNameList’, while an array contains the new boundary reaction rates (’value’). The type of bound can be set to lower bound (’l’), upper bound (’u’). Alternatively, both bounds can be changed (’b’).

CAUTION Using the functions in the COBRA Toolbox, it is very easy to change reaction constraints but it is sometimes difficult to keep track of all the changes. In fact, one of the most common reasons for errors in simulation is that reaction constraints are not correctly set. Therefore, it is important to have an expectation of the results before running a simulation to avoid erroneous conclusions. It is recommended that the constraints are checked by copying the model reaction abbreviations AND lower and upper bounds into excel. For most models, this is the easiest way to see where problems are with the constraints. Similarly, copying calculated solution(s) into Excel may be of help. The COBRA Toolbox has a function that lists all constrained reactions that are greater than a minimal value (’MinInf’) and smaller than a maximal value (’MaxInf’):

```
>> PrintConstraints(model, MinInf, MaxInf)
```

Additionally, there is a function available that lists all reactions and their flux values in a solution:

```
>> printFluxVector(model,fluxData)
```

3.4.4 Network evaluation = ’Debugging mode’

Step 33| Test if network is mass- and charge balanced. In **Step 9**, the reaction stoichiometry was determined for each network reaction. Here, we will test the correctness and consistency of the reaction stoichiometry by verifying the mass- and charge-balancing. The Toolbox function CheckMassChargeBalance can be used to determine

stoichiometrically unbalanced reactions. All, or a subset, of the network reactions can be given as input ('RxnList') along with the model structure ('model'):
>> [UnbalancedRxns] = CheckMassChargeBalance(model,RxnList)
In case there are unbalanced reactions, the script returns a structure containing the name of the unbalanced reaction and which elements are unbalanced (UnbalancedRxns). Looking at the reaction equations and the charged formula for each metabolite will help to balance the reactions.
Normally there are two common errors causing unbalanced reactions:

1. Missing proton and/or water.
   (a) If a proton as substrate is missing, a proton donor may be necessary (e.g., NADH, NADPH). This will require a literature search to identify a candidate proton donor.
   (b) If a water molecule is missing, keep in mind that after adding water to the equation the proton and oxygen will need to be balanced.
2. Stoichiometric coefficient of at least one metabolite is wrong. Repeat **Step 9**.

Also refer to the information provided in **Step 9** and Figure 3.2 for mass- and charge-balancing of network reactions.
CAUTION A few network reactions are always unbalanced. These reactions include the biomass reaction, demand, sink and, exchange reactions.

Step 34| Identify metabolic dead-ends. At this point, the first iteration of manual curated reconstruction is finished. It is expected that the network contains a significant number of gaps, i.e., missing reactions and functions. We recommend performing a first gap analysis at this stage of the reconstruction process as it will ease the subsequent computation and reduce the number of "bugs" in the model. Use
>> [Gaps] = AnalyzeGaps(model)
to identify gaps. The function will return a list of all metabolites ('Gaps') that are only produced ('Product') or consumed ('Substrate') in the network. Copy this gap list into an excel sheet where information and references can be easily added for each dead-end metabolite.

Step 35| Identify candidate reactions to fill gaps. This step will require a literature search and may include re-annotation of a genome to find candidate genes and reactions to fill the gap (see Table 3.1 and 8 for some example tools). Use KEGG [130] maps, biochemical textbooks, or other available biochemical maps to identify the metabolic 'environment' of the dead-end metabolite. If the genome annotation of the target organism is present in KEGG [130], one can highlight the dead-end metabolite on the map. This may give an indication of which enzyme(s) may be able to produce or synthesize the dead-end metabolite and thus provide a good starting point for literature and/or genome search.

CRITICAL STEP Gap-filling is a tricky business. In some cases, a gap should be filled to ensure that the model is functional, i.e., biomass precursor synthesis or a certain physiological function can be simulated. In other cases, filling a gap may enable the model to perform a function that the organism is not able to do (see Figure 3.18 for some examples). In general, if no information supports the existence of a particular gap reaction, the gap should only be filled if it is required for the model's functionality. In such cases, the confidence score should be set to 1, which corresponds to "modeling purpose" only, and allows retrieving these low confidence reactions readily, if desired.

CAUTION In **Step 13**, we highlighted that enzymes which are listed in biochemical databases to catalyze multiple reactions should be included in the reconstruction with care and that it should be noted if evidence for all of the reactions could be found. Some of the identified dead-end metabolites will originate from such secondary reactions of these "multitasking" enzymes. Closing these gaps may affect the predictive potential of the reconstruction, therefore, only gaps should be filled which are required for network functionality (e.g., biomass precursor synthesis) or which have supporting data

Step 36| Add gap reactions to the reconstruction. If experimental and/or annotation data support gap reactions or they are needed for modeling purposes, the reaction(s) should be added to the reconstruction by repeating **Steps 6 through 16**.

CAUTION Keep in mind that adding new reactions to the network may cause new gaps. Therefore, when adding reactions you should make sure that all metabolites are connected to the network.

Step 37| Add notes and references to dead-end metabolites. Each dead-end metabolite should be documented. For those dead-ends that remain, the collected information and references should be added to the reconstruction information. In the simplest case, the note should distinguish between knowledge and scope gap (Figure 3.18).

CRITICAL STEP The more detailed and carefully the gap filling steps are done (**Steps**

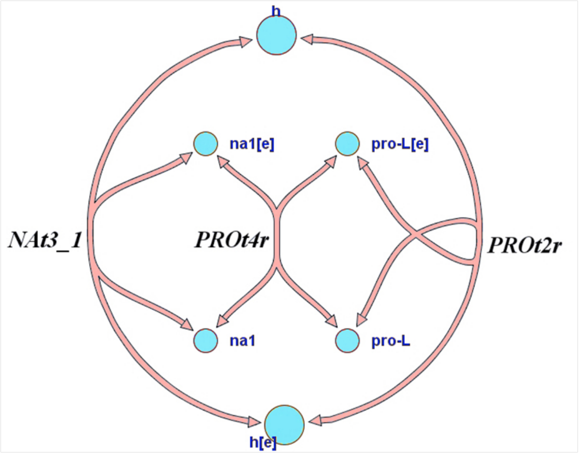


Figure 3.16: Example of a stoichiometrically balanced cycle (SBCs) found in the *Helicobacter pylori* reconstruction [280]. More complex SBCs can be found elsewhere [206, 220]. h - proton, na1 - sodium, pro-L - L-proline, [e] - extracellular. PROt2r - L-proline reversible transport via proton symport, PROt4r - proline transporter, NAt3_1 - sodium proton antiporter (H:NA is 1:1).

34 to 36) the easier and faster the debugging process will be.

Step 38| Add missing exchange reactions to model. The gap filling process may have resulted in the inclusion of further transport reactions. Exchange reactions thus need to be added to the reconstruction. Repeat **Step 20**.

The reconstruction content was evaluated so far without doing any simulations. The next steps will involve testing the model to ensure i) functionality and ii) comparable properties with the target organism.

Step 39| Set exchange constraints for a simulation condition. Determine an environmental condition in which most network evaluation tests should be carried out initially ('standard condition'). In *E. coli*, this condition could be minimal medium (M9) supplemental with glucose while oxygen is present. Use

```
>> model = changeRxnBounds(model,rxnNameList,value,boundType)
```

to set the constraints. Reactions whose bounds should be changed are listed in 'rxnNameList'. The new value for each reaction is contained in the array 'value'. Finally, the type of constraint has to be defined in the list 'boundType'. The possible types are: 'l' for lower bound, 'u' for upper bound, and 'b' if both reaction bounds should be set to the specified value.

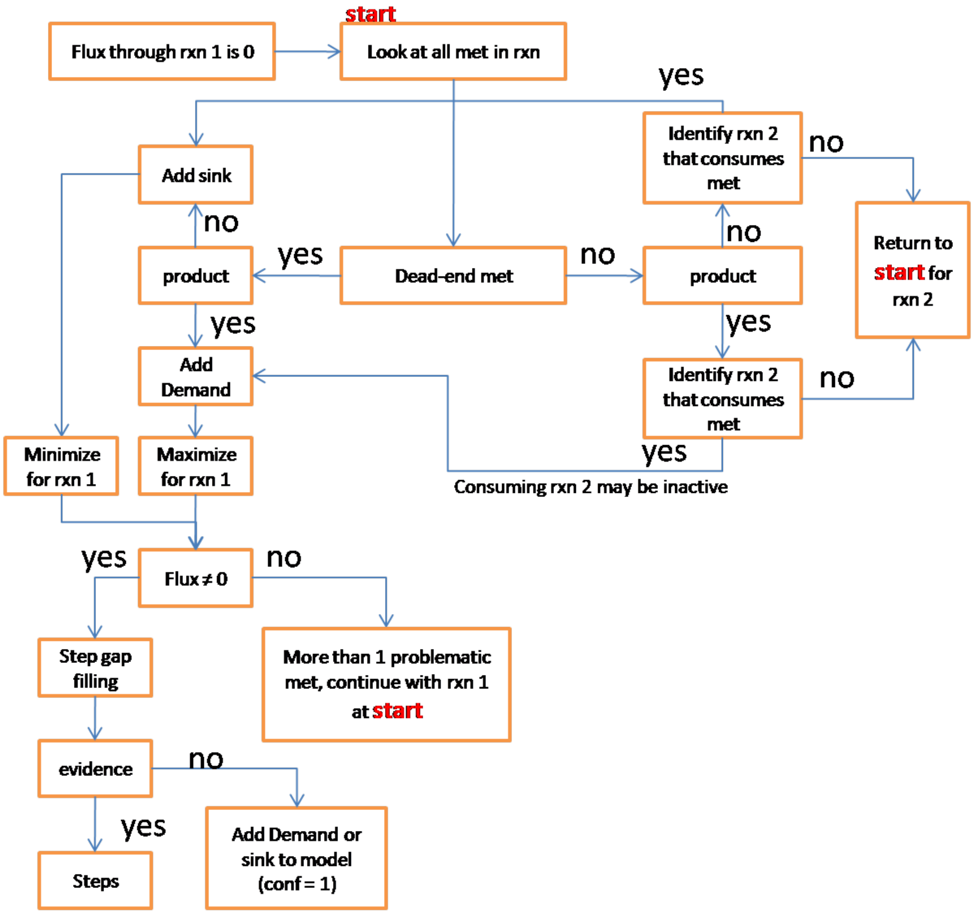

start
Flux through rxn 1 is 0
Look at all met in rxn
yes
Add sink
Identify rxn 2 that consumes met
no
Return to start for rxn 2
no
no
product
yes
Dead-end met
no
product
yes
yes
Add Demand
Identify rxn 2 that consumes met
no
Minimize for rxn 1
Maximize for rxn 1
yes
Consuming rxn 2 may be inactive
yes
Flux ≠ 0
no
Step gap filling
More than 1 problematic met, continue with rxn 1 at start
evidence
no
yes
Steps
Add Demand or sink to model (conf = 1)

Figure 3.17: **Flow chart on debugging network reactions that cannot carry flux.** 'rxn ' stands for reaction. 'conf' stands for confidence score. 'met' stands for metabolite.

Step 40| Test for stoichiometrically balanced cycles (optional, Figure 3.16). Stoichiometrically balanced cycles (SBCs), or Type III extreme pathways, are formed by internal network reactions and can carry fluxes despite closed exchange reactions (closed system). These SBCs are artifacts of metabolic reconstructions due to insufficient constraints (e.g., thermodynamic constraints and regulatory constraints). Recent efforts have concentrated on dealing with these SBCs ([220]). Note that SBCs are not futile cycles. Here, we will mainly try to identify, and in some cases eliminate, the SBCs, as they do not directly affect the model's predictions.

1. **Test for Type III pathways.** Therefore, use the following function:
   >>TestForTypeIIIPathways(model,ListExch);
   A list of indices of the exchange reactions in the S matrix ('ListExch') has to be provided to the function. These exchange reactions will be set to zero and then the flux variability of the closed model is calculated. This function requires that X3.exe is in the working directory. The function will return if there are Type III pathways in the model.

2. **Output if Type III pathways found.** If Type III pathways have been identified, there are two output files: one file ('ModelTestTypeIII_myT3.txt') has all Type III pathways as a matrix, where the rows are the different pathways and the columns correspond to the network reaction (in the same order as given in 'ModelTestTypeIII_myRxnMet.txt'). Note that the extreme pathway package converts network reactions into elementary reactions (i.e., irreversible reactions). A second file ('ModelTestTypeIII_myT3_Sprs.txt') contains the Type III pathways in a sparse format, which is easier to analyze by hand.

3. **Identify Type III pathways.** Note that reversible reactions form Type III pathways as well. In general, you are looking for Type III pathways that contain three or more reactions. It is possible that multiple, complicated Type III pathways exist in the model. Listing the corresponding reaction formulas or even drawing a map might be helpful to understand how the reactions form the loop(s). Examples for simple or more complex Type III pathways in metabolic networks can be found in [220] or [206].

4. For each Type III pathway, analyze for every participating reaction:

   (a) Is the directionality correct (see **Step 10**)?

   (b) Is the reaction falsely included?

   (c) If none of the reactions or reaction directions can be corrected based on experimental or thermodynamic information, one can try to iteratively limit the directionality of the loop reactions. A more elaborate procedure has been described elsewhere [220].

   (d) After eliminating a reaction direction or a deletion of a reaction, repeat the Type III pathway analysis. Also, make sure that the removal of directionality or reaction does not affect growth.

   CAUTION Keep in mind that such a change to the network is a hypothesis and may cause problems under different simulation conditions (e.g., environmental conditions).

5. Adjust directionality for all reactions identified in **Step 40.v**, note the change and reason.

TROUBLESHOOTING

Step 41| Re-compute gap list.

>>[Gaps] = AnalyzeGaps(model)

Again, the list 'Gaps' will contain remaining gaps in the network. It will be helpful to have an overview of the remaining dead-end metabolites. (See **Steps 34 to 37** for more details).

The following steps will test if the model can or cannot grow. This means that we will test for qualitative behavior but ignore the correctness of the predicted growth rate.

Step 42| Test if biomass precursors can be produced in standard medium (set in **Step 31**). In **Step 23** the composition of the biomass reaction was determined. Here, we will test for the ability to produce each individual biomass component.

1. Obtain the list of biomass components:

   >> [BiomassComponent, BiomassFraction] = PrintBiomass(model, BiomassNumber)

   where the biomass reaction index is provided with 'BiomassNumber'. The function returns all biomass components ('BiomassComponent') and their corresponding fractions in the array 'BiomassFraction'. It also prints the results in the command window.

2. Add demand function for each biomass precursor ('metaboliteNameList'):
   >> [modelNew,rxnNames] = addDemandReaction(model, metaboliteNameList);
   Note that 'metaboliteNameList' should be identical to 'BiomassComponent', obtained in i). The new model is returned ('modelNew'), which has additional demand reactions for every precursor whose abbreviations are listed in 'rxnNames'.

3. For each biomass component i, perform the following test:
   (a) Change objective function to the demand function ('rxnName'):
   >> modelNew = changeObjective function(modelNew, rxnName);
   (b) Maximize ('max') for new objective function (Demand function)
   >> FBAsolution = optimizeCbModel (modelNew,'max'); The structure 'FBAsolution' contains the optimal solution vector ('FBAsolution.x') and also the value for the objective reaction ('FBAsolution.obj').
      - Case 1: model can produce biomass component (FBAsolution.obj > 0) → proceed with next biomass component
      - Case 2: model cannot produce biomass component (FBAsolution.obj = 0)
         - Identify reactions that are mainly responsible for synthesizing the biomass component.
         - For each of these reactions, following wire diagram given in Figure 3.17.

CRITICAL STEP The overall performance of the model in standard medium condition is determined and, in some cases, corrected. This step needs great care since there may be many possible ways of filling a gap.

CRITICAL STEP This step is also the most likely in which reactions are added to the network with the tag "modeling purposes" only (confidence score of 1). Be careful with such reactions as too many of them may change the overall properties of the network (in this or other simulation conditions).

CRITICAL STEP Added demand and sink reactions represent hypotheses for missing functions that can be tested by experiments.

CRITICAL STEP Comparing dead-end metabolites identified in this step with the list generated in **Steps 34 through 37** will accelerate the debugging process.

CAUTION Keep in mind that adding new reactions to the network may cause new gaps. Therefore, when adding reactions you should make sure that all metabolites are connected to the network.

Step 43| Test if biomass precursors can be produced in other growth media. Repeat **Step 42**.

CRITICAL STEP In this step, the correctness of the network content is evaluated in respect to all known growth conditions of the target organism. This includes all known carbon, nitrogen, sulfur, and phosphor sources.

CRITICAL STEP Physiological information is of great value to determine all growth conditions. For example, Gutnick *et al.* have tested about 600 compounds and have found that 100 can serve as carbon-or nitrogen source for *Salmonella typhimurium* [100]. The model should be able to produce biomass in the majority of these instances.

CRITICAL STEP Not all known conditions may be reproduced by the model - this is not a problem as it represents a starting point for experimental studies to identify missing metabolic functions. However, great attention should be given to collecting and documenting those cases and thus to enable other researchers to pursue them.

Step 44| Test if model can produce known secretion products. If such information is available, they can be used to further refine the model. The first question is if the model can produce the secretion product(s) given a substrate, while the subsequent question is if the ratio between the by-products is correct.

1. To answer the first question: can the by-product(s) be secreted from a given substrate?

   (a) Set the constraints to the desired medium condition (e.g., minimal medium + carbon source). For changing the constraints use the following function:
   >> model = changeRxnBounds(model,rxnNameList,value,boundType)
   Reactions whose bounds should be changed are listed in 'rxnNameList'. The new value for each reaction is contained in the array 'value'. Finally, the type of constraint has to be defined in the list 'boundType'. The possible types are: 'l' for lower bound, 'u' for upper bound, and 'b' if both reaction bounds should be set to the specified value.

   (b) If the model shall be required to grow in addition to producing the by-product,

set the lower bound (boundType = 'l') of the biomass reaction ('rxnNameList ') to the corresponding value ('value').
>> model = changeRxnBounds(model,rxnNameList,value,boundType);

(c) Change the objective function to the exchange reaction of your secretion product:
>> model = changeObjective(model,rxnNameList,objectiveCoeff)
The reaction(s) that should be set as the objective function is given by 'rxnNameList'. They will receive a corresponding coefficient 'objectiveCoeff'.

(d) Maximize ('max') for the new objective function (as a secretion is expected to have a positive flux value, see Figure 3.2):
>> FBAsolution = optimizeCBModel(model,'max');

(e) If the product can be produced (FBAsolution.obj > 0), proceed with the next by-product.

(f) If the product cannot be produced (FBAsolution.obj = 0), the corresponding pathway is missing or incomplete and thus gap analysis must be performed (**Steps 34 to 37**).

2. To answer the second question: Can a certain ratio of by-products be produced?

   (a) First you should verify that both by-products can be produced independently. See ii).

   (b) Set the constraints to the desired medium condition (e.g., minimal medium + carbon source). For changing the constraints use the following function:
   >> model = changeRxnBounds(model,rxnNameList,value,boundType)

   (c) Add a row to the S matrix (see Figure 3.7 for an example of a S matrix) to couple the by-product secretion reactions:
   >> modelNew = AddRatioReaction(model, ListOfRxns, RatioCoeff)
   The two reactions that should be set to a certain ratio are listed in 'ListOfRxns'. Their ratio is given in 'RatioCoeff' by listing the corresponding coefficients in this array. For example, 1:2 is given as [1 2].

   (d) If the model is required to growth while producing the by-product, set the lower bound of the biomass reaction to the corresponding value.
   >> model = changeRxnBounds(model,rxnNameList,value,boundType);

(e) Change the objective function to the exchange reaction of one of your secretion products:
>> model = changeObjective(model,rxnNameList,objectiveCoeff)

(f) Maximize for the new objective function (as a secretion is expected to have a positive flux value, see Figure 3.2):
>> FBAsolution = optimizeCBModel(model,'max');

(g) If the product can be produced (FBAsolution.obj > 0), the second by-product can be produced in the defined ratio.

(h) If the product cannot be produced (FBAsolution.obj = 0, or problem is infeasible), i.e., the ratio cannot be matched. The debugging is less straight-forward in this case as multiple reasons may apply. One very likely reason is that the organism (or cell) in the experimental condition under which the ratio was determined did not grow optimally. However, if you set in ii.d) a lower bound on the growth rate that may cause the discrepancy (due to competition for, e.g., carbons in by-products and biomass reaction). You could try to set this bound lower. Alternatively, some more elaborate tools that are currently not in the COBRA Toolbox can be used to identify missing genes/reactions (Table 3.8).

Step 45| Check for blocked reactions. Reactions that cannot carry any flux in any simulation conditions are called blocked reactions. These reactions are directly or indirectly associated with dead-end metabolites, which cannot be balanced and give rise to so-called blocked compounds [149]. It is good to be aware of those reactions, especially if one expects different results in a simulation (e.g., false-negative analysis of single gene deletion). Furthermore, one might decide to fill some more gaps based on these results. The easiest way to determine blocked reactions is by performing flux variability analysis which is implemented in the function FindBlockedReaction:

1. Change simulation conditions to rich medium or open all exchange reactions:
   >> model = changeRxnBounds(model,rxnNameList, value,boundType)
   Note that the value of the exchange reactions ('rxnNameList') does not matter as this step is testing a qualitative not quantitative property. Therefore, one can set the value to - infinity (e.g., -1000) and + infinity (e.g., +1000). Since we are changing upper and lower bound the boundType is 'b'.

Table 3.8: **List of more elaborate, published tools that have been developed for metabolic reconstructions and that can be used to refine and expand the content**. Many of these tools propose additional metabolic functions to the reconstruction to match observed phenotypes and thus propose new hypothesis that can be readily used to verify experimentally. Note that these tools are not implemented in the COBRA Toolbox and should be obtained directly from the references.

Tool Name	**Description**
ADOMET [142]	Gap filling: Identification of candidates genes based on local structure of the metabolic network using gene clustering, phylogentic information and others.
BOSS [94]	Alternate cellular objectives: identifies candidate objective reaction that can be an existing reaction, a combination of existing reactions, or a previously uncharacterized reaction. However this reaction must be linear and stoichiometric. The framework relies on a reconstructed network and experimental data.
GapFill [149]	Gap filling: proposes changes or additions to the network to connect blocked metabolites to the remainder of the network
GapFind [149]	Gap filling: systematic identification of metabolites along blocked reactions (blocked metabolites)
GrowMatch [150]	An automated method for reconciling in silico/in vivo growth predictions.
OptFind [39]	Alternate cellular objectives: uses experimental data, a reconstructed network and a scoring system to identify a network reaction that is likely to be a component of the cellular objective function
SMILEY [229]	Gap filling: proposes possible missing functions given a metabolic model and physiological, growth data

2. Run analysis for blocked reactions. The function returns a list of blocked reactions ('BlockedReactions').

   ```
   >> BlockedReactions = FindBlockedReaction(model)
   ```

3. Depending on the function of the blocked reaction, one might be interested in "connecting" the reaction to the remaining network. Therefore, follow the diagram in Figure 3.17.

Step 46| Compute single gene deletion phenotypes (Figure 3.19). Analysis of false positive and false negative predictions will help to further refine the network content if the information is available or provides a basis for experimental studies otherwise.

1. Gene deletion: use the following function in the COBRA Toolbox:

   ```
   >> [grRatio,grRateKO,grRateWT] = singleGeneDeletion (model, method, ge-neList)
   ```

   This function allows the use of different methods ('method') for optimization, e.g., FBA, minimization of metabolic adjustment (MOMA) [252], or linear MOMA [21]. The list of genes that shall be deleted is given by 'geneList'. If no gene list is given or the string is empty, all genes in the reconstruction will be deleted and tested for growth capabilities of the knock-out mutant. The function calculates the growth rate of the wild-type strain ('grRateWT') of each deletion strain ('grRateKO') as well as the relative growth rate ratios ('grRatio').

2. Compare with experimental data.

CRITICAL STEP The evaluation of inconsistencies will lead to further reconstruction refinement. Repeat the gap analysis as necessary (**Steps 34 to 37**).

Step 47| Test for known incapabilities of the organism. So far we compared whether the model could reproduce growth on certain substrate, secrete a particular by-product, etc. In this step it should be tested if known incapabilities of the organism can also be reproduced by the model. For example, *Helicobacter pylori* is known to be autotroph for certain amino acids, subsequently, their lack in the medium should decrease *in silico* growth [280].

CRITICAL STEP It is important to use those "negative" data (incapabilities) as well and correct for errors. Error cases can be removed by analyzing the confidence score associated with the reactions along the pathway. In the example of *H. pylori*, this would be the biosynthetic reactions leading to amino acid synthesis [280]. In a more algorithmic approach, a single reaction deletion study can be carried out (see **Step 46** for definition of the variables):

```
>> [grRatio,grRateKO,grRateWT,hasEffect,delRxns,fluxSolution] = singleGeneDeletion
(model);
```

and the results can be analyzed in terms of which deletions disable growth. This smaller subset of reactions needs to be manually evaluated. Note that the deletion of a single function may not be sufficient when alternate pathways exist in the network.

CAUTION Missing incapabilities may not only be caused by falsely added reactions in the metabolic network, but may be a consequence of missing regulatory information. Literature may provide the necessary data.

The following steps will test if the model can predict the correct growth rate or other quantitative properties.

Step 48| Compare predicted physiological properties with known properties. In **Step 47**, known, qualitative incapabilities were compared with experimental data. Here, the network is tested for known capabilities. Figure 3.20 illustrates some examples for the kind of tests that can be performed. Clearly the nature of the tests depends on the available experimental data.

CRITICAL STEP Discrepancies may lead to further reaction refinements (e.g., stoichiometry) or even identification of wrongly entered reactions.

Step 49| Test if the model can grow fast enough. Optimize for biomass reaction in different medium conditions and compare with experimental data.

1. **If the model does not grow at all.** Check your boundary constraints. If these are correct, it is possible that the simulated condition does not support growth (compare with experimental data) or your network is incomplete. In the latter case, return to **Steps 34 to 37** to identify missing links in the network.

2. **If the model does not grow fast enough.** Check your boundary constraints. If these are correct, the possibilities of error modes are quite numerous. It is advised to verify the constraints applied to the model. Use the function which lists all constrained reactions that are greater than a minimal value (’MinInf’) and smaller than a maximal value (’MaxInf’):

```
>> PrintConstraints(model,MinInf, MaxInf);
```

Too slow growth means that at least one precursor of the biomass function cannot be synthesized sufficiently. This implies that the model's biomass production is carbon-, nitrogen-, oxygen-, sulfur-, or phosphate-limited. Since there are generally less active uptake reactions than biomass precursors, it is faster to test if any of the medium components are growth limiting. Therefore, increase the uptake rate ('value') of one substrate ('rxnNameList ') at a time by using:

```
>> model = changeRxnBounds(model,rxnNameList,value,boundType)
```

and setting the bound type to lower bound 'l' ('boundType'). Then, maximize for biomass. If the biomass reaction value increases, it means that this compound is limiting. This gives you a hint as to, where in the network something must be missing. Figure 3.20 shows an example of the *P. putida* network [187] that is not able to grow as fast as reported experimentally *in silico* when toluene is the carbon source. *in silico* analysis suggested that oxygen is rate limiting and that more oxygen-efficient reactions are missing in the network.

3. **Reduced cost.** Linear Programming (LP) problems have two parameters, shadow price and reduced cost, which can be used to characterize the optimal solution. While shadow prices are associated with each network metabolite, reduced costs are associated with each network reaction. The reduced cost signifies the amount by which the objective function (e.g., growth rate) would increase when the flux rate through a chosen reaction was increased by a single unit [225]. Analyses of the reduced costs associated with uptake rates in the growth limited conditions may indicate which compound is limiting. Furthermore, the analysis of reduced cost can be used to identify candidate reactions through which an increased flux would result in a higher growth rate. Use

   ```
   >> FBAsolution = optimizeCbModel(model,osenseStr,primalOnlyFlag)
   ```

   Set primalOnlyFlag to 'false' to get the reduced cost returned with the optimal solution. When maximizing the objective function 'osenseStr' will be 'max' while minimization is defined by 'min'.

CAUTION Whether this discrepancy can be resolved by iterative network refinement depends on the specific case, and thus no general solution can be proposed. As in the case of *P. putida*'s oxygen restriction, such error cases can lead to further experimental investigation which will ultimately increase our biological insight and the reconstructions' quality.

Step 50| Test if the model grows too fast. Optimize for biomass reaction in different medium conditions and compare with experimental data.

1. Again the first check one should do is to verify that the constraints on the model are as expected. Use the function which lists all constrained reactions that are greater than a minimal value ('MinInf') and smaller than a maximal value ('MaxInf'):
   `>> PrintConstraints(model,MinInf, MaxInf);`

2. When the predicted growth rate is higher than expected, it can indicate that constraints are missing. Knowledge about your model and the expected flux map is crucial for identifying the errors. In the following, we propose some possible tests which may indicate where the error is:

   (a) In the worst case, proton shuttling reactions may be present that circumvent the ATP synthetase (e.g., due to a futile cycle). Note that this is only the case in aerobic growth conditions. Such shuttling reactions may be enabled by many reversible transport reactions. Using **Step 40**, reactions associated with such loops can be identified. Also, looking at the flux through the oxidative phosphorylation may indicate if it is used under the aerobic condition or not.

   (b) **Single reaction deletion.** Assume that there is one reaction that enables the model to grow too fast. In this case, a single reaction deletion study will push you towards the right solution. Use the following function by setting the 'method' to 'FBA' and the 'rxnList' should contain one or more reactions to be deleted. If all network reactions shall be tested 'rxnList' does not need to be defined:
   `>> [grRatio,grRateKO,grRateWT] = singleRxnDeletion(model, method, rxnList)`
   The function will return the wild-type growth rate ('grRateW'), the growth rate of the reaction deleted network ('grRateKO'), and the relative growth rate ratio ('grRatio'). However, it is most likely that multiple reactions contribute to this observation and thus they are not identified by this method.

   (c) **Reduced cost.** (See **Step 49** for explanation). The reduced cost analysis can be used to identify those reactions that have a reduced growth rate. Use:

>> FBAsolution = optimizeCbModel(model,osenseStr,primalOnlyFlag)
Set primalOnlyFlag to 'false' to get the reduced cost returned with the optimal solution. When maximizing the objective function 'osenseStr' will be 'max' while minimization is defined by 'min'.

(d) As indicated earlier, reaction directionality may play a role in the fast growth. Therefore, changing the reaction directionality may help. Make sure that only those reactions which are known to produce ATP are allowed for ATP synthesis, while all other reactions are set irreversible (ATP utilization). Similarly, reactions using quinones as electron acceptor should not run reversibly. This might cause problems and may allow circumventing the electron transport chain. These examples are very specific to a model and problem, and no general rule for corrections can be proposed.

CAUTION Changes to the model may be condition-specific and should be well documented.

3.4.5 Data assembly and Dissemination

Step 51| Print Matlab model content. The final reconstruction should be made available to the research community in at least 2 formats: 1. as a spreadsheet containing all information collected during the reconstruction process (as shown in Figure 3.14); and 2. in SBML format which is a transportable format of the models and can be used with other modeling tools. To export the reconstruction from Matlab into Excel format, use:
>> writeCBmodel(model,format, FileName)
where 'format' is 'xls'. To export a model in SBML format, use the same function but change the format to 'sbml'. The output file name is defined by 'FileName'. **CAUTION** Note that the SBML format will not contain all identifiers, references and notes. It is therefore crucial to distribute the reconstruction in a different format. Ideally, the reconstruction content is made available through a web page, such as BiGG [247], which facilitates queries.

Step 52| Add gap information to the reconstruction output. In **Steps 34 to 37** information regarding the remaining and resolved network gaps were collected and should be associated with the output of the final reconstruction (e.g., in Excel format).

3.5 Timing

Step 1| **through 4**| (draft reconstruction): days to a week. **Step 5**|(collection of experimental data): ongoing throughout the reconstruction process **Step 6**| **through 28**| (reconstruction refinement): months to a year (if debugging and gap filling is done along the way) **Step 23**| **through 25**| (biomass determination): days to weeks, depending on data availability **Step 28**| (growth requirements): days to weeks, depending on data availability **Step 29**| **through 32**| (conversion): days to a week. **Step 33**| **through 50**| (network evaluation/debugging): week to months. **Step 51**| **and 52**| (Data assembly): days to weeks, depending how much and in which format data was collected.

All COBRA Toolbox functions described in this protocol finish with a couple of seconds to some few hours on a newer personal computer (Intel Core 2 Duo 6600 2.4 GHz with 4Gb of memory running Windows Vista).

3.6 Troubleshooting

Step 29| See installation instructions of the COBRA Toolbox [21] for details on how to install and setup Matlab, SBML and COBRA Toolbox. **Step 30**| The script may fail during the loading of the model from the xls files. Check:

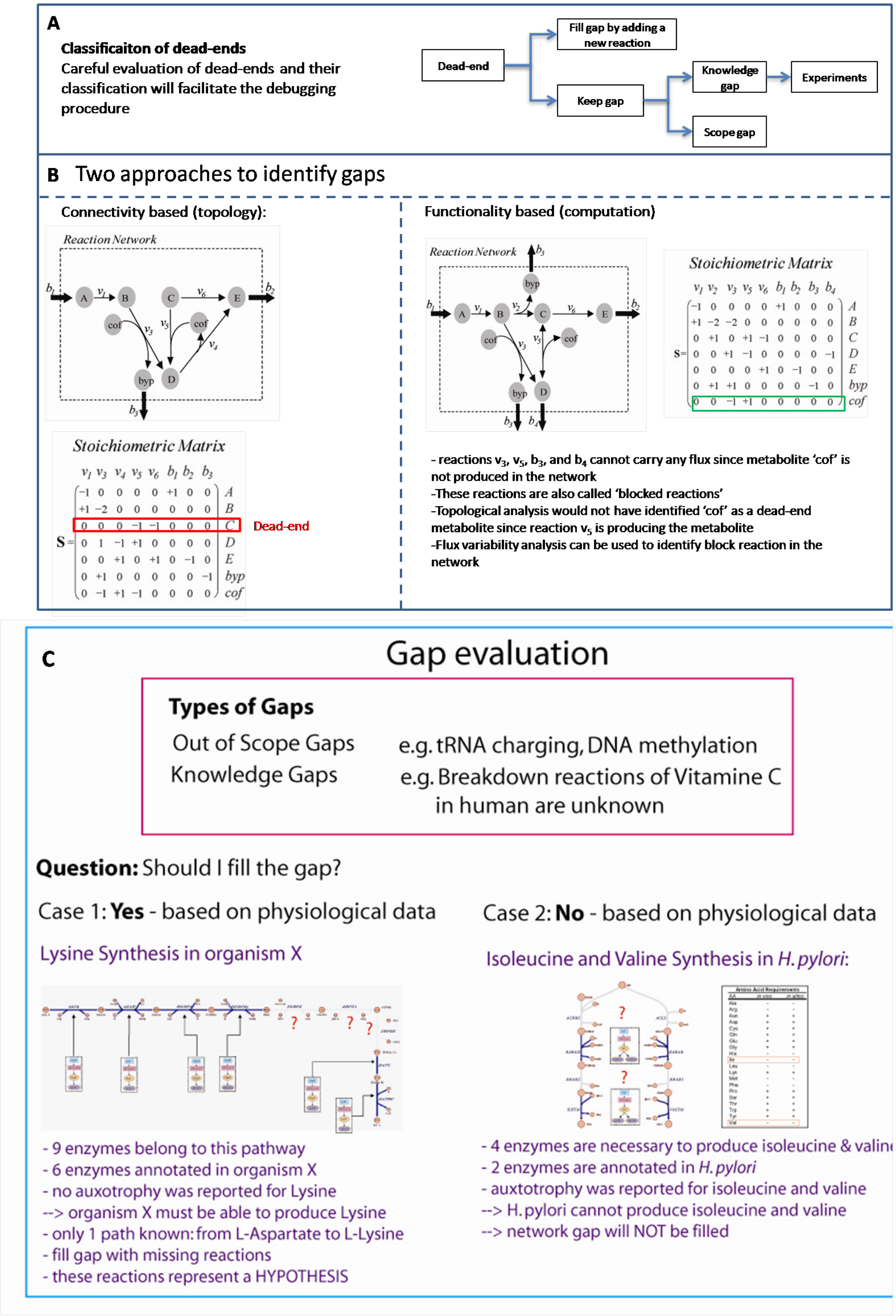
- if headers are correct (Figure 3.14)
- if all necessary information is available
- if metabolic reaction is written correctly ' example; if multiple spaces in the reaction, the script does not work. Separator for left hand side and right hand side can be $-->$, $->$, <==>, <=>
- Mixing number and string can cause problems as well. See Ecoli_core.xls as example on how the input file should look like.

Step 40| Make sure that you are working in the directory were the X3.exe script was copied to. The .expa file produced by the function must be in the same directory as X3.exe.

3.7 Anticipated Results

This protocol will result in a reconstruction that covers most of the known metabolic information of the target organism and represents a knowledge database. This reconstruction can be used as a resource for information (query tool), high-throughput data mapping (context for content), and a starting point for mathematical models.

The text of this chapter, in part or in full, is a reprint of the material as it appears in I. Thiele and B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction, submitted, 2009. I was the primary author of this publication and the co-author participated and directed the research which forms the basis for this chapter.



Classificaiton of dead-ends
Careful evaluation of dead-ends and their classification will facilitate the debugging procedure

B Two approaches to identify gaps

Connectivity based (topology):

$$\mathbf{S}=\begin{array}{c}\begin{array}{cccccccc} v_1 & v_3 & v_4 & v_5 & v_6 & b_1 & b_2 & b_3 \end{array}\\ \left(\begin{array}{cccccccc} -1 & 0 & 0 & 0 & 0 & +1 & 0 & 0 \\ +1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & +1 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 & +1 & 0 & -1 & 0 \\ 0 & +1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & +1 & -1 & 0 & 0 & 0 & 0 \end{array}\right)\end{array}\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ byp \\ cof \end{array}$$

Functionality based (computation)

$$\mathbf{S}=\begin{array}{c}\begin{array}{ccccccccc} v_1 & v_2 & v_3 & v_5 & v_6 & b_1 & b_2 & b_3 & b_4 \end{array}\\ \left(\begin{array}{ccccccccc} -1 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 \\ +1 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & +1 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 \\ 0 & +1 & +1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & +1 & 0 & 0 & 0 & 0 & 0 \end{array}\right)\end{array}\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ byp \\ cof \end{array}$$

- reactions v_3, v_5, b_3, and b_4 cannot carry any flux since metabolite ‘cof’ is not produced in the network
-These reactions are also called ‘blocked reactions’
-Topological analysis would not have identified ‘cof’ as a dead-end metabolite since reaction v_5 is producing the metabolite
-Flux variability analysis can be used to identify block reaction in the network

Gap evaluation

Types of Gaps

Out of Scope Gaps	e.g. tRNA charging, DNA methylation
Knowledge Gaps	e.g. Breakdown reactions of Vitamine C in human are unknown

Question: Should I fill the gap?

Case 1: **Yes** - based on physiological data

Lysine Synthesis in organism X

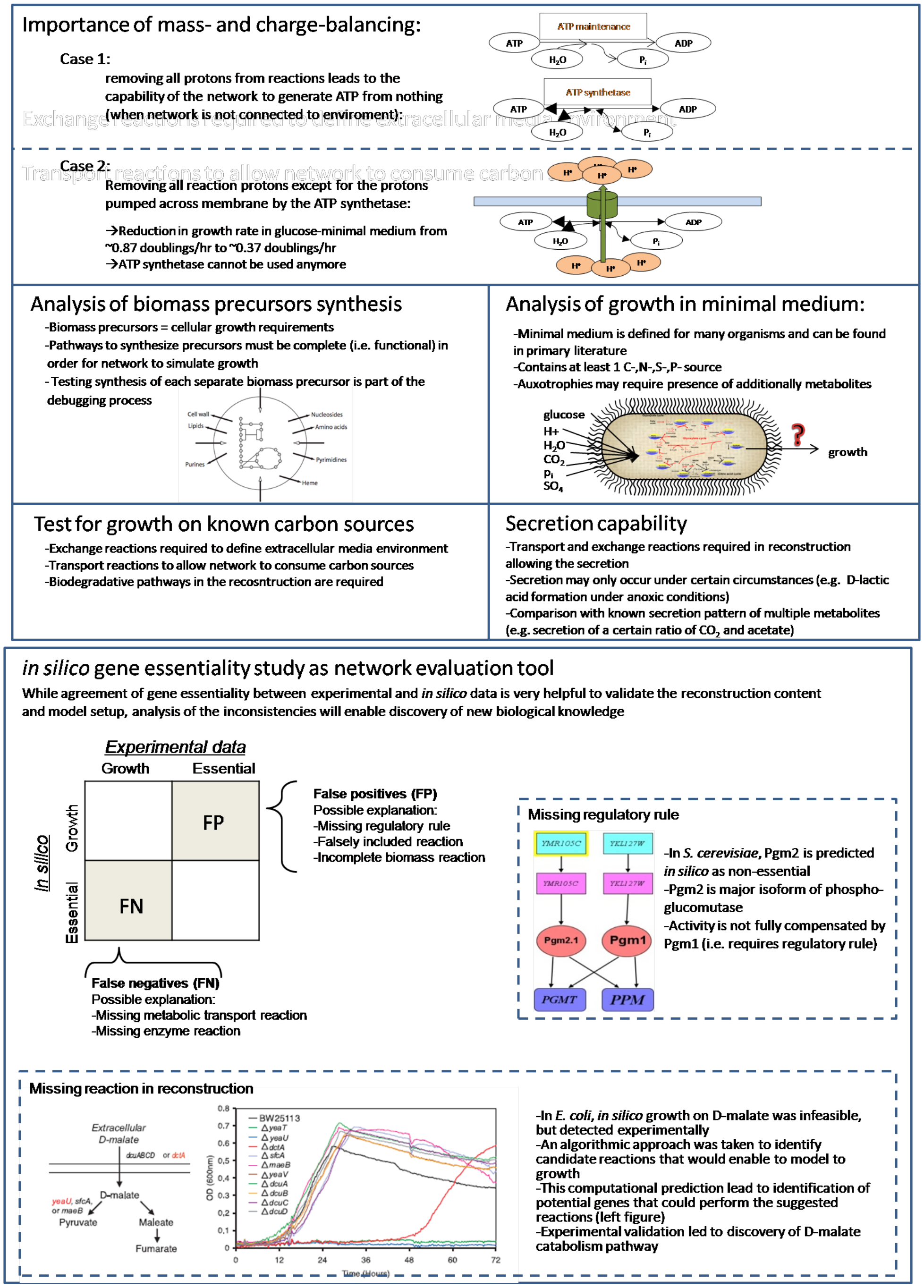
- 9 enzymes belong to this pathway
- 6 enzymes annotated in organism X
- no auxotrophy was reported for Lysine
--> organism X must be able to produce Lysine
- only 1 path known: from L-Aspartate to L-Lysine
- fill gap with missing reactions
- these reactions represent a HYPOTHESIS

Case 2: **No** - based on physiological data

Isoleucine and Valine Synthesis in *H. pylori*:

- 4 enzymes are necessary to produce isoleucine & valin
- 2 enzymes are annotated in *H. pylori*
- auxtotrophy was reported for isoleucine and valine
--> H. pylori cannot produce isoleucine and valine
--> network gap will NOT be filled

Figure 3.18: Gap analysis.



Importance of mass- and charge-balancing:

Case 1:

removing all protons from reactions leads to the capability of the network to generate ATP from nothing (when network is not connected to enviroment):

ATP maintenance
ATP
ADP
H_2O
P_i
ATP synthetase
ATP
ADP
H_2O
P_i

Case 2:

Removing all reaction protons except for the protons pumped across membrane by the ATP synthetase:

→Reduction in growth rate in glucose-minimal medium from ~0.87 doublings/hr to ~0.37 doublings/hr
→ATP synthetase cannot be used anymore

H^+
H^+
H^+
ATP
ADP
H_2O
P_i
H^+
H^+
H^+

Analysis of biomass precursors synthesis

-Biomass precursors = cellular growth requirements
-Pathways to synthesize precursors must be complete (i.e. functional) in order for network to simulate growth
- Testing synthesis of each separate biomass precursor is part of the debugging process

Cell wall
Lipids
Purines
Nucleosides
Amino acids
Pyrimidines
Heme

Analysis of growth in minimal medium:

-Minimal medium is defined for many organisms and can be found in primary literature
-Contains at least 1 C-,N-,S-,P- source
-Auxotrophies may require presence of additionally metabolites

glucose
H+
H_2O
CO_2
P_i
SO_4
?
growth

Test for growth on known carbon sources

-Exchange reactions required to define extracellular media environment
-Transport reactions to allow network to consume carbon sources
-Biodegradative pathways in the recosntruction are required

Secretion capability

-Transport and exchange reactions required in reconstruction allowing the secretion
-Secretion may only occur under certain circumstances (e.g. D-lactic acid formation under anoxic conditions)
-Comparison with known secretion pattern of multiple metabolites (e.g. secretion of a certain ratio of CO_2 and acetate)

in silico gene essentiality study as network evaluation tool

While agreement of gene essentiality between experimental and *in silico* data is very helpful to validate the reconstruction content and model setup, analysis of the inconsistencies will enable discovery of new biological knowledge

Experimental data
Growth
Essential
in silico
Growth
Essential
FP
FN

False positives (FP)
Possible explanation:
-Missing regulatory rule
-Falsely included reaction
-Incomplete biomass reaction

False negatives (FN)
Possible explanation:
-Missing metabolic transport reaction
-Missing enzyme reaction

Missing regulatory rule

YMR105C
YKL127W
YMR105C
YKL127W
Pgm2.1
Pgm1
PGMT
PPM

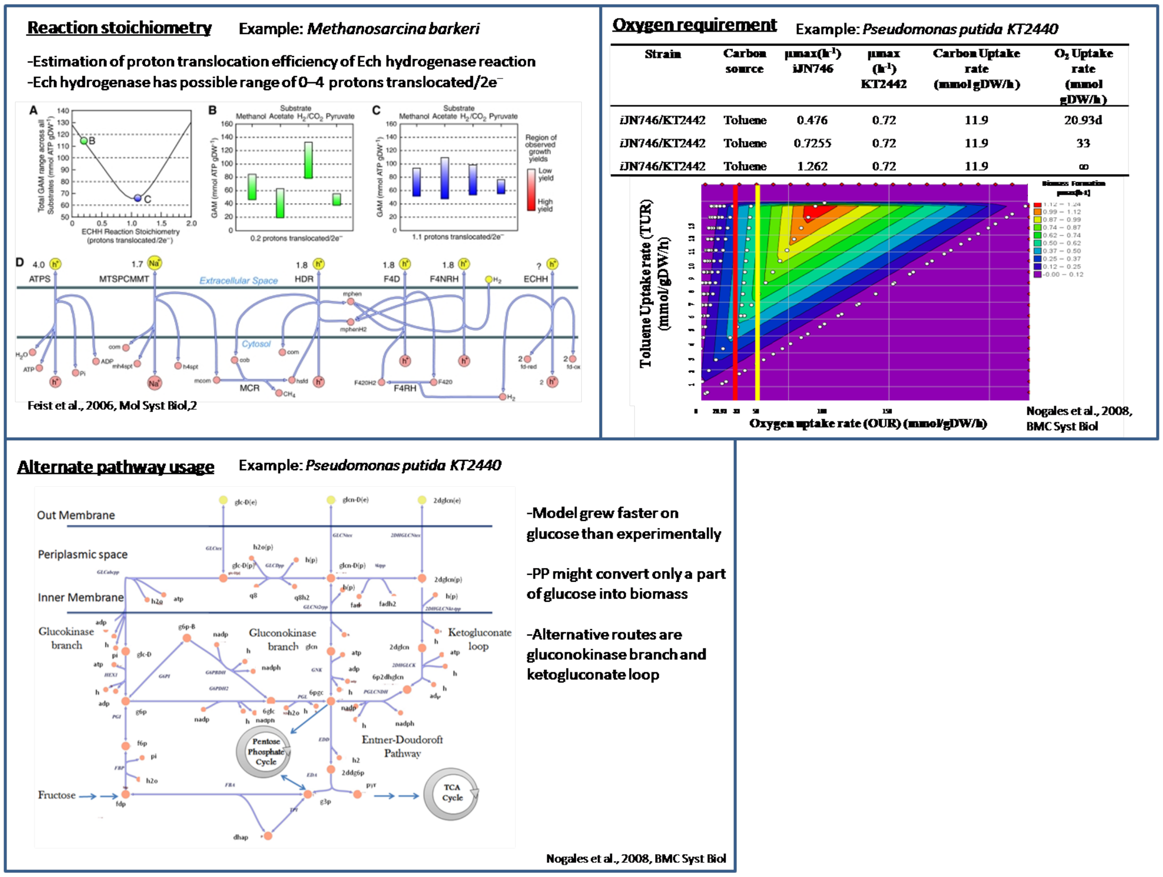
-In *S. cerevisiae*, Pgm2 is predicted *in silico* as non-essential
-Pgm2 is major isoform of phospho-glucomutase
-Activity is not fully compensated by Pgm1 (i.e. requires regulatory rule)

Missing reaction in reconstruction

Extracellular
D-malate
dcuABCD or dctA
D-malate
yeaU, sfcA, or maeB
Pyruvate
Maleate
Fumarate
OD (600nm)
Time (Hours)
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
12
24
36
48
60
72
BW25113
ΔyeaT
ΔyeaU
ΔdctA
ΔsfcA
ΔmaeB
ΔyeaV
ΔdcuA
ΔdcuB
ΔdcuC
ΔdcuD

-In *E. coli*, *in silico* growth on D-malate was infeasible, but detected experimentally
-An algorithmic approach was taken to identify candidate reactions that would enable to model to growth
-This computational prediction lead to identification of potential genes that could perform the suggested reactions (left figure)
-Experimental validation led to discovery of D-malate catabolism pathway

Figure 3.19: Network evaluation.



Reaction stoichiometry Example: *Methanosarcina barkeri*

-Estimation of proton translocation efficiency of Ech hydrogenase reaction

-Ech hydrogenase has possible range of 0–4 protons translocated/2e⁻

Feist et al., 2006, Mol Syst Biol,2

Oxygen requirement Example: *Pseudomonas putida KT2440*

Strain	Carbon source	μmax(h^{-1}) iJN746	μmax (h^{-1}) KT2442	Carbon Uptake rate (mmol gDW/h)	O_2 Uptake rate (mmol gDW/h)
*i*JN746/KT2442	Toluene	0.476	0.72	11.9	20.93d
*i*JN746/KT2442	Toluene	0.7255	0.72	11.9	33
*i*JN746/KT2442	Toluene	1.262	0.72	11.9	∞

Nogales et al., 2008, BMC Syst Biol

Alternate pathway usage Example: *Pseudomonas putida KT2440*

-Model grew faster on glucose than experimentally

-PP might convert only a part of glucose into biomass

-Alternative routes are gluconokinase branch and ketogluconate loop

Nogales et al., 2008, BMC Syst Biol

A
B
C
D
Total GAM range across all Substrates (mmol ATP gDW⁻¹)
ECHH Reaction Stoichiometry (protons translocated/2e⁻)
GAM (mmol ATP gDW⁻¹)
Substrate
Methanol Acetate H2/CO2 Pyruvate
0.2 protons translocated/2e⁻
1.1 protons translocated/2e⁻
Region of observed growth yields
Low yield
High yield
ATPS
MTSPCMMT
HDR
F4D
F4NRH
ECHH
Extracellular Space
Cytosol
MCR
F4RH
Toluene Uptake rate (TUR) (mmol/gDW/h)
Oxygen uptake rate (OUR) (mmol/gDW/h)
Biomass Formation μmax(h-1)
1.12 – 1.24
0.99 – 1.12
0.87 – 0.99
0.74 – 0.87
0.62 – 0.74
0.50 – 0.62
0.37 – 0.50
0.25 – 0.37
0.12 – 0.25
-0.00 – 0.12
Out Membrane
Periplasmic space
Inner Membrane
Glucokinase branch
Gluconokinase branch
Ketogluconate loop
Pentose Phosphate Cycle
Entner-Doudoroff Pathway
TCA Cycle
Fructose

Figure 3.20: Physiological properties that can be important for network evaluation. This evaluation is very case-specific and depends on the reconstructed organism, available experimental data and open questions. For instance, the stoichiometry for the Ech dehydrogenase reaction of translocating protons was not known in *Methanosarcina bakeri* [81]. Another example is the increased oxygen requirement *in silico* of *Pseudomonas putida* when grown on toluene as carbon source [187]. It was hypothesized that an alternative, more oxygen efficient degradation pathway may exist in *P. putida*, since no increased oxygen uptake was measured *in vivo* [187].The last example also involves *P. putida*, which was found to grow faster *in silico* than experimentally observed when glucose was the major carbon source [187]. The simulation result mapped onto the reconstruction map suggested that *P. putida* is using mainly two alternative degradation routes, while the model favored to glycolytic pathway. This discrepancy indicated missing regulation and/or capacity information [187].

Chapter 4

State-of the art reconstructions of cellular networks

4.1 Available metabolic reconstructions

Genome-scale metabolic network reconstructions represent biochemical, genetic, and genomic (BiGG) knowledge bases for target organisms [229]. They effectively represent two-dimensional (2D) genome annotations: that is, all the nodes and links that comprise a cellular network [204]. Reconstructions enable the conversion of biological knowledge into a mathematical format and subsequent computation of physiological properties. As such, network reconstructions enable the investigation of the mechanisms underlying the genotype-phenotype relationship. They are a common denominator in systems biology and thus represent community property and interest [108].

To date, genome-scale metabolic network reconstructions have been published for more than 30 organisms and this number is likely to increase significantly in the coming years (Figure 4.1). The metabolic reconstruction process is well established and has recently been reviewed [229, 79, 88] (see also Chapter 3).

As mentioned earlier, the reconstruction process is iterative. Analysis of the mathematical model may identify missing BiGG information that needs to be added to the reconstruction and further evaluation of the literature may be needed. Clearly, the reconstruction also requires continuous maintenance by updating and expanding its content as new experimental information and knowledge becomes available.

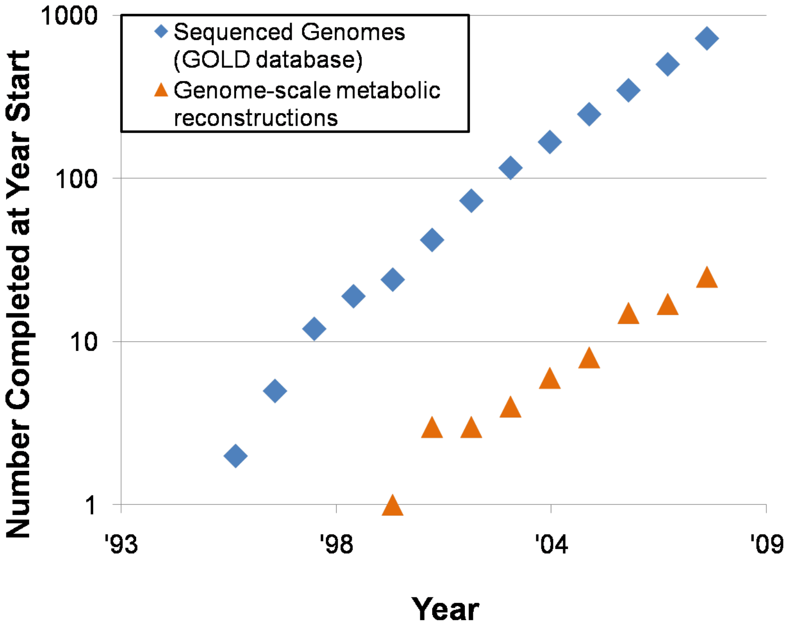


Figure 4.1: **Growth of genome sequences and genome-scale metabolic reconstructions.** The number of network reconstruction has grown exponentially, which is comparable with the pace genome sequences have appeared.

As for genome annotation, the network reconstruction process is ongoing and iterative. For instance, the metabolic network reconstruction for *Escherichia coli* has been updated and refined through six iterations [80], which lead to the most comprehensive metabolic reconstruction available to-date. The metabolic reconstructions for Baker's yeast [87, 179, 67, 148] and for the human pathogens *Helicobacter pylori* [281, 248] and *Haemophilus influenza* [249, 72] have undergone a similar iterative development.

4.1.1 Metabolic reconstructions of *Escherichia coli*

Over the last 20 years the metabolic network of *E. coli* was successively reconstructed (Table 4.1). While the first constraint-based reconstructions focused mainly on *E. coli*'s central metabolism [171, 287, 288, 290], the reconstruction, published by the time *E. coli*'s complete genome was sequenced in 1997 [29], accounted for 26% of metabolic genes [216] (Table 4.1).

Over the next five years the percentage grew to nearly 80% of the annotated metabolic genes [232]. This reconstruction captured more metabolic pathways in *E. coli* and it represented many reactions more accurately. For instance, improvements over pre-

	Year	Genes	R	M	T	Reference
Majewski & Domach[a]	1990	24	14	17	121	[171]
Varma *et al.*	1993-1995	250	146	118	216[b]	[287, 288, 290]
Pramanik & Keasling	1997	306	300	289	75	[216]
Edwards *et al.*	2000	695	720	436	159	[73]
Covert & Palsson[c]	2002	149	113	63	48	[52]
Reed *et al.*	2003	904	929	625	66	[232]
Feist *et al.*	2007	1261	2077	1039		[78]

Table 4.1: **19-year history of reconstruction of the *E. coli*'s metabolic network.** R = Reactions. M = Metabolites. T = Times cited. [a] Core network. [b] A total of 3 papers. [b]Regulated network.

vious reconstructions were i) the usage of quinones in the electron transport chain, ii) expanded carbon source utilization pathways, iii) higher level of curation due to the assurance of both charge and elemental balancing and iv) a larger number of characterized transport systems and their encoding genes. With these characteristics, the model set a new standard in metabolic network reconstructions with its wealth of information incorporated as well as its proven predictive potential in adaptive evolution, metabolic engineering, and understanding of *E. coli*'s physiology (See [78] for more details and references).

In early 2007, a more recent reconstruction of *E.coli*'s metabolism was published that accounts for 1260 genes [78], and thus covering almost 30% of all protein-coding genes in *E. coli*. The latest reconstruction has an increased scope compared to preceding networks and represents the lipid and the lipopolysaccharide biosynthesis reactions more accurately. The reconstruction is compartmentalized into the cytoplasm, periplasm, or extracellular space. Another advance is the alignment of the the reconstruction with the EcoCyc database [133], which provided expanded coverage for the network and content mappings for further computational analysis.

This latest metabolic reconstruction covers almost all known metabolic functions occurring in an *E. coli* cell. At this stage, the reconstruction can be used in a prospective manner, e.g., by analyzing the remaining knowledge gaps and employing molecular and biochemical approaches to identify the corresponding genes. It is expected that an update of the reconstruction will include expansion of the content beyond metabolism. This thesis represents a first step toward a cell-scale model of *E. coli*.

4.2 Reconstruction jamborees

Here, we discuss why and how a community should come together to produce a 2D reconstruction for a target organism.

Concept of a reconstruction jamboree and its rationale: Given the rapidly growing interest in genome-scale reconstruction and modeling, parallel reconstruction efforts for the same target organism have occurred that have resulted in alternative metabolic networks reconstructions for a number of organisms (Table 4.2). This occurrence is unfortunate as it duplicates efforts and creates tension in the field. In comparison to the examples listed above where an iterative refinement of content and scope lead to multiple publications, these parallel reconstructions may be very different in content and format due to differences in reconstruction approaches, literature interpretation and domain expertise of the reconstruction group.

A metabolic reconstruction is specific to the target organism and should summarize all the relevant and available knowledge. Therefore, reconciliation of existing alternative reconstructions is desired and necessary. The development of consensus network reconstructions necessitates a collective community effort to formalize such networks. This need has led to the concept of a 2D annotation (or a reconstruction) jamboree, in analogy to the 1D genome annotation jamborees that lead to community driven genome annotation process.

To date, 2D annotation jamborees have been launched for three target organisms, namely, *Saccharomyces cerevisiae* (Baker's yeast) [108], *Salmonella typhimurium*, and human. It is important to establish standards and criteria that guide the jamboree teams (Figure 4.2). The structured, organized manner will ensure that the consensus reconstructions will be of high use for the research community and guarantee its longevity. A network reconstruction jamboree should have the following goals:

1. Reconcile and refine currently available knowledge and, if available, multiple existing metabolic network reconstructions;
2. Continuously update, re-evaluate and refine the network content; and
3. Form a basis to expand the consensus reconstruction to include more cellular functions, such as transcription and translation, transcriptional regulation, signaling.

Table 4.2: **This table lists existing metabolic reconstruction for the same organisms but by different research groups.** M = Metabolites, R = Reactions. [*] c = cytoplasm, e = extraorganism, p = periplasm, m = mitochondrion, n = nucleus, x = peroxisome, r = endoplasmatic reticulus, v= vacuole, g = golgi apparatur. [**] Number does not include exchange reactions. [a] D. Bumann, personal communication. [b] These three yeast reconstructions have been published as part of iterative reconstruction process, considering recent advances and knowledge of yeast biology and modeling techniques. [c] *i*LL672 was based on *i*FF708 and is thus a parallel reconstruction to *i*ND750. [d] *i*IN800 was based on *i*FF708.

Organism	Genes	Version	Genes	M	R	Compartments*	Reference
Clostridium acetobutylicum ATCC 824	3,848		474	422	552	2 (c,e)	[254]
			432	479	502	2 (c,e)	[157]
Mycobacterium tuberculosis H37Rv	4,402	GSMN-TB	726	739	849	2 (c,e)	[28]
		iNJ661	661	828	939	2 (c,e)	[122]
Pseudomonas putida KT2440	5,350	*i*NJ746	746	911	950	3 (c,p,e)	[187]
		iJP815	815	888	877**	2 (c,e)	[221]
Salmonella typhimurium LT2		*i*RR1083	1,083	N/A	1,087	2 (c,e)	[224]
			1,222	1,084	2,439	3 (c,p,e)	[a]
Staphylococcus aureus N315	2,588	*i*SB619	619	571	641	2 (c,e)	[19]
		*i*MH551	551	604	712	2 (c,e)	[106]
Saccharomyces cerevisiae Sc288	6,183	*i*FF708[b]	708	584	1,175	3 (c,e,m)	[87]
		*i*ND750[b]	750	646	1,149	8 (c,e,m,x,n,r,v,g)	[67]
		*i*LL672[c]	672	636	1,038	3 (c,e,m)	[148]
		*i*IN800[d]	800	1013	1446	3(c,e,m)	[191]
		*i*MM904[b]	904	713	1,412	8 (c,e,m,x,n,r,v,g)	[179]

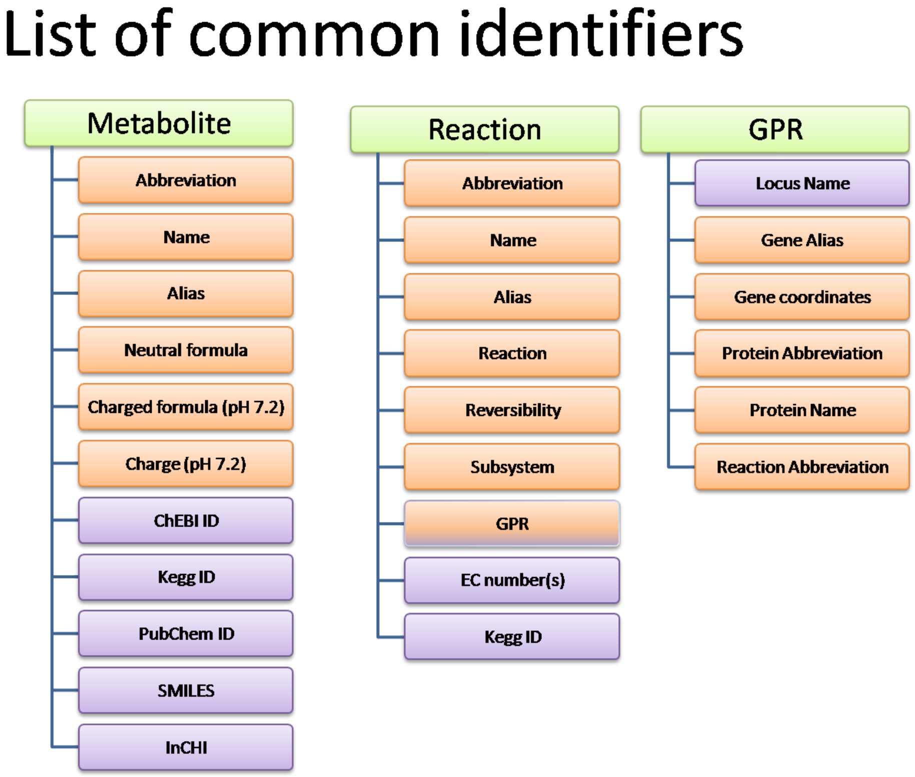

List of common identifiers
Metabolite
Abbreviation
Name
Alias
Neutral formula
Charged formula (pH 7.2)
Charge (pH 7.2)
ChEBI ID
Kegg ID
PubChem ID
SMILES
InCHI
Reaction
Abbreviation
Name
Alias
Reaction
Reversibility
Subsystem
GPR
EC number(s)
Kegg ID
GPR
Locus Name
Gene Alias
Gene coordinates
Protein Abbreviation
Protein Name
Reaction Abbreviation

Figure 4.2: **List of information that needs to be associated with each metabolite and each reaction in the consensus reconstruction.** In orange are highlighted the information that link to the reconstruction. The purple information links the metabolites and reactions to other databases. Gene-protein-reaction (GPR) associations represent a link to the reconstruction as well as to other databases.

These goals are most efficiently achieved with a community approach that assembles experts of different areas and provides a platform for regular meetings and workshops. The 2D jamboree will foster collaborations as well as to inform the community about the properties, content, and capabilities of the consensus reconstruction to ensure its broad use for different biomedical and biotechnological applications.

Information that needs reconciliation: Currently, a 2D jamboree can be laid out for metabolic reconstructions. Common differences between metabolic reconstructions include scope, content, and terminology to describe chemical entities or reactions. At least three areas of metabolic reconstructions need detailed attention by a jamboree team, which include metabolites, metabolic reactions, and the gene-protein-reaction (GPR) associations.

Metabolic 2D annotation jamborees: To date, reconstruction jamborees have been successfully launched for three organisms. The first reconstruction jamboree was organized for the model organism *S. cerevisiae* and resulted in a consensus reconstruction through the joined efforts of experts in *S. cerevisiae* biology and modeling [108]. Subsequently, reconstruction jamborees were launched for the human and the *S. typhimurium* LT2. A workflow was developed for this reconstruction jamboree (Figure 4.3) and it should serve as a template for future reconstruction jamborees.

In Closing A 2D annotation jamboree provides a forum for bringing researchers together to build an organism-specific BiGG knowledge base, and for fostering ensuing collaboration and scientific communication. Ideally, a jamboree should be held regularly, e.g., every other year, depending on the size of the community around the target organism and the availability of new data (e.g., biochemical, genetic, proteomic, metabolomic) as well as incorporating more cellular functions (e.g., signaling pathways, transcriptional regulation, etc.) to the reconstruction. This condition-specific data incorporation will ensure that the consensus reconstruction will serve as starting point for question- and condition-specific models as well as that new experimental evidence, which may be derived from the reconciliation, is captured and incorporated. It is desirable that a reconstruction process becomes a community effort. A well-crafted and executed reconstruction jamboree should accelerate the understanding of the systems biology of the target organism as well as provide the platform for targeted experimental investigation for biological discovery, understanding

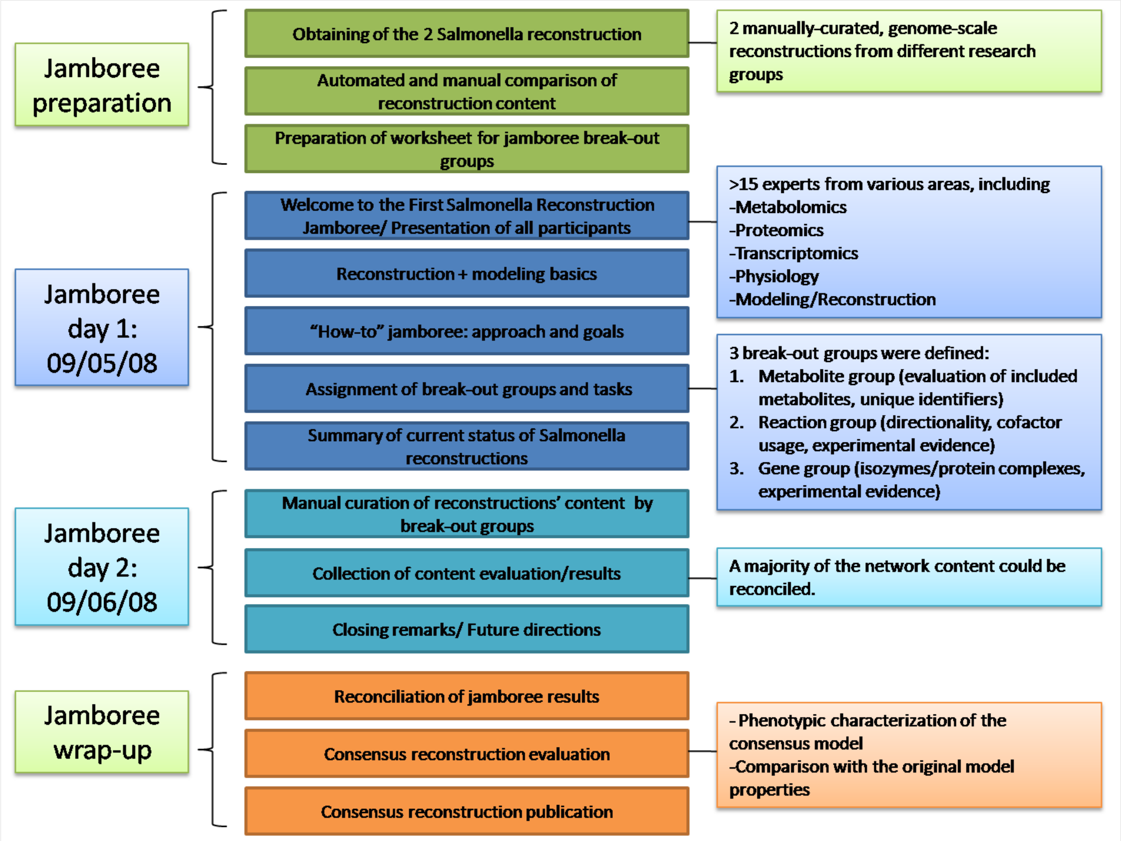

Jamboree preparation
Obtaining of the 2 Salmonella reconstruction
Automated and manual comparison of reconstruction content
Preparation of worksheet for jamboree break-out groups
2 manually-curated, genome-scale reconstructions from different research groups
Jamboree day 1: 09/05/08
Welcome to the First Salmonella Reconstruction Jamboree/ Presentation of all participants
Reconstruction + modeling basics
"How-to" jamboree: approach and goals
Assignment of break-out groups and tasks
Summary of current status of Salmonella reconstructions
>15 experts from various areas, including
-Metabolomics
-Proteomics
-Transcriptomics
-Physiology
-Modeling/Reconstruction
3 break-out groups were defined:
1. Metabolite group (evaluation of included metabolites, unique identifiers)
2. Reaction group (directionality, cofactor usage, experimental evidence)
3. Gene group (isozymes/protein complexes, experimental evidence)
Jamboree day 2: 09/06/08
Manual curation of reconstructions' content by break-out groups
Collection of content evaluation/results
Closing remarks/ Future directions
A majority of the network content could be reconciled.
Jamboree wrap-up
Reconciliation of jamboree results
Consensus reconstruction evaluation
Consensus reconstruction publication
- Phenotypic characterization of the consensus model
-Comparison with the original model properties

Figure 4.3: **Workflow for the *Salmonella* reconstruction jamboree conducted on September 5th and 6th, 2008, at the University of Iceland.** This workflow should serve as a template for organization of future metabolic reconstruction jamborees as it highlights important steps and features of the consensus reconstruction.

and synthesis.

4.3 Reconstruction of other cellular networks

The bottom-up reconstruction process is best developed for metabolic networks but the same underlying principles have been applied to reconstruct signaling networks [209, 161], and in this thesis, to reconstruct a transcriptional and translational network [277]. A pseudostoichiometric approach to represent transcriptional regulatory networks has been recently proposed [95] and applied to a subset of transcriptional regulatory rules found in *E. coli* [93].

4.3.1 Reconstruction of signaling networks

Signaling networks have been successfully constructed based on stoichiometric representation of the underlying biochemical transformations. The human B-cell JAK-STAT signaling network was the first to be reconstructed [209], followed later on by the Toll-like receptor signaling network [161]. While initially extreme pathway analysis was the computational tool of choice, more recent work illustrated how FBA can be employed to investigate the properties of the network [161, 58].

In contrast, no bacterial signaling networks have been stoichiometrically reconstructed yet. The predominant signal transfer method in bacteria is via a so-called two-component signaling pathway (TCP). TCPs are present in nearly all prokaryotes, with some organisms having as many as 200. The TCPs have been found to mediate the response to a wide range of signals and stimuli including nutrients, cellular redox states, changes in osmolarity, quorum sensing, antibiotics and more. The TCP system also occurs in certain eukaryotes like protozoa and higher plants.

A TCP consists of a sensor histidine kinase (HK) and the cognate response regulator (RR). The activation of the HK leads to the autophosphorylation on a conserved histidine residue followed by the transfer of the phosphoryl group to the RR. The phosphorylation of the RR occurs on an aspartate residue within its receiver domain which normally activates an attached output domain. Therefore, the basic chemical reactions are:

1. Autophosphorylation: HK-His + ATP → HK-His P + ADP

2. Phosphotransfer: HK-His P + RR-Asp → HK-His + RR-Asp P
3. Dephosphorylation: RR-Asp P + H2O → RR-Asp + Pi

Many different output domains have been identified but in many cases they are DNA-binding domains. Therefore, the phosphorylation of the RR triggers a transcriptional response.

E. coli has 29 known HKs and 32 RRs that respond to a variety of environmental stimuli, including nitrogen, oxygen, and phosphate limitations and osmolarity [178]. It is expected that the TCP network will be reconstructed in near future for *E. coli.* Such reconstruction will further highlight missing information, such as environmental stimuli, and cross-talk between HKs and RRs. External signals known to impact transcription in microorganisms include carbon source, amino acid, and electron acceptor availability, as well as pH level, and heat and cold stress. *E. coli* has been predicted to have 314 TFs [214] and on the basis of the primary literature 1468 regulatory interactions have been identified [255]. These numbers of regulatory interactions are most likely to be underestimates, but they give an indication of the order of magnitude of the regulatory network reconstruction task.

4.3.2 Reconstruction of transcriptional regulatory networks

The predictive potential of metabolic models have been proven to be high, for example, ranging between 70% and 90% accuracy in predicting growth phenotypes of knockout mutants. Covert *et al.* showed that the prediction accuracy can be further improved by including regulatory information in the metabolic model of *E. coli* accounting for 104 transcription factors and 904 metabolic genes [51]. The regulatory rules were encoded as boolean rules allowing the distinction of on/off in gene expression. Dual perturbation experiments were performed to analyze this model, *i*MC1010 [51]. A systematic approach of reconstructing and interrogating the integrated network of *E. coli* led to the novel characterization of multiple regulatory rules, and an expansion of a genome-scale TRN, based on a model-driven analysis of multiple high-throughput data sets.

More recently, a pseudo-stoichiometric formalism was proposed, which allows the representation of the Boolean rules in matrix format [95].

4.3.3 Reconstruction of transcription and translation

Transcription and translation represent key cellular events as they synthesis the proteins required for all cellular processes. Prior to this thesis work, a coarse-grained, stoichiometric formalism was developed and applied to small-scale gene networks [5]. This work, however, did not synthesize the machinery required to produce functional gene products.

Alternative, non-stoichiometric, kinetic models have been proposed. Abstract models of protein synthesis have been created including non-sequence dependent models with genome-scale models [282, 272] and mechanistically detailed kinetic, but not genome-scale, models [65, 213]. Furthermore, detailed kinetic models have been developed for individual genes and operons and the proteins for which they encode, including the lac operon [300] and the trp operon [243, 257] in *E. coli*.

Therefore, this thesis is the first effort to stoichiometrically represent the required synthesis reactions for a large number of an organism's genes. Furthermore, this thesis worked shows that the approach scalable, expandable and can be integrated with other cellular processes, such as metabolism.

4.3.4 Reconstruction of integrated networks

The Covert *et al.* presented an integrated model for metabolism and regulation [51]. In 2008, two parallel studies were published showing the integration of metabolism, regulation and signaling processes into one model. While the approach was different in these two studies, they both were small-scale and did not explicitly account for the proteins [53, 155].

In this thesis, a scalable reconstruction and analysis framework was developed and applied that allows the integration of multiple cellular functions (See Chapter 8).

The text of this chapter, in part, is a reprint of the material as it appears in I. Thiele and B. Ø. Palsson, 2D genome annotation jamborees: A community effort in systems biology, submitted, 2009. I was the primary author of this publication and the co-author participated and directed the research which forms the basis for this chapter.

Chapter 5

From Biology to computers - Representation of biological processes in computable format

This chapter describes in detail the conversion of the biological processes, necessary for macromolecular synthesis, into a computer readable format. While the stoichiometric representation of metabolic reactions is well established for most core pathways, the representation of other cellular processes, such as signaling, macromolecular synthesis, and regulation, is still subject of research. The conversion includes the identification of the stoichiometric coefficients of each participating component (e.g., metabolite, protein, RNA) as well as mass-and charge balancing consideration. While the first one involves the review of scientific literature, the latter one requires the determination of the elemental formula for each component. This information is then used to formulate template reactions, since the macromolecular synthesis reaction are similar for all macromolecules of a kind. A total of 37 template were formulated based on the information collected in this chapter. More details on individual components or template reactions can be obtained from the supplemental material in Thiele *et al.* [277].

In 1958, Francis Crick first enunciated the central dogma of molecular biology [54] describing the flow of information from DNA to RNA to protein. This chapter will illustrate the different processes involved in this information transfer, in particular for *E. coli* and how it can be converted into mass-balanced reactions. The next chapters will then deal with the reconstruction of this dogma and what can be learned from it.

5.1 Transcription: From DNA to RNA

Transcription units (TU) *E. coli*'s genome is organized in a set of operons, or TUs (Figure 5.1). An operon is a group of genes that are co-transcribed and form a single mRNA molecule (polycistronic mRNA). Every operon has at least one promoter sequence, where the RNA polymerase can bind, and a terminator sequence that defines where the transcription stops (RNA polymerase dissociates from the DNA). A TU, in contrast, is associated with a specific promoter and terminator. Therefore, an operon can be "covered" by multiple TUs. For computational purposes, it is important to give each combination of promoter/terminator a unique identifier. Figure 5.1 illustrates two examples of operons in *E. coli*. Both of these operons have multiple transcription factors, i.e., proteins that may enhance or repress transcription initiation. Subsequently, multiple TUs exist for each operon.

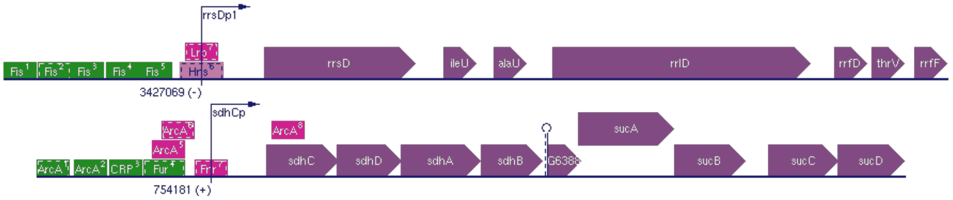


Figure 5.1: **Examples of transcription units in *E. coli*. A.** *E. coli* has seven operons encoding the ribosomal RNA, so called rrn operons. This schematic shows one of these operons. They all have the same organization: 16S rRNA - spacer tRNA - 23S rRNA - spacer tRNA - 5S rRNA - spacer tRNA. The terminal tRNA does not occur in all operon. **B.** This sdhCDAB-sucABCD operon encodes for two important multi-protein complexes involved in central metabolism: succinate dehydrogenase and 2-ketoglutarate dehydrogenase. Taken from Ecocyc [133]

When modeling transcription, it is important to consider this property of bacterial genomes then the behavior and predictions of the model will differ significantly if no co-expression of genes is required. The 'E-matrix' accounts for this property by transcribing TUs rather than genes. A total of 303 genes, involved in transcriptional and translational machinery, were encoded by 249 TU, while 12 further genes had no TU assignment in EcoCyc [133]. (See also Chapter 6). A total of 423 gene products were synthesized, but only 303 of these gene products are directly involved in the cellular processes captured by the

'E-matrix'. Due to the TU organization of the network, 120 gene products were synthesized although they were not within the scope of the 'E-matrix' (i.e., metabolic or regulatory gene products). These 120 gene products were not connected to the other network components since their function was either out of scope or unknown, but corresponding demand functions were included.

Alternate transcripts and overlapping open reading frames. The regions between open reading frames (ORFs) in an operon is called intergenic region and may varying lengths in *E. coli.* In fact, the longest intergenic region included in 'E-matrix' is 28 nucleotides long (b4167-b4168). Thus, these long intergenic regions within an operon have two energy costs associated: i) synthesis cost and ii) degradation cost. Another feature that needed to be considered during the reconstruction is the presence of overlapping ORFs, which have been reported and studied by numerous groups [198, 3, 57, 238, 212]. In this case, the transcription occurs as polycistronic mRNA, but the second (or overlapping) gene is assumed to be cleaved after the stop codon of the upstream gene (see below for more details). Again, since the translation occurs *in vivo* on the polycistronic mRNA, the overlapping does not affect the functionality of the second gene product but may affect the half-life time of the mRNA.

Furthermore, the transcription is a bit different for stable RNA (i.e., ribosomal RNA and tRNA) operons and protein coding operons as they require the presence of different proteins. These differences are highlighted in the following sections.

rRNA operons. Application of the steady state mass conservation assumption requires that input fluxes of every component are balanced by output fluxes. Consequently, only those reactions whose substrates and products can be balanced within the network can carry a non-zero flux. In the 'E-matrix' there is one reaction, the transcription of the rRNA operon rrnD (b3272-b3278), which is stoichiometrically unbalanced. In contrast to the other rRNA operons, the transcription of this operon produces two 5S rRNA (rrfD,b3272 and rrfF, b3274), while only one 23S rRNA and one 16S rRNA is produced. However, ribosomes contain only one copy of each rRNA type. In terms of mass conservation, the transcription of this operon is infeasible since both of the 5S rRNAs cannot be incorporated into one ribosome, and accumulation of rrfD is not allowed by the steady-state assumption. Since this operon also encodes for three tRNA molecules, namely ileU (b3277), alaU (b3276), and thrV (b3273), we created a sink reaction for rrfD 5S rRNA. The operon structure for the rrfD 5S rRNA has been experimentally determined and sequenced by Duester and Holmes [69].

Transcription initiation. A main player is the DNA-dependent RNA polymerase, which consists of four proteins: two α subunits, one β subunit and a β' subunit (holoenzyme). For its function, the RNA polymerase requires the covalent binding of another protein: a sigma factor. *E. coli* has seven different sigma factors, which are required for different genes and thus may activate transcription under distinct environmental conditions (see Table 2.4). Transcription initiation occurs if the RNA polymerase (with sigma factor bound) binds to the promoter site on the operon. This DNA-protein complex is called closed complex and its formation is reversible. In the next step, the RNA polymerase unwinds the DNA stretch it clasps around and forms the so called open complex. The unwound DNA stretch is an approx. 12 bp wide bubble [185]. The formation of the open complex is relatively slow. In a third step, the length of the bubble is extended to 18 bp and the transcription of the first 16 bp of the operon occurs [117, 186]. This initial transcription may be reversible, resulting in abortion products and re-starting of the RNA polymerase.

Transcription elongation. Once the RNA polymerase leaves the promoter site, the transcription elongation produces a growing nascent transcript and so called elongation factors are required to help over potential bumps (e.g., pausing sites) on the DNA. In general, there are different sets of transcription elongation factors depending on specific (pausing) sites in the DNA template. Thus, the actual set of involved transcription factors might differ slightly between ORFs. We define three different sets of transcription factors: i) stable RNA encoding ORF, ii) ORF with annotated/verified Rho-dependent transcription termination, and iii) all remaining ORF. Information about sigma factors for each gene was obtained from EcoCyc [133]. If no information was available, sigma 70 is assumed to be required for transcription. In general, the reconstruction does not account for abortive products during transcription initiation, pausing sites, errors in transcription, folding, etc. It also does not account for arrested RNA polymerase that can either resume or abort the transcription; however, both factors necessary for these actions (GreA, GreB) were included in the transcription reactions. Furthermore, the sigma factor leaves the RNA polymerase after transcription initiation. The elongation factors include:

- GreA and GreB, which are required for efficient transcription [32, 33, 76, 197].
- NusA, which increases the duration of pausing of RNA polymerase at pausing sites, but also protects nascent mRNA from cleavage [236, 169].
- Mfd (transcription repair factor), which reactivates or recycles stalled or arrested RNA polymerases during elongation. Its action is ATP-dependent which was not explicitly modeled [253, 211, 239], since the overall cost per ORF is unknown.

On a given operon, not all of these factors may be required but we included them in all transcription elongation reactions as no large-scale data sets are available listing Gre- and Nus-protein-dependent genes.

The transcription elongation of stable RNA operons requires GreA, GreB, NusA, NusB, NusG, S10, S4, L3, L4, L13 [284]:

- NusG allows Rho-dependent termination [236].
- NusB is required for antitermination in rrn operon as well as NusA, NusG, RpsJ(aka NusE, S10) and one unknown factor.
- NusB and RpsJ (aka NusE, S10) bind directly to a nucleotide sequence called boxA.
- NusA increases the duration of pausing of RNAP at pausing sites, but also protects nascent mRNA from cleavage [236, 169]. For rho-dependent antitermination, NusA may help to bind ribosome on nascent mRNA [236]. NusA was found to be non-essential in rho-mutant [236].

Transcription termination. Different strategies for transcription termination are known including attenuation and rho-dependent termination.

In the reconstruction, we only consider rho-dependent and rho-independent termination, as attenuation seems to be a feature of few operons. The rho protein, a hexamer, winds around the nascent transcript, a ATP-dependent process, until it is close to the DNA - RNA polymerase - mRNA complex. Its presence destabilizes the complex and the RNA polymerase falls off the DNA strand eventually.

Rho-dependent termination. There are specific sequences in mRNA that are recognized by Rho. These regions are accessible if no ribosome blocks it (low secondary structure of mRNA) and the RNA polymerase pauses at specific pausing sites. Thus, transcription is terminated when mRNA is not sufficiently transcribed. In addition to Rho, NusG is necessary for termination. *E. coli*'s transcription termination protein Rho is a hexamer (a trimer of dimers) [258, 266, 265]. Three ATP are required per Rho-hexamer [265].

Cleavage of polycistronic mRNA. The transcription resulted in the synthesis of a mono- or polycistronic mRNA depending if one or more genes were encoded by the operon. As mentioned earlier, in bacteria there is no spatial and temporal separation between transcription and translation. In the reconstruction, polycistronic mRNA is always cleaved by the action of RNase III prior to translation. This assumption allows different translation frequencies/levels for the gene products in polycistronic mRNAs [212]. RNase P has also been reported to be responsible for cleavage of polycistronic mRNAs [162]. However, in order to reduce the number of total network reactions only RNase III is considered for cleavage in the reconstruction [233, 162]. *In vivo*, the cleavage of polycistronic mRNA may destabilize the mRNA products, although stabilization has been observed in some cases [184]. Furthermore, TUs with overlapping codons (stop-start codon juxtaposition) are cleaved as well, whereby a full length transcript was created for the upstream (5') gene and a shorter transcript for the downstream gene (3'). However, the translation on the shorter transcript used the full-length sequence such that the protein sequence of the gene product is complete. For more details about stop-start codon juxtaposition refer to the following studies: [198, 3, 57, 238, 212] and the review from Normark *et al.* [192]. The different frequencies of translation and half life time of cleavage products could be represented using constraints on the reaction fluxes. Decay of cleavage products occurs independent of cleavage of polycistronic mRNA. Cleavage products end with a 3' mono-phosphate group, which may affect the overall half-life time of cleavage products.

Cleavage of stable RNA The processing of rRNA occurs prior to ribosome formation. Some of the posttranslational modifications as well as cleavage of rRNA require the association to ribosomal proteins. Furthermore, these factors have been shown to be involved in ribosome maturation based on 30S binding to these factors [41, 159]. In the reconstruction, these factors are only involved in ribosome maturation and not in rRNA cleavage/ modification, which takes place prior to the ribosomal assembly.

For monocistronic tRNA operons: a generic RNase ('RNase_Gen') represents the possible cleavage by alternative RNases: RNase II, RNase D, RNase RNase BN, RNase T, RNase PH. Two studies showed that the presence of one of the RNases is sufficient to cleave the tRNA operon transcription products [235, 139].

For polycistronic stable RNA operons (which include tRNAs and rRNAs): a generic RNase ('RNase_Gen') represents the possible cleavage by alternative RNases: RNase II, RNase D, RNase RNase BN, RNase T, RNase PH. Again, it was observed that the presence of one RNases is sufficient to cleave the stable operon transcription products [235, 139]. Neidhardt *et al.* report the existence of RNase_m23, RNase_m16, RNase_m5, which are necessary for the trimming of rRNA; however, the corresponding genes are unknown [184]. The *E. coli* gene for RNase PH (b3643) is a pseudogene [237]. Although 'RNase_Gen' is in the reconstruction, no gene encodes for the RNase_PH protein. Furthermore, although some literature is reporting the presence of RNase PH in *E. coli*, no information is available for the *E. coli* K12 MG1665 strain.

To summarize the last sections, the transcription yields in mono- and polycistronic mRNA and stable RNA. Polycistronic RNA is subsequently cleaved in its constituents by a number of proteins (RNases). Note that the cleavage of polycistronic mRNA is done for modeling reasons and does not occur as such in the cell, since in many cases the cleavage of a polycistronic mRNA is believed to destabilize the transcript and thus accelerate the mRNA degradation. In the reconstruction, however, this step greatly reduces the number of network reactions by limiting number of combinations of protein factors and RNA. As we see in the following section, many ribosomes can be on a transcript at a time depending on the ribosome affinity to the mRNA.

5.2 Translation

In vivo, the translation rate depends on various factors, such as binding affinity of the ribosome to the Shine-Dalgarno sequence, tertiary mRNA structure, and mRNA degradation rate, resulting in variable transcription efficiency under different growth conditions, e.g., environmental stresses [184]. In many cases, the transcription unit (operon) structure provides a translational coupling of the gene, where the translation of the downstream gene is dependent on the translation of the upstream gene [192]. One reason for this is that the coding region of the first gene may contain the Shine-Dalgarno Sequence necessary for the translation of the second gene. Such dependency was not modeled in the 'E-matrix'.

Translation initiation. The 30S initiation complex consists of 30S ribosomal subunit, mRNA, $fMet - tRNA_f^{Met}$, and IF1, IF2-GTP, and IF3. IF3 assisted by IF1 promotes the dissociation of vacant 70S ribosomes and thus provides the pool of free ribosomal subunits. The binding of 50S ribosomal subunit to the initiation complex leads to the release of IF1 and IF3. After GTP hydrolysis, IF2-GDP is also released and the translation initiation complex is ready for translation elongation. A peptidyl-tRNA bound to EF-Tu is presented to the ribosome. Beside the start codon AUG, the following two codons are also recognized and translated as formyl-methionine if they occur as first codon in the mRNA sequence: TTG and GTG [283].

Ribosomal binding to mRNA. The translation initiation of an mRNA occurs relatively frequently, depending on the binding strength to the Shine-Dalgarno sequence and other factors, leading to multiple ribosomes per mRNA. It has been found that the minimum distance between two ribosomes is 17 amino acids [132]. This variation in mRNA occupancy is modeled in the E-matrix by three translation reactions per mRNA: i) translation with one ribosome per mRNA, ii) translation with maximal possible number of ribosomes per mRNA (i.e., 17 amino acids distance), and iii) translation with half of the maximal possible ribosome number per mRNA (i.e., about 34 amino acids distance). Thus, for each mRNA there are three sets of translation reactions differing in the number of ribosomes per mRNA and polypeptide products. This simplification reduces the number of possible combinations as well as allowing adjustments of the translational rate based on translation initiation, if such data is available.

Translation elongation Three translation elongation factors are involved in translation elongation: EF-Tu, EF-Ts, and EF-G.

The ribosome can fit three tRNA molecules: one in the A site, one in the P site, and a third one in the E site. The A site accepts the presented aminoacyl-tRNA, while the P site is occupied by an aminoacyl-tRNA to which a partially completed peptide chain is attached. The E site is occupied by the tRNA that is about to exit the ribosome [185]. The anticodons of the tRNA are bind to codons on the mRNA on all three sites.

EF-Tu is a GTP-dependent protein that binds to charged (i.e., aminoacyl-) tRNA molecules and presents it to the ribosome. This ternary complex (tRNA-EF-Tu-GTP) enters the A site, the anticodon binds to the codon on the mRNA. This happens under hydrolysis of the GTP bound to EF-Tu, while p_i. The action of EF-Ts is required to remove EF-Tu from the complex. The formation of the peptidyl bond between the amino group of the tRNA in the A site and the peptide chain attached to the tRNA in the P site is catalyzed by the 50S subunit, and it does not require energy. The uncharged tRNA is moved from the P site to the E site and the ribosome moves a codon further on the mRNA sequence (in 3' direction) and a new elongation cycle begins. This translocation is assisted by a third elongation factor, EF-G, under hydrolysis of a second GTP [185].

EF-Ts acts as a catalyst in the displacement of the GDP from the EF-Tu-GDP complex and allows the binding of GTP so that the ternary complex EF-Tu-GTP-aminoacyl-tRNA can be formed. The crystal structure of the complex has been solved in which no ions have been reported beside Mg^{2+} of EF-Tu [138]. The EF-Tu-EF-Ts complex has potentially a 1:1 stoichiometry. EF-Tu has higher affinity to GDP than to GTP. EF-Ts stimulates the dissociation of EF-Tu and GDP by formation of an tertiary complex: EF-Tu-GDP-EF-Ts. Subsequently, GDP is released. GTP binds to the binary complex EF-Tu-EF-Ts. This tertiary complex dissociates to EF-Tu-GTP and EF-Ts [99].

Translation termination. There are three release factors (RF1, RF2, RF3) and ones ribosomal release factor (RRF). The stop codon recognition is not achieved by a tRNA but by two stecified proteins: RF1 and RF2. They have overlapping stop codon recognition (RF1 - UAG and UAA; and RF2 - UGA and UAA). The binding of RF1 or RF2 to the ribosome triggers the hydrolysis of peptidyl-tRNA. RF3 is a GTP-binding protein, which catalyzes the release of RF1 or RF2 and thus accelerates the transition from translation termination to ribosome recycling. Finally, RRF acts together with EF-G and RF3 to dissociate the post-termination complex by GTP hydrolysis. While RF1 and RF2 are essential in *E. coli*, RF3 is non-essential but necessary for optimal translation, especially under stress conditions, [222]. The deacylated tRNA still present on the 30S particle is displaced by IF3 allowing the recycling of 30S subunit.

The functional protein RF2 has a glutamine residue methylated at position 252, which increases the translational termination rate [63]. Furthermore, it has been found that RF1 reads far more often the UAA stop codon than RF2 in *E. coli* [63].

For simplicity in the reconstruction, the 70S ribosome is release from the translation termination reaction (free from nascent polypeptide). The 70S ribosome reacts with IF3 and IF1, which leads to the release of the 50 S subunit. The ribosome release factor (RRF) forms a stable complex with 70S ribosome. The action of EF-G releases RRF from 70S ribosome, stimulated by GTP [144].

Codon usage and tRNA assignment. The anticodon and codons for each of the 86 tRNA species was obtained from EcoCyc [133] and the Riley annotation [237]. Since some of the tRNAs can read more than one codon or have overlapping functions with other tRNAs due to the wobble position, template tRNA were created for those cases. For example, $tRNA^{gltT}$, $tRNA^{gltU}$, $tRNA^{gltV}$, and $tRNA^{gltW}$ have an UUC anticodon and can translate the codons GAA and GAG. Instead of having this variation on level of the translation reactions, which would lead to a combinatorial explosion of the number of network reactions, a generic tRNAglt1 specie was used for the translation reactions. (See also Table 8.7 and 8.8). Additionally, four reactions were created by converting each $tRNA^{glt}$ species into $tRNA^{glt1}$ (irreversibly). This simplification enabled a dramatic reduction in the number of network reactions, while conserving the intrinsic property of redundant codon reading.

5.3 Protein Maturation.

Polypeptides released from the translation termination complex have a formyl-methionyl group bound to the N'-terminus. While the formyl-group is removed for all polypeptides by the peptide deformylase (Def, b3287), the methionyl-group is removed only from some polypeptides depending on the amino acids that follow. This latter information was also obtained from CyberCell [269]. CyberCell [269] lists the sequence for matured proteins, which may have the N'-terminal methionine or even signal sequences removed. Although the origin of these sequences is often not clear (experimental determination or computational prediction), the data were incorporated into the 'E-matrix'. Prior to incorporation the matured sequences were compared with predicted amino acid sequences from the corresponding nucleotide sequences to eliminate potential errors in the database. The removal of signal sequences is enzyme independent in the reconstruction and no action, e.g., protein export, is associated with the removal since these two processes were outside the defined scope of the 'E-matrix'.

The methionine aminopeptidase (Map, b0168)is responsible for the removal of the N'-terminal methionine in all polypeptides.

5.4 Metallo-ions incorporation.

Metallo-ions. Many *E. coli*'s proteins need metallo-ions for correct folding and/or function. This information was obtained from primary crystallization literature together with the structures deposited in the Protein Database (PDB, [25]). In some cases, additional experimental studies were available, which tested the protein function under various concentrations of metallo-ions. In those cases, if no favored cation was identified, Mg^{2+} was assumed to be the incorporated metallo-ion. Furthermore, some of the metallo-ions, mainly cations, are only involved in the reaction mechanism but are not covalently bound to the protein, and thus may leave the protein after termination of the reaction. These metallo-ions were not incorporated in the proteins of the reconstruction. Wilson *et al.* proposed that the metallo-ions are incorporated into the proteins prior to the protein folding [299], which was implemented in the reconstruction.

Iron-sulfur-cluster biogenesis. A number of features were included in the reconstruction such as metallo-ion binding of tRNA, rRNA, and proteins. Some of *E. coli*'s proteins have $[4Fe\text{-}4S]^{2+}$-clusters incorporated, which often function as iron- or oxygen-sensor or are involved in oxidation-reduction reactions [22]. These iron-sulfur clusters are formed outside of the target protein and is transferred to the corresponding proteins by an IscU dimer (b2529) [2, 137, 194, 1, 173, 127, 301]. Following extensive perusal of the primary literature regarding iron-sulfur-cluster biogenesis, a mechanism was used for this reconstruction, which summarized the current consensus rather than copying a specific proposed mechanism (see below and Table 5.1). There does not seem to be a general consensus on the biogenesis of the iron-sulfur-cluster in the scientific community. In the current reconstruction, two of the biogenesis reactions are not balanced because the formation of $[2Fe\text{-}2S]^{2+}$ and $[4Fe\text{-}4S]^{2+}$ requires an electron acceptor. Kato *et al.* [137] proposed glutathione as electron acceptor, however, glutathione may involve a proton transfer, which would require an additional acceptor. In short, since an appropriate electron transfer mechanism could not be found, the two reactions had to remain unbalanced. The two chaperones, HscA (b2526) and HscB (b2527), were not included in the reconstruction since their functions in the iron-sulfur-cluster biogenesis has not been completely elucidated [303, 273, 111, 151, 256].

[2Fe-2S]_TRANSF. It is possible that Fe^{2+} instead of Fe^{3+} is bound by IscA. However, the literature is not very clear. Regardless which iron species get bound to IscA and transferred to IscU in the iron-sulfur cluster, a reducing agent would be necessary to transfer some of the electrons from the irons in order to produce $[4Fe\text{-}4S]^{2+}$ cluster. Since there is no clear information available regarding such agent (although Kato *et al.* suggested that glutathione might participate in one of the final cluster forming reactions [137], Fe^{3+} and Fe^{2+}sulfur clusters are combined here, to produce $[4Fe\text{-}4S]^{2+}$ cluster with charge balanced reactions.

$[4Fe-4S]$**_FORM.** IscU_dim_$[4Fe-4S]$ is used to transfer $[4Fe-4S]$ to iron-sulfur cluster proteins [1, 2, 127, 173, 194, 301].

IscA_TETRA and IscU_dim_S-SH_FORM2, IscA binds two iron per tetramer [64]. Ding *et al.* proposed it to be a scaffold protein but it cannot replace IscU in IscU-cells (in *Azotobacter vinelandii*) since IscU deletion is lethal [127]. Hence, it is not clear whether IscA is a scaffold protein or a Fe-donor [64].

IscU_dim_S-SH_FORM1. IscS and IscU form a 1:1 complex in which IscS and IscU are linked with each other through a disulfide bond. This bond is only formed in presence of L-cys. IscS has to have IscS-(SH)2 formed for complex formation with IscU [151, 137]. To increase the turnover rate of desulfurase reaction, IscU must be dissociated from IscS immediately (after S transfer). In Kurihara *et al.*, they achieved this with addition of DTT but propose that this might be the funciton of the two chaperones encoded in the same operon (HscA and HscB), which have been shown to interact with IscU [151]. In addition, a reducing agent is necessary, which has been proposed to be gluthatione [137].

Other players: It seems to be clear that cysteine provides the sulfur for the cluster via IscS; however, the Fe^{2+} source is not clear since iron is toxic to the cells, so that its cytoplasmic, soluble concentration is low [12]. In addition, *E. coli* has a number of iron-binding proteins, which remove free iron. YggX (b2962) is another protein that has

Table 5.1: **List of reactions included in the E-matrix that synthesize iron-sulfur-cluster.** fe2 = Fe^{2+}; h = proton; trdox = oxidized thioredoxin; trdrd = reduced thioredoxin; nadp = nicotinamide adenine dinucleotide phosphate; nadph = nicotinamide adenine dinucleotide phosphate (reduced).

Reaction abbreviation	**Reaction name**	**Reaction**	**Reference**
[2Fe-2S]_TRANSF	formation of [2Fe-2S] cluster	2 IscU_dim_2Fe(2)-2S → 1 IscU_dim_[2Fe-2S]2 + 2 IscU_mono	[1, 2, 127, 137, 173, 194, 301]
[$2Fe - 2S$]_FORM	transfer of 2 Fe2 from IscA (tetramer) to 2 IscU (monomer)	1 IscA_tetra_Fe(2) + 2 IscU_mono_S-SH → 1 IscA_tetra + 1 IscU_dim_2Fe(2)-2S	[1]
[$4Fe - 4S$]_FORM	formation of [$4Fe - 4S$] cluster	1 IscU_dim_[$2Fe - 2S$]2 → 1 IscU_dim_[$4Fe - 4S$]	[1, 2, 127, 173, 194, 301]
IscA-Fe_FORM	2 Fe2 bind to IscA (tetramer)	1 IscA_tetra + 2 fe2 + 1 h + 1 nadph + 1 trdox → 1 IscA_tetra_Fe(2) + 1 nadp + 1 trdrd	[64]
IscA_TETRA	IscA tetramer formation	4 IscA_mono → 1 IscA_tetra	[1, 56, 64, 127, 137, 151, 173, 259, 285, 301]
IscU_dim_S-SH_FORM2	formation of IscU S-S bound	1 IscS_IscU_cplx → 1 IscS_dim_S-H + 1 IscU_mono + 1 IscU_mono_S-SH	[1, 56, 64, 127, 137, 151, 173, 259, 285, 301]

been shown to be able to bind iron [97, 98]; however, a recent report did not manifest the ability of YggX to bind iron [199]. Another candidate is CyaY (b3807), a frataxin homolog and tetramer in solution, which can bind 6 to 26 Fe^{3+} ions when there is an excess of intracellular iron [1, 34]. The deletion of CyaY does not have any apparent effect on biogenesis of iron-sulfur cluster in *E. coli* [160] and in yeast [68], but *Salmonella enterica* strains lacking CyaY show defects *in vivo* in Fe-S cluster metabolism [291]. Since the function of CyaY has not been completely elucidated, we did not include this gene product into the reconstruction.

Three mechanism for iron-sulfur cluster formation have been proposed [153].

1. Iron binds to IscU, then the sulfur is transfered from IscS to IscU (based on observation of a stable iron-IscU complex in the case of *Thermotoga maritima* IscU [194]). However, there is no experimental evidence that the addition of sulfur atoms to an iron-loaded IscU gives rise to a cluster. For reasons probably related to structural differences between IscU proteins [127], IscU from *E. coli* and *Azotobacter vinelandii* do not bind iron [259, 1, 64].

2. Sulfur binding occurs first, and is followed by iron binding. This is supported by the finding that sulfur transferred from IscS to IscU through transpersulfuration reactions is effective [259, 285]. However, there is no evidence for iron-sulfur cluster formation upon the addition of iron to sulfur-containing forms of IscU either [194]. Further work is needed to show that this mechanism is possible.

3. The frataxin homolog CyaY binds Fe^{3+} and forms a complex with IscS, which is bound to IscU. A cysteine molecule persulfates IscS with release of alanine. Using a second cysteine as electron donor for Fe^{3+} reduction, Fe2+ is transferred to IscS. And IscU eventually contains the [2Fe-2S] cluster [153]. More details can be found in the reviews in Layer *et al.* [153] and Mansy *et al.* [173].

Yang *et al.* [301] studied the iron-sulfur cluster formation under physiologically relevant conditions. They found that IscU is the preferred scaffold protein when iron, L-cysteine, and IscS are present. When L-cysteine is not present in incubation solution, IscA acts as an iron chaperon, which binds 'free' iron. The iron binding in IscA appears to prevent the formation of inaccessible ferric hydroxide under aerobic conditions, subsequent addition of L-cysteine mobilizes the iron center in IscA and transfers the iron for the iron-sulfur cluster assembly in IscU even under aerobic conditions.

Reasons against the CyaY model: Li *et al.* found that deletion of CyaY did not affect cellular iron content and growth behavior [160]. This is in agreement with an IscU deletion being lethal since IscA is delivering iron but does not function as cluster assembly scaffold protein. However, CyaY may act au lieu of IscA. The iron binding capacity/affinity differs between the references. While Layer *et al.* [153] listed a number of references showing a high iron binding capacity for CyaY, Yang *et al.* [301] cited references, which claim poor binding affinities for CyaY. However, this model does not contradict the models rejected by Layer *et al.* [153].

5.5 Protein Folding.

The folding of nascent polypeptides is achieved via three distinct pathways: i) spontaneous folding; ii) through the DnaKJ-GrpE system; and iii) through GroEL/ES chaperones [174, 104, 226]. Two recently published large-scale datasets [60, 140] were used for the assigning the folding pathway to the individual polypeptides. If no information was available, spontaneous protein folding was assumed. The template protein folding reactions were derived from various primary and review literature.

Trigger Factor. In the *E. coli* cytosol, nascent polypeptides interact first with trigger factor (TF) [61, 110, 275], that binds to the ribosome at proteins L23/L29 near the polypeptide exit site [147, 164]. Thus, TF displayed its shielding function only in its ribosome-bound state and specifically for nascent chains still connected to the peptidyl-transferase center [112]. TF has a naturally low affinity to nascent polypeptide, however, when bound to the ribosome, it can interact with polypeptide and protect it from protease digestion (degradation). Nascent polypeptide (unfolded) that leaves ribosome is not longer TF associated, and thus not protected by TF for degradation. The nascent polypeptide may be prefolded by action of trigger factor that is constantly associated with 50S ribosomal subunit, is then released from the ribosome. The deformylase and the methionine aminopeptidases act on nascent polypeptide (see above), then the DnaK system or GroEL/S folds the polypeptide to the functional protein.

DnaKJ-GrpE system. DnaK is a monomer [250]. It was found that DnaK binds to zinc in a global study of *E. coli*s zinc-binding proteins [136], however, no further evidence could be found in literature. DnaJ has 2 zinc binding centers one of which is crucial for the interaction with DnaK [166]. This might be the reason why Katayama *et al.* identified DnaK as zinc binding protein. DnaK has ATP bound; the first binding of ATP is assumed to be spontaneous because no protein for ATP transfer to DnaK has been reported. DnaK and GrpE have a stoichiometry of 1:2 [250, 102]. DnaK/J-GrpE dependent folding reactions were based on Fig. 3 in Hartl and Hartl [104]. TF and DnaK/J/GrpE system can be knockout individually but combined deletion is lethal.

GroEL/ES chaperones Unlike TF and DnaK, the cylindrical chaperonin complex GroEL and its cofactor GroES are absolutely essential in *E. coli* and act posttranslationally in the folding of a subset of cytosolic proteins (10% of total), most of which are below 60 kDa in size [89, 104].

Most chaperonin substrates are medium-size proteins, between 25 and 60 kDa. This observed size distribution suggests that very small proteins do not need the protected environment of the chaperonin cavity to fold. Conversely, proteins too large to fit are presumably composed of smaller individual domains that can fold co-translationally [89].

In general, a polypeptide has to go through multiple rounds of GroEL/ES folding until it reaches its final conformation. For this, it is released and recaptured by GroEL/E [104]. The folding requires about 100 ATP for Rhodanse and DHFR [174]. Since the amount of required ATP might differ from protein to protein we decided to ac-

Table 5.2: Template reactions for DnaK/J-GrpE dependent folding.

Reaction abbreviation	**Reaction name**	**Reaction**
xxx_fold_KJE_1	xxx_m folding: KJE mediated	1 xxx_m(_ions) + 1 DnaK_mono.ATP + 1 DnaJ_dim → 1 xxx_m_DnaKJ_complex
xxx_fold_KJE_2	xxx_m folding: KJE mediated	1 xxx_m_DnaKJ_complex + 1 GrpE_dim + 1 h2o −¿ 1 xxx_DnaK_GrpE_complex + 1 pi + 1 h + 1 DnaJ_dim_inact + 1 adp
xxx_fold_KJE_3	xxx_m folding: KJE mediated	1 xxx_DnaK_GrpE_complex + 1 atp −¿ 1 xxx_mono + 1 DnaK_mono.ATP_inact + 1 GrpE_dim_inact

Table 5.3: Template reactions for GroEL/ES dependent folding.

Reaction abbreviation	**Reaction name**	**Reaction**
xxx_fold_GroEL/ES_1	xxx_m folding: GroEL/ES mediated; polypeptide is going in GroEL/ES complex	1 xxx_m(_ions) + 1 GroEL.(7)ADP.cisGroES + 1 transGroES_hepta + 7 atp → 1 xxx_m_GroEL.(7)ATP.transGroES + 1 cisGroES_hepta + 7 adp
xxx_fold_GroEL/ES_2	xxx_m folding: GroEL/ES mediated; folding of polypeptide under ATP hydrolysis	xxx_m_GroEL.(7)ATP.transGroES + 7 h2o → 1 xxx_GroEL.(7)ADP.transGroES + 7 h + 7 pi
xxx_fold_GroEL/ES_3	xxx_m folding: GroEL/ES mediated; release of native protein	xxx_GroEL.(7)ADP.transGroES → 1 xxx_mono + 1 GroEL.(7)ADP.transGroES

count only for the cost of one round of folding. Hartl and Hayer-Hartl provide a very clear and comprehensive reaction mechanism in their review [104]. (See also Table 5.3).

5.6 mRNA degradation.

The mRNA degradation in the 'E-matrix' is carried out solely by the degradosome. The reconstruction degradosome consists only of the components that have been found to be necessary or essential for its action (Eno, Pnp, RNase_E, RhlB) but not accessory factors (such as DnaK, GroEL, PPK, PAP, S1) [286, 47, 42, 143]. No degradation of stable RNA was modeled, since tRNAs and rRNAs are known to be highly stable under normal growth conditions and are believed to be insensitive to the decay processes that lead to the fast turnover of short half-lived mRNAs [182, 23, 27]. Functions of different RNase identified in *E. coli* that were not included in the 'E-matrix' either because they are outside the scope of the reconstruction or their function is not well established:

- **RNase I** is responsible for stable RNA decay, especially 23S rRNA [92, 82].
- **RNase M** is mutated form of RNase I [267].
- **RNase LS** was not included because little is presently known about its precise function, although it is believed to be involved in mRNA decay [201].
- **RNase T** is responsible for tRNA turnover [305].
- **RNase R**: Cheng *et al.* [44] showed that RNAse R is important in mRNA decay. It seems that this RNase acts on mRNAs with high secondary structures (REP elements) and that it can replace PNPase action. However, PNPase, together with other degradosome proteins, can digest mRNAs with (complex) secondary structures. It has been found that RNase R amount increase with stress conditions [44] . PNPase/RNase R both need a longer sequence before structural elements to dock on and to do degrade. Hence, it is thought that those RNA fragments get a poly(A) tail by PAP I and PAP II. It seems that RNase II has a rather protective function than degradative. Furthermore, one ribonuclease of RNase II, RNase R, PNPase can be deleted. The double deletion of RNase II and RNase R was found to be viable but double deletions of PNPase/RNaseII or PNPase/RNase R were lethal. RNase R also degraded stable RNA (whereby tRNA is poor substrate, but defectuous tRNA was a good substrate [44]). Deutscher *et al.* [62] said that the function of RNase II and

RNase R will have to be re-evaluated. Hence, the reconstruction does not account for RNase R action. PNPase, together with the other degradosome proteins, seems to be sufficient under normal growth conditions.

Polyadenylation of mRNA was not included since its effect is not sufficiently established. While poly(A)-tail on some mRNAs lead to a prolonged half-life time (protective cap), it seems to be a degradation sign on other mRNAs [184, 244, 101, 121, 44].

5.7 tRNA and rRNA processing.

The tRNA and rRNA modification reactions were created sequentially, allowing the representation of each modification reaction within the network. The tRNA modification positions were obtained from a tRNA database [262]. Each tRNA sequence obtained from the genome sequence and its genome coordinates was aligned to the tRNAs listed in the tRNA database, since different nomenclature was used in the databases. For three *E. coli* tRNA, namely ProK (b3545), ProL (b2189), and ProM (b3799), no entries were available in the tRNA database and no reports of identified modifications could be found. *S. typhi* modifications were used for these three tRNAs (tRNA database entries RP1700, RP1701, and RP1702). The positions for rRNA modifications were obtained from the RNA modification database [165, 241, 175]. The formula for each modified nucleotide (tRNA and rRNA) was calculated based on its structure found in [184], RNA modification databases [165, 241, 175], and primary literature. The modification reactions were obtained based on primary literature. In some cases, a consensus mechanism could not be derived from the literature, thus, the most popular mechanism was chosen for the reconstruction.

Chapter 6

Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery

Metabolic network reconstructions represent valuable scaffolds for 'omics' data integration and are used to computationally interrogate network properties. However, they do not explicitly account for the synthesis of macromolecules (i.e., proteins and RNA). Here, we present the first genome-scale, fine-grained reconstruction of *E. coli*'s transcriptional and translational machinery, which produces 423 functional gene products in a sequence-specific manner and accounts for all necessary chemical transformations. Legacy data from over 500 publications and three databases were reviewed and many pathways were considered, including stable RNA maturation and modification, protein complex formation and iron-sulfur cluster biogenesis. This reconstruction represents the most comprehensive knowledge base for these important cellular functions in *E. coli* and is unique in its scope. Furthermore, it was converted into a mathematical model and used to: 1) quantitatively integrate gene expression data as reaction constraints, and 2) compute functional network states, which were compared to reported experimental data. For example, the model predicted accurately the ribosome production, without any parameterization. Also, *in silico* rRNA operon deletion suggested that a high RNA polymerase density on the remaining rRNA operons is needed to reproduce the reported experimental ribosome numbers. Moreover, functional protein modules were determined and many were found to contain gene products from multiple subsystems highlighting the functional interaction of these proteins. This genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery presents a milestone in Systems Biology since it will enable quantitative integration of 'omics' datasets and thus to study the mechanistic principles underlying the genotype-phenotype relationship.

6.1 Introduction

High-throughput experimental technologies enable the production of heterogeneous data, such as expression profiles and proteomic data, for almost any organism of interest. A detailed mathematical representation of the *in vivo* cellular network is required to obtain a holistic understanding of cellular processes from these data sets and to quantitatively integrate them into a biological context. One such approach is the bottom-up network reconstruction, which builds manually networks in a brick-by-brick manner using genome annotation and component-specific information (e.g., biochemical characterization of enzymes) [229, 80]. This reconstruction procedure is well established for metabolic reaction networks and has been applied to many organisms, including Human [66], *Saccharomyces cerevisiae* [67, 148], *Leishmania major* [43], *Escherichia coli* [78], *Helicobacter pylori* [280], *Pseudomonas aeruginosa* [195], and *Pseudomonas putida* [187].

These bottom-up metabolic networks differ from other network reconstructions as they are tailored to the genomic content of the target organism and built manually using biochemical, physiological, and other experimental information in addition to the genome annotation. Hence, these reconstructions can be thought of as biochemically, genetically, and genomically structured (BiGG) knowledge bases [206]. The reconstruction and modeling procedure is a 4-step process:

- obtaining a draft reaction list based on genome annotation and biochemical databases,
- refinement of reaction list using experimental information (e.g., from literature),
- conversion of the reaction list (reconstruction) into a computable format and application of systems boundaries to define condition-specific models, and
- the evaluation and validation of the model content using various mathematical methods (see also [206, 229, 278, 80]).

By iterating step 2 to 4, reconstructions that are self-consistent within their defined scope can be generated.

Metabolic network reconstruction have demonstrated to be useful in at least five areas of applications [80]: i) biological discovery [229], ii) phenotypic behavior [280], iii) bacterial evolution [85], iv) network analysis [6], and v) metabolic engineering [210]. This wide range of applications of the metabolic reconstructions is possible because they can be readily converted into predictive, condition-specific models. Unlike more traditional approaches to modeling metabolism, the constraint-based modeling approach (COBRA) requires few, if any, parameters [218, 206]. The stoichiometric information encoded in the reconstruction (i.e., reaction list) can be represented mathematically as a stoichiometric matrix, S, where the rows correspond to the components and the columns correspond to the reactions (Figure 6.1).

While the COBRA approach has been successfully applied to metabolic networks, the same principles and assumptions can be also employed to reconstruct and model other cellular functions, such as signaling [207, 58, 161], regulation [95], and protein synthesis [5]. In this study, we extended and refined earlier work by Allen *et al.*, which proposed a stoichiometric formalism to model protein synthesis and illustrated it on some *E. coli* genes and operons [5]. We created a more detailed, gene-specific representation of the transcriptional and translational processes, which explicitly accounts for the sequence-specific synthesis of DNA, mRNA, and proteins. This reconstruction enables quantitative integration of high-throughput data such as gene expression, proteomic, and mRNA degradation data. Moreover, proteins are produced in high copy numbers in growing cells; thus, any quantitative mechanistic modeling and analysis of high-throughput data needs to account for the synthesis cost associated with these molecules.

Numerous studies have been published that investigate protein synthesis using kinetic models [177, 176, 270, 304, 272]. These models are generally tailored to the questions they address making it difficult to readily apply them for modified problems. Since stoichiometric relationships are a common requisite for any type of mechanistic modeling, organism-specific BiGG knowledge bases can be used as templates to derive problem-specific, mechanistic models (Figure 6.1). In fact, network stoichiometry is a dominant feature of kinetic models as well [123]. Thus, network reconstruction serves as a platform

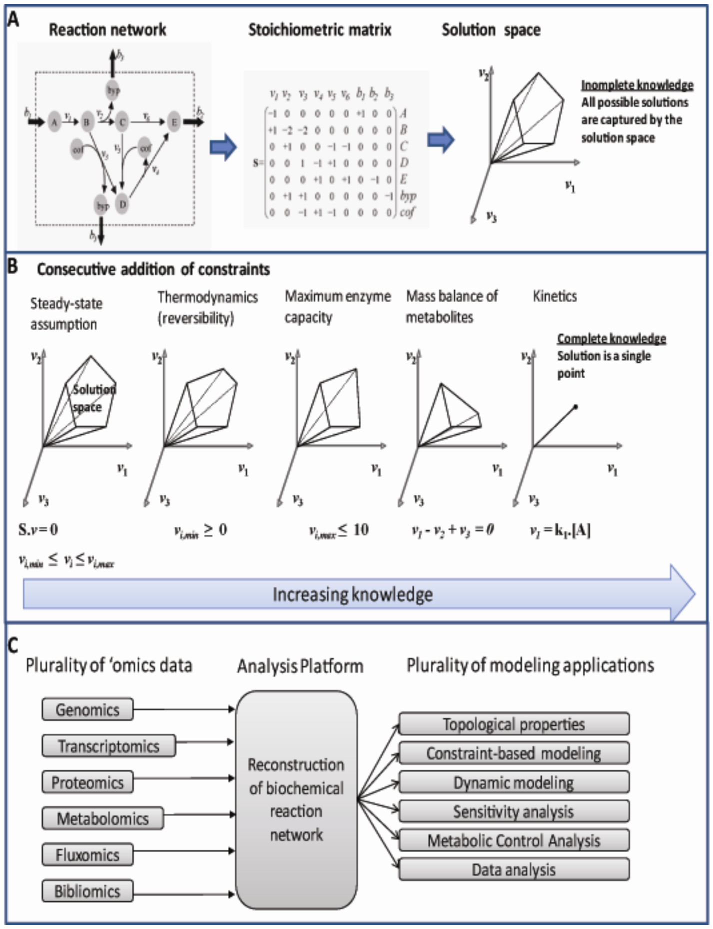

A
Reaction network
Stoichiometric matrix
Solution space
Inomplete knowledge
All possible solutions are captured by the solution space
B
Consecutive addition of constraints
Steady-state assumption
Thermodynamics (reversibility)
Maximum enzyme capacity
Mass balance of metabolites
Kinetics
Complete knowledge
Solution is a single point
Solution space
S.v=0
$v_{i,min} \leq v_i \leq v_{i,max}$
$v_{i,min} \geq 0$
$v_{i,max} \leq 10$
$v_1 - v_2 + v_3 = 0$
$v_1 = k_1 \cdot [A]$
Increasing knowledge
C
Plurality of 'omics data
Analysis Platform
Plurality of modeling applications
Genomics
Transcriptomics
Proteomics
Metabolomics
Fluxomics
Bibliomics
Reconstruction of biochemical reaction network
Topological properties
Constraint-based modeling
Dynamic modeling
Sensitivity analysis
Metabolic Control Analysis
Data analysis

Figure 6.1: **Overview of constraint-based reconstruction and analysis. A.** Schematic illustration of the conversion of a biochemical reaction network into a mathematical format (stoichiometric matrix, S). Since there are normally less columns (reactions) than rows (metabolites) there does not exist a single solution but rather a steady-state solution space containing all possible solutions. **B.** The successive addition of constraints will shrink the solution space by eliminating biologically infeasible steady-state solutions. Complete knowledge would reduce the steady-state solution space to a single solution. Since complete knowledge is not available for the majority of biochemical reaction networks the investigation of properties and capabilities of the solution space is very useful. **C.** This graphic illustrate the central role of reconstruction of biochemical networks to systems biology and how they serve as a foundation for many applications and problem-specific models.

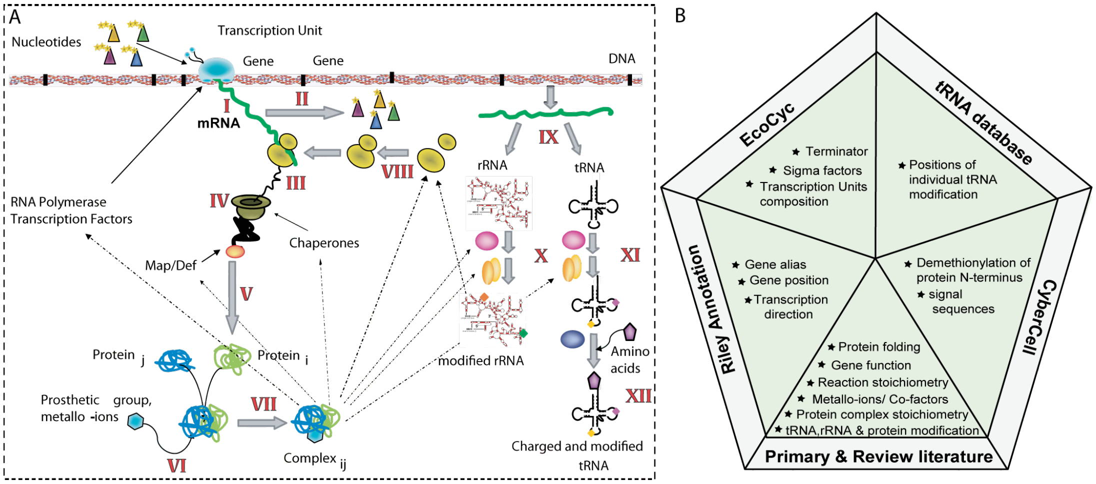

A
Nucleotides
Transcription Unit
Gene
Gene
DNA
I
mRNA
II
III
VIII
IX
rRNA
tRNA
RNA Polymerase
Transcription Factors
IV
Chaperones
Map/Def
V
X
XI
Protein j
Protein i
modified rRNA
Amino acids
Prosthetic group, metallo -ions
VII
XII
VI
Complex ij
Charged and modified tRNA
B
EcoCyc
tRNA database
Riley Annotation
CyberCell
Primary & Review literature
★ Terminator
★ Sigma factors
★ Transcription Units composition
★ Positions of individual tRNA modification
★ Gene alias
★ Gene position
★ Transcription direction
★ Demethionylation of protein N-terminus
★ signal sequences
★ Protein folding
★ Gene function
★ Reaction stoichiometry
★ Metallo-ions/ Co-factors
★ Protein complex stoichiometry
★ tRNA,rRNA & protein modification

Figure 6.2: **Content of the 'E-matrix'. A.** Schematic representation of the network components and reactions is shown. In addition to the macromolecular synthesis of RNA and proteins, rRNA and tRNA processing reactions were included in the reconstruction. I: Transcription; II: mRNA degradation; III: translation; IV: protein maturation, V: protein folding; VI: metallo-ion binding; VII: protein complex formation; VIII: ribosome assembly; IX: RNA processing; X: rRNA modification; XI: tRNA modification; XII: tRNA charging (see Table 6.2 for complete list of subsystems.**B.** The pentagram shows the five main data sources incorporated in the 'E-matrix': EcoCyc [133], CyberCell [269], and tRNA DB [262], the revised genome annotation [237], and the genome sequence (m56, [29]).

for steady-state and kinetic modeling (Figure 6.1).

In this study, we present a new generation of network reconstructions, which directly account for the synthesis of individual mRNA and proteins (Figure 6.2A). We named the mathematical representation of this reconstruction the Expression matrix, or 'E-matrix', since it encodes the expression of mRNA and proteins. All network reactions were formulated to account for gene-specific and *E. coli*-specific details, such as nucleotide composition, operon association, and sigma factor usage. Furthermore, we used information from three databases and more than 500 scientific publications to formulate mechanistically detailed and accurate reactions. This reconstruction is the first comprehensive database detailing the available information for these cellular functions and can thus be deemed a knowledge base. After conversion of the 'E-matrix' reconstruction into condition-specific models corresponding to different doubling times, we were able to accurately predict the ribosome production reported in literature, without any parameterization. Furthermore, we show that the 'E-matrix' can be used to study the effect of rRNA operon deletion. Our results predict that a high density of RNA polymerases is required on the remaining rRNA operons, to achieve the reported ribosome numbers. Finally, we show that proteins used in the 'E-matrix' could be grouped into functional modules which lead to a more simplified view of the network.

6.2 Methods

Reconstruction procedure. The reconstruction process of any biological network depends heavily on the quality of the genome annotation and the amount of experimental data available for an organism [229]. The transcriptional and translational machinery of *E. coli* was selected for reconstruction, as *E. coli* is one of the most extensively studied organisms [183]. We aimed to create a high-resolution reconstruction that would accurately account for the cellular processes necessary to produce functional gene products of this machinery (Figure 6.2).

The manual reconstruction of the transcriptional and translational machinery of *E. coli* was performed in an algorithmic manner (Figure 6.2). First, the identification of its key components in the genome annotation resulted in an initial component list. Then, the functional roles of these key components were identified and translated into stoichiometrically accurate reactions using textbooks, reviews, and primary literature (Figure 6.2). This step led to the identification of components missing in the initial component list, which were subsequently added to the network. In the end, this reconstruction approach led to the identification of 228 proteins and 109 RNA species, which are directly involved in one or more subsystems (Figure 6.2).The synthesis reactions for every network component were created using template reactions because each of the similarity of reactions between the network components. These template reactions were carefully formulated and derived from primary and review literature. While they combine linear steps (e.g., elongation of nascent mRNA during transcription) they separate key reactions and known rate limiting steps (e.g., separation of transcription initiation and elongation). This enables the incorporation of different sets of constraints but also enables the reduction of the network size by combining linear template reactions. Hence, this step-wise representation captures key event in cellular processes and can be directly used to understand their pathway/reaction mechanism at a high resolution.

Template reactions. Once the main factors involved in the various processes, or subsystems, were identified the reactions carried out by one or an ensemble of these components were defined based on up-to-date literature. For a majority of the network reactions, we used the fact that the reactions were very similar for every gene or gene product. For example, the transcription initiation and elongation involves the RNA polymerase and transcriptional factors such as NusA and NusB for all genes. Subsequently, the reaction formulation of most of the network reaction could be done based on template reactions (see Chapter 5). The template-based network reconstruction was performed using the scripting language, Perl (http://www.perl.com/). Each template reaction as well as protein complex formation reactions were generated manually based on legacy data.

The basis for the reconstruction is the genome sequence, m56 [29], the most current gene coordinates from [237], and the transcription unit definition provided by EcoCyc (downloaded version 10.6, [133]). This information, in addition to the other data resources, were used to i) calculate formula and charge for each mRNA and protein species; ii) individually adjust the template reaction, i.e., the number of each NTP needed for the transcription; and iii) enable the transcription of operons rather than genes. The transcriptional and translational reactions were formulated for all gene product involved in the machinery as well as for genes being part of their operons, e.g. only one of four operon genes is included in the machinery but all three other genes have to be transcribed, and translated, to enable the formation of the 'E-matrix' RNA product.

Mass and charge balancing. For each network component the corresponding chemical formula, charge state at pH 7.2 [229], and molecular weight were calculated. Almost all (99.5%) of the reactions in the 'E-matrix' are mass and charge-balanced. The remaining reactions were left unbalanced for two reasons: i) unknown electron acceptor e.g. iron-sulfur-cluster biogenesis; or ii) alternate precursors with different formulae, e.g., transcription product from different overlapping transcription units, which were combined in subsequent reactions to reduce the overall number of network reactions.

The systems boundaries of the 'E-matrix' were defined by adding 76 exchange reactions for amino acids, NTP, and other metabolic components. Furthermore, demand reactions were added for each protein gene product.

Iterative network reconstruction and QC/QA. A comprehensive, iterative quality control/ quality assurance procedure (QC/QA) ensured that the resulting network has similar properties and capabilities as *E. coli.* This QC/QA procedure included the mass- and charge balancing of most network reactions,gap analysis, and testing for the production of every network component and its intermediate form. Hence, this reconstruction follows the quality control standards developed for metabolic network reconstructions [229].

After an initial reaction list was created, as described above, a network gap analysis was performed. This procedure was an iterative process, as for metabolic network reconstructions [229], during which further components and reactions were added (Figure 6.2). Multiple iterations helped ensure completeness of the network within the predefined scope. Furthermore, flux balance analysis calculations were carried out to verify that every network component could be produced and consumed. Therefore, a demand function for every network component was independently added to the network and maximized using linear programming. These quality control tools ensured that the final network was comprehensive and functional. Only one network gap remained, which is the ribonuclease PH (RNase_PH) whose gene was found to be a pseudogene in *E. coli* MG1655 genome [237]. Experimental studies characterized this gene product but in different *E. coli* strains [163, 271, 200].

Constraint-based modeling. The mathematical model of the 'E-matrix' was represented by a stoichiometric matrix, S (m rows x n columns), where m is the number of components and n is the number of reactions [229]. Reactions within the network were mass-balanced and assumed to be at steady state such that $S \cdot v = 0$, where v is flux vector. Additional upper, $v_{i,max}$, and lower, $v_{i,min}$, bounds were applied in form of $v_{i,min} \leq v_i \leq v_{i,max}$ on each reaction i. The lower limits were set to zero for irreversible reactions. The unit for each reaction flux was defined to be $\frac{nmol}{g_{DW} \cdot doubling\ time(min)}$, if not stated differently.

Simulation constraints. The upper bounds on exchange reactions for NTPs and amino acids were constrained for all simulation conditions, while the lower bounds remained unconstrained. The fractional contribution of NTPs and amino acids were calculated based on experimental data [185] and scaled by RNA and protein content found at each doubling time. The upper bounds of stable RNA transcription initiation reactions were constraint based on experimental data [183] using the following formula: $v_{rRNA,max} = (\frac{genes}{cell}) \cdot i_{rrn} \cdot T_D$, where i_{rrn} is the rRNA transcription initiation rate and T_D the doubling time (see Chapter 2).Note that these transcription limitation constraints accounted for the different gene dosage caused by multiple replication forks and different rRNA operon positions on the chromosome (Table 6.1). The mRNA degradation rates were calculated using expression data in LB medium and mRNA half-life times [26] with $v_{degradation,max,i} = [mRNA]_i \cdot max(\frac{ln2}{T_{\frac{1}{2},LB,i}}, \frac{ln2}{T_{\frac{1}{2},M9,i}})$, assuming a total number of 4,600 mRNA per cell at 30 min doubling time [183]. The lower bound ($v_{degradation,min,i}$)was set to be 0. Since the expression data as well as the total mRNA number have experimental errors, the upper bound on each reaction flux had to be relaxed by multiplying each mRNA concentration with a factor of ten. The upper bound on mRNA recycling, or CONV2 reactions, were constrained using the following formula: $v_{CONV2,max,i} = [mRNA]_i \cdot T_D \cdot \frac{r_{elo}}{(\frac{L_{mRNA,i}}{3})}$, where T_D is the doubling time (s), $L_{mRNA,i}$ is the length of mRNA i, and r_{elo} is the translation elongation rate at T_D . This later set of reactions accounts for multiple translation rounds of an mRNA transcript between synthesis and degradation.

Ribosome production rate. The exchange flux rates and the transcription initiation rates of ribosomal RNA operons were constrained as described above. At each doubling time, the ribosome production rate (DM_rib_50) was chosen as objective function, and the maximal possible production rate under the given set of constraints was calculated using linear programming.

***in silico* rRNA operon deletion.** This analysis was carried out as illustrated in Figure 6.4. First, the transcription initiation rates were applied as constraints to all rRNA operons for the different doubling times (as described above). Using flux balance analysis (FBA) [289, 71] we optimized for ribosome production (DM_rib_50). For the strains deficient in one rRNA operon, we deleted each operon separately by setting the maximal possible transcription initiation rate to 0 ($v_{rRNA,max,i} = 0 \frac{nmol}{g_{DW} \cdot hr}$), which corresponds the deletion of the reaction from the network. We optimized again for the ribosome production. For multiple rRNA operon deficient strains, all possible combinations of rRNA operon deletion were considered (Table 6.1), leading to the error bars in Figure 6.4. The compensation factors were chosen arbitrarily (1.5, 2, 2.5, and 4) and multiplied to all active rRNA operons in the mutant strains. Note that the unit for these simulations was $\frac{nmol}{g_{DW} \cdot hr}$.

Table 6.1: **List of rRNA transcription units and their basic characteristics.** This information was obtained from the most recent genome annotation [237]. * Transcription unit names are listed as given by EcoCyc [133]. Promoter name is given in parenthesis. The gene number per cell (gene dosage) was calculated as described in Chapter 2. ^{a} Doubling time in minutes

TU*	Names	Alias	Strand	Coordinates (in bp)	$\frac{Genes}{cell}^a$					
					30	90	100	60	40	24
TU0-1181 (P1)	b3851-b3855	rrsA-ileT-alaT-rrlA-rrF	forward	4,033,554 - 4,038,659	4.49	2.07	1.92	2.37	3.24	6.17
TU0-1182 (P1)	b3968-b3971	rrsB-gltT-rrlB-rrfB	forward	4,164,682 - 4,169,779	4.24	2.01	1.87	2.29	3.10	5.77
TU0-1186 (P1)	b4007-b4010	rrsE-gltV-rrlE-rrfE	forward	4,206,170 - 4,211,182	4.17	1.99	1.85	2.27	3.06	5.64
TU0-1189 (P1); TU0-1190 (P2)	b0201-b0205	rrsH-ileV-alaV-rrlH-rrfH	forward	223,771 - 228,875	3.15	1.72	1.62	1.93	2.45	4.00
TU0-1187 (P1); TU0-1188 (P2)	b2588-b2591	rrsG-gltW-rrlG-rrfG	complement	2,727,638 - 2,724,210	2.81	1.62	1.54	1.80	2.25	3.49
TU0-1191 (P1); TU0-1192 (P2)	b3272-b3278	rrsD-ileU-alaU-rrlD-rrfD-thrV-rrfF	complement	3,425,243 - 3,421,564	3.79	1.90	1.77	2.15	2.84	5.02
TU0-1183 (P1); TU0-1184 (P2)	b3756-b3759	rrsC-gltU-rrlC-rrfC	forward	3,939,831 3,944,842	4.67	2.12	1.95	2.42	3.35	6.48

Flux variability analysis. Flux variability analysis was performed as described by Mahadevan [170] using linear programming. Briefly, for every network reaction the minimal and maximal solution was determined by successively defining each network reaction as objective function. The lower bound of the ribosome production rate (DM_rib_50) was constrained to $v_{min} = 0.75 \cdot v_{max}$.

Correlation of protein utilization. The pair-wise correlations between protein component recycling reactions (PROT_RECYCL) were determined in LB-medium using linear programming. The maximal reaction flux for reaction A was determined and its upper and lower bound was set to be the maximal flux value. The minimal and maximal reaction flux for reaction B was determined under this new set of constraints. The same procedure was repeated for the minimal flux rate through reaction A. The same approach was repeated for reaction B with respect to reaction A. This method resulted in pair wise dependency plots for all recycling reactions. The area of feasible flux rates was determined using a convex hull algorithm [16] and scaled by the maximal flux rates for each reaction. The reaction correlation was defined to be 1 minus the area between two network reactions.

All calculation were performed using MatLab(The MathWorks, Inc, Natick, MA) and TomLab (TomLab Optimization, Inc, Pullman, WA).

Availability: This knowledge-base is freely available at
http://bigg.ucsd.edu/E-matrix

6.3 Results and Discussion

The 'central dogma' of molecular biology was first enunciated by Crick in 1958 and dealt with the transfer of sequential information from DNA to RNA to proteins [54]. The machinery necessary to conduct this information transfer was reconstructed in this study on a genome-scale, i.e., all known components in *E. coli* were considered. The 'E-matrix' encodes for all known reactions, which synthesize the components of the macromolecular synthesis machinery, in a mechanistically detailed fashion.

6.3.1 Legacy data.

The 'E-matrix' reconstruction was based on *E. coli*-specific information derived from more than 500 primary and review publications, three databases, and the revised genome annotation [237] (Figure 6.2B). This detailed information enabled the sequence-specific formulation of synthesis reactions, at high resolution, for every network component, namely DNA, mRNA, proteins, protein complexes, and metabolites. The reconstructed network accurately represents all known reactions required to produce the active, functional components of the transcriptional and translational machinery in *E. coli* (Figure 6.2A).

6.3.2 Reconstruction approach.

The manual reconstruction of the 'E-matrix' was performed in an algorithmic manner by first identifying key components in the genome annotation. The functional roles of these key components were determined and then translated into stoichiometrically accurate reactions using multiple data sources (Figure 6.2B). A total of 303 components (proteins and RNA) were found to be directly involved in one or more subsystems, which represent groups of functionally related transformation pathways (Table 6.2. In this reconstruction linear transformation steps, e.g., elongation of nascent mRNA during transcription, were combined into a single reaction, while key reactions and known rate limiting steps were kept as separate reactions, e.g., transcription initiation and elongation. This representation captures key events in cellular processes and can be directly used to understand their reaction mechanisms at a high resolution.

A comprehensive, iterative quality control/quality assurance (QC/QA) procedure ensured that the resulting network had similar properties and capabilities as *E. coli*. This QC/QA procedure included gap analysis, testing for the production of every network component, and mass- and charge-balancing of more than 99% of the network reactions. Hence, the 'E-matrix' reconstruction follows the quality control standards developed for metabolic network reconstructions [229].

6.3.3 Unique properties of the 'E-matrix'.

This reconstruction is unique in the depth and breadth of information included as well as an advancement of other transcriptional and translational networks currently available [177, 176, 270, 304, 272]. It is also the largest reconstructed network to date,

Table 6.2: **Reactions per subsystems.** The numbers I to XII correspond to the numbering shown in Figure 6.2A.

Number	Subsystem	Reactions
I	Transcription	783
II	mRNA degradation	628
III	Translation	6,812
IV	Protein maturation	628
IX	RNA processing	122
V	Protein folding	570
VI	Metallo-ion binding	128
VII	Protein complex formation	87
VIII	Ribosomal assembly	13
X	rRNA modification	864
XI	tRNA modification	1,597
XII	tRNA charging	177
XIII	Aminoacyl-tRNA synthetase charging	33
XIV	Charging EF-Tu	4
XV	Cleavage polycistronic mRNA	222
XVI	Demands	302
XVII	Exchange reactions	76
XVIII	Iron-sulfur cluster biosynthesis	6
XIX	Iron-sulfur cluster incorporation	6
XX	Protein modification	12
XXI	Protein recycling	148
XXII	Ribosomal protein modification	21
XXIII	rRNA formation	38
XXIV	Sinks	35
XXV	Transcription regulation	261
XXVI	Transport	76
XXVII	tRNA activation (EF-TU)	45
	Total number of reactions	13,694

with 11,991 components and 13,694 reactions (Table 6.3). The 'E-matrix' accounts for all known gene products necessary to produce the active components of the machinery itself, and is therefore self-contained. Furthermore, sequence-dependent synthesis reactions were carefully formulated to incorporate known reaction stoichiometry including protein-substrate complex intermediates, metallo-ions and cofactors. Necessary modifications of stable RNA and proteins were also considered. Additionally, the transcription reactions were formulated in terms of transcription units rather than genes, providing a biologically accurate representation of operon organization in bacterial genomes. These reactions can be readily extended to account for the production of other gene products such as metabolic enzymes or transcription factors. Lastly, this framework facilitates future integration of the 'E-matrix' reconstruction with the metabolic and regulatory network of *E. coli.*

Table 6.3: **Overview of the 'E-matrix' content.** * involved refers to those gene products that are functionally involved in 'E-matrix' processes compared to genes that were included because of co-transcription with involved genes.

Number of transcription units	249
Number of genes (involved*)	423 (303)
Number of genes with/ without transcription unit	411/12
Number of components (with/ without genes)	337 (303/34)
-tRNA	86
-rRNA	22
-miscellaneous RNA	1
-involved* proteins (with/ without genes)	228(194 /34)
Number of subsystems	27
Number of reactions	13,694
-Number of demand reactions	302
-Number of exchange reactions	76
Number of network components	11,991
Number of references	+500

6.3.4 'E-matrix' versus available databases.

The 'E-matrix' is distinguished from available online databases, such as KEGG[131] and EcoCyc [133], as all transcriptional, translational, and modification reactions were defined in a sequence dependent manner for every included *E. coli* gene. This task was achieved by determining the nucleotide and amino acid composition of each DNA, RNA and protein from the genome sequence, respectively. Furthermore, we determined the elemental composition of these macromolecules and mass balanced all network reactions. In contrast, KEGG [131] and EcoCyc [133] list mainly generic reactions using gene- and organism independent terms such as 'DNA', 'protein', and 'RNA'. Subsequently, they contain only a subset of the synthesis reactions present in the 'E-matrix'. Furthermore, neither of these databases can be directly converted into a comprehensive, self-consistent mathematical format that permits rigorous computational characterization of network fluxes. Another difference between the 'E-matrix' and these databases is the extent of mechanistic detail incorporated into the 'E-matrix', such as rRNA and tRNA modification reactions, iron-sulfur cluster formation, chaperone-dependent protein folding and protein complex formation.

6.3.5 Knowledge gaps.

The transcriptional and translational machinery is essential for cellular growth. Considering the wealth of information available for *E. coli*, it was surprising to discover numerous knowledge gaps, or missing information, during the reconstruction process. For example, reaction mechanisms for some RNA modifications and iron-sulfur cluster biogenesis were either poorly understood or a general consensus on the mechanistic details was lacking. For instance, 15% of the included proteins had no gene annotation and their existence was suggested in the literature solely based on identification of modified proteins or stable RNA. Furthermore, there are three metabolites with unknown metabolic transformations. One of these metabolites is $preQ_0$, a precursor of $preQ_1$, which is important for the queuosine formation in some tRNA (position G^{34}). This precursor is formed from GTP and it has been suggested that two ribose units of two GTP molecules contribute to the formation of three carbons in $preQ_0$ (C_5,C_6, and cyano carbon) but further information is missing [260, 268]. The two other missing metabolites are byproducts of the formation of uridine-5-oxyacetic-acid at position U^{34} in some tRNA. It has been suggested that chorismate acts as precursor for this nucleotide modification, however, such reaction would release two metabolites with formulae of $C_{10}H_8O_5$ and $C_9H_9O_4$, which have not been characterized yet [260, 268]. All of the knowledge gaps were highlighted in the reconstruction and associated with notes about currently available information, which will hopefully promote their elucidation as it has been the case for some of the metabolic knowledge gaps in *E. coli* [229].

6.3.6 Network topology.

The 'E-matrix' has a relatively 'linear structure' with only few components participating in multiple reactions since a majority of network components are only transferred from one reaction to another. This linearity is a dominant feature of the 'E-matrix' and it is less evident in metabolic reconstructions due to their much higher connectivity. Analysis of the component connectivity of the 'E-matrix' showed that the highest connected components are protons, water, and orthophosphate, which participate in 44%, 39%, and 32% of reactions, respectively. These compounds are also found to have the highest connectivity in metabolic networks [20]. In contrast to metabolic networks, ATP and ADP were not the next most highly connected but rather GTP and GDP, which participated in the numerous translational reactions. While the ATP requirement for cellular functions is accounted for in the biomass reaction of metabolic reconstructions, the high GTP requirement is not generally considered [78].

The conversion of a network reconstruction into a mathematical model can be achieved, analogously to metabolic networks [229], by defining system boundaries and applying condition-dependent constraints on exchange and intracellular reactions (Figure 6.1) [219, 229]. Therefore, experimental data can be used to constrain the set of feasible network fluxes in a physiologically relevant manner. In the following section, we will illustrate the use of condition-specific models that were derived from the 'E-matrix' reconstruction.

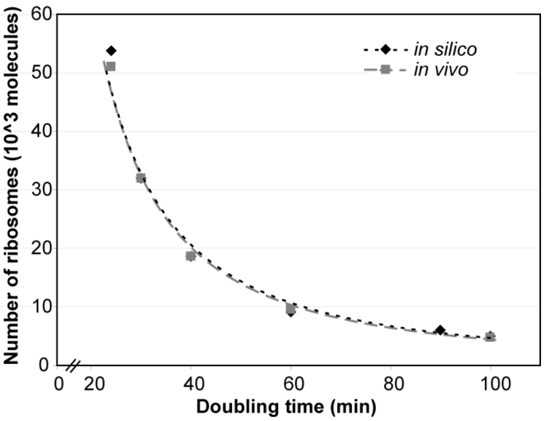


Figure 6.3: **Comparison of *in vivo* [183] and *in silico* maximal number of ribosomes at different doubling times.** Two sets of constraints were applied to the models: uptake rates for amino acids and NTPs, and maximal possible rates on stable RNA transcription initiation (see text for more details).

6.3.7 Validation of the 'E-matrix' functionality - Ribosome production.

Cell growth is directly correlated with the protein synthesis capacity and thus with the number of active ribosomes [188]. Accordingly, we used the model's ribosome production capability as an indicator of its ability to support growth. For every growth rate, the uptake rates for NTP and amino acids as well as the transcription initiation rates of the rRNA operons were quantitatively constrained based on experimental data [183]. The *in silico* computed ribosome production capabilities showed very good agreement with the reported *in vivo* ribosome production capabilities [183] for all investigated doubling times (Figure 6.3), indicating that the capabilities of the reconstruction were very similar to those of an *E. coli* cell. This overlap between experimental data and predictions was somewhat expected as the constraints used, i.e., stable RNA transcription initiation rates as upper constraints for the rRNA operons (see Material & Methods), were dominant (governing) constraints. Thus, these results validated the predictive capability of the reconstructed network. Moreover, our results show that: i) the network is capable of reproducing experimentally reported ribosome number given the uptake constraints, and ii) an increase in transcription initiation rate would lead to an increase of ribosome production (see also Figure 6.4B). This latter result implies that the regulation of rRNA synthesis, which is outside the scope of this reconstruction, plays a significant role in determining the transcription rate [189].

6.3.8 The effect of *in silico* rRNA operon deletions on ribosome production.

The *E. coli* genome contains seven rRNA operons, which have similar structures (16S rRNA, tRNA, 23S rRNA, tRNA, 5S rRNA, and, in some cases, tRNA). Generally, it is assumed that rRNA operon redundancy in *E. coli* and other species, has evolved to provide high levels of ribosomes and thus to support rapid growth rates [190]. However, there is experimental evidence that rRNA operon multiplicity is rather required for rapid adaptation to changes in physiological conditions [49, 264]. In fact, it has been shown that the presence of only one rRNA operon on the chromosome is sufficient for synthesis of 56% of the wild-type rRNA concentration [10] and the deletion of multiple rRNA operons had only small effect on growth rate and ribosome content [49, 50, 10]. Subsequently, it was experimentally observed that the remaining rRNA operons were able to compensate for the loss by increasing the transcriptional rate [49].

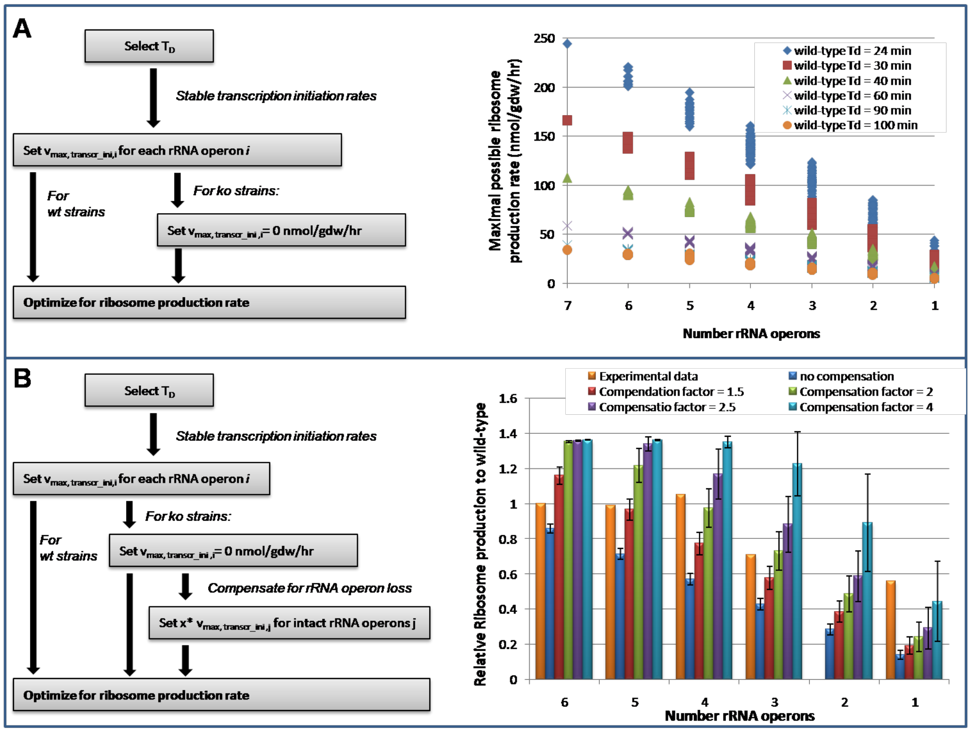


Figure 6.4: **rRNA operon deletion study. A.** Analysis of the effect of rRNA operon deletion to the ribosome production capability of the network. As expected, the ribosome production rate decreased with decreasing number of available rRNA operons. All possible combinations of operon deletions were considered resulting in different maximal possible ribosome production rates for a given number of remaining rRNA operons. This is due the gene dosage effect since multiple replication forks are present at higher growth rates. **B.** Experimental data (orange bars, [10, 49]) suggested much higher ribosome production than we determined in A. This compensation is achieved by increasing the transcription rate of the remaining rRNA operon. We tested different possible compensation factors and compared the results with the experimental data. The error bars are again caused by different combination of rRNA operons.

Since the early days of the development and application of COBRA methods, *in silico* gene deletion analysis has been productively used to evaluate the consequences of gene deletions to metabolism and cellular growth [74, 87, 51, 280]. Here, we used the same approach to evaluate the consequences of rRNA operon multiplicity to the ribosome production capabilities of the 'E-matrix' by *in silico* operon deletion analysis. First, we set the stable RNA transcription initiation rates based on doubling time as reported in Neidhardt *et al.* [185], and optimized for ribosome production using linear programming. Subsequently, we created single and multiple *in silico* knockout mutants by deleting the rRNA operons and optimized again for ribosome production (Figure 6.4). Since the maximal possible rRNA transcription rates were set to the reported rates, we observed a linear decrease in ribosome production for all tested doubling times (Figure 6.4). This result was expected as the stable RNA transcription initiation rates were found to be the governing constraints (see above). Therefore, this simulation setup did not allow for the compensation of rRNA operon loss.

To simulate this compensation, we multiplied the transcription initiation rate of each rRNA operon with various scaling factors and re-computed the maximal possible ribosome production rate (see Figure 6.4, and Materials & Methods). Comparison with experimental data [49, 10] showed that similar compensation could be obtained *in silico* by using a transcriptional compensation factor. The compensation factor had to be increased *in silico* when multiple rRNA operons were deleted. To compare the calculated compensation factor with experimental data, we converted the measured number of RNA polymerases (RNAP) per operon in rRNA operon deficient strains [49] into compensation factors by diving them with the reported RNAP binding frequency in the wild-type [185]. These experimental compensation factors in good agreement with our *in silico* results (data not shown). Surprisingly, it was found experimentally that strains with only one intact rRNA operon can still produce 56% of wild-type rRNA [10]. This situation would correspond to an *in silico* compensation factor of 4 and thus, to approx. 150 RNAP bound to the remaining rRNA operon. Since the average length of an rRNA operon is 5100 nucleotides, this high number of bound RNAP corresponds to a RNAP every 34 nucleotides. Such an increase in RNAP density on the operon could be achieved by increasing the transcription elongation rate and/or modulating the frequency of RNAP binding to the promoter [49]. It is not known which regulatory elements could lead to such an increase in rRNA transcription; however, Condon *et al.* found the ppGpp concentration, responsible for the stringent response under amino acid starvation, unaltered [49]. Gaal *et al.* showed that rRNA synthesis is regulated by NTPs, which stabilize the open complex of RNAP and P1 promoter of an rRNA operon. The formation of the open complex is necessary for successful transcription initiation [90]. Feedback inhibition is also controlling the rRNA synthesis, where an excess of ribosomes might regulate the transcriptional rate [189]. In agreement with our predictions, experimental data have shown an increase in ribosomal content for some rRNA deficient strains (Figure 6.4) [49]. Furthermore, different rRNA operon knockout combinations resulted in large differences in compensation due to different gene dosage depending on the positions of the various operons on the chromosome (Figure 6.4, Table 6.1). We did not determine the growth rates of the knockout strains as such calculation would require to assume the same correlation between doubling time and ribosome production as is present in wild-type *E. coli* (Figure 6.2). Our results suggest that the transcriptional initiation rate, and thus ribosome production rate, will be limited by competition for precursors, especially NTPs (data not shown). This agrees with the experimental observation that an increase in rRNA operon number will reduce the overall transcription initiation rate and thus maintain a constant rRNA content in the cell [293]. However, many complex regulatory mechanisms, which are outside the scope of the current model, are known to control ribosome production [189, 90]. The incorporation of regulation with the current model should lend further insight into the nature of rRNA operon multiplicity.

6.3.9 Integration of '-omics' data into 'E-matrix'.

An overall aim of this reconstruction effort was to create a stoichiometric representation of mRNA and protein synthesis machinery that allows the integration with experimental data. Interrogation of the data-constraint model would allow the investigation of the remaining network capabilities (Figure 6.5A). Here, we incorporated successively experimental data sets into the model as constraints, and investigated the resulting network capabilities. More specifically, we used the difference between minimal and maximal flux rate for each reaction (flux span) as a measure of constraint stringency. We successively integrated three different datasets (Figure 6.5):

- First, we constrained the upper bounds of exchange reactions in the 'E-matrix' to uptake rates corresponding to LB-medium conditions (Figure 6.5B). This set of constraints was not sufficient to eliminate biologically irrelevant solutions since, for in-

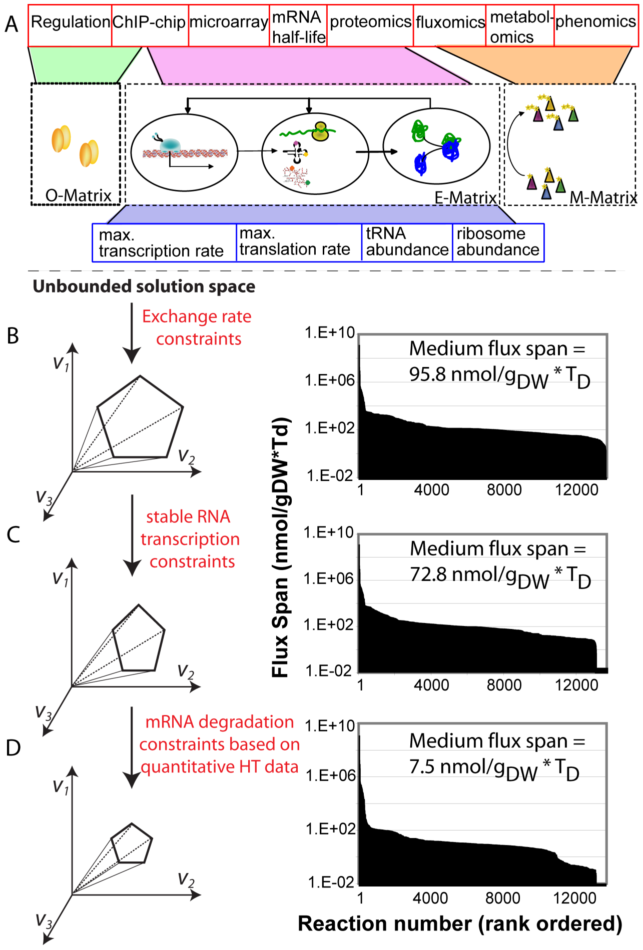

A
Regulation
ChIP-chip
microarray
mRNA half-life
proteomics
fluxomics
metabolomics
phenomics
O-Matrix
E-Matrix
M-Matrix
max. transcription rate
max. translation rate
tRNA abundance
ribosome abundance
Unbounded solution space
B
Exchange rate constraints
Medium flux span = 95.8 nmol/gDW * TD
C
stable RNA transcription constraints
Medium flux span = 72.8 nmol/gDW * TD
D
mRNA degradation constraints based on quantitative HT data
Medium flux span = 7.5 nmol/gDW * TD
v1
v2
v3
1.E+10
1.E+06
1.E+02
1.E-02
1
4000
8000
12000
Flux Span (nmol/gDW*Td)
Reaction number (rank ordered)

Figure 6.5: **Integration of "-omics" data into 'E-matrix' as reaction constraints. A.** This schema illustrates the types of high-throughput data (HT, red boxes) or low-throughput data (LT, blue boxes) that can be directly integrated with the 'E-matrix' as it accounts for the different macromolecules measured in these data sets. In contrast, the integration of regulatory information would require the formulation of the regulatory network in matrix format ('Operon' or 'O'-matrix). Furthermore, the metabolic network, here represented as 'M-matrix', would enable the mapping of fluxomic, metabolomic and phenomic data. **B-D.** Absolute flux span in 'E-matrix' while incorporating successively more complex constraints (see text for more details). **B.** LB-medium specific constraints were applied on exchange reactions. **C.** The upper bounds of stable RNA transcription initiation reactions were constrained. **D.** Additional constraints on upper bound of mRNA degradation flux rates were applied.

stance, the model was able to produce up to 45,000 ribosomes while approximately 30,000 ribosomes were observed experimentally [183].

- Second, further constraints were applied on the stable RNA transcription initiation rates based on low-throughput data [183] to exclude physiologically infeasible stable RNA transcription rates (Figure 6.5C). However, the maximal flux rates for synthesis reactions of most network mRNAs were still found to be too high when compared to expression data [26].
- Finally, we used high-throughput data, namely gene expression data from LB medium [26] and mRNA half life times [26], to further constrain the network. Numerical values for mRNA degradation rate, specific to each sequence of mRNA, were calculated based on these two data sets and applied as upper bounds on the mRNA degradation reactions in the network. This last set of constraints had a significant effect on the overall flux span, which highlights the importance of mRNA transcription constraints on the set of feasible solutions (Figure 6.5D).

A qualitative evaluation of mRNA expression in Boolean terms (on/off) - as used in metabolic modeling [51] - did not result in significant reduction of the size of the solution space (data not shown). Despite the mRNA degradation reaction constraints, many protein synthesis reactions still achieved high flux values. This result is consistent with the fact that low numbers of transcripts can be sufficient to synthesize high numbers of proteins and hence, the translation reactions can carry large flux rates. Thus, the application of quantitatively accurate proteomic data could greatly help to further constrain the set of feasible steady-state solutions.

6.3.10 Defining functional modules.

Correlated reaction sets (co-sets) have been calculated for metabolic networks to obtain insight into the network structure and properties [38, 280]. Here, we applied the same concept to the 'E-matrix' to identify functional coupling between proteins. In the reconstruction, every protein is associated with a recycling reaction representing its overall utilization rate in the cell. It can be expected that proteins whose utilization rates are perfectly correlated based on stoichiometry would show similar pattern of protein expression, but not necessarily of gene expression, under different environmental conditions. A total of 14 multi-protein modules (or co-sets) were identified accounting for 91 out of

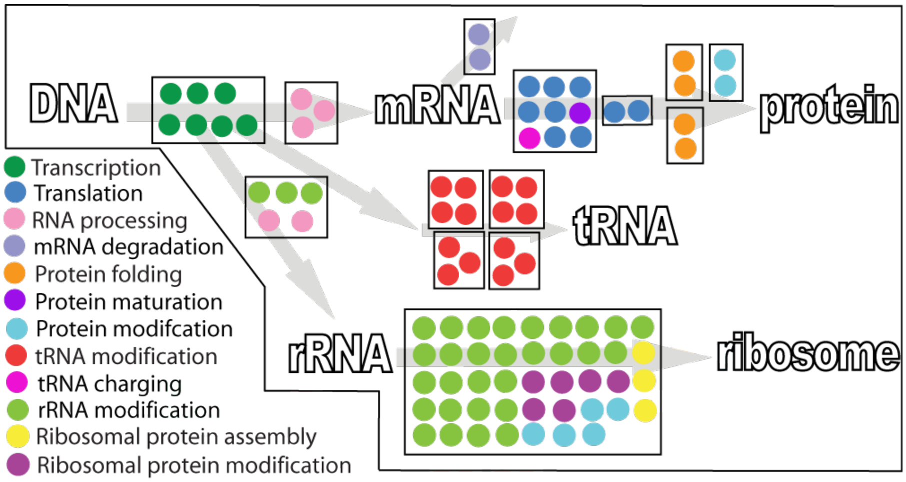

DNA
mRNA
protein
tRNA
rRNA
ribosome
Transcription
Translation
RNA processing
mRNA degradation
Protein folding
Protein maturation
Protein modifcation
tRNA modification
tRNA charging
rRNA modification
Ribosomal protein assembly
Ribosomal protein modification

Figure 6.6: Schematic representation of calculated functional modules with associated proteins and their canonical assignments. Functional modules that consist of one protein are not shown.

153 proteins or protein complexes. Interestingly, many modules contained proteins from different subsystems, which were assigned based on classical pathway designation (Figure 6.6). Hence, our calculations suggest that some canonical pathway assignments may not necessarily represent the functional relationships between the proteins in the cell (Figure 6.6). Furthermore, no direct correlation between the calculated functional modules and protein-protein interaction data [40, 9]could be observed (data not shown). In contrast, stoichiometrically coupled changes of translation initiation factor 1 (IF-1) and ribosomes [55] observed experimentally, suggest that our calculated functional modules are biologically relevant. As more accurate quantitative proteomic data becomes available the functional modules reported herein should be useful in interpretation of this data and help resolve missing gene annotations.

6.3.11 Integration with other cellular functions.

The scope of the 'E-matrix' was limited to the reactions required for synthesis of *E. coli*'s transcriptional and translational machinery, which can account for 50% of the dry weight in fast growing cells [185]. Subsequently, the synthesis and maintenance of this machinery places significant material and energy demands for biosynthetic precursors from metabolism. In the 'E-matrix', these precursors are provided via exchange reactions. As a next step, one could imagine replacing these exchange reaction with the stoichiometric matrix for the metabolic network of *E. coli* [78] ('M-matrix', Figure 6.5A).This integration would allow the direct assessment of the metabolic demand that the transcriptional and translational machinery imposes on a cell. Moreover, integration of the transcriptional regulation of individual operons would enable a more accurate determination of the genotype - phenotype relationship ('O-matrix', Figure 6.5A). Thus the genome-scale integrated network, or 'OME-matrix', would account for three major cellular processes and may capture more than 2,000 of *E. coli*'s gene. Recently, two studies proposed approaches to integrate different cellular processes [53, 155] but no genome-scale representation is available yet.

6.4 Conclusion

In this study, we present the first, mechanistically and chemically detailed, genome-scale network reconstruction of the transcriptional and translational machinery of *E. coli.* Biochemical components, reaction formulation, and quality control measures analogous to metabolic network reconstructions were used to incorporate bibliomic data from the last 50 years into one reconstruction (Figure 6.2).The corresponding knowledge base can be queried online (http://bigg.ucsd.edu/ E-matrix). This stoichiometric reconstruction represents a first step towards modeling this complex cellular function, and will require iterative refinement as new data becomes available. By describing the stoichiometric relationships between the components involved in transcription and translation, this reconstruction enables the quantitative integration of disparate '-omics' data into a computational model (Figure 6.5). We demonstrated that low- and high-throughput data can be readily integrated and used as constraints on model reactions and the subsequent reduction of the feasible set of reaction fluxes results in physiological relevant predictions (Figure 6.5B-D)Furthermore, we showed that the computational model can be used to accurately predict ribosome production under different growth conditions (Figure 6.3). The deletion of single or multiple rRNA operons from the 'E-matrix' predicted that a high density of RNA polymerases is required on the remaining rRNA operons to achieve the reported ribosome numbers (Figure 6.4B). Computational analysis of the 'E-matrix' can provide further insight into the topologically local and global relationship between proteins in terms of functional modules (Figure 6.6).

This 'E-matrix' reconstruction ushers in a new generation of cellular network models that account quantitatively for mRNA and proteins. The 'E-matrix' offers the potential to i) serve as a platform for integrated, numerical analysis of heterogeneous, quantitative high-throughput datasets; ii) increase our understanding of the relationship between mRNA and protein abundance; iii) be integrated with metabolism by extending the transcriptional and translational reactions to metabolic genes; iv) be integrated with regulatory events by formulating regulatory rules for the genes of the 'E-matrix' and extending the transcriptional and translational reactions to transcription factors; and v) enable computation of the material and energetic cost of macromolecular synthesis. These capabilities are important milestones in moving towards a more comprehensive genome-scale *in silico* model of all cellular processes in *E. coli*. Furthermore, the underlying reconstruction methodology can be readily extended and applied to other prokaryotes. Such extension could lead to further insight into conserved and unique features of the transcriptional and translational machinery of prokaryotes.

The history of *E. coli* metabolic reconstructions now spans more than 17 years, with numerous iterative reconstruction refinements and applications superseding initial expectations [80]. The reconstruction of transcriptional and translational machinery *E. coli*, and other prokaryotes, will have the same impact on systems biology, especially when integrated with metabolism, regulation, and condition-specific high-throughput data sets (Figure 6.5A). This work represents hence a crucial step towards the important and ambitious goal of whole cell modeling [113].

The text of this chapter, in full, is a reprint of the material as it appears in I. Thiele, N. Jamshidi, R.M.T. Fleming, B.Ø. Palsson, Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization, PLoS Comp. Biol., 2009. I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

Chapter 7

Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional & translational machinery

The constraint-based reconstruction and analysis (COBRA) approach has recently been extended to describe *Escherichia coli*'s transcriptional and translational (tr/tr) machinery. The current stoichiometric, steady-state formulation does not explicitly represent enzymes in biochemical reactions. Here, we introduce the concept of reaction coupling to represent the dependency between protein synthesis and utilization. These additional coupling constraints lead to a significant contraction of the feasible set of steady-state fluxes. The subset of alternate optimal solutions (AOS) consistent with maximal ribosome production was calculated, and the majority of tr/tr reactions were active for all of these AOS showing that the network has a low degree of redundancy. Furthermore, all calculated AOS contained the qualitative expression of at least 92% of the known essential genes. Principal component analysis of AOS demonstrated that energy currencies (ATP, GTP, and phosphate) dominate the network's capability to produce ribosomes. Additionally, we identified regulatory control points of the network, which include the transcription reactions of sigma factor 70 (RpoD), of degradosome components (Rne, Pnp), and of the protein chaperone (GroS). These reactions contribute significant variance between AOS. These results shows that COBRA can be applied to gain insight into the systemic properties of *E. coli*'s transcriptional and translational machinery.

7.1 Introduction

Kinetic models of transcription [213, 152, 35], translation [152, 105, 65, 176], and the cell cycle [4] have been formulated with systems of ordinary differential equations. These models describe the temporal changes in concentration accompanying production, degradation, transport, or modification of the molecules in the network. While this modeling approach has been shown to be very useful and mechanistically insightful for small scale *E. coli* networks, such as the trp operon [257, 243] and lac operon [300], it cannot be readily applied for large-scale networks due to the paucity of experimentally measured kinetic parameters.

Constraint-based reconstruction and analysis (COBRA) can be used to model biological systems without the use of kinetic parameters. In this approach, the network is formulated as a set of linear equations describing the biochemical transformations taking place within a cell. Models are constructed in a bottom-up fashion based on available genomic, biochemical, and bibliomic data describing the known biochemical transformations of a particular cellular function in a target organism [79, 229]. If available, information about reaction rates may be incorporated into this COBRA approach as constraints (bounds) on network reactions [219, 229]. This reconstruction approach has been well established for metabolism (see reviews [88, 229, 278, 79] and Chapters 3, 4) and is in wide use [80]. More recently, the COBRA approach has been extended for other cellular functions such as signaling [209, 161], transcriptional regulation [95], and protein synthesis [277] (Chapters 5 and 6).

Flux balance analysis (FBA) is a constraint-based optimization approach, in which the flux through a particular stoichiometric model reaction is optimized while ensuring that biological and physico-chemical constraints are obeyed. FBA relies on linear programming to find the optimal solution of an objective function, which maximizes or minimizes a particular flux. However, depending on the properties of the model, the identified solution may not be unique meaning that there may be an infinite number of different flux vectors giving an identical optimal objective value (Figure 7.1).

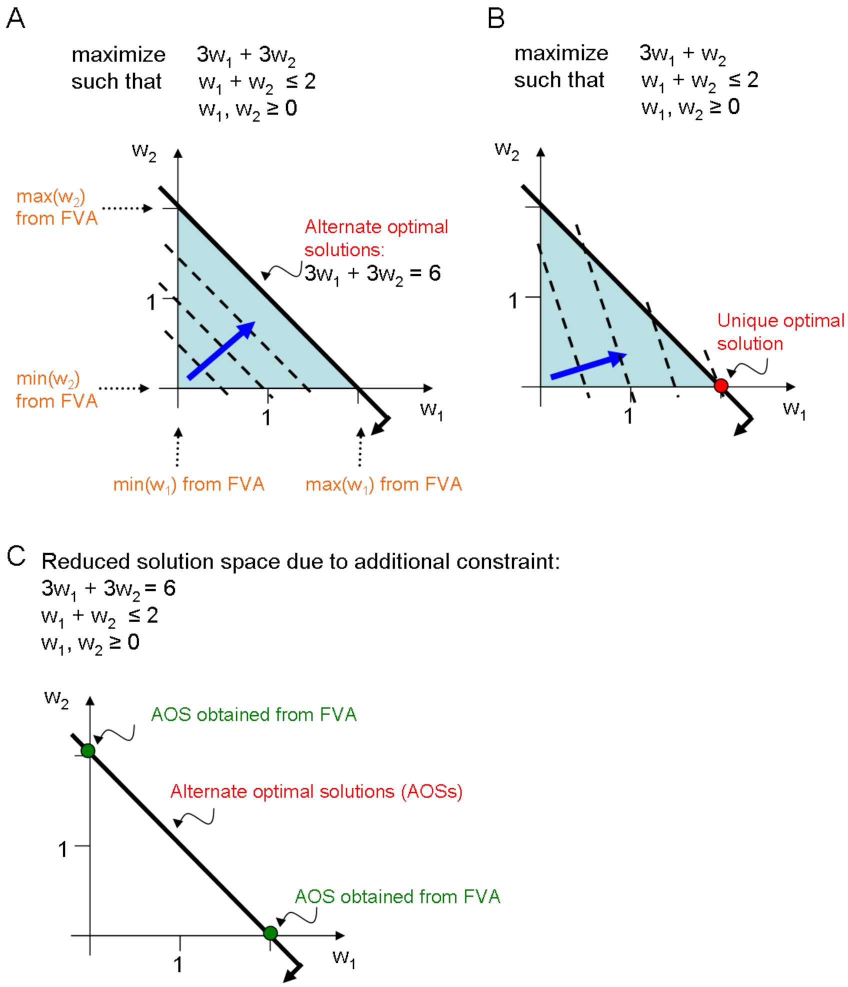

A
maximize $3w_1 + 3w_2$
such that $w_1 + w_2 \leq 2$
$w_1, w_2 \geq 0$
w_2
max(w_2) from FVA
Alternate optimal solutions:
$3w_1 + 3w_2 = 6$
1
min(w_2) from FVA
1
w_1
min(w_1) from FVA
max(w_1) from FVA
B
maximize $3w_1 + w_2$
such that $w_1 + w_2 \leq 2$
$w_1, w_2 \geq 0$
w_2
1
Unique optimal solution
1
w_1
C
Reduced solution space due to additional constraint:
$3w_1 + 3w_2 = 6$
$w_1 + w_2 \leq 2$
$w_1, w_2 \geq 0$
w_2
AOS obtained from FVA
Alternate optimal solutions (AOSs)
1
AOS obtained from FVA
1
w_1

Figure 7.1: **A-C.** Schematic illustration of alternate optimal solutions (AOS), unique solutions, and results of flux variability analysis (FVA) on a linear toy problem.

In the context of metabolic models, these flux vectors are called alternate optimal solutions (AOS) or equivalent phenotypic states [156, 170, 230]. The presence of AOS in constraint-based models was realized in the early 90s when FBA was applied to biologically realistic networks [287]. Consider the example shown in Figure 7.1A. An infinite number of AOS lays on the line with optimal value for the objective function $3w_1 + 3w_2$, whereas the lie vector for each AOS is different. Therefore, all AOS cannot be determined but a representative subset of AOS can be calculated. Different mathematical methods have been used to determine subsets of AOS, e.g., vertex enumeration [230, 292] or flux variability analysis (FVA, [67]). Challenges associated with computing AOS in genome-scale metabolic networks are due to redundant, alternate pathways [170]. Reed *et al.* calculated subsets of AOS for *E. coli*'s metabolic network, which differ in at least one active reaction, at different growth environments and determined correlated reaction sets [230]. This computation is very time consuming. In this study, we will use FVA to determine AOS that correspond to the a subset of extreme points of the steady-state solution space associated with the tr/tr process in *E. coli*. In Figure 7.1C, such extreme points are highlighted. Recently, we reconstructed the first genome-scale network of the transcriptional and translational (tr/tr) machinery that accounts for the synthesis and function of all known components involved in RNA and protein synthesis in *E. coli* [277] (Chapter 6).

This comprehensive reconstruction, named the expression or E-matrix, accounts for the sequence-specific synthesis reactions matrix of 423 functional gene products, including rRNAs, tRNAs, ribosomes, and RNA polymerases. It is well known that the growth rate of *E. coli*, and other organisms, directly correlates with the cellular abundance of its protein synthesis machinery [188]. While the E-matrix does not account for metabolism, it contains exchange reactions, which supply the network with precursors (i.e., amino acids, nucleotide triphosphates (NTP)) and remove metabolic by-products from the network (i.e., nucleotide monophosphates (NMP), orthophosphate) [277]. Defining systems boundaries around protein synthesis, these exchange reactions can be used to determine the dependency between tr/tr and metabolism *in silico* under various environmental conditions. In this study, we determine AOS of the E-matrix, characterize their properties and compare the *in silico* expressed genes with experimental gene essentiality data [14].

Table 7.1: **Overview of the '*E*-matrix' content.** For more details, see Chapter 6 and Thiele *et al.* [277]

Number of transcription units	249
Number of genes	423
Number of components (with/ without genes)	337 (303/34)
-tRNA	86
-rRNA	22
-miscellaneous RNA	1
-proteins (with/ without genes)	228(194 /34)
Number of subsystems	27
Number of reactions	13,694
-Number of demand reactions	302
-Number of exchange reactions	76
Number of network components	11,991

7.2 Material and Methods

Reconstruction We used the recently published reconstruction of *E. coli*'s transcriptional and translational machinery, the *E*-matrix [277]. Briefly, 13,694 reactions and 11,991 components (i.e., metabolites, proteins, RNA molecules, and intermediate complexes) describe the sequence-specific synthesis reactions and cellular functions of 423 known gene products involved in this protein synthesis machinery (Table 7.1). Gene products include 86 tRNAs, proteins such as ribosomes (with rRNA incorporated), RNA polymerase, transcription and translation factors. Note that transcription regulators were not accounted for in the *E*-matrix. A more detailed description of the network content can be found in Thiele *et al.* [277] and Chapter 6.

Constraint-based modeling The *E*-matrix reconstruction can be converted into a mathematical format as stoichiometric matrix, $S \in \Re^{m,n}$, where each row corresponds to a network component and each column corresponds to a network reaction. By definition, the stoichiometric coefficients for substrates are negative numbers, while products are positive coefficients. For the analysis of the network properties, we assume that the system is at steady-state, therefore

$$S \cdot v = \frac{dx}{dt} = 0 \tag{7.1}$$

where $\frac{dx}{dt}$ is the rate of change in concentration of a component x over time, which is zero in steady state.

The E-matrix is under-determined, since there are more variables (reactions) than equations (mass-balances). Therefore, a unique solution to this set of linear equations does not exist (Figure 7.1). The addition of further inequalities (e.g., reaction rates) may reduce this set of feasible solutions.

Network constraints Other constraints may include the directionality of a reaction, v_i, based on thermodynamic information (e.g., the ATP-dependent phosphorylation of glucose to glucose-6-phosphate is effectively irreversible) or environmental constraints for the availability of a nutrient in the medium (e.g., restricting glucose to be the sole carbon source by constraining all uptake fluxes for other carbon sources to be zero). By changing the set of inequality constraints applied to the model, different subsets of the steady-state feasible set are obtained and their differences can be studied using mathematical tools.

Network boundaries The inputs to the E-matrix are biosynthetic precursors, such as amino acids and NTPs, which are provided to the network via exchange reactions. In the E-matrix, by-products of protein synthesis, such as nucleotide monophosphate and orthophosphate are also removed from the system [277]. For every protein and tRNA species, a demand function was created to mimic the requirement of that component for growth. These reactions were introduced as the steady-state assumption does not allow the accumulation of intracellular components but cell doubling includes a doubling of the proteome. Hence, these demand reactions represent the newly produced proteome of the *in silico* cell.

Objective function The demand reaction of ribosomal 50S subunit production (DM_rib_ 50) was chosen as an objective function for the model, since the ribosome content of the cell is correlated to the growth rate [188]. The optimization problem is formulated as follows:

$$max\ c^T \cdot v \tag{7.2}$$

$$s.t.\ S \cdot v = \frac{dx}{dt} \equiv 0 \tag{7.3}$$

$$v_{i,min} \leq v_i \leq v_{i,max}\ for\ all\ i\ reactions \tag{7.4}$$

Simulation constraints To model the E-matrix corresponding to different doubling times, we calculated the maximal possible stable RNA transcription initiation rates based on the data given in Neidhardt *et al.* [183] (Table 7.2).

Table 7.2: **List of *E. coli* cell specific parameters used for calculation in this work.** See also Table 3 in [183] and Chapter 2.

Doubling time (min)	24	30	40	60	90	100
initiation rate at rrn genes, i_{rrn} ($\frac{initiations}{min \cdot genes}$)	58	39	23	10	5	4
cell mass, z ($\frac{\mu g_{DW}}{10^9\,cells}$)	865	641	433	258	162	148

The total transcription initiation rate for stable RNA gene i is given by

$$v_{transcription\_initiation_i} = i_{rrn} \cdot g_i \tag{7.5}$$

where i_{rrn} is the initiation rate per ribosomal RNA copy ($initiation\, min^{-1}\, gene^{-1}$) (Table 3 in [183], see also Table 7.2). To accounts for the gene dosage effect, we multiplied i_{rrn} by g_i ($gene\, cell^{-1}$), which is the gene copy number. The number of gene copies depends on the number of replication forks, which creates multiple copies of the chromosome within one cell. Therefore, the copy number of the same gene depends on its genome position ($m_i^{'}$ and doubling time (t). g_i is given by:

$$g_i = 2^{\frac{(D \cdot (1 - m_i^{'}) + C)}{t}} \tag{7.6}$$

where D is the time necessary to replicate the chromosome ($D = 0.3314 \cdot t + 32.564$, t in minutes), C is lag time between chromosome replications ($C = 0.0898 \cdot t + 21.238$,$t$ in minutes) and t is the doubling time (in minutes).

The total transcription initiation rate of stable RNA can be converted into a $nmol\, g_{DW}^{-1}\, h^{-1}$ rate, by multiplying Eq. 7.5 by the scaling factor

$$F = \frac{1}{z} \cdot \frac{t}{N_A} \cdot 10^9 \tag{7.7}$$

where N_A is the Avogadro number ($6.022 \cdot 10^{23}\, molecules\, mol^{-1}$), z is the mass per cell ($\frac{\mu g_{DW}}{10^9\, cells}$), and t is time-scale factor (60 in this case).

Flux variability analysis and flux span Given a set of constraints, flux variability analysis (FVA) [170] can be used to assess the network flexibility and network redundancy. In this study, we fixed the ribosome production rate to its maximal value ($v_{DM\_rib50,min} = v_{DM\_rib50,max} = max$, based on Table 7.2). Then, every network reaction i was minimized and maximized. The flux span of a network reaction i is given by $\|v_{i,max} - v_{i,min}\| = span_i$.

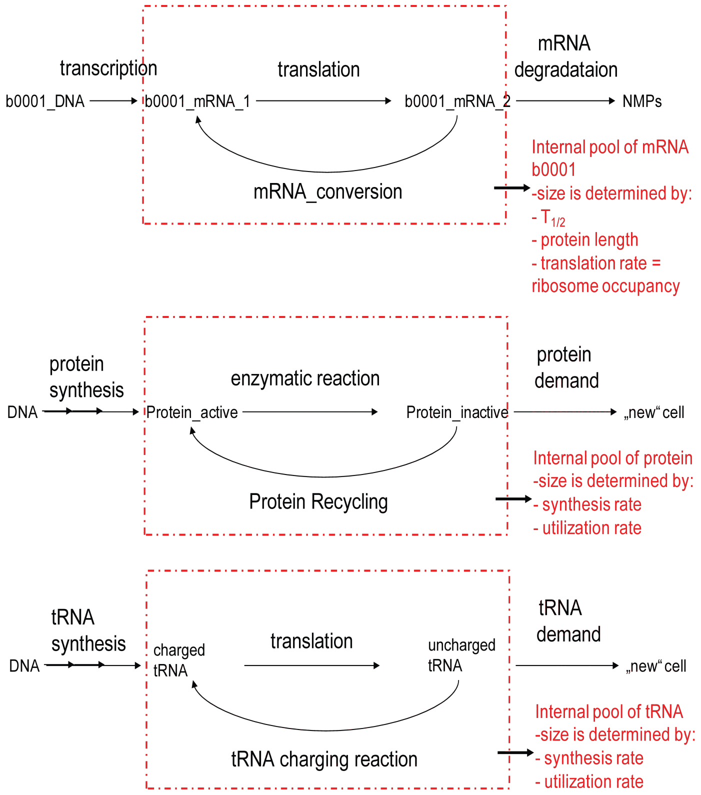

transcription
translation
mRNA
degradataion
b0001_DNA
b0001_mRNA_1
b0001_mRNA_2
NMPs
mRNA_conversion
Internal pool of mRNA
b0001
-size is determined by:
- $T_{1/2}$
- protein length
- translation rate =
ribosome occupancy
protein
synthesis
enzymatic reaction
protein
demand
DNA
Protein_active
Protein_inactive
„new" cell
Protein Recycling
Internal pool of protein
-size is determined by:
- synthesis rate
- utilization rate
tRNA
synthesis
translation
tRNA
demand
DNA
charged
tRNA
uncharged
tRNA
„new" cell
tRNA charging reaction
Internal pool of tRNA
-size is determined by:
- synthesis rate
- utilization rate

Figure 7.2: **Schematic representation of the mRNA and protein pools present in the *E*-matrix.** In contrast to metabolic networks, the tr/tr network requires that component pools are added to ensure that the network functions are similar known *in vivo* features. For example, a mRNA molecule serves as a template for many protein synthesis reactions before it gets degraded. If no pool of mRNA is present and each molecule is only used once, the overall transcription rate, and thus energetic costs, supersedes the experimental measured data. By introducing loops and appropriate constraints one can represent different pool sizes of the components.

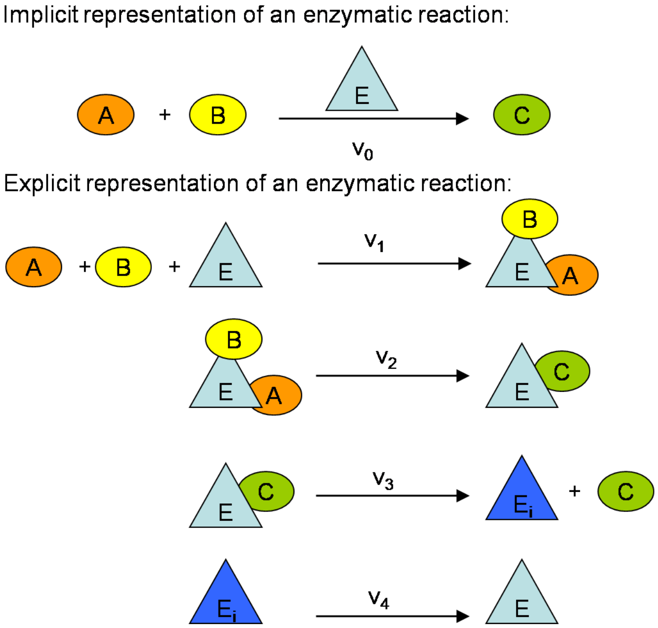

Implicit representation of an enzymatic reaction:
A
+
B
E
C
v_0
Explicit representation of an enzymatic reaction:
A
+
B
+
E
v_1
B
E
A
v_2
C
v_3
E_i
+
C
v_4

Figure 7.3: **Schematic representation of the participation of tr/tr enzymes in network reactions.** An enzyme is unchanged after the reaction is finished. In canonical network formulations, enzyme reaction participation is implied but not explicitly modeled. The tr/tr network produces enzymes, hence, the explicit incorporation of enzymes in their catalyzed reactions is desired. As consequence, the reaction needs to be reformulated and additional constraints have to be added (see text) to ensure that the enzyme is required for its catalyzing reaction. The same approach applied if the reactant E is a tRNA molecule or a protein.

Alternate optimal solutions Alternate optimal solutions (AOS) were determined using FVA. The FVA was carried out as described above and every solution vector of each FVA calculation was stored.

Principal component analysis of alternate optimal solutions In order to identify the sets of reactions that account for the greatest variance in flux between different simulation conditions, we used principal component analysis. Principal component analysis involves a mathematical procedure that transforms a number of possibly correlated vectors into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variance in the data as possible, and each succeeding component accounts for as much of the remaining variance as possible. Principal component analysis of a set of flux vectors can be thought of as revealing the low dimensional projection of a this set in a way which best explains the variance within this set. If a set of flux vectors as a set of coordinates in an n-dimensional flux vector space (1 dimension per flux), then principal component analysis reveals the lower-dimensional projection, a "shadow" of this object when viewed from the coordinates of the object itself, not from the perspective of the vector space in which it is placed.

In the same way, we can take a set of steady state alternative optimal flux vectors, $\mathbf{P} \in \mathbb{R}^{n,N}$, in the nullspace of the stoichiometric matrix, $\mathbf{S} \cdot \mathbf{P} = \mathbf{0}$, which lie in an n-dimensional flux vector space, but use principal component analysis to reveal the intrinsically significant axes which account for the variation within this set. First, we calculate the flux covariance matrix, $\mathbf{C} \in \mathbb{R}^{n,n}$, where the covariance between two fluxes is given by

$$\mathbf{C}_{i,j} = \frac{\sum_{k=1}^{N} (\mathbf{P}_{i,k} - \overline{\mathbf{P}_i})(\mathbf{P}_{j,k} - \overline{\mathbf{P}_j})}{N}$$

with $\overline{\mathbf{P}}_i$ denoting the average flux of reaction i over all N flux vectors. Singular value decomposition of the covariance matrix gives

$$\mathbf{C} = \mathbf{U} \cdot \Sigma \cdot \mathbf{V}^T$$

where $\mathbf{U} = \mathbf{V}$ since $\mathbf{C}$ is a square diagonally symmetric matrix. Each row of $\mathbf{V}$ contains a components, or singular vectors, of the covariance matrix. Each singular vector gives the direction of an intrinsic axis, which is linearly independent from all other intrinsic axes. The principal components are given by the eigenvectors which correspond to the largest eigenvectors. As such, the principal components give the intrinsic axes which account for the largest variation in the set of AOS. The standard deviation for each principal component may be calculated by taking the square root of the singular values, the diagonal entries in Σ [17]. Principal component analysis of the covariance matrix is mathematically equivalent to principal component analysis of the alternate optima themselves, but the former is computationally more efficient. Principal component analysis was carried out on the alternate optimal for $t = 90min$ doubling time.

Formulation of general coupling constraints Typically network reconstructions do not stoichiometrically represents reactants that are both substrates and products in the same reactions. Their involvement is implicit and not explicitly represented in the reaction. An example is an enzyme in an metabolic reaction. However, in the E-matrix, proteins are explicitly included in the reactions they catalyze as illustrated in Figure 7.3. The four explicit reactions (v_1 to v_4) are equivalent to the reaction (v_0) in the implicit formulation. It follows that the synthesis of the recycled reactant E is not essential to permit steady-state flux through v_1 to v_4 as it is recycled by the last reaction (v_4). Subsequently, the conversion of $A + B \rightarrow C$ will occur regardless if the model is synthesizing E. Additional constraints are needed to enforce the synthesis of E if its set of explicit reaction(s) is active in a particular steady-state. We require the condition

$$if\ v_{synthesis,E} > 0\ then\ v_4 > 0 \tag{7.8}$$

where $v_{synthesis,E}$ is the synthesis reaction rate of reactant E. Furthermore, it would be desirable to relate the flux through reaction v_4 and the synthesis of E with some proportionality,

$$v_4 \propto v_{synthesis,E} \tag{7.9}$$

even though the exact proportion factor can only be approximated within bounds (see below).

Since reactant E may be required in multiple reactions, the flux through the recycling reaction (v_4) will be their sum. Subsequently, choosing v_4 for Eq. 7.8 and 7.9 ensures that the synthesis rate of E will be greater than zero if any network reaction that utilizes E is active.

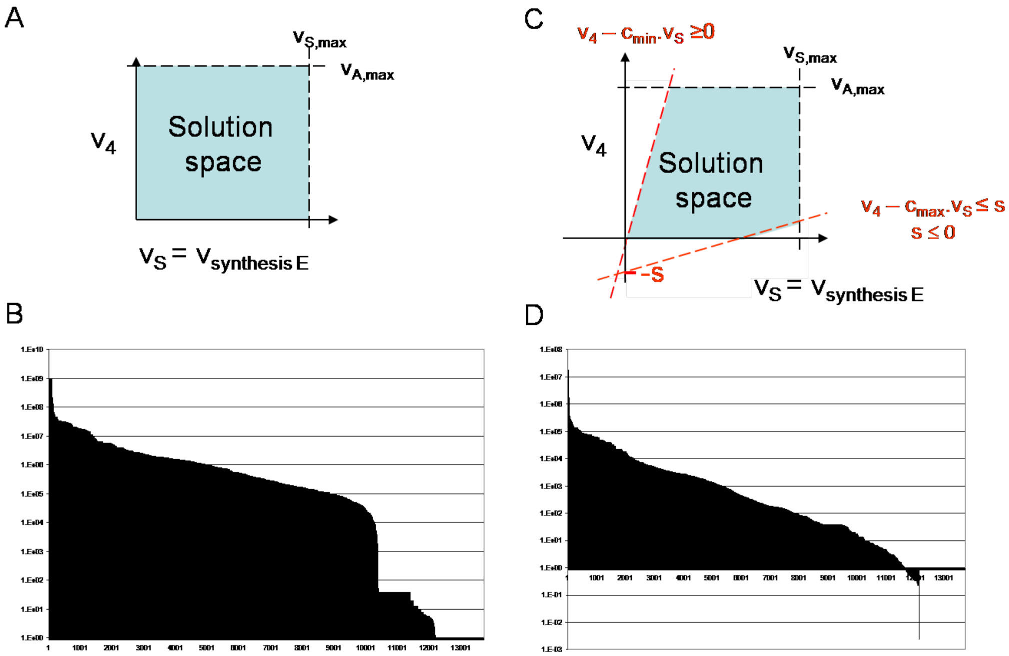


Figure 7.4: **Illustration of reduction of flux span and therefore size of solution space achieved by coupling of network reactions. A, B.** Uncoupled model. **C, D.** Coupled model contains a set of 1,056 coupling constraints between 528 network reactions. Note that the flux span corresponds to the variability of each network reaction while producing maximum rate of ribosomes. The simulation condition correspond to doubling time *t*=90 min.

The relationships expressed in Eq. 7.8 and 7.9 can be represented in a linear fashion with:

$$v_4 - c_{min} \cdot v_{synthesis,E} \geq 0 \tag{7.10}$$

$$v_4 - c_{max} \cdot v_{synthesis,E} \leq s,\ s \leq 0 \tag{7.11}$$

where c_{min} and c_{max} are the bounds on the proportion factor (deemed 'coupling coefficients'). Note that Eq. 7.11 ensures that a higher flux through v_4 raises the lower bound on the synthesis reaction $v_{synthesis,E}$. Furthermore, s can be used to loosen the formulation given in Eq. 7.8 by allowing the synthesis of reactant E without being used in the model up to its value. In this study, however, we set s to be zero, since we intended to determine AOS in which all synthesized reactants are used. Linear inequality coupling constraints retain the numerically scalable character of flux balance analysis.

A schematic representation of the coupling constraints can be found in Figure 7.4. In the E-matrix, there are three sets of reactions that require coupling: i) transcription and translation, ii) translation and protein utilization, and iii) tRNA synthesis and tRNA utilization (Figure 7.2). In each case, the inequalities are the same as Eqs. 7.10 and 7.11 but the definition of the coupling coefficients depends on the nature of the coupled reactions.

The following sets of reactions require coupling constraints (see also Figure 7.2):

1. Transcription and translation: mRNA degradation reactions (e.g., b0001_mRNA_degr1) were coupled to the corresponding mRNA conversion reactions (e.g., b0001_mRNA_CONV2)
2. Translation and protein utilization: protein demand reactions (e.g., DM_AlaS_mono), which allow the accumulation of proteins in the network, and the corresponding protein recycling/utilization reactions (e.g., AlaS_mono_RECYCL) were coupled.
3. tRNA synthesis and tRNA utilization: tRNA charging reactions (e.g., ala1_tRNA_CHARG), representing the tRNA utilization, were coupled with the corresponding tRNA formation reactions (e.g., alaT_to_ala1).

Coupling transcription and translation At steady-state, the rate of mRNA synthesis $v_{synthesis,i}$ (transcription) is equal to the rate of mRNA degradation $v_{degradation,i}$, which is given by

$$v_{synthesis,i} = v_{degradation,i} = k_{degradation,i} \cdot [mRNA]_i = \frac{ln2}{T_{\frac{1}{2},i}} \cdot [mRNA]_i \tag{7.12}$$

where $[mRNA]_i$ is the cellular concentration of mRNA ($molecules\, cell^{-1}$), and $T_{\frac{1}{2},i}$ is the half-life time of mRNA i ($seconds$).

Since the E-matrix genes are transcribed in terms of transcription units, we will couple the mRNA degradation reaction ($v_{degradation,i}$) with the corresponding recycling reaction ($v_{CONV2,i}$)(Figure 7.2). This reaction recycles an mRNA_2 compound released from a translation reaction into an mRNA_1 compound, which is used in translation reactions. This recycling enables the re-utilization of a single transcript for multiple translation rounds before degradation. The mRNA recycling reaction forms a cycle together with the translation reactions (Figure 7.2). This cycle allows the representation of an internal 'mRNA pool' corresponding to the steady-state concentration of the mRNA, which can be used for quantitative integration of gene expression data on transcript abundance in future studies.

Definition of tr/tr coupling factor

In this section, we derive a meaningful coupling factor ($c_{min,i}, c_{max,i}$) between mRNA degradation reaction ($v_{degradation,i}$) with the utilization reaction ($v_{CONV2,i}$)(Figure 7.2).

$$v_{CONV2,i} - c_{max,i} \cdot v_{degradation,i} \leq s, s \leq 0 \tag{7.13}$$

where $v_{CONV2,i} = v_{translation,i}$.

The translation flux is the product of translation rate and mRNA concentration:

$$v_{translation,i} = k_{translation,i} \cdot [mRNA]_i \tag{7.14}$$

In a cell of $t = 60$ minutes doubling time, consider an mRNA i with the following properties:

- the cellular concentration is $[mRNA]_i = 10 \frac{molecules}{cell}$, and
- the half-life time is $T_{\frac{1}{2},i} = 300\, sec$
- mRNA i encodes for a protein i of length

$$L_{p,i} = 330\, aa \tag{7.15}$$

The translation rate of a ribosome is

$$r_{tl} = 16\frac{aa}{sec} \tag{7.16}$$

at doubling time $t = 60$ min [183]. From Eq. 7.15 and 7.16 it follows that one ribosome bound to one mRNA molecule can produce one protein i in ≈ 21 seconds, since

$$k_{translation,i} = \frac{L_{p,i}}{r_{tl}} \frac{sec}{ribosome \cdot protein \cdot mRNA} \tag{7.17}$$

Using the half-life time of mRNA i ($T_{\frac{1}{2},i}$) and the cellular concentration ($[mRNA]_i$), it follows that

$$v_{translation,i} = \frac{T_{\frac{1}{2},i}}{\frac{L_{p,i}}{r_{tl}}} \cdot [mRNA]_i \frac{protein\ molecules}{ribosome \cdot cell \cdot doublingtime} \tag{7.18}$$

Therefore, one mRNA i can result in maximal 14 nascent proteins i during its half-life time of $T_{\frac{1}{2},i} = 300sec$.

However, multiple ribosomes can bind on a mRNA with a minimum spacing of $r_{space} = 17\,aa$ [132]. Therefore,

$$v_{translation,i} = \frac{T_{\frac{1}{2},i} \cdot \frac{L_{p,i}}{r_{space}}}{\frac{L_{p,i}}{r_{tl}}} \cdot [mRNA]_i \ \frac{protein\ molecules}{ribosome \cdot cell \cdot doublingtime} \tag{7.19}$$

Since the mRNA concentration is constant in the cell, at steady-state, we have to multiply Eq. 7.19 by $\frac{t}{T_{\frac{1}{2},i}}$

$$v_{translation,i} = \frac{T_{\frac{1}{2},i} \cdot \frac{L_{p,i}}{r_{space}}}{\frac{L_{p,i}}{r_{tl}}} \cdot \frac{t}{T_{\frac{1}{2},i}} \cdot [mRNA]_i \ \frac{protein\ molecules}{ribosome \cdot cell \cdot doublingtime} \tag{7.20}$$

where t is the doubling time in minutes.

Using scaling factor F (Eq 7.7) and dividing by t, we obtain

$$v_{translation,i} = F \cdot \frac{r_{tl}}{r_{space}} \cdot [mRNA]_i \tag{7.21}$$

where $v_{translation,i}$ is in $\left[\frac{nmol}{g_{DW} \cdot h}\right]$.

Subsequently, the translation of the ten copies of mRNA i can result in up to 1680 proteins i per ribosome and up to 31,920 proteins i, if 19 ribosomes are bound to each mRNA i, during a doubling time $t = 60$ min.

To obtain the $v_{degradation,i}$ in the same unit, Eq. 7.12 needs to be converted:

$$v_{degradation,i} = F \cdot \frac{ln2}{T_{\frac{1}{2},i}} \cdot [mRNA]_i \tag{7.22}$$

where $v_{degradation,i}$ is in $\left[\frac{nmol}{g_{DW} \cdot h}\right]$.

Under the steady-state assumption, we can equate Eq. 7.21 and 7.22 to obtain:

$$\frac{r_{space}}{r_{tl}} \cdot v_{translation,i} = \frac{T_{\frac{1}{2},i}}{ln2} \cdot v_{degradation,i} \tag{7.23}$$

Since the recycling reaction rate ($v_{CONV2,i}$) is equal to the translation reactions rate for mRNA i in the network, it follows that:

$$v_{CONV2,i} = \frac{r_{tl}}{r_{space}} \cdot \frac{T_{\frac{1}{2},i}}{ln2} \cdot v_{degradation,i} \tag{7.24}$$

Subsequently, the coupling factor $c_{max,i}$ between the degradation and translation rate is:

$$c_{max,i} = \frac{r_{tl}}{r_{space}} \cdot \frac{T_{\frac{1}{2},i}}{ln2} \tag{7.25}$$

where r_{tl} is the translation rate at a given doubling time (in $\left[\frac{aa}{seconds \cdot ribosome}\right]$); r_{space} is the spacing between 2 consecutive ribosomes (in $\left[\frac{aa}{ribosome}\right]$), and $T_{\frac{1}{2},i}$ is the half-life time of mRNA i (seconds).

The minimum coupling factor $c_{min,i}$ was determined assuming 1 ribosome bound per transcript:

$$c_{min,i} = \frac{r_{tl}}{L_{P,i}} \cdot \frac{T_{\frac{1}{2},i}}{ln2} \tag{7.26}$$

where $L_{P,i}$ is the length of the protein i (in aa).

Coupling protein synthesis and utilization & tRNA synthesis and utilization The protein and tRNA synthesis reactions were coupled to their utilizing reactions in a similar fashion. However, an arbitrary number of 10^5 was chosen for the coupling factor ($c_{max,i}$)as the interpretation of this factor is quite different from the mRNA recycling. Since most proteins and tRNAs are assumed to be stable in the time-scale of an average cell's doubling time, protein and tRNA degradation was ignored. However, the turnover rate of a protein or tRNA is limited and depends on the individual species. The coupling factor represents such turnover limitation as it enforces the synthesis of more protein/tRNA if they are highly used in the network.

In total, 1,056 additional inequality constraints (628 on mRNA, 120 on tRNA, and 308 on protein synthesis) were added to the E-matrix resulting in a problem size of 13,047 equality and inequality constraints and 13,726 variables (reactions). This additionally constrained E-matrix ('$E_{coupled}$-matrix') was used throughout the chapter unless stated differently.

7.3 Results

Comparison of flux span with and without flux coupling We expected a significant reduction in the size of the steady-state solution space in the $E_{coupled}$-matrix. To assess the change in solution space size, we determined the flux span of the E-matrix reactions and of the $E_{coupled}$-matrix (Figure 7.4). We found that the coupling constraints reduced the mean flux span by three orders of magnitude (from $1.1 \cdot 10^7 \pm 9.2 \cdot 10^7$ $\frac{nmol}{g_{DW} \cdot h}$ in the E-matrix to $6.76 \cdot 10^4 \pm 1.38 \cdot 10^6$ $\frac{nmol}{g_{DW} \cdot h}$ in the $E_{coupled}$-matrix) (Figure 7.4). The same tendency was observed when the median flux span was compared (from $3.04 \cdot 10^5$ $\frac{nmol}{g_{DW} \cdot h}$ to $1.99 \cdot 10^2$ $\frac{nmol}{g_{DW} \cdot h}$). Thus, the coupling constraints shrank significantly the size of steady-state feasible set.

Ribosome production in the $E_{coupled}$-matrix Ribosomes are required for the synthesis of proteins involved in other cellular functions such as metabolism, cell division or transcriptional regulation. Ribosome synthesis is one of the main tasks of the machinery encoded in E-matrix. Furthermore, ribosome production rate is correlated with the growth rate [188]. Since the additional constraints may have altered the $E_{coupled}$-matrix ribosome production capabilities, we recomputed the values corresponding to various doubling times (data not shown). We found that the computed ribosome values were in good agreement with the published experimental data [183] and the *in silico* production capabilities of the E-matrix [277]. These results ensured that the ribosome production capacity was not affected by the coupling constraints.

AOS for maximal ribosome production AOS are flux vectors that have the same optimal value for an objective function but differ in their distributions of flux (Figure 7.1) [170, 230]. Here, we used flux variability analysis (FVA) to enumerate all AOS that produced maximally ribosomes and have an optimal (minimal or maximal) value for at least one other network reactions. This FVA-derived subset of AOS thus corresponded to extreme (or boundary) AOS. The characteristics of the AOS of four different models, corresponding to doubling times of $t = 24$,$t = 60$,$t = 90$, and $t = 100$ minutes, were determined.

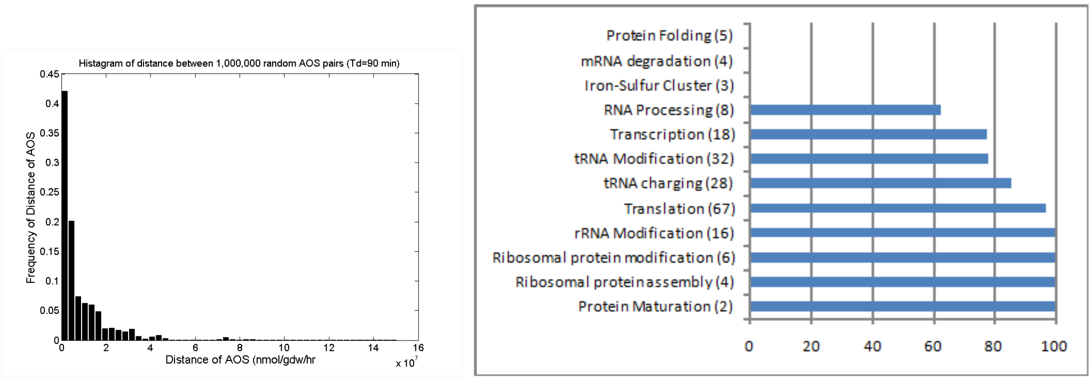


Figure 7.5: **Properties of AOS in $\boldsymbol{E}$-matrix. A.** Distance between AOS. To assess the overall distance between the set of AOS we computed the distance between 10^6 randomly chosen AOS pairs. The distance between the majority of AOS is small indicating that the optimal vertices are in geometric proximity (doubling time t=90). **B.** Percentage of genes being expressed in all AOS per subsystem. Number of genes per subsystem is given in parenthesis. x axis is in percentage.

Average distance of alternate optima solutions Since the FVA-derived AOS represent only a subset of all possible AOS, we computed the average Euclidean distance between AOS. The distance between two AOS also represents a measure of how evenly they are distributed in the solution space. We compared the distance of 10^6 pairs of AOS (Figure 7.5A). As expected, the AOS were not evenly distributed, however, interestingly, many AOS were relatively close to each other. Although we did not determine the volume of the steady-state solution space, as it is a computationally challenging calculation, these results together with the flux span analysis indicate a relatively confined space.

Principal component analysis (PCA) of alternate optimal solutions PCA is an objective non-parametic, analytical method in wide use for a variety of applications including signal processing [180], and more recently mRNA expression analysis [7, 115, 114].

Here, we used PCA to investigate the effective dimensionality of the $E_{coupled}$-matrix and furthermore, to determine the number of branch points, or control points, in the gene expression system of *E. coli* tr/tr machinery. For the entire network, the first ten modes (z scores) could reconstruct 90% of the variance between AOS (Figure 7.6A). The first four 'eigen-reactions' correspond to 1) diphosphate exchange, proton exchange, and water exchange; 2) diphosphate exchange, proton exchange, orthophosphate exchange, and water exchange; 3) GTP exchange, GDP exchange, and water exchange; and 4) ATP exchange, orthophosphate exchange, and water exchange. Interestingly, these four 'eigen-reactions' were dominated by exchange and transport reactions of energy currency into and out of the $E_{coupled}$-matrix. These results indicate that ribosome production, together with synthesis of other tr/tr components, are mainly controlled by the energy state of the cell.

To investigate the set of tr/tr genes that were likely to correspond to key control points, we performed the PCA on the subset of mRNA synthesis reactions. We found that 75 modes were necessary to recover 90% of the information content in the AOS for the 314 protein coding genes (Figure 7.6B). This result was quite different to the PCA analysis of the entire network where ten modes were sufficient to recover the majority of information content in the AOS. The first 'eigen-reaction' was dominated by the expression of sigma factor 70 (b3067, RpoD), which is the primary sigma factor during exponential growth, targeting RNA polymerase sigma 70 to a wide range of promoters that are essential for normal growth [126]. The second 'eigen-reaction' consisted of the gene synthesis reaction for b1084 (Rne), b3164 (Pnp), b4142 (GroS), and b2794 (QueF). The first two genes are part of the multi-protein complex degradosome, which is responsible for mRNA degradation in *E. coli.* GroS is part of the protein folding complex GroEL/S, which helps to fold larger proteins [104]. QueF is a protein involved in the synthesis of pre_Q0, a precursor to queuosine that is an important modified nucleotide in *E. coli*'s tRNA. The next four 'eigen-reactions' also consisted of these genes. In addition, b0884 (InfA), the protein chain initiation factor 1 (IF1), contributed to the fourth 'eigen-reaction' and b2573 (RpoE) contributed to the sixth 'eigen-reaction'. This later gene corresponds to sigma factor 24, which drives transcription of a number of genes whose functions revolve around heat shock and mis-folded proteins. Taken together, the first six modes of the genes expression reactions recovered about 60% of the information content and the corresponding 'eigen-reactions' consisted of the main players involved in transcription, translation, mRNA degradation, and protein folding. Based on the proposed interpretation of the 'eigen-reactions' as key control points [17] it is to be expected that the gene expression of these seven genes is highly regulated in *E. coli.* In fact, preliminary analysis of the regulatory rules in two main databases [141, 242] for *E. coli* genes indicate that there are at least 30 transcriptional regulators involved in controlling the synthesis of tr/tr genes under different environmental conditions.

Length and reaction participation of alternate optima solutions Metabolic networks are known for their redundancy as it increases the flexibility and fitness of the cell to sudden environmental changes [217, 280]. For the E-matrix, a certain rigidity is expected as the majority of the associated functions have only one coding gene in the genome. When optimizing for ribosome synthesis rate the number of active reactions in the AOS can be used as a measure of network flexibility. We found that on average about 6,500 reactions ($\approx$ 50%) were active per AOS, i.e., they had a non-zero flux value. 3,800 of these 6,500 reactions were active in all AOS in a simulation condition. Overall, a set of 3,616 reactions was active in all AOS under all simulated conditions. Additional 1,048 reactions were active in 95% of the AOS under all simulation conditions.

This high number of active reactions is a consequence of the linear structure of the transcriptional and translational network [277]: A gene is transcribed into mRNA; its mRNA is then either degraded or used as a template for translation into a protein, which catalyzes one or more biochemical transformation along this path. In contrast, metabolic networks have more interconnections with numerous alternative (redundant) pathways. Subsequently, an average of approx. 30% of the reactions present in *E. coli*'s metabolic network were found to be active per AOS [230]. Furthermore, 37% of the metabolic reactions are used in any AOS in the environmental conditions tested, thus, they are irrelevant for optimal growth rate [230]. This observation was quite different to our observation of inactive reactions in the $E_{coupled}$-matrix. These results illustrate the fundamental differences in topology and redundancy between the networks of these two important cellular functions.

Essential genes are expressed in all AOS The $E_{coupled}$-matrix accounts for a total of 314 protein coding genes, many of which are directly involved in processes of the macromolecular machinery [277]. First, we analyzed how many genes were expressed in all AOS. We found that at a doubling time $t = 90$ minutes, 227 genes (73%) were expressed in all AOS (required genes), while only two genes were not expressed in any AOS. These two genes, b4292 (fecR) and b4293 (fecI, sigma 19), are part of the same operon and hence co-expressed in the network. The transcription factor sigma 19 was not expressed in any AOS as none of the included genes have sigma 19-dependent transcription [141, 277]. In fact, sigma 19 seems to have few genomic binding sites in *E. coli* (B.K. Cho, personal communications). 85 of 314 (27%) genes were transcribed in many but not all AOS. We compared the required genes with *in vivo* essentiality data [14]. *E. coli* has 303 essential genes (in rich medium) [14], 99 of these genes were present in the *E*-matrix network and 91 of these essential genes were required genes in all simulated conditions (doubling times of 24, 60, 90, and 100 minutes). The remaining 136 genes expressed in all AOS but not *in vivo* essential, since the current formulation of the tr/tr reactions requires the presence of the proteins. *In vivo* the absence of a protein may cause some phenotypic changes but may not be lethal.

Table 7.3: *in vivo* essential *E. coli* genes that were not expressed in all AOS for maximal ribosome production (doubling times of 90 minutes).

Gene name	Gene	Function	AOS Participation (%)
b2563 (acpS)	holo-[acyl-carrier-protein] synthase 1	Metabolism	2.44
b2614 (grpE)	heat shock protein	Protein folding	3.31
b1092 (fabD)	malonyl-CoA-[acyl-carrier-protein] transacylase	Metabolism	4.46
b1093 (fabG)	3-oxoacyl-[acyl-carrier-protein] reductase	Metabolism	4.46
b0188 (tilS)	tRNA(Ile)-lysidine synthetase	tRNA modification	5.49
b2573 (rpoE)	RNA polymerase, sigma 24 (sigma E) factor	Transcription	74.52
b2779 (eno)	enolase	Metabolism	75.64
b4142 (groS)	Cpn10 chaperonin GroES, small subunit of GroESL	Protein Folding	99.80

Only eight *in vivo* essential genes were not active in all AOS (Table 7.3). Four of these essential genes were metabolic genes that were co-expressed with genes involved in the synthesis machinery. Since the *E*-matrix does not account for metabolism, no gene essentiality was expected and this disagreement can be neglected. The remaining five genes were involved in different processes of the synthesis machinery (Table 7.3). RpoE (b2573) is the minor sigma factor (sigma E) in *E. coli*, which responds to heat shock and other stress situations. In the E-matrix only four transcription units are dependent on sigma E transcription: TU00512 (b1909 (leuZ), b1910 (cysT), and b1911 (glyW)) encoding for three tRNA genes; TU-8392 (b2893 (dsbC), b2892 (recJ), b2891 (prfB)); TU-8397 (b3181 (greA)); and TU-8398 (b3201 (lptB), b3202 (rpoN), b3203 (hpf), b3204 (pstN), b3205 (yhbJ). However, since sigma E is known to have about 70 binding sites on the *E. coli*'s genome (B.K. Cho, personal communication), it is very likely that the E-matrix did not account for essential functions dependent on sigma E transcription. In contrast, GroS is the smaller subunit of the GroEL/ES chaperone that is responsible for correct folding of larger proteins. Many of the E-matrix proteins can be folded spontaneously, DnaK/J-GrpE chaperone dependent, and/or GroEL/ES dependent. The corresponding information was included based on two large-scale experimental studies identifying targets specific for these chaperones [140, 60]. The overlapping action of DnaK/J-GrpE chaperone and GroEL/S chaperone explains the missing essentiality in the $E_{coupled}$-matrix.

The last false negative predictions included proteins for a tRNA modification (Table 7.3), which modifies the nucleotide at position 34 in ileX and ileY-tRNA (conversion of cytidine into lysidine) [261, 119]. These two tRNA recognize the same codon (ATA), which was less frequently used in the E-matrix associated genes compared to the genome (unpublished data), which may explain its non-essentiality in our calculations.

The analysis of the percentage distribution of genes expressed in all AOS, regardless the simulation condition, over the different subsystems showed that the majority of genes were synthesized in all but three subsystems (Figure 7.5B). The protein folding subsystem consists of five gene products, which have overlapping functions as described above. Interestingly, the genes encoding for the subunits of the degradosome (b2779 (eno), pnp (b3164), and b3780 (rhlB)) and the oligoribonuclease (b4162, orn) were not synthesized in all AOS. The degradosome is the network's only pathway to degrade mRNA's. However, these genes were expressed in at least 75% of the AOS under the simulated conditions. This result indicates that the coefficients used for coupling these reactions may not be tight enough.

The gene products involved in RNA processing (rnb, rnd and elaC) were not synthesized in all AOS as their functions overlap in the E-matrix. Only two genes of the

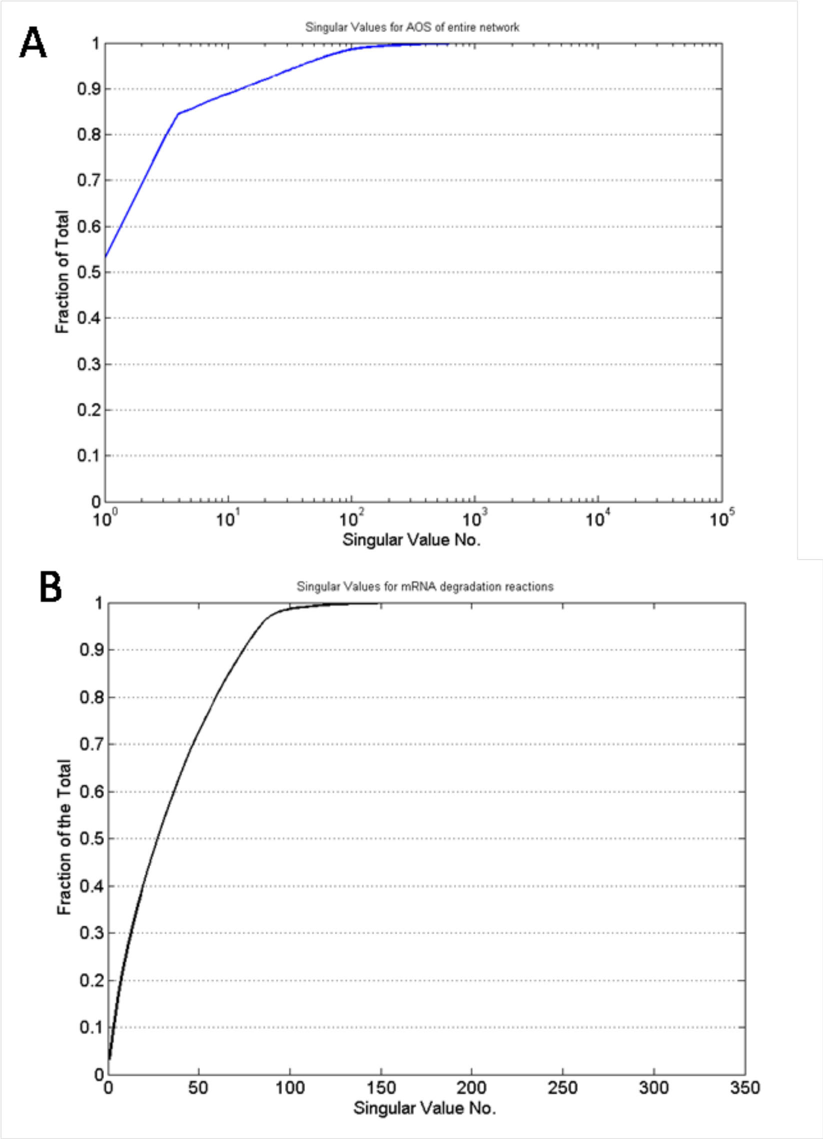


Figure 7.6: **Principal component analysis (PCA). Z-scores of the entire '$E_{coupled}$-matrix' network (A) and of the gene expression reactions (B).** The PCA analysis was performed on the set of AOS (doubling time t=90min).

translation subsystem were dispensable (prfB and tufB), however both genes were produced in at least 98.8% and 91% of the AOS under the simulated conditions, respectively. Furthermore, two genes associated with the tRNA-charging subsystem were dispensable: yadB is an alternate glutamyl-tRNA synthetase to gltX, which is essential, and lysS, a lysine-tRNA synthetase that is an alternative to lysU.

7.4 Conclusion

In this study, we introduced the concept of coupling constraints for the transcriptional and translational machinery of *E. coli* and illustrated their effects on the steady-state solution space. The addition of these coupling constraints permitted the representation of mRNA, tRNA, and protein pools in the network, which may be used for integration of quantitative transcriptomic and proteomic data. We calculated alternate optimal solutions (AOS) and found that they are directly subject to the constraints applied to the network. This means that while some of these AOS might be eliminated if an additional or tighter constraint is added to the model (Figure 7.1), a subset of the AOS have biological meaning and directly represent the flexibility and redundancy inherent to biological systems.

This functional redundancy directly contributes to the robustness of cells to slight perturbations or changes in the network [280]. For instance, the existence of multiple AOS could correspond to silent phenotypes [223], which implies that the same overall cellular performance can be obtained using different metabolic reactions [86]. Adaptively evolving *E. coli* to grow faster on lactate yielded in many different mutations in evolved strains, indicating that evolutionarily equivalent optimal genotypes exists for the resulting *E. coli* phenotype to grow optimally on this medium. In contrast, *E. coli* strains evolved on glycerol exhibited only few different mutations, many of which reappeared in different strains [109]. These experimental results indicate that AOS may be indeed biological relevant and may be inherent to the dynamics and heterogeneity of cellular populations. Certainly, more experimental evidence is necessary to verify and establish such population dynamics but we are confident that cellular modeling will play a key role in identifying and strengthening such relationships.

The text of this chapter, in full, is a reprint of the material as it appears in I. Thiele, R.M.T. Fleming, A. Bordbar, J. Schellenberger, B.Ø. Palsson, Functional characterization of alternate

optimal solutions of *Escherichia coli*'s transcriptional and translational machinery, *to be submitted*. I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

Chapter 8

An integrated model of macromolecular synthesis and metabolism of *Escherichia coli*.

In the previous chapters, existing methods and approaches to reconstruct metabolic networks of various organisms, and we showed that the underlying approach can be expanded to reconstruct stoichiometric networks of other cellular functions. The last two chapters concentrated on transcription and translation (tr/tr). Here, we will show that it is possible to combine the metabolic model of *Escherichia coli* with the tr/tr network in a meaningful manner and that the resulting Metabolism - Expression ('ME') matrix can be employed to investigate cellular properties that could not be accessed with the individual networks.

8.1 Introduction

Cell-scale modeling is one of the great goals of computational biology. In fact, in 2002 an international *Escherichia coli* alliance was formed with the aim to generate data and tools necessary to formulate a whole cell computer representation of this bacteria [113]. Bottom-up network reconstructions have been developed for metabolism [78, 66, 281], signaling networks [209, 161] and more recently for macromolecular synthesis [277]. These network reconstructions are created based on an organism's genome and biochemical information and represent two dimensional annotations of genomes [205]. A detailed description and reviews about metabolic networks can be found elsewhere [229, 79, 70]. (See also Chapter 3 and 4.)

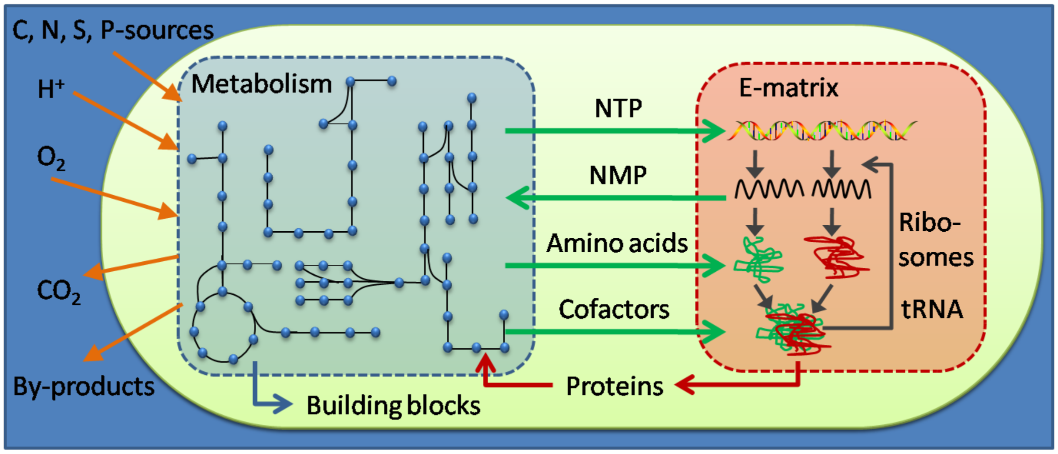


Figure 8.1: **Functional synergy between the metabolic network and the macromolecular synthesis network in *E. coli*.**

The metabolic reconstruction of *E. coli* is the most comprehensive network reconstruction available, accounting for the functions of almost 30% of the open reading frames (ORF) in *E. coli*'s genome [78]. We recently constructed the first genome-scale, stoichiometric network of the transcriptional and translational (tr/tr) machinery of *E. coli* [277]. This latter reconstruction covers the function of 303 gene products, including ribosomal proteins, RNA polymerase, tRNA and rRNA. It represents the synthesis reactions of all known components necessary to produce themselves.

Here, we integrate these two reconstructions into a Metabolic-Expression ('ME') matrix reconstruction that accounts for the synthesis and function of almost 2,000 *E. coli* genes (Figure 8.1). The reconstruction and modeling was done using the constraint-based reconstruction and analysis (COBRA) method [206]. We show that the models derived from ME-matrix reconstruction allow us to address a new biotechnological and biomedical questions that have not been modeled yet, including codon usage, protein engineering and prediction of cellular proteome. This ME-matrix represents a mile-stone towards cell-scale modeling and sets stage in modeling techniques to achieve this ambitious goal in near future.

To-date, only few examples of integrated networks of cellular functions have been published, including i) a metabolic-regulatory network using metabolic reconstruction and transcriptional regulatory network in form of Boolean expressions, for *E. coli* [51]; and ii) two metabolic-signaling-regulatory models [53, 155]. However, these integrated functional networks do not explicitly account for the proteins (enzymes and regulators) and they employ other modeling tools than COBRA (e.g., ordinary differential equations or Boolean logic). Therefore, the presented integrated network is unique and the first of its kind.

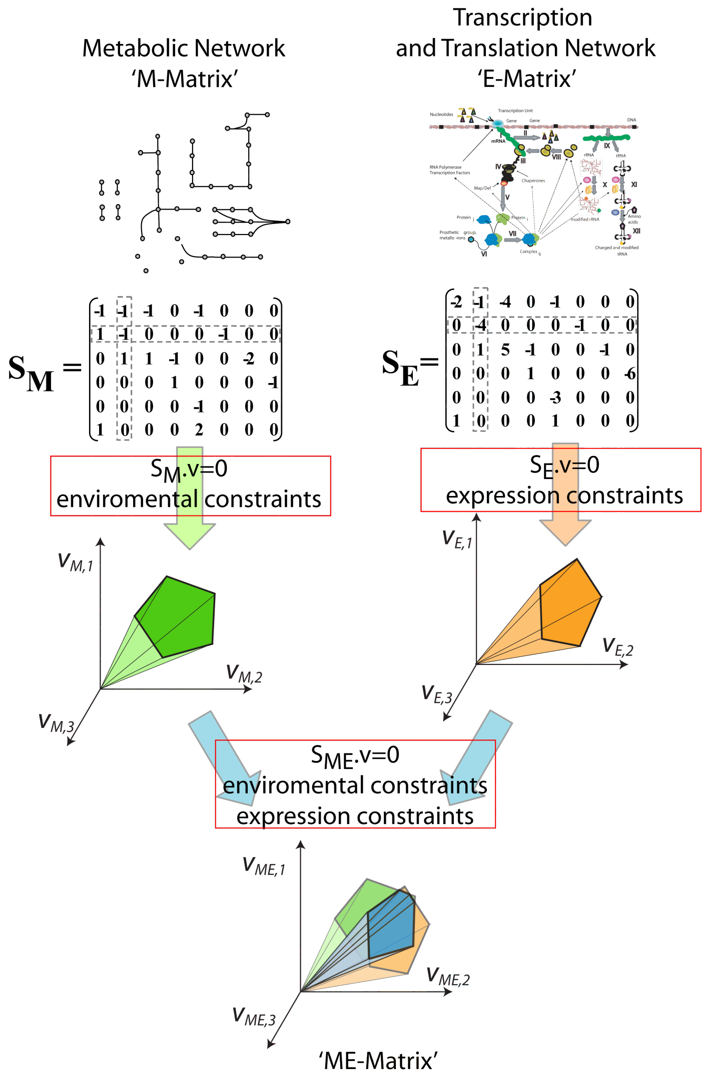

Metabolic Network
'M-Matrix'
Transcription
and Translation Network
'E-Matrix'
Protein j
Protein i
Prosthetic group, metallo -ions
Complex ij
modified rRNA
Amino acids
Charged and modified tRNA
Map/Def
Chaperones
$S_M =$
-1 -1 -1 0 -1 0 0 0
1 -1 0 0 0 -1 0 0
0 1 1 -1 0 0 -2 0
0 0 0 1 0 0 0 -1
0 0 0 0 -1 0 0 0
1 0 0 0 2 0 0 0
$S_E =$
-2 -1 -4 0 -1 0 0 0
0 -4 0 0 0 -1 0 0
0 1 5 -1 0 0 -1 0
0 0 0 1 0 0 0 -6
0 0 0 0 -3 0 0 0
1 0 0 0 1 0 0 0
$S_M.v=0$
enviromental constraints
$S_E.v=0$
expression constraints
$v_{M,1}$
$v_{M,2}$
$v_{M,3}$
$v_{E,1}$
$v_{E,2}$
$v_{E,3}$
$S_{ME}.v=0$
enviromental constraints
expression constraints
$v_{ME,1}$
$v_{ME,2}$
$v_{ME,3}$
'ME-Matrix'

Figure 8.2: **Schematic depiction of the integration of the metabolic network (M-matrix) and transcriptional/translational network (E-matrix) into a combined, integrated network (ME-matrix) with changed solution space.**

8.2 Materials and Methods

8.2.1 Constraint-based reconstruction and modeling appro-ach

The reconstructed biochemical network is often represented in a tabular format, listing all network reactions and metabolites in a human-readable manner (see Chapter 3 and [229, 79] for details). The conversion into a mathematical, or computer-readable format, can be done automatically by parsing the stoichiometric coefficients from the network reaction list (e.g. using the COBRA toolbox [21]). The mathematical format is called a stoichiometric matrix, or S matrix, in which the rows correspond to the network metabolites and the columns represent the network reactions. For each reaction, the stoichiometric coefficients of the substrates are listed with a minus sign in the corresponding cell of the matrix, while the product coefficients are positive numbers, by definition. The resulting size of the S matrix is m x n, where m is the number of metabolites and n the number of network reactions. Mathematically, the S matrix is a linear transformation of the flux vector $v = (v_1, v_2, ..., v_n)$ to a vector of time derivatives of the concentration vector $x = (x_1, x_2, ..., x_m)$ as $\frac{dx}{dt} = S \cdot v$. At steady-state, the change in concentration as a function of time is zero; hence, it follows: $\frac{dx}{dt} = S \cdot v = 0$. The set of possible flux vectors v that satisfy this equality constraint might be subject to further constraints by defining $v_{i,min} \leq v_i \leq v_{i,max}$ for reaction i. In fact, for every irreversible network reaction i, the lower bound was defined as $v_{i,min} \geq 0$ and the upper bound was defined as $v_{i,max} \geq 0$. Exchange reactions, which supply the network with nutrients or remove secretion products from the medium, were defined for all known medium components. The uptake of a substrate by the network was defined by a flux rate $v_i < 0$ and secretion of a by-product was defined to be $v_i > 0$ for every exchange reaction i. An exchange reaction is represented in the reaction is as follows: e.g., D-glucose exchange: Ex_glc-D: 1 glc-D →. Note that this exchange reaction is unbalanced. Exchange (uptake) reactions define the presence of media components as if one would add metabolites into an *in silico* flask. Finally, the application of constraints corresponding to different environmental conditions (e.g., minimal growth medium) or different genetic background (e.g., enzyme-deficient mutant) allow the transition from biochemical network reconstruction to condition-specific model. Note that the network reconstruction is unique to the target organism (and defined by its genome) while it can give rise to many different models by applying condition-specific constraints. All flux rates, , except biomass formation, are given in $\frac{nmol}{gDW \cdot h}$.

8.2.2 Reconstruction of the 'ME'-matrix

Metabolic reconstruction The metabolic reconstruction of *E. coli*, *i*AF1260, was obtained in SBML format (Ec_iAF1260_flux1.xml), from http://systemsbiology.ucsd.edu/In_Silico _Organisms/E_coli/E_coli_SBML) and imported into Matlab (Mathworks Inc.) using the COBRA Toolbox [21]. *i*AF1260 accounts for 1,260 *E. coli* genes and 2,077 reactions, including 1,339 unique metabolic reactions, 690 transport reactions, and 304 exchange reactions [78]. 1,294 reactions have gene-protein-reaction associations. iAF1260 accounts for 1,039 unique metabolites. 1,148 unique, functional proteins are accounted for, including 167 multigene complexes and 346 isozymes [78].

Macromolecular synthesis reaction The tr/tr machinery reconstruction, 'E-matrix', was also imported into Matlab (E_matrix.mat) [277]. Detailed information about the E-matrix can be found in Chapter 6.

Construction of transcription and translation reactions for metabolic enzymes The integration of the E-matrix with the iAF1260 requires that all metabolic enzymes (1260 gene products) are synthesized by the network. Therefore, we used the template reactions for transcription, translation, mRNA degradation, etc. as well as the gene information (e.g., transcription unit assignment from EcoCyc [133], gene coordinates and gene direction from Riley *et al.* [237]) (see Table 8.1 for a complete list). The formulation of the reactions was done in an automated fashion,using a Perl scripting language, as described elsewhere [277].

Protein Complex formation Information about protein complex formations was obtained from iAF1260, which describes the relationship between gene products and metabolic reactions in terms of Boolean logic [78]. This information was complemented with protein complex formation information obtained from EcoCyc [133] and primary literature. Protein complex formation reactions for multimeric proteins were formulated manually, assuming that all subunits bind simultaneously.

Metallo-ions and prosthetic groups Information about metallo-ion and/or prosthetic groups were obtained from EcoCyc [133], protein structures of *E. coli* enzymes and primary literature. The information was manually assembled, while the network reactions were formulated based on the template reactions (see Thiele *et al.* [277] and Chapter 5 for details).

Creation of ME-matrix The tr/tr reactions for all metabolic genes were added to the E-matrix by adding additional rows and columns for the new components and reactions, respectively.

Table 8.1: **Information used for synthesis reactions of *E. coli*'s metabolic genes.**

Information	Subsystem/Reaction	Source/Reference
Transcription unit	Transcription	EcoCyc [133]
Gene coordinate, direction	Transcription	Riley *et al.* [237]
Gene function	Metabolism	*i*AF1260 [78]
Protein information	Protein complex formation	*i*AF1260 [78], EcoCyc [133], primary literature
Metallo-ion	Protein maturation	EcoCyc [133], protein structure, primary literature
Prosthetic group	Protein maturation	EcoCyc [133], protein structure, primary literature

The integration of iAF1260 and this extended E-matrix was done computationally by creating a non-redundant reaction list containing the union of both reconstructions. Functional overlap between the two networks exists on two points: i) exchange reactions of the E-matrix and the metabolic synthesis reactions; and ii) the metabolites incorporated by the E-matrix into RNA and proteins that are also consumed by the biomass reaction of the metabolic network (Figure 8.1 and 8.2). This intermediate matrix was deemed metE-matrix, in which the metabolic reactions are unaltered, i.e., they do not include the catalyzing enzymes.

The metabolic reactions were reformulated to include their enzymes in a subsequent step, which resulted in the ME-matrix. This was done using Matlab.

The merged matrix involves adding enzymes, enzyme complexes, and inactive enzymes as metabolites to the metE-matrix.

A reaction (G6PP) such as:

- $G6P \ + H_2O \ \rightleftharpoons \ Glc - D \ + \ P_i$

The preceding equation can be changed by adding enzymatic complexes. First, information is collected about the reaction (G6PP), specifically:

- Gene Loci = b0822
- Gene = ybiV

- Protein = YbiV

Second, the original reaction is converted to the following (notice name change):

- $G6PP\_A:\ G6P\ +\ H_2O\ +\ YbiV\_mono\ \rightleftharpoons\ YbiV\_G6P\_cplx$

Third, new reactions are added at the end of the reaction list. They are the following:

- $G6PP\_B:\ YbiV\_G6P\_cplx \rightarrow YbiV\_Glc - D\_cplx$
- $G6PP\_C:\ YbiV\_Glc - D\_cplx \rightarrow Glc - D\ +\ P_i\ +\ YbiV\_mono\_inact$
- $G6PP\_DREC:\ YbiV\_mono\_inact \rightarrow YbiV\_mono$

If the reaction is reversible (which G6PP is) the reverse reactions are:

- $G6PP\_E:\ Glc - D\ +\ P_i + YbiV\_mono \rightleftharpoons YbiV\_Glc - D\_cplx\_R$
- $G6PP\_F:\ YbiV\_Glc - D\_cplx\_R \rightarrow YbiV\_G6P\_cplx\_R$
- $G6PP\_G:\ YbiV\_G6P\_cplx\_R \rightarrow G6P\ + H_2O\ +\ YbiV\_mono\_inact$

If the equation occurred in the periplasm ([p]) or endoplasm ([e]), transport reaction would have also been included. This reaction is in the cytoplasm, not requiring transport reactions, however if they were, hypothetical transport reactions would be the following:

- If in the periplasm:
  - $YbiV\_export[p]:\ YbiV\_mono \rightleftharpoons YbiV\_mono[p]$
- If in the endoplasm:
  - $YbiV\_export[p]: YbiV\_mono \rightleftharpoons YbiV\_mono[p]$
  - $YbiV\_export[e]: YbiV\_mono[p] \rightleftharpoons YbiV\_mono[e]$

When the reaction is in either the periplasm or endoplasm, the $YbiV\_mono$ would be $YbiV\_mono[p]$ and $YbiV\_mono[e]$, respectively.

All lower and upper bounds are set to the -inf and +inf, unless the reaction is only in the forward direction (0 +inf).

The aforementioned example assumes only one gene to one protein. There are three other possibilities.

First is the "OR" case. Two or more different genes can code to a protein that can facilitate the same reaction. In this case, each gene is treated as its own reaction as shown above. Therefore if the G6PP reaction could be created by $YbiV$ and some other protein ($XxxY$), the script would create the reactions listed above and also repeats the process with $XxxY$ reactions. In this instance, the naming convention for reactions is changed. Instead of using $G6PP\_A$, $G6PP\_YbiV\_A$ and $G6PP\_XxxY\_A$ are used to differentiate between the different proteins.

The second instance is the "AND" case. Multiple genes code multiple proteins that must form a complex to facilitate the reaction. In this case, an additional reaction is added known as the complex formation reaction. Suppose $YbiV$ and $XxxY$ are both required to facilitate $G6PP$. A complex formation reaction would be created:

- $YbiV\_XxxY\_cplx\_FORM : \ YbiV \ + \ XxxY \rightleftharpoons YbiV\_XxxY\_cplx$
- This new complex would then be used in the reactions above replacing $YbiV\_mo-no$.

The third instance is the combination of both the "OR" and "AND" case. The rules laid out above are then used to combine the two.

Coupling constraints In traditional stoichiometric networks, proteins and mRNAs are not explicitly modeled and do not account for the representation of molecule concentrations. We have developed a new set of constraints added to the E- and ME-matrix reconstructions, which allows the representation of pools for each network component. (See Thiele *et al.* [276] and Chapter 7 for more details).

Simulation constraints used Experimental measurements of substrate and oxygen uptake rates were applied on the exchange reactions (see Table 8.4). Note that the unit of the ME-matrix is $\frac{nmol}{g_{DW} \cdot h}$. Therefore, the listed rates had to be multiplied by a factor of 10^6. In addition, the maximal reaction rates of stable RNA synthesis were constrained as described in [277] and Chapter 6. The rates were calculated based on the experimentally measured growth rates. Similarly, the maximal reaction rates of mRNA synthesis were constrained using the same approach but changing the mRNA transcription elongation rate according to the data listed in Table 2.3. In all simulation, the non-growth associated maintenance (ATPM) requirement was set to $v_{min,ATPM} = v_{max,ATPM} = 8.39 \cdot 10^6 \ \frac{nmol}{g_{DW} \cdot h}$ as defined in Feist *et al.* [78].

Adjustment of biomass The amino acid and growth associated maintenance (GAM) of the *E. coli* biomass reaction in the ME-matrix was adjusted to account for the cost of synthesis of the machinery and proteins in the ME-matrix. After performing a sensitivity analysis for these two parameters (see Result section), we adjusted the biomass reaction to account for 50% of the amino acid content and 50% of the GAM of the biomass reaction in the metabolic reconstruction, iAF1260. The adjusted biomass reaction was used in all simulations if not noted differently.

Biomass yield The growth rates were determined at different substrate uptake rates (SUR) by setting the lower and upper bound on the corresponding exchange reaction to the tested value (e.g., $v_{min} = v_{max} = -1 \cdot 10^6 \, \frac{nmol}{g_{DW} \cdot h}$). We then maximized the biomass reaction.

Growth comparison with Biolog and iAF1260 Biolog data were downloaded from the website (http://biolog.com) for *E. coli* K12 MG1655. A total of 170 tested compounds were in the reconstruction. The different environments were simulated by adding compounds to a base medium and allowing oxygen to be consumed ($v_{min} = -18.5 \cdot 10^6 \, \frac{nmol}{g_{DW} \cdot h}$ and $v_{max} = 0 \, \frac{nmol}{g_{DW} \cdot h}$).

The base medium allowed the free uptake of the following compounds by setting their corresponding lower bound to $v_{min} = -1 \cdot 10^6 \, \frac{nmol}{g_{DW} \cdot h}$:
EX_h2s(e), EX_ca2(e), EX_cl(e),EX_co2(e), EX_cobalt2(e), EX_cu2(e), EX_fe2(e), EX_fe3 (e), EX_h2o(e), EX_h(e), EX_k(e), EX_mg2(e), EX_mn2(e), EX_mobd(e), EX_na1(e), EX _tungs(e), EX_zn2(e), EX_cbl1(e).

Furthermore, the maximal possible transcription rates for each stable RNA transcription unit and for each protein coding gene were limited assuming a doubling time of 24 minutes (which provides an upper bound), since we have no information about growth rates for the different growth conditions tested in the Biolog data.

The ribosome production rate (DM_rib_50) and the biomass reaction (Ec_biomass _iAF1260_core_59p81M) were unbounded. Each nutrient was added to the base medium by setting the corresponding uptake rate to $v_{min} = -10 \cdot 10^6 \, \frac{nmol}{g_{DW} \cdot h}$ in the case of carbon sources, and $v_{min} = -20 \cdot 10^6 \, \frac{nmol}{g_{DW} \cdot h}$ in the case of nitrogen, phosphorus, and sulphur sources. Default elemental sources were as follows: D-glucose as carbon source, ammonium ion (NH_4) as nitrogen source, orthophosphate (p_i) as phosphorus source, and SO_4 as sulfur source. The sources were added to the base medium, when the corresponding source was not tested for. The growth results for iAF1260 were obtained from [78].

Single gene deletion study of metabolic enzymes The ME-matrix accounts for 1,260 metabolic genes. We tested the *in silico* growth phenotype of the single gene deficient strain in aerobic glycerol minimal medium and compared them with recently published experimental study [129] and with the *in silico* single knockouts of iAF1260 (results were taken from [78]).

Performing a single gene deletion study in the ME-matrix is a little different to the single deletion study in metabolic networks, because (i) proteins are explicit part of the metabolic reactions and (ii) transcription may occur with other genes (if co-expressed in a transcription unit) and thus coupling constraints would cause all genes in the transcription unit to not be expressed. Therefore, all translation initiation reactions for the gene are identified (e.g., 'tl_ini_bxxx') and the corresponding lower and upper bounds were set to zero. Then, all coupling constraints were identified and removed (see above and Chapter 7 for more details on coupling constraints). We then maximized for the biomass reaction in the *in silico* knockout strain. The same procedure was repeated for all 1,260 metabolic genes.

tRNA deletion For the deletion of tRNA genes a similar approach was employed as for single gene deletion. First, we identified all tRNA synthesizing reactions, set their lower and upper bound to zeros, and removed subsequently all coupling constraints. A total of 106 reactions were deleted accounting for the synthesis of 86 tRNA molecules (more reactions than tRNA exists due to the overlapping codon recognition - Table 8.7 and 8.8).

Creation of *in silico* strain library

Strains with biased codon usage The biased strains were generated using the following algorithm:

```
Input: model, sequence for each gene in model, number of iterations m

Output: model_biased
```

Algorithm:

```
1. Choose randomly a codon, $c_1$
2. Identify possible synonymous codons: $c_s = \{c_1 = c_{s1}, c_{s2}, , c_{sk}\}$
3. Choose randomly one codon from $c_s$: $c_{si}$
4. Replace all instances of $c_1$ with $c_{si}$
5. Update ME-matrix for all genes based on new gene sequence:
    (a) Transcription reactions
    (b) mRNA degradation reactions
    (c) Translation reactions (tRNA molecule will be updated based on codon recognition)
6. Repeat 1 through 5 m times
```

Strains with equilibrated codon usage The equilibrated strains were produced as follows:

Input: model, sequence for each gene in model, number of iterations m

Output: model_eq

Algorithm:

```
1. Initialize vector codon= zeros, which will count the occurrences of different codons in the genome
2. Define a random order of genes to start step 3
3. For each gene $i$ of the model genes
    (a) For each codon $c_{s,j}$ in gene sequence $i$
    (b) Identify possible synonymous codons: $c_{s,j} = \{c_1 = c_{s1}, c_{s2}, , c_{sk}\}$
    (c) Choose codon $c_{s,j}$ from $c_{s,j}$ with lowest usage in vector codon
    (d) Replace $c_{s,j}$ with $c_{s,j}$ in gene sequence $i$
    (e) Update codon
4. Update ME-matrix for all genes based on new gene sequence:
    (a) Transcription reactions
    (b) mRNA degradation reactions
```

   (c) Translation reactions (tRNA molecule will be updated based on codon recognition)

5. Repeat 1 and through 4 m times

Note that each strain has its own ME-matrix, which contains the alterations.

GC content The GC content of the individual strains was calculated by counting the instances of guanine and cytosine residues in the 1,823 protein coding genes included in the ME-matrix. The genome sequence used for this analysis was version m56, [29].

Entropy In order to quantify the degree of synonymous codon bias in a sequence we computed the synonymous codon entropy [302]. We used the entropy function since it reaches a maximum when all codons have equal probability of coding for their respective amino acids. Conversely, the entropy reaches its minimum when each amino acid is exclusively coded for by one of its possible codons. The synonymous codon entropy, H_{synnon}, is defined as

$$H_{synnon} = -\frac{\sum_{a=1}^{20}\left(N_a\left(\sum_{c=1}^{64} p_{ac}\ln p_{ac}\right)\right)}{\sum_{a=1}^{20} N_a}$$

where p_{ac} is the probability that amino acid a is encoded by codon c, and ln denotes the natural logarithm. If no amino acid is not coded for by a particular codon, $p_{ac} = 0$, then we use the definition $0 = 0 \ln 0$. Here we weight the contribution to the total synonymous codon entropy by the number of each particular amino acid, N_a, within a sequence. This means that a rare amino acid with highly biased synonymous codon usage does not overly effect the total entropy of a sequence if the remainder of the common amino acids have relatively unbiased codon usage. Since we wish to compare the synonymous codon bias between genes, we normalize the total by the total number of amino acids in a sequence, $\sum_{a=1}^{20} N_a$. If we wish to calculate the total entropy for a set of genes then we simply sum up the synonymous codon entropy for each gene's sequence, then divide by the total number of genes. Therefore, the total synonymous codon entropy is comparable between different sequences, such as mutant biased, wild type, and mutant equilibrated strains, which have low, medium and high total synonymous entropy, respectively.

Linear programming method Many linear programming (LP) problems have multiple optimal solutions (alternate optimal solutions), which have the same optimal value but differ in the flux values for the individual network reactions (see also Chapter 7).

Different LP methods exist, with the simplex methods [227] being the one most frequently used. The simplex method searches the optimal solution by moving along the boundaries of the solution space. Subsequently, many reaction fluxes have boundary flux values (mostly zero). The Barrier method, in contrast, penalizes solutions that are close to the boundaries [227]. As a consequence, solution points identified with the Barrier method have more non-zero flux values. Here, we used the barrier method as LP methods if not stated differently.

Numerical tests Calculating with the ME-matrix is rather time-consuming and numerically challenging due to the matrix's stiffness (see Results section). Therefore, it is required to test each computed point if it lies within the solution space (i.e., test if $S \cdot v = 0$ is true). We evaluated every solution for feasibility status returned by the solver and the associated error ($S \cdot v = 0$).

All simulation were carried out in Matlab (Mathwork, Inc.) using Tomlab (Tomlab, Inc.) as numerical analysis interface (for linear programming).

8.3 Results & Discussion

The integrated reconstruction of *E. coli*'s metabolic, transcription and translational network is the most comprehensive and complex biochemical model available to-date. The range of possible applications is just emerging. Here, we will present the content of the reconstruction, which biological functions have been accounted for as well as the gene coverage. Furthermore, we will compare predicted properties with experimental data and the available metabolic reconstruction of *E. coli.* Finally, we illustrate possible applications of this integrated model and how it can be used to obtain further insight in biological properties of *E. coli* and governing constraints in biological systems.

8.3.1 Properties of the reconstruction

The integrated reconstruction of metabolism and gene expression (transcription and translation) will be referred to the ME-reconstruction and its mathematical format will be called ME-matrix. The ME-reconstruction accounts for 1,260 metabolic genes, 303 macromolecular synthesis machinery genes, and 375 genes that are co-expressed with the former listed genes. A total of 1,823 protein coding genes and 115 RNA coding genes are

Table 8.2: **Functional coverage of the ME-matrix.** Distribution of clusters of orthologous groups (COG) is shown for a total of 2,806 *E. coli* genes, of which 1,436 are in the ME-matrix. There are no *E. coli* genes associated with the following COG: Chromatin structure and dynamics (B), Nuclear structure (Y), Cytoskeleton (Z), Extracellular structures (W)

COG Category	ME genes	non-ME genes
Amino acid transport & metabolism (E)	253	65
Carbohydrate transport & metabolism (G)	182	82
Energy production & conversion (C)	181	57
Inorganic ion transport & metabolism (P)	137	50
Translation, ribosomal structure & biogenesis (J)	129	30
Coenzyme transport & metabolism (H)	114	21
Cell wall/membrane/envelope biogenesis (M)	113	82
General function prediction only (R)	106	245
Nucleotide transport & metabolism (F)	70	4
Lipid transport & metabolism (I)	56	20
Transcription (K)	49	142
Posttranslational modification, protein turnover, chaperones (O)	47	74
Function unknown (S)	42	248
Replication, recombination & repair (L)	30	110
Secondary metabolites biosynthesis, transport & catabolism (Q)	28	25
Signal transduction mechanisms (T)	24	81
Defense mechanisms (V)	11	30
Intracellular trafficking, secretion, & vesicular transport (U)	11	85
Cell cycle control, cell division, chromosome partitioning (D)	10	21
Cell motility (N)	1	79
RNA processing & modification(A)	1	0

captured in the ME-reconstruction along with their synthesis reactions resulting in active, functional gene products.

Functional coverage The functional coverage of the genes included in the ME-recon-struction may be best accessed by looking at the distribution of clusters of orthologous groups (COG) [274]. A total of 2,806 *E. coli* genes has a COG function assigned, of which 1,436 are in the ME-reconstruction. The remaining 496 ME-genes have no COG information and thus cannot be considered for functional coverage analysis. As expected the majority of the central metabolic functions are captured by the ME-matrix (Table 8.2). 142 genes of the Transcription category are not included in the ME-matrix, as it does not accounts for transcriptional regulation yet. Similarly, genes of the replication and signal transduction categories are missing.

This evaluation shows that the overall functional coverage of the ME-reconstruction is well within its scope (i.e., metabolism and macromolecular synthesis) and highlights the remaining functions to be included to obtain a more complete cell-scale representation of *E. coli*. Having accounted for almost half of the functional gene products in *E. coli*, we have now a biochemical model that covers many of the known functions and characteristics of *E. coli*.

Properties of the components Most notably, the ME-reconstruction accounted for information regarding protein complex formation, metallo-ion requirement, and necessary prostethic groups of enzymes. The metabolic reconstruction provided information regarding the gene-protein-reaction (GPR) associations that encode in Boolean terms, which genes encode what metabolic functions. While the GPR capture multiprotein complexes, it does not contain any information regarding homomers. This information was manually retrieved from literature, databases, and protein structures (see Table 8.1). A total of 495 protein complex formation reactions were added manually. Furthermore, 305 proteins containing metallo-ions including iron-sulfur clusters (as $[Fe_4S_4]^{2+}$ and $[Fe_2S_2]^{2+}$, depending on preference), magnesium, etc (Table 8.3). Note that only those metallo-ions were included that were reported to be covalently bound to the protein. If no information about the number of associated ions could be found, we assumed one ion per monomer. Moreover, the ME-reconstruction accounts for 11 different kinds of prosthetic groups in 99 proteins (Table 8.3).

Table 8.3: **Metallo-ions and prostethic groups included in the ME-matrix.**

Name	Abbr.	# of proteins
Metallo-ions		
Calcium	Ca2	4
Cobalt	cobalt2	7
Cupper	Cu2	2
Iron(II)	Fe2	13
iron-sulfur-cluster	$[Fe_2S_2]^{2+}$	34
iron-sulfur-cluster	$[Fe_4S_4]^{2+}$	2
Kalium	K	12
Magnesium	Mg2	173
Manganese	Mn2	27
Molybdate	Mo	1
Nickel	Ni2	2
Zinc	Zn2	44
Prostethic groups		
2'-(5"-triphosphoribosyl)-3'-dephospho-CoA	2tpr3dpcoa	1
Adenosylcobalamin	adocbl	1
Biotin	btn	1
Flavin adenine dinucleotide (oxidized)	fad	21
Flavomononucleotide (oxidized)	fmn	9
Flavomononucleotide (reduced)	fmnh2	1
Nicotinamide adenine dinucleotide	nad	6
Nicotinamide adenine dinucleotide phosphate	nadp	2
Pyridoxal 5'-phosphate	pydx5p	42
Protoheme	pheme	6
Siroheme	sheme	2
Thiamine diphosphate	thmpp	7

These latter information have not been considered in any other reconstruction or mathematical model, to our knowledge, for these cellular functions.

Links between metabolism and macromolecular synthesis The subsystems, or biological functions, covered by the ME-reconstruction is the sum of the subsystems of the parental reconstructions. The functional synergy between the two networks is schematically illustrated in Figure 8.1. The E-matrix part produce the enzymes the M-matrix requires to catalyze the synthesis of amino acids and nucleotide triphosphates, which are in turn required by the E-matrix to produce the functional gene products. This co-dependency creates tight constraints between the two matrices which will govern the overall possible functional states as we will see in the following.

Another important property of this integrated network is that the magnitude of flux rate of RNA and protein synthesizing reactions is different to the flux rates of metabolic reactions. The flux unit in most metabolic reconstructions is $\frac{mmol}{g_{DW} \cdot h}$. In contrast, the unit of most macromolecular synthesis reaction is $\frac{nmol}{g_{DW} \cdot h}$. This due to the fact that many precursors (i.e., amino acids and nucleotide triphosphates) are needed to form one macromolecule (Figure 8.3).

As a consequence, calculations with the ME-matrix are numerically challenging due to its size and stiffness (i.e., large numbers and large reaction rates). In particular, the addition of coupling constraints (see below) make calculation with the ME-matrix time-consuming (i.e., to solve a LP problem) and cumbersome (i.e., numerical accuracy). To-date, the only Matlab interface to numerical analysis able to handle this matrix is Tomlab, which we used in all simulations.

Coupling constraints The conversion of a reconstruction to a mathematical model normally consists of the definition of the systems boundaries, the addition of exchange and demand reactions and the application of condition-specific constraints on exchange and/or intracellular reactions (See Chapter 3). The same steps were undertaken to convert the ME-reconstruction into the ME-matrix. However, an additional step needed to be performed, which added further constraints to the model (Figure 7.2 and 7.3). These constraints are called coupling constraints and link (or "couple") the flux through a biosynthetic flux, v_s, (e.g., transcription) with the corresponding utilization reaction(s), v_u, (e.g., translation). The formulation of the constraints ensure that if the biosynthetic flux is zero

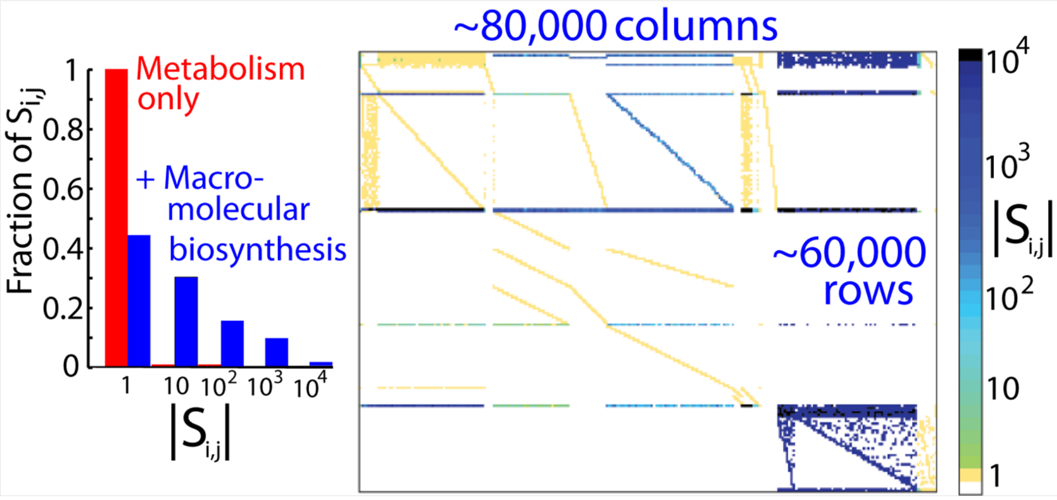


Figure 8.3: **Stoichiometric coefficients in the ME-matrix. Left:** Histogram of log10 magnitude of stochiometric coefficients. iAF1260 (red), Merged matrix (red) The coefficients are spread over 5 orders of magnitude because of 1. the wide range of reaction rates in metabolism versus macromolecular synthesis, and 2. the wide difference in number of biochemical moieties within different biochemical species. **Right:** Merged matrix showing magnitudes of coefficients using log10 colorbar (far right). Colorbar: $S_{ij} = 1$ (light orange), up to $S_{ij} = 10{,}000$ (black).

then the utilization flux has to be zero as well. An upper bound on the coupling constraint ensured that if an utilization reaction carries a high flux that the biosynthetic flux is higher as well. This requires the network to produce more gene products if they are highly used and thus represents a limit on enzyme capacity. The coupling constraints and their consequences to the steady-state solution space have been discussed in detail in Chapter 7 and [276].

8.3.2 Model Validation

In the following section, we will use experimental data, retrieved from literature, and the predictions of the metabolic reconstruction, iAF1260, to evaluate and validate the ME-matrix predictive capability. We found that the predictive potential of the ME-matrix is comparable with iAF1260 and in some cases improvements of the predictions could be observed.

Adjustment of biomass The ME-matrix accounts for the synthesis of almost half of the functions encoded in *E. coli*'s genome. Subsequently, the biomass reaction, which accounts for precursors to the macromolecular building blocks, needs to be adjusted for

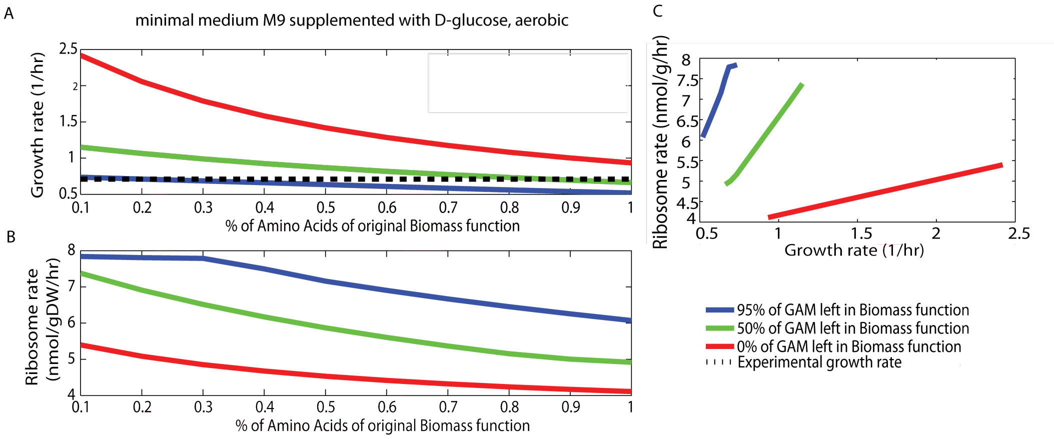


Figure 8.4: **Sensitivity analysis.** We tested the sensitivity of the predicted growth rate as a function of the remaining amino acid requirement in the biomass function and as a function of the remaining growth associated maintenance (GAM) that is left in the biomass function. The experimentally observed growth rate is shown with the dotted line. Since the ME-matrix covers about 1,900 of 4,400 *E. coli* genes, we decided to allocate 50% of the amino acid requirements and the 50% of the GAM for the ME-matrix genes and gene products. This plot also highlights that finetuning of these two parameters will be important to obtain accurate predictions in growth rate.

the fraction of amino acids (AA) and nucleotide triphosphates (NTP) use for synthesis of ME-matrix proteins and RNAs. Therefore, we carried out a sensitivity analysis to identify the right parameters to such that the model achieved the experimentally observed growth rates (Figure 8.4 and Table 8.4). Two main parameters were considered: (i) the fraction of amino acids required in the biomass reaction and (ii) the fraction of the growth associated maintenance (GAM) requirement. The later one is included in the biomass reaction to account for the energy necessary to synthesize RNA and proteins (in terms of ATP hydrolysis)[78] (Chapter 3). Note that we did not alter the fraction of NTPs since their overall contribution is relative small in the biomass reaction. We found that a good overlap between *in silico* and *in vivo* growth rate was achieved when the biomass reaction was adjusted to 50% of the amino acid requirement and 50% of the GAM. Finetuning these two parameter may lead to an improvement of quantitative growth rate predictions.

Accurate prediction of growth rate

The ME-matrix can predict quantitative growth phenotypes given experimentally measured substrate and oxygen uptake rates (Table 8.4). The experimental data were obtained from the literature and correspond to wildtype strains in multiple environmental

Table 8.4: **Comparison of predicted and experimentally determined growth rates.** SUR = substrate uptake rate. OUR = oxygen uptake rate. O^+ = aerobic. O^- = anaerobic. WT = wildtype strain. EV = evolved strain. The experimental data were obtained from [51, 84].

Condition		SUR (min/max) $\frac{mmol}{g_{DW} \cdot h}$	OUR (min/max) $\frac{mmol}{g_{DW} \cdot h}$	ME-matrix $(\frac{1}{hr})$	iAF1260 $(\frac{1}{hr})$	*in vivo* $(\frac{1}{hr})$
Glucose	O^+	8.7/10	14/18.5	0.688	0.886	0.71
	O^-	16.5/17.5	0/0	0.359	0.391	0.48
Glycerol	WT	7/8	11/13	0.281	0.412	0.22
	EV	14.5/15.5	16/17	0.637	0.788	0.5
Lactate	WT	13/14.5	13.5/18	0.316	0.579	0.23
	EV	17.5/18.5	18/20	0.383	0.674	0.5

conditions (i.e., minimal medium supplemented with glucose, glycerol, or lactate in aerobic and anoxic conditions). Furthermore, the wildtype cells were evolved on minimal medium supplemented with glycerol or lactate and after 60 days of evolution (with optimal growth as selection pressure) the substrate and oxygen uptake rates were measured [84, 51]. We compared the ME-matrix predictions with optimal growth rates calculated with iAF1260. We found that in many cases the metabolic network predicted too high growth rates, while the ME-matrix growth rates were often below the experimentally measured ones (Table 8.4). This is mainly caused by the parameters used for remaining amino acids and GAM in the biomass reaction of the ME-matrix. As the results of the sensitivity analysis showed (Figure 8.4), these two parameters play a key role in accurate growth rate prediction and will require adjustment to match the measured growth rates.

Reduced cost Reduced cost is a parameter of linear programming (LP) problems, which is associated with each network reaction (v_i) and represents the amount by which the objective function (e.g., growth rate) could be increased when the flux rate through this reaction was increased by a single unit [225]. Reduced cost is often used to analyze the obtained optimal solution and evaluate alternate solutions from the original solution [225]. In this study, we use the reduced cost analysis to identify constraining reaction rates in the model. Therefore, we analyzed the reduced cost of the four simulated conditions (see Table 8.4, only wildtype data were used). We found that the transcription initiation reactions of the rRNA operons had the greatest reduced cost associated in all four conditions. This result was somehow expected, as ribosome synthesis rate and biomass production are competing for resources. Curiously, the transcription initiation reactions of two further transcription units were associated with significant reduced cost:

1. A 15 gene operon (TU0-941, b0095-b0081).
2. A 2 gene operon (TU00334, b3642 (pyrE, orotate phosphoribosyltransferase) and b3643 (rph, RNase PH, pseudogene))

These two reactions appear in the reduced cost analysis as their reaction rates are at the upper bound (which were set based on mRNA transcription elongation rates and gene dosage, see Material & Methods for more details). More interestingly, the second operon is well known from evolution experiments on D-lactate of *E. coli*, where seven out of 11 evolved strains showed a 85 bp deletion in b3643 which is believed to increase the transcription rate of the second gene (b3642) (T. Conrad, personal communication).

Improvement of gene deletion analysis We used the ME-matrix to determine *in silico* growth phenotypes for single gene knockout strains in glycerol minimal medium. We considered only the 1,260 metabolic genes and compared the predictions with the *in vivo* essential genes [129, 14]. Joyce *et al.* evaluated the essentiality of 904 metabolic genes *in vivo* and *in silico*, using a metabolic network of *E. coli* [232].

We found 979 non-essential genes which were also reported non-essential in the two studies. A total of 132 essential genes agreed with the *in vivo* essentiality. The ME-matrix improved the prediction of seven essential genes (Table 8.5), which were non-essential *in silico* when the metabolic network was used alone [129].

Table 8.5: **Improved prediction of gene essentiality** These genes were correctly predicted to be essential in the ME-matrix compared to the metabolic network used in Joyce *et al.* [129]

Blattner #	Locus Name	Function
b0052	PdxA	4-hydroxy-L-threonine phosphate dehydrogenase, NAD-dependent
b2320	PdxB	erythronate-4-phosphate dehydrogenase
b2564	PdxJ	pyridoxine 5'-phosphate synthase
b0003	ThrB	homoserine kinase
b0004	ThrC	threonine synthase
b3926	GlpK	glycerol kinase
b0052	PdxA	4-hydroxy-L-threonine phosphate dehydrogenase, NAD-dependent

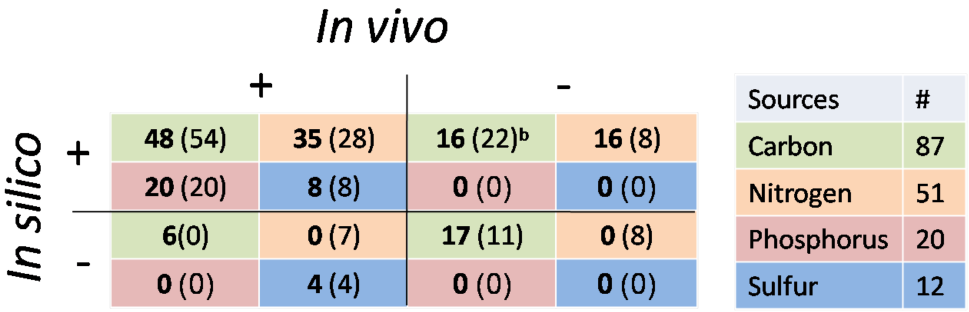
The ME-matrix predicted 39 essential genes, which are non-essential. While 86 genes are required for growth *in vivo* while the model genes are non-essential. These genes span 11 metabolic subsystems (Table 8.6).

Table 8.6: **Remaining false positives predictions of gene essentiality**

Subsystem #	Gene names (Blattner #)
Alanine and Aspartate Metabolism	aspC(b0928)
Alternate Carbon Metabolism	yhfE(b3385)
Arginine and Proline Metabolism	dtu(b3359), argB(b3959), argC(b3958)
Cell wall	kdsC(b3198), mepA (b2328)
Cofactor and Prosthetic Group Biosynthesis	coaE(b0103), coaA(b3974), yadA(b0159), pabC(b1096), pdxY(b1636), ydiB(b1692), thiJ(b2103), thiH(b3990), thiB(b3991), thiF(b3992), thiA(b3993), thiC(b3994), thiS(b4407), gapB(b2927), cysG (b3368), aroK(b3390), dapF(b3809), hemC(b3997), ubiC(b4039)
Glycolysis/Gluconeogenesis	glpX(b3925)
Methionine Metabolism	luxS(b2687)
Pentose Phosphate Pathway	rpe(b3386)
Transport	argG(b3172), btuB(b3966), yrbG(b3196), zupT(b3040), pnuC(b0751), aqpZ(b0875), glpF (b3927)
tRNA	alaS(b2697), glyS(b3559)
Valine, Leucine, & Isoleucine Metabolism	ilvE(b3770)

Comparison of Biolog data. Biolog data were used to compare with predicted growth phenotypes of the ME-matrix. The results are shown in Figure 8.5. Overall the ME-matrix predicted 128 of 170 growth phenotypes correctly (75%). Moreover, the ME-matrix shows improved prediction in 14 cases compared to iAF1260 but worsen the prediction in 11 cases. In particular, the ME-matrix was able to use all 51 tested nitrogen sources for growth. 48 of the 87 tested carbons supported growth *in silico* and *in vivo.*

While it is valuable to know the number of correct growth phenotypes, the analysis of the false negative and false positive results is more interesting as this may lead to new biological discoveries. False positive predictions (e.g., model can grow while no growth was



		In vivo			
		+		-	
In silico	+	**48** (54)	**35** (28)	**16** (22)^{b}	**16** (8)
		20 (20)	**8** (8)	**0** (0)	**0** (0)
	-	**6**(0)	**0** (7)	**17** (11)	**0** (8)
		0 (0)	**4** (4)	**0** (0)	**0** (0)

Sources	#
Carbon	87
Nitrogen	51
Phosphorus	20
Sulfur	12

Figure 8.5: **Comparison of *in vivo* growth phenotype predictions with *in silico* calculation.** ()The results for iAF1260 are given in parenthesis.

observed experimentally) hint towards missing regulation. We identified 32 of those case, where half of them were on carbon sources and the other half on nitrogen sources. The false positive growth on nitrogen may also be caused by the fact, that the carbon source is not known which was used in the Biolog data. Interestingly, the ME-matrix corrected six cases of false-positive predictions compared to iAF1260. This result illustrates the further confined solution space of the ME-matrix. In contrast, false negative predictions indicate missing links in the network. Since no link was removed from iAF1260, growth conditions which did not support growth of the ME-matrix, but of iAF120, were caused by the additional constraints (e.g., stoichiometric synthesis constraints or coupling constraints). For instance, four carbon sources did not supported growth of the ME-matrix anymore. One of these carbon sources is formate, which was reported as weak growth *in vivo* and in *i*AF1260 [78]. We tested growth of ME-matrix at various formate uptake rates but no growth could be observed. Two of the other false negatives are glucose-1-phosphate and fructose-6-phosphate, which did not supported growth in the previous *E. coli* metabolic reconstruction. Further analysis will be required to elucidate why the ME-matrix is not able to grow on these two media.

Taken together, our results show that the growth phenotype of the ME-matrix is comparable with the metabolic reconstruction of *E. coli*. This result was somehow expected as the metabolic reconstruction served as baseline for the ME-matrix.

8.3.3 Novel Application

One exciting aspect of the ME-matrix is to investigate phenotypic properties which cannot be addressed with conventional metabolic (or integrated) networks. In the following section, we will concentrate on three aspects and discuss further, candidate applications in the conclusion section.

Biomass Yield calculation

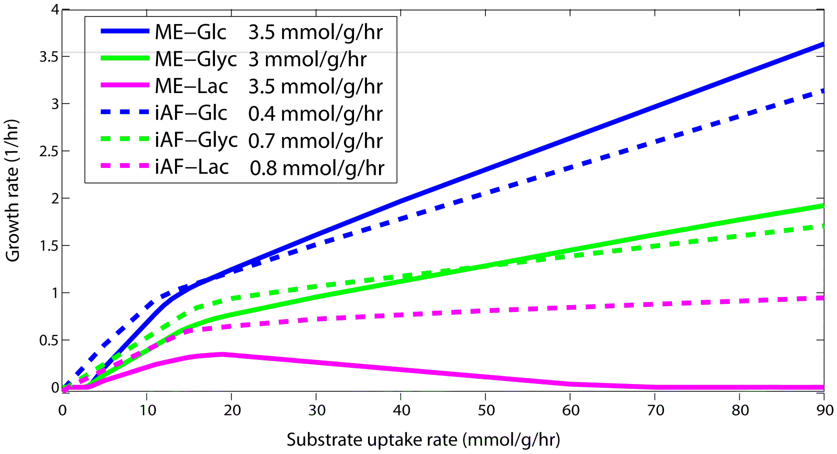


Figure 8.6: **The growth of the ME-matrix and iAF1260 as a function of substrate uptake rate (SUR).** The minimal SUR necessary to support growth (for non-growth associated and growth associated maintenance cost) is given in the legend.

Model reproduce successfully increased requirement for biomass for higher growth rates Biomass yields can be computed based on stoichiometric needs for biomass components. Growth and yields become equivalent outcomes. If the non-growth associated maintenance is included then a time dimension appears and growth and yield are no longer equivalent.

Overall the biomass yield is smaller in the ME-matrix than in iAF1260. Moreover, the required substrate uptake rate (SUR) is much higher in the ME-matrix than in iAF1260 (by about 10 times) (Figure 8.6, legend). The growth curve of these two reconstructions are comparable. Interestingly, with increasing D-lactate uptake rate in the ME-matrix shows a reduced growth rate (Figure 8.6). This reduction in growth was caused by the surplus of D-lactate as it takes energy to remove it which cannot be used anymore for growth.

tRNA essentiality *E. coli* has 86 tRNA genes which have overlapping codon recognition (see Table 8.7 and 8.8). We tested for the *in silico* essentiality of the individual tRNA genes by deleting each gene and maximizing for biomass production.

Table 8.7: **Codons recognition by tRNA in the ME-matrix. Part I.**

	generic tRNA	codon	Amino acid
alaT, alaU, alaV	ala1-tRNA	gct	ala-L
alaT, alaU, alaV	ala1-tRNA	gca	ala-L
alaT, alaU, alaV	ala1-tRNA	gcg	ala-L
alaW, alaX	ala2-tRNA	gcc	ala-L
argQ, argV, argY, argZ	arg1-tRNA	cgt	arg-L
argQ, argV, argY, argZ	arg1-tRNA	cgc	arg-L
argQ, argV, argY, argZ	arg1-tRNA	cga	arg-L
	argU-tRNA	aga	arg-L
	argW-tRNA	agg	arg-L
	argX-tRNA	cgg	arg-L
asnU, asnV, ansW	asn1-tRNA	aac	asn-L
asnU, asnV, ansW	asn1-tRNA	aat	asn-L
aspT, aspU, aspV	asp1-tRNA	gac	asp-L
aspT, aspU, aspV	asp1-tRNA	gat	asp-L
	cysT-tRNA	tgc	cys-L
	cysT-tRNA	tgt	cys-L
glnU, glnW	gln1-tRNA	cag	gln-L
glnV,glnX	gln2-tRNA	caa	gln-L
gltT, gltU, gltV, gltW	glu1-tRNA	gaa	glu-L
gltT, gltU, gltV, gltW	glu1-tRNA	gag	glu-L
glyV, glyW, glyX, glyY	gly1-tRNA	ggc	gly
glyV, glyW, glyX, glyY	gly1-tRNA	ggt	gly
	glyT-tRNA	gga	gly
	glyU-tRNA	ggg	gly
	hisR-tRNA	cac	his-L
	hisR-tRNA	cat	his-L
ileT, ileU, ileV	ile1-tRNA	atc	ile-L
ileT, ileU, ileV	ile1-tRNA	att	ile-L
ileX, ileY	ile2-tRNA	ata	ile-L
leuO, leuQ, leuT, leuV, leuW	leu1-tRNA	ctg	leu-L
	leuU-tRNA	ctc	leu-L

Table 8.8: **Codons recognition by tRNA in the ME-matrix. Part II.**

	generic tRNA	codon	Amino acid
	leuU-tRNA	ctt	leu-L
	leuW-tRNA	cta	leu-L
leuX, leuZ	leu2-tRNA	ttg	leu-L
	leuZ-tRNA	tta	leu-L
lysQ, lysT, lysV, lysW, lysY, lysZ	lys1-tRNA	aaa	lys-L
lysQ, lysT, lysV, lysW, lysY, lysZ	lys1-tRNA	aag	lys-L
metT, metU	met1-tRNA	atg	met-L
pheU, pheV	phe1-tRNA	ttc	phe-L
pheU, pheV	phe1-tRNA	ttt	phe-L
proK, proM	pro1-tRNA	ccg	pro-L
	proL-tRNA	ccc	pro-L
proL, proM	pro2-tRNA	cct	pro-L
	proM-tRNA	cca	pro-L
serW, serX	ser1-tRNA	tcc	ser-L
serT, serW, serX	ser2-tRNA	tct	ser-L
	serT-tRNA	tca	ser-L
serT, serU	ser3-tRNA	tcg	ser-L
	serV-tRNA	agc	ser-L
	serV-tRNA	agt	ser-L
thrT, thrV	thr1-tRNA	acc	thr-L
thrT, thrU, thrV	thr2-tRNA	act	thr-L
	thrU-tRNA	aca	thr-L
thrU, thrW	thr3-tRNA	acg	thr-L
	trpT-tRNA	tgg	trp-L
tyrT, tyrU, tyrV	tyr1-tRNA	tac	tyr-L
tyrT, tyrU, tyrV	tyr1-tRNA	tat	tyr-L
valT, valU,valX, valY, valZ	val1-tRNA	gta	val-L
valT, valU,valX, valY, valZ	val1-tRNA	gtg	val-L
valV, valW	val2-tRNA	gtc	val-L
valT, valU, valW, valX, valY, valZ	val3-tRNA	gtt	val-L
metV, metW, metY, metZ	fmet-tRNA	atg	start

Of all 86 tRNA genes, we found the following seven tRNAs essential in glucose minimal medium and glycerol minimal medium: argU (b0536), argW (b2348), argX (b3796), cysT (b1910), leuU (b3174), leuW (b0672) and leuZ (b1909). All other tRNA were to be dispensable, in fact, their deletion did not significantly reduced the biomass production rate. The FVA analysis confirmed these results (data not shown). This result was expected due to the functional overlap of the tRNAs. Interestingly, these essential tRNA genes include those tRNA, which were found limiting in codon-biased strains (see below), which suggests that there essentiality is caused by high usage for their respective codons in the genome. In fact, leucine and arginine are one of the most abundant amino acids in *E. coli* (data not shown). In contrast, cysteine has only one tRNA recognizing its two codons, and cysteine appears in many proteins, though it is not very abundant in the proteome. Cysteines are needed to form S-S bonds, which are important for protein tertiary structure.

Changing growth phenotype by alternating the codon usage

Creation of an *in silico* strain library To test the impact of codon usage on the functional properties of *E. coli*, we generated 15 strains with identical gene content and location as *E. coli* wildtype, but with altered codon usage. In ten of these strains, 100 randomly chosen codons were replaced in all ME-matrix genes by one of the possible synonymous codons. (See Materials & Methods). This replacement lead to a more biased codon usage. In addition, we generated five strains which have an equilibrated codon usage. The change of codon usage was introduced to ME-matrix by (i) adapting the nucleotide triphosphate requirements in the corresponding transcription reactions, (ii) changing the nucleotide monophosphates released in the mRNA degradation reactions, and (iii) updating the tRNA species according to the new codons (Table 8.7 and 8.8). Each strain has its own ME-matrix. Note that the start codon as well as the stop codons were not modified in the strains.

Properties of the *in silico* strains and their genomes:

1. **Codon usage** While the codon usage is almost perfectly correlated in the equilibrated strains (Figure 8.7), the codon usage in the wildtype strain and the biased strains is very distinct. In fact, the codon usage in the biased strains is not correlated at all (based on Pearson correlation), except for strain B7 and B10 which have an $R^2 \approx 0.8$ (Figure 8.7). It is also notable the codon usage in the wildtype strain is only

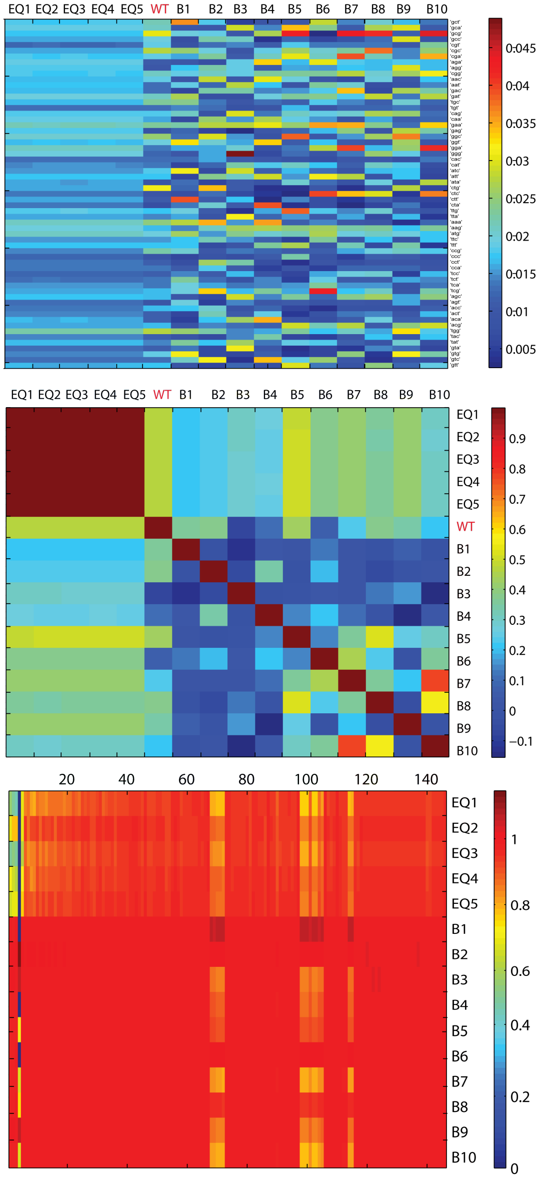
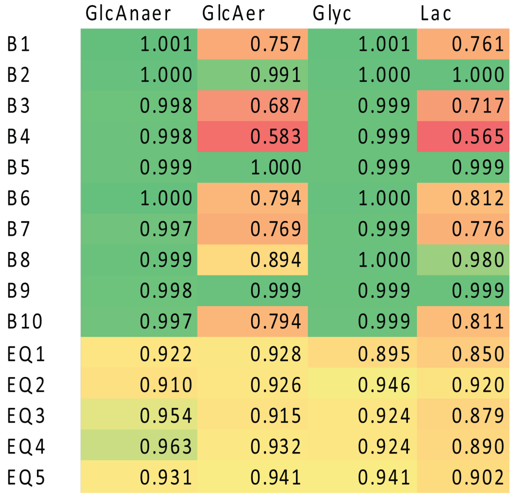


Figure 8.7: Heatmap illustrating the usage of the 61 different codons, including start codon, in the wildtype, the biased strains, and the equilibrated strains.

weakly correlated to the codon usage in the biased strains and at most moderately correlated to the codon usage in equilibrated strains.

2. **Growth phenotypes of strains** We then tested the growth performance of the strains using (i) experimental data on substrate and oxygen uptake rates (SUR and OUR, respectively), and (ii) qualitative growth data obtained from Biolog (see Materials and Methods).

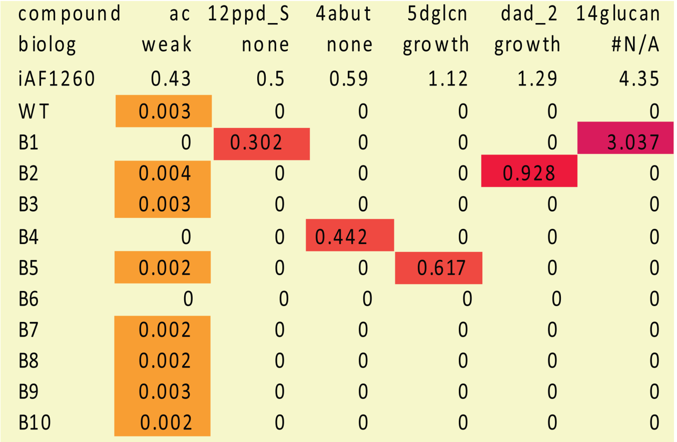


	GlcAnaer	GlcAer	Glyc	Lac
B1	1.001	0.757	1.001	0.761
B2	1.000	0.991	1.000	1.000
B3	0.998	0.687	0.999	0.717
B4	0.998	0.583	0.999	0.565
B5	0.999	1.000	0.999	0.999
B6	1.000	0.794	1.000	0.812
B7	0.997	0.769	0.999	0.776
B8	0.999	0.894	1.000	0.980
B9	0.998	0.999	0.999	0.999
B10	0.997	0.794	0.999	0.811
EQ1	0.922	0.928	0.895	0.850
EQ2	0.910	0.926	0.946	0.920
EQ3	0.954	0.915	0.924	0.879
EQ4	0.963	0.932	0.924	0.890
EQ5	0.931	0.941	0.941	0.902

Figure 8.8: **Relative growth rates of biased strains when real SUR and OUR were chosen as constraints.** See Table 8.4 for constraints and *in silico* growth rates of wildtype strain.

First, we investigated the consequences of change codon usage on the growth performance when experimentally measured OUR and SUR are applied as constraints. The growth performance of the wildtype is comparable with experimentally measured growth rates. Interestingly, no change in growth performance was observed for the biased strains when glycerol was carbon source. Similarly, the growth rate was not changed in glucose minimal medium, anaerobic conditions. In contrast, we observed up to 50% reduction in growth rates of some biased strains in glucose and lactate minimal medium, aerobic conditions (Figure 8.8). The equilibrated strains showed in all conditions a reduced growth performance (data not shown).

In a second step, we investigated the growth performance of the strains in the different nutrient conditions that have been tested by Biolog (see above). The results were rather surprising.



compound	ac	12ppd_S	4abut	5dglcn	dad_2	14glucan
biolog	weak	none	none	growth	growth	#N/A
iAF1260	0.43	0.5	0.59	1.12	1.29	4.35
WT	0.003	0	0	0	0	0
B1	0	0.302	0	0	0	3.037
B2	0.004	0	0	0	0.928	0
B3	0.003	0	0	0	0	0
B4	0	0	0.442	0	0	0
B5	0.002	0	0	0.617	0	0
B6	0	0	0	0	0	0
B7	0.002	0	0	0	0	0
B8	0.002	0	0	0	0	0
B9	0.003	0	0	0	0	0
B10	0.002	0	0	0	0	0

Figure 8.9: Growth rates biased strains under different growth conditions. Some biased strains were capable to grow where the wildtype strain (ME-matrix) was not able to grow. Note that iAF1260 were able to produce biomass in all cases. 12ppd_S = (S)-Propane-1,2-diol; 4abut = 4-Aminobutanoate; 5dglcn = 5-Dehydro-D-gluconate; dad_2 = Deoxyadenosine; 14glucan = 1,4-alpha-D-glucan

(a) Same growth capabilities as wildtype were observed with one exception: some of the strains were unable to grow on acetate as carbon source. Note that the wildtype growth rate was very low (˜0.003 $\frac{1}{h}$) and the Biolog data reported weak growth of *E. coli* K12 MG1655. Similarly, the growth rate of the *in silico* strains were very low, although biased strain B2 showed an increase in growth rate of 15% compared to wildtype (0.0035 $\frac{1}{h}$).

(b) Depending on strain and nutrient source, we observed reduction in growth rate as low as 60% (Figure 8.7). This is remarkable considering that we "only" replaced codons by synomous codons. These results clearly indicate competition for resources (i.e., for available charged tRNAs). In general, the reduction in growth rate was more pronounced in equilibrated strains than in biased strains.

(c) We identified five cases, in which a mutant was able to grow while the wildtype was unable to grow (Figure 8.7 and 8.9). Since no links were added to the network in the strains ME-matrices, these results mean that the wildtype ME-matrix was unable to grow due to constraints on tRNA availability. Changing codon usage allowed to alleviate these constraints. (Note iAF1260 was able to growth on all of these instances (Figure 8.9).) These results indicate that the transcription unit assignment is incorrect in the current ME-matrix, requiring the co-expression of low expressed genes (with mainly low frequency codons) and highly expressed genes (with highly abundant codons).

Furthermore, none of the equilibrated strains showed growth phenotype where the wildtype strain was incapable of growing.

Overall, our *in silico* results suggest that growth on certain nutrient is codon usage dependent while growth on other media was supported equally well for all strains. The first observation is in agreement with a computational study, which suggested that the codon usage is dependent on environmental conditions [297]. Willenbrock *et al.* derived this hypothesis by clustering the codon adaptation index (CAI) of more than 300 organisms and identified shared environmental niches of the organisms within a cluster [297]. Here, we showed with a computational model that changing the codon usage can have a dramatic effect on the growth performance of *E. coli* in a significant number of environments (Figure 8.8A).

The next question is what are the governing constraints? Or in other words, what happened? How can it be that the *in silico* strains perform differently in the different environmental conditions?

GC content *E. coli*'s genes (wildtype) have a guanosine-cytosine (GC) content of 53%. We analyzed (i) if the synonous codon replacement may have led to a changed GC content and (ii) if such change correlates with the observed *in silico* growth phenotype. The results suggest that *in silico* strains with lower GC content cannot achieve as high growth rates as the wildtype or other strains with higher GC content (Figure 8.10A). Furthermore, the data suggests that the maximal possible growth rate is lower for *in silico* strains with higher GC content than for the wildtype *E. coli* strain. This implies that *E. coli*'s GC content is optimal for high growth rates under the tested conditions. However, we have not sufficient data points to verify this hypothesis and in a future analysis more mutants with variable GC content need to be created.

Entropy Another approach to assess the extend of changes to the gene sequence that we introduced by biasing or equilibrating the codon usage is via an entropy. This measure reflects how biased or unbiased the codon usage of the 1,823 genes in the strains is compared to a random distribution. Hereby, it counts that the higher the value the more random is the sequence. As expected, the equilibrated strains had the highest entropy,

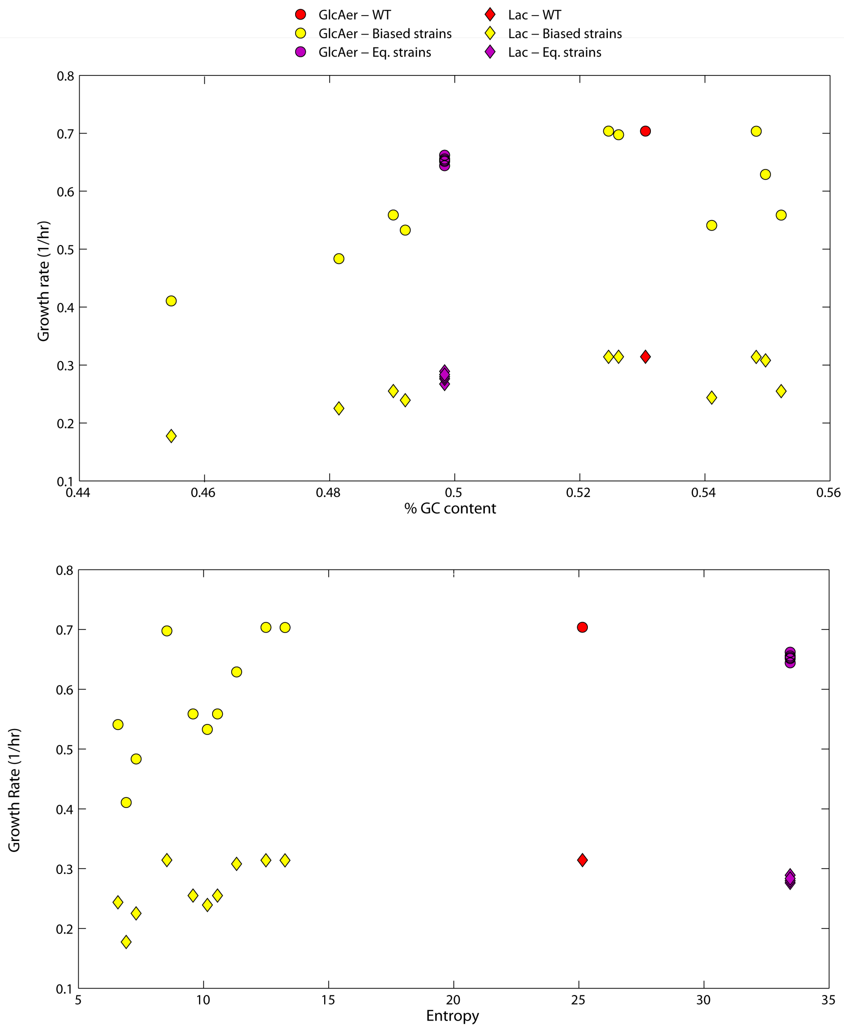


Figure 8.10: **Genome parameters affecting *in silico* strain growth rates. A.** Growth rate versus GC content. **B.** versus genome entropy. The 16 *in silico* strains are shown with their predicted growth rates in glucose minimal medium (GlcAer), aerobic conditions and lactate minimal medium, aerobic conditions (LacWT). Eq. strains = equilibrated strains.

while the biased strains had a lower entropy than the wildtype strain (Figure 8.10B). No obvious correlation between entropy value and maximal achievable growth rate could be observed, except that high entropy has a reducing effect on the growth rate (Figure 8.10B).

So far we investigated two properties of the genome in order to identify the cause for the observed *in silico* reduction in growth rate. However, our parameters were not highly correlated with the predicted growth rate. In the following section, we will analyze a network property associated with the optimal solution, the reduced cost.

Reduced cost In Section 8.3.2, we investigated the reduced cost associated with the optimal growth solution of the wildtype strain. As mentioned, the network reactions with largest reduced cost were mainly ribosomal RNA operon transcription reactions. Here, we analyzed the reduced cost associated with all strains when they were maximized for biomass production in different environmental conditions.

For the biased strains in glycerol medium and anaerobic, glucose medium, the reactions with highest reduced cost were, as expected, the ribosomal RNA transcription reactions similar to the wildtype (data not shown). This result shows that the ribosomal RNA transcription is still growth rate limiting. In contrast, in lactate medium and aerobic, glucose growth condition, we identified tRNA transcription reaction associations with highest reduced cost (Table 8.9). While the ribosomal RNA transcription reaction had moderate reduced cost values, the reduced costs associated with the tRNA transcription reactions are significant. A reduced cost of 2 units (e.g., $\frac{mmol}{g_{DW} \cdot h}$), means that the objective reaction (here biomass production) would increase by 2 units if the flux through the reaction is increased by 1 unit. This means that an increase by one unit through the tRNA transcription reactions, the growth rate of the wildtype strain could be basically restored. However, one unit increase in transcription rate would be a substantial increase.Furthermore, it is interesting that one single tRNA operon is responsible for the observed reduction in growth rate. The majority of strains were impaired in $tRNA^{leuU}$ (Table 8.9). Even strain B2, which had a fairly mild phenotype (Figure 8.8), was tRNA synthesis limited. Interestingly, two different tRNA operons were constraining the growth of B2 in the two tested medium conditions.

The reduced cost of the reactions in the equilibrated strains is a bit different (data not shown). Many different tRNA transcription reactions had a significant reduced cost in all four tested conditions. However, none of the reduced cost were as large as for the biased strains, which confirms that these strains have multiple deficiencies, i.e., shortages in tRNA supply.

Table 8.9: **Network reactions, which have high reduced cost associated with maximal growth solutions in the biased *in silico* strains.** GlcAer = glucose minimal medium, aerobic condition. LacAer = D-lactate minimal medium, aerobic condition. The reduced cost is given in $\frac{mmol}{g_{DW} \cdot h}$

Strain		tscr_ini_ TU00518_stab b3174 $tRNA^{leuU}$	tscr_ini_ TU00500_stab b2348 $tRNA^{argW}$	tscr_ini_ TU00512_stab b1909 - b1911 $tRNA^{leuZ}$, $tRNA^{cysT}$, $tRNA^{glyW}$	tscr_ini_ TU00507_stab b0664 - b0673 $tRNA^{glnX}$, $tRNA^{glnV}$, $tRNA^{metU}$, $tRNA^{glnW}$, $tRNA^{glnU}$, $tRNA^{leuW}$, $tRNA^{metT}$
B1	GlcAer	0.307			
B2	LacAer	0.143			
B2	GlcAer		0.057		
B3	GlcAer			0.413	
B3	LacAer			0.186	
B4	GlcAer				0.376
B4	LacAer				0.205
B6	GlcAer	0.312			
B6	LacAer	0.154			
B7	GlcAer	0.295			
B7	LacAer	0.146			
B8	GlcAer	0.084			
B8	LacAer	0.035			
B10	GlcAer	0.312			
B10	LacAer	0.154			

The analysis of the reduced cost yielded in sufficient insight into what caused the change in growth phenotype, i.e., limitation of available tRNA. In the next section, we will analyze the nominal and optimal tRNA abundance.

8.4 Conclusion

In this study, we presented the first genome-scale reconstruction combining macromolecular synthesis and metabolism for *E. coli*. This work will set stage for reconstructing such integrated networks for other organisms. We showed that predictive capability of the deemed ME-matrix is comparable with its parent metabolic network. Furthermore, we illustrated novel questions and applications that can be addressed with this matrix. Most interestingly, the ME-matrix allows to assess the impact of codon usage on the functional steady-states the network can achieve.

Our results suggested that *E. coli*'s codon usage and GC content is optimized for growth on glucose and lactate, while suboptimal for growth on glycerol and glucose under anoxic conditions. This sub-optimality may explain why we did not see any changes in growth performance of the biased strains. In contrast, our results clearly show that codon bias is necessary for fast growth of *E. coli* due to competition for tRNA molecules (with the current genome organization).

Other applications will include i) expansion of the reconstruction by adding further cellular functions (e.g., transcriptional regulation), ii) study of genome evolution through minimal network determination, iii) protein engineering.

The text of this chapter, in full, is a reprint of the material as it appears in I. Thiele, R.M.T. Fleming, A. Bordbar, R. Que, B.Ø Palsson, An integrated model of macromolecular synthesis and metabolism of *Escherichia coli.*, *in preparation.* I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

Chapter 9

Conclusion: Towards whole-cell modeling

9.1 Constraint-based reconstruction and analysis approach as tool of choice

The motivation for this thesis was to develop a formalism that allows the construction and analysis of whole-cell models. Advances in constraint-based reconstruction and analysis (COBRA) techniques for metabolism showed that COBRA can be used as a powerful and insightful tool despite incomplete knowledge about metabolism. Five years ago other researchers had undertaken very important steps to create reconstructions of other cellular functions, including signaling [208] and protein synthesis [5]. It seemed obvious to strive the effort towards representing macromolecular synthesis in COBRA format and to combine different cellular networks for an integrated analysis. The latter one has been published recently in two smaller scale efforts [53, 155].

COBRA has the following compelling properties that make it a promising framework for cell-scale modeling:

1. No complete knowledge about the cellular function modeled is required. COBRA is based on reaction stoichiometry, which can be readily defined for most cellular processes. (See also Section 1.2.)
2. COBRA reconstructions represent knowledge-bases as they capture, structure, and summarize all available information regarding the cellular process in the target organism.

3. Condition-specific information and experimental data (e.g., enzymatic turnover rates, reaction rates) can be added to the stoichiometric format to reduce the set of feasible steady-state flux solutions.
4. The COBRA approach assumes the modeled system to be in steady-state, but this assumption allows to investigate properties, such as gene essentiality and by-production capacity, which may be more difficult to address with kinetic models.
5. Although current reconstructions do not account for regulation, the derived models have a high predictive accuracy. For example, more recent metabolic reconstructions correctly predict the phenotypic outcome of gene deletion in 70 to 90% of the time [281, 78, 196, 195].
6. *In silico* studies showed that stoichiometry imposes significant constraints on network capabilities, which could not be identified otherwise, [280].
7. Linear optimization approaches can be readily applied, which makes this modeling approach very scalable. Having a scalable formalism is crucial if one desires to represent multiple cellular processes.

Taken together, sufficient arguments exists as to why COBRA is the right approach for whole-cell modeling.

9.2 Challenges in reconstructing non-metabolic functions

Genome-scale metabolic reconstructions have been generated since more than 19 years. Within this thesis work, the reconstruction procedure of metabolic reconstructions was further refined and captured as a standard operation procedure that will hopefully ensure the quality and comparability of reconstructions (Chapter 3). The experience gained with the metabolic reconstructions was used to formulate all cellular processes involved in the synthesis of the protein machinery necessary to produce functional RNA and protein molecules in *E. coli*. This reconstruction was done in analogy to the procedures developed for metabolism. However, since the transcriptional/translational (tr/tr) network was the first of its kind, the biochemical reactions needed to be formulated for all necessary transformations based on literature. In comparison, the majority of the metabolic reactions used in metabolic network reconstructions can be obtained from databases, such as

KEGG [130] and Brenda [18], and textbooks. While the tr/tr reactions are generally described in molecular biology textbooks, these normally do not describe all needed details for stoichiometric representation, so that information from different literature sources needed to be evaluated (Chapter 5). In some cases, the mechanism of a cellular process is still not known or multiple, parallel models exist. In those cases, a consensus mechanism needed to be derived from publications that captures key features and components (metabolites, proteins). In short, a new way of reconstruction approach needed to be developed along with quality control measures to ensure completeness and consistency of the reconstruction (Chapter 6). At its end, an algorithmic description of the reconstruction process of tr/tr networks was obtained [79], which will facilitate further reconstruction efforts. In fact, the developed procedure can be readily applied for other bacteria (see below).

Once reconstructed, further methodology needed to be developed that considered the properties, such as explicit representation of proteins in biochemical reactions, of the tr/tr network. The addition of so-called coupling constraints allows to model and predict tr/tr properties that are consistent with cellular properties, e.g., a biochemical reaction is only active if its involved proteins are produced in the model (Chapter 7). These constraints were necessary since steady-state assumption states implies that the change of component concentration over time is zero, or with other words, the sum flux through synthesizing reactions of a components equals the sum flux of consuming reactions. As a consequence, the presence of an enzyme in the tr/tr reactions would not be required, without the coupling constraints, as they leave any biochemical reaction unchanged. Thus, this formulation of coupling constraints was crucial for simulating network properties beyond topology (Chapter 7).

9.3 Integration of metabolism and macromolecular synthesis - A stiff matrix

Finally, the tr/tr network of *E. coli* was integrated with the available metabolic reconstruction (Chapter 8). This reconstruction is the first of its kind, capturing the functions of almost 2,000 *E. coli* genes. The scale of stoichiometric coefficients and the resulting different scale of flux rates (i.e., $\frac{nmol}{g_{DW} \cdot h}$ for macromolecular synthesis reactions and $\frac{mmol}{g_{DW} \cdot h}$ for metabolic reactions) creates computational and numerical challenges, which are worsen by the use of coupling constraints (Chapter 8). On the long run, new linear programming solvers will need to be developed that are able to deal with such large, stiff matrices as integration with additional cellular processes will impose further challenges on currently available solvers. Furthermore, new COBRA techniques will have to be develop, which are able to deal with these large-scale matrices, e.g., by minimizing the number of necessary linear optimizations to obtain desired answers. Many of the available COBRA tools do not use optimized algorithms to answer questions. For example, one desired to know which and how many reactions in the network cannot carry any flux. Currently, flux variability analysis is used for finding the correct answer, which requires minimizing and maximizing the flux through every network reaction. A more optimized approach would look at each computed flux vector and exclude those reactions from analysis that carry a non-zero flux in at least one of the vectors. This approach would significantly reduce the number of linear optimization necessary to identify all blocked reactions and thus require less computation time. Within this thesis a 'omics'-toolbox was developed which provides more than 100 Matlab functions allowing to calculate, simulate, and manipulate integrated biochemical networks.

9.4 Applications of integrated models

Bacillus species are frequently used are for industrial production of natural and recombinant proteins. Traditionally, protein expression systems are developed and optimized in experimental settings which are costly and time-consuming. Comprehensive, integrated models of metabolism, regulation and RNA and protein synthesis in Bacillus could be used to design protein expression systems *in silico*. COBRA modeling could be used to over-express natural and recombinant proteins *in silico* and to define mutants that are able to produce the protein optimally.

Therefore, computational approaches for assessing the network changes caused by high-copy plasmid expression and flux re-direction need to be developed. Current *in silico* engineering techniques rely on bi-level optimization problems (e.g., OptKnock) that identify a set of candidate genes to be deleted in order to growth couple the (metabolic) by-product secretion. Due to the size of integrated networks, new techniques need to be developed. This may be mainly achieved by using linear optimization and convex analysis.

Creation of a minimal organism Craig Venter and others are on the way to generate minimal organisms, which are able to replicate (although they grow mostly on rich media). It is expected that these designed organisms will contain less than 500 genes, which is approximately the genome size of *Mycoplasma genitalium.*

Existing reconstruction technologies and integrated network of *E. coli* could be used to propose the genome composition of an *in silico* minimal organism. This *in silico* organism would of course not account for replication but the number of necessary genes) could be estimated. Such minimal organism is expected to have many auxotrophies (i.e., requirements of nutrients in the medium). Investigating the minimal genome content as a function of auxotrophies will enable to design in silico a suite of organism whose genome content depends on the growth environment and task (e.g., producing a particular by-product).

9.5 What's next

The goal of a whole cell model is clearly not reached by combining tr/tr with metabolism nonetheless it was an important undertaken. The other cellular functions to be included into such a model are listed below. However, in terms of energy cost the ME-matrix covers a bulk part of cellular energy requirements since it covers proteins synthesis. The remaining cellular functions require less of the cell. The following cellular functions are missing in a "truly" whole-cell model of *E. coli*:

- Replication of the bacterial chromosome
- Transcriptional regulation
- Enzyme regulation
- Signaling pathways, which consists mainly of two-component systems in most bacteria
- Flagellum

Chromosome replication. This cellular process could be readily reconstructed in a similar approach as the tr/tr network. It is expected that sufficiently information about the components and the biochemical reactions is available in textbooks and primary literature.

Transcriptional regulation. The reconstruction of transcriptional regulatory network will require more detailed experiments to identify which genes is expressed under which environmental and/or genetic condition(s). Genome-scale approaches are now available thanks to microarray, chromatin immunoprecipitation (ChIP) chip and sequencing technologies. However, the elucidation of complex rules (e.g., gene x is expressed if stimuli y is present, z is absent and protein u is bound) will require the generation and analysis of many datasets in different environmental and genetic conditions, which may not be available in near future. Nevertheless, I believe that the COBRA approach will be a suitable method to encode and simulate consequences of regulatory events. Recently, a pseudo-stoichiometric formalism (R-matrix) has been developed to encode the Boolean regulatory relationships between genes and environmental cues [95, 93]. Unfortunately, this formalism cannot readily be integrated with the E-matrix or the ME-matrix. Furthermore, this formalism is still limited to Boolean representation (On/Off) of regulatory rules while their nature may rather be stochastic. More research will be necessary to develop a formalism for transcriptional regulation that can be integrated with the ME-matrix. I anticipate that the coupling constraints which were used for coupling synthesis and utilization (Chapter 7) will be useful for encoding regulatory rules. The most difficult Boolean rule to represent with constraints may be A XOR B since this relation is inherently non-convex. Stochastic approaches may facilitate the implementation of those complicated rules but may not eliminate the non-convex nature of the relation ship. The aforementioned R-matrix formalism dealt with the XOR relationship by creating negated versions of A and B. Taken together, the integration of the ME-matrix with transcriptional regulation will require further advancement in modeling technologies.

Enzyme regulation. In comparison to transcriptional regulation, the enzyme regulation can be readily reconstructed since it has been studies intensively in the last 50 years, or so, and a significant body of literature is available. Regarding modeling, the implementation of enzyme regulation is much easier than transcriptional regulation. It is expected that the appropriate use of coupling constraints will enable to accurately represent different forms of enzymatic regulation.

Signaling pathways. Human signaling networks have been reconstructed using COBRA [208, 161]. Bacterial signaling networks are mainly two-component systems, consisting of a histidine-kinase (HK) and a response regulator (RR) (Chapter 4). Thus, the reconstruction technology is well established. However, further experimental data are necessary to assist the reconstruction of bacterial signaling networks as many of the environmental cues which activate the HK are not known or have tags like ”low iron concentration in medium” which cannot be translated into a stoichiometric reaction. Some experimental data exists, which may be further facilitated by structural biology/genomics reporting the number of binding sites of the sensor protein for the sensed molecule.

Flagellum The flagellum is a long, thin filament that allows bacteria to move in their environment. The flagellar movement is an energy (ATP) driven process which can be estimated from experiments. The self-assembly of the flagellum is a complex process as the flagellum consists of three main parts: an engine, a propeller and a universal joint that connects the them. However, recent reviews describe the assembly process in detail and the corresponding genes are known [8, 45]. Therefore, it is expected that the reconstruction can be obtained quickly and an integration with other cellular processes should be, in principle, straight forward.

Taken together, it appears that a whole-cell reconstruction and model of *E. coli* may be obtained within the next five to ten years, depending on the wealth of information generated. Transcriptional regulation may be the most challenging of the remaining cellular processes to reconstruct but high-throughout technologies are continuously improving.

tr/tr for all Now that reconstruction and modeling techniques have been established for a significant number of other cellular processes than metabolism, integrated networks can be readily generated for other bacteria. For instance, the tr/tr reconstruction approach, which used template reactions to generate the network reactions, can be adopted to reconstruction the tr/tr machinery for other bacteria. Having the metabolic networks available for those bacteria will quickly yield in integrated networks.

In contrast, it is possible that a whole-cell networks will first be generated for simpler organisms than *E. coli*, which have not as many transcription factors. These organisms may include *Mycoplasma genitalium* or *Helicobacter pylori.*

Chapter 10

Glossary

Bibliome The collection of primary literature, review literature and textbooks on a particular topic.

Biochemical, genetic and genomic (BiGG) knowledge base A structured genome-scale metabolic network reconstruction which accounts for and incorporates knowledge about the genomic, proteomic, and biochemical components and relationships in a network reconstruction for a particular organism or cell.

Biomass function A pseudo-reaction representing the stoichiometric consumption of metabolites necessary for cellular growth (i.e., to produce biomass). When this pseudo-reaction is placed in a model, a flux through it represents the in silico growth rate of the organism or population.

Blocked reactions Network reactions that cannot carry any flux in any simulation condition are called blocked reactions. Generally, these blocked reactions are caused by missing links in the network.

Convex space A multi-dimensional space in which a straight line can be drawn from any two points in the space, without leaving the space.

Constraint-based reconstruction and analysis (COBRA) - A set of approaches for constructing manually curated, stoichiometric network reconstructions and analyzing the resulting models by applying equality and inequality constraints and computing functional states. In general mass conservation and thermodynamics (for directionality) are the fundamental constraints. Additional constraints reflecting experimental conditions and other biological constraints (such as regulatory states) can be applied. The analysis approaches generally fall into two classes: biased and unbiased methods. Biased methods involve the application of various optimization approaches which require the definition of an objective function. Unbiased methods do not require an objective function.

Dead-end metabolite A metabolite that is only produced or consumed in the network.

When the consumption reaction(s) of a metabolite is not known or outside the scope of the reconstruction it can be represented by this unbalanced, intracellular reaction (e.g., 1 A →).

Exchange reactions These reactions are unbalanced, extra-organism reactions that represent the supply to or removal of metabolites from the extra-organism ”space”. (See Box 3 for a pictorial description).

Extreme pathways (ExPa’s) ExPa’s are a unique and minimal set of flux vectors which lie at the edges of the bounded null space. Biochemically meaningful steady-state solutions can be obtained by nonnegative linear combination of ExPa’s.

Flux-balance analysis (FBA) The formalism in which a metabolic network is framed as a linear programming optimization problem. The principal constraints in FBA are those imposed by steady state mass conservation of metabolites in the system.

Futile cycles Stoichiometrically unbalanced cycles, which are associated with energy consumption.

Gene-protein-reaction association (GPRs) A mathematical representation of the relationships between gene loci, gene transcripts, protein sub-units, enzymes, and reactions using logical relationships (AND, OR).

Genome-scale The characterization of a cellular function/system on its genome scale, i.e., incorporation/consideration of all known associated components encoded in the organism’s genome.

Genome-scale model (GEM) A GENRE can be converted into a mathematical form (i.e., an in silico model) and used to computationally assess phenotypic properties (reviewed in [219]).

Genome-scale network reconstruction (GENRE) An organism-specific BiGG knowledge base is the basis for a GENRE. The term GENRE applies to a particular organism, for example, GENRE of Escherichia coli (for which more than four consecutive reconstruction have been created by refining and expanding the predecessor). A GENRE contains a list of all the known (and some predicted) chemical transformations that are believed to take place in the particular network (e.g., metabolic, transcriptional regulatory network, etc.).

Flux variability analysis (FVA) FVA is a frequently used computational tool for investigating more global capabilities under a given simulation condition (e.g., network redundancy). Therefore, every network reaction will be chosen as an objective function and the minimal and maximal possible flux value through the reaction is determined by minimizing and maximizing the objective function.

Knowledge base A specific type of reconstruction that also accounts for the following information: molecular formulae, subsystem assignments, gene-protein-reaction associations, references to primary and review literature, and additional pertinent notes.

Linear programming (LP) A class of optimization problems in which a linear objective function is maximized or minimized under a series of linear equality and inequality constraints.

Network gap A reaction that is missing in the network (e.g., connecting 2 dead-end metabolites, transport reaction).

Network reconstruction An assembly of the components and their interconversions for an organism, based on the genome-annotation and the bibliome.

Objective function A function that is maximized or minimized in optimization problems. In FBA, the objective function is a linear combination of fluxes. For prokaryotes and simple eukaryotes grown in the laboratory under controlled conditions, the biomass function is often used as the objective function.

P/O ratio This ratio represents the number of ATP molecules (P) which are formed per oxygen atom (O) consumed during respiration.

Sink reaction When the synthesis reaction(s) of a metabolite is not known or outside the scope of the reconstruction its discharge can be represented by this unbalanced, intracellular reaction (e.g., 1 A $\leftrightarrow$)

Type III extreme pathway These stoichiometric balanced cycles (SBC) are a subset of ExPa's that are only composed of intracellular reactions, i.e., that all exchange reactions (i.e., systems boundaries) have zero flux.

Bibliography

[1] S. Adinolfi, F. Rizzo, L. Masino, M. Nair, S. R. Martin, A. Pastore, and P. A. Temussi. Bacterial IscU is a well folded and functional single domain protein. *European Journal of Biochemistry*, 271(11):2093–2100, 2004.

[2] J. N. Agar, C. Krebs, J. Frazzon, B. H. Huynh, D. R. Dean, and M. K. Johnson. IscU as a scaffold for iron-sulfur cluster biosynthesis: sequential assembly of [2Fe-2S] and [4Fe-4S] clusters in IscU. *Biochemistry*, 39(27):7856–7862, 2000.

[3] S. Aksoy, C. L. Squires, and C. Squires. Translational coupling of the TrpB and TrpA genes in the *Escherichia coli* tryptophan operon. *Journal of Bacteriology*, 157(2):363–367, 1984.

[4] L. Alberghina and L. Mariani. Analysis of a cell cycle model for *Escherichia coli*. *Journal of Mathematical Biology*, 9(4):389–398, 1980.

[5] T.E. Allen and B. Ø. Palsson. Sequenced-based analysis of metabolic demands for protein synthesis in prokaryotes. *Journal of Theoretical Biology*, 220(1), 2003.

[6] E. Almaas, B. Kovacs, T. Vicsek, Z. N. Oltvai, and A. L. Barabasi. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427(6977), 2004.

[7] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–10106, 2000.

[8] D. Apel and M. G. Surette. Bringing order to a complex molecular machine: The assembly of the bacterial flagella. *Biochimica et Biophysica Acta - Biomembranes*, 1778(9):1851–1858, 2008.

[9] M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H.C. Huang, A. Hirai, K. Tsuzuki, Nakamura S., M. Altaf-Ul Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, Kitagawa M., M. Tomita, S. Kanaya, C. Wada, and H. Mori. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Research*, 16(5):686–691, 2006.

[10] T. Asai, C. Condon, J. Voulgaris, D. Zaporojets, B. Shen, M. Al-Omar, C. Squires, and C. L. Squires. Construction and initial characterization of *Escherichia coli* strains with few or no intact chromosomal rRNA operons. *Journal of Bacteriology*, 181(12):3803–3809, 1999.

[11] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1), 2000.

[12] F. Aslund and J. Beckwith. The thioredoxin superfamily: redundancy, specificity, and gray-area genomics. *Journal of Bacteriology*, 181(5):1375–1379, 1999.

[13] R. K. Aziz, D. Bartels, A. A. Best, M. Dejongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. The RAST server: Rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75–75, 2008.

[14] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2:2006.0008–2006.0008, 2006.

[15] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[16] C.B. Barber, D.P. Dobkin, and H.T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4), 1996.

[17] C. L. Barrett, T. Y. Kim, H. U. Kim, B. Ø. Palsson, and S. Y. Lee. Systems biology as a foundation for genome-scale synthetic biology. *Current Opinion in Biotechnology*, 17(5), 2006.

[18] J. Barthelmes, C. Ebeling, A. Chang, I. Schomburg, and D. Schomburg. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Research*, 35(Database ISSUE):D511–D514, 2007.

[19] S. A. Becker and B. Ø. Palsson. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, 5(1):8–8, 2005.

[20] S. A. Becker, N. D. Price, and B. Ø. Palsson. Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics*, 7(111), 2006.

[21] S.A. Becker, A.M. Feist, M.L. Mo, G. Hannum, B. Ø. Palsson, and M.J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nature Protocols*, 2(3), 2007.

[22] H. Beinert, R. H. Holm, and E. Munck. Iron-sulfur clusters: nature's modular, multipurpose structures. *Science*, 277(5326):653–659, 1997.

[23] F. Ben-Hamida and D. Schlessinger. Synthesis and breakdown of ribonucleic acid in *Escherichia coli* starving for nitrogen. *Biochimica et Biophysica Acta*, 119(1):183–191, 1966.

[24] S. Benthin, J. Nielsen, and J. Villadsen. A simple and reliable method for the determination of cellular RNA content. *Biotechnology Techniques*, 5:39–42, 1991.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[26] B. E. Bernstein, E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides, and S. L. Schreiber. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences*, 99(13), 2002.

[27] D. A. Bessarab, V. R. Kaberdin, C. L. Wei, G. G. Liou, and S. Lin-Chao. RNA components of *Escherichia coli* degradosome: evidence for rRNA decay. *Proceedings of the National Academy of Sciences*, 95(6):3157–3161, 1998.

[28] D. J. Beste, T. Hooper, G. Stewart, B. Bonde, C. Avignone-Rossa, M. E. Bushell, P. Wheeler, S. Klamt, A. M. Kierzek, and J. McFadden. GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biology*, 8(5):R89–R89, 2007.

[29] F. R. Blattner, 3rd Plunkett, G., C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–1474, 1997.

[30] I.G. Boneca, H. Reuse, J.C. Epinat, M. Pupin, A. Labigne, and I. Moszer. A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Research*, 31(6):1704–1714, 2003.

[31] R. Bonneau, M. T. Facciotti, D. J. Reiss, A. K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M. H. Johnson, J. C. Bare, W. Longabaugh, M. Vuthoori, K. Whitehead, A. Madar, L. Suzuki, T. Mori, D. E. Chang, J. Diruggiero, C. H. Johnson, L. Hood, and N. S. Baliga. A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131(7):1354–1365, 2007.

[32] S. Borukhov, A. Polyakov, V. Nikiforov, and A. Goldfarb. GreA protein: a transcription elongation factor from *Escherichia coli. Proceedings of the National Academy of Sciences*, 89(19):8899–8902, 1992.

[33] S. Borukhov, V. Sagitov, and A. Goldfarb. Transcript cleavage factors from *E. coli. Cell*, 72(3):459–466, 1993.

[34] F. Bou-Abdallah, A. C. Lewin, N. E. Le Brun, G. R. Moore, and N. D. Chasteen. Iron detoxification properties of *Escherichia coli* bacterioferritin. attenuation of oxyradical chemistry. *Journal of Biological Chemistry*, 277(40):37064–37069, 2002.

[35] H. Bremer, P. Dennis, and M. Ehrenberg. Free RNA polymerase and modeling global transcription in *Escherichia coli. Biochimie*, 85(6), 2003.

[36] C. Brooksbank, G. Cameron, and J. Thornton. The european bioinformatics institute's data resources: towards systems biology. *Nucleic Acids Research*, 33(Database ISSUE):D46–D53, 2005.

[37] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9):5136–5141, 2003.

[38] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14(2):301–312, 2004.

[39] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–657, 2003.

[40] G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in *Escherichia coli. Nature*, 433(7025):531–537, 2005.

[41] G. O. Bylund, L. C. Wipemo, L. A. Lundberg, and P. M. Wikstrom. RimM and RbfA are essential for efficient processing of 16S rRNA in *Escherichia coli. Journal of Bacteriology*, 180(1), 1998.

[42] A. J. Carpousis. The *Escherichia coli* RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. *Biochem Soc Trans*, 30(2):150–155, 2002.

[43] A. K. Chavali, J. D. Whittemore, J. A. Eddy, K. T. Williams, and J. A. Papin. Systems analysis of metabolism in the pathogenic trypanosomatid *Leishmania major. Molecular Systems Biology*, 4:177–177, 2008.

[44] Z. F. Cheng and M. P. Deutscher. An important role for RNase R in mRNA decay. *Molecular Cell*, 17(2):313–318, 2005.

[45] F. F. V. Chevance and K. T. Hughes. Coordinating assembly of a bacterial macromolecular machine. *Nature Reviews Microbiolog.*

[46] S. R. Chhabra, K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima. Journal of Biological Chemistry*, 278(9):7540–7552, 2003.

[47] G. A. Coburn, X. Miao, D. J. Briant, and G. A. Mackie. Reconstitution of a minimal RNA degradosome demonstrates functional coordination between a 3' exonuclease and a DEAD-box RNA helicase. *Genes & Development*, 13(19):2594–2603, 1999.

[48] S. J. Coles, N. E. Day, P. Murray-Rust, H. S. Rzepa, and Y. Zhang. Enhancement of the chemical semantic web through the use of inchi identifiers. *Org Biomol Chem*, 3(10):1832–1834, 2005.

[49] C. Condon, S. French, C. Squires, and C. L. Squires. Depletion of functional ribosomal RNA operons in *Escherichia coli* causes increased expression of the remaining intact copies. *EMBO Journal*, 12(11):4305–4315, 1993.

[50] C. Condon, D. Liveris, C. Squires, I. Schwartz, and C. L. Squires. rRNA operon multiplicity in *Escherichia coli* and the physiological implications of rrn inactivation. *Journal of Bacteriology*, 177(14):4152–4156, 1995.

[51] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. Ø. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, 2004.

[52] M. W. Covert and B. Ø. Palsson. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *Journal of Biological Chemistry*, 277(31):28058–28064, 2002.

[53] M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044–2050, 2008.

[54] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–163, 1958.

[55] H. S. Cummings and J. W. Hershey. Translation initiation factor IF1 is essential for cell viability in *Escherichia coli*. *Journal of Bacteriology*, 176(1), 1994.

[56] J. R. Cupp-Vickery, H. Urbina, and L. E. Vickery. Crystal structure of IscS, a cysteine desulfurase from *Escherichia coli*. *Journal of Molecular Biology*, 330(5):1049–1059, 2003.

[57] A. Das and C. Yanofsky. Restoration of a translational stop-start overlap reinstates translational coupling in a mutant TrpB'-TrpA gene pair of the *Escherichia coli* tryptophan operon. *Nucleic Acids Research*, 17(22):9333–9340, 1989.

[58] M. S. Dasika, A. Burgard, and C. D. Maranas. A computational framework for the topological analysis and targeted disruption of signal transduction networks. *Biophysical Journal*, 91(1), 2006.

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

[60] E. Deuerling, H. Patzelt, S. Vorderwulbecke, T. Rauch, G. Kramer, E. Schaffitzel, A. Mogk, A. Schulze-Specking, H. Langen, and B. Bukau. Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Molecular Microbiology*, 47(5):1317–1328, 2003.

[61] E. Deuerling, A. Schulze-Specking, T. Tomoyasu, A. Mogk, and B. Bukau. Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature*, 400(6745):693–696, 1999.

[62] M. P. Deutscher. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research*, 34(2):659–666, 2006.

[63] V. Dincbas-Renqvist, A. Engstrom, L. Mora, V. Heurgue-Hamard, R. Buckingham, and M. Ehrenberg. A post-translational modification in the GGQ motif of RF2 from *Escherichia coli* stimulates termination of translation. *EMBO Journal*, 19(24):6900–6907, 2000.

[64] B. Ding, E. S. Smith, and H. Ding. Mobilization of the iron centre in IscA for the iron-sulphur cluster assembly in IscU. *Biochemical Journal*, 389(Pt 3), 2005.

[65] D. A. Drew. A mathematical model for prokaryotic protein synthesis. *Bulletin of Mathematical Biology*, 63(2):329–351, 2001.

[66] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.

[67] N. C. Duarte, M. J. Herrgard, and B. Palsson. Reconstruction and validation of *Saccharomyces cerevisiae i*ND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14(7):1298–1309, 2004.

[68] G. Duby, F. Foury, A. Ramazzotti, J. Herrmann, and T. Lutz. A non-essential function for yeast frataxin in iron-sulfur cluster assembly. *Human Molecular Genetics*, 11(21):2635–2643, 2002.

[69] G. L. Duester and W. M. Holmes. The distal end of the ribosomal RNA operon rrnd of *Escherichia coli* contains a trna1thr gene, two 5S rRNA genes and a transcription terminator. *Nucleic Acids Research*, 8(17):3793–3807, 1980.

[70] M. Durot, P. Y. Bourguignon, and V. Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1):164–90, 2009.

[71] J. S. Edwards, M. Covert, and B. Palsson. Metabolic modeling of microbes: the flux-balance approach. *Environmental Microbiology*, 4(3):133–140, 2002.

[72] J. S. Edwards and B. Ø. Palsson. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Journal of Biological Chemistry*, 274(25):17410–17416, 1999.

[73] J.S. Edwards and B. Ø. Palsson. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences.*, 97(10), 2000.

[74] J.S. Edwards and B. Ø. Palsson. Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*, 1(1), 2000.

[75] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating proteins in the cell using targetp, signalp and related tools. *Nature protocols*, 2(4):953–971, 2007.

[76] D. A. Erie, O. Hajiseyedjavadi, M. C. Young, and P. H. von Hippel. Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science*, 262(5135):867–873, 1993.

[77] T. Escherich. Die Darmbakterien des Neugeborenen und S
uglings. *Fortschritte der Medizin*, 3:515–522, 1885.

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

[79] A. M. Feist, M. J. Herrgard, I. Thiele, J. L. Reed, and B. Ø. Palsson. Reconstruction of biochemical networks in microbial organisms. *Nature Reviews Microbiology*, in press, 2009.

[80] A. M. Feist and B. Ø. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature Biotechnology*, 26(6), 2008.

[81] A. M. Feist, J. C. M. Scholten, B. Ø. Palsson, F. J. Brockman, and T. Ideker. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Molecular Systems Biology*, 2(2006.0004), 2006.

[82] T. J. Fiedler, H. A. Vincent, Y. Zuo, O. Gavrialov, and A. Malhotra. Purification and crystallization of *Escherichia coli* oligoribonuclease. *Acta Crystallographica Section D: Biological Crystallography*, 60(Pt 4):736–739, 2004.

[83] R. M. T. Fleming, I. Thiele, and B. Ø. Palsson. Quantitative assignment of reaction directionality in constraint-based models of metabolism. submitted(2009).

[84] S. S. Fong, A. R. Joyce, and B. Ø. Palsson. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Research*, 15(10):1365–1372, 2005.

[85] S. S. Fong and B. Ø. Palsson. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nature Genetics*, 36(10):1056–1058, 2004.

[86] S.S. Fong, J.Y. Marciniak, and B. Ø. Palsson. Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 using a genome-scale *in silico* metabolic model. *Journal of Bacteriology*, 185(21):6400–6408, 2003.

[87] J. Forster, I. Famili, P.C. Fu, B. Ø. Palsson, and J. Nielsen. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2):244–253, 2003.

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

[89] J. Frydman. Folding of newly translated roteins *in vivo*: The role of molecular chaperones. *Annual Reviews in Biochemistry*, 70(1):603–647, 2001.

[90] T. Gaal, M. S. Bartlett, W. Ross, Jr. Turnbough, C. L., and R. L. Gourse. Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science*, 278(5346):2092–2097, 1997.

[91] J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. Brinkman. Psortb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Trends in Parasitology*, 21:617–623, 2005.

[92] S. Ghosh and M. P. Deutscher. Oligoribonuclease is an essential component of the mRNA decay pathway. *Proceedings of the National Academy of Sciences*, 96(8):4372–4377, 1999.

[93] E. Gianchandani, A. R. Joyce, B. Ø. Palsson, and J. A. Papin. Functional states of the *Escherichia coli* transcriptional regulatory system at the genome-scale. *submitted*, 2009.

[94] E. P. Gianchandani, M. A. Oberhardt, A. P. Burgard, C. D. Maranas, and J. A. Papin. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics*, 9:43, 2008.

[95] E. P. Gianchandani, J. A. Papin, N. D. Price, A. R. Joyce, and B. Ø. Palsson. Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Computational Biology*, 2(8):e101–e101, 2006.

[96] H. Ginsburg. Caveat emptor: limitations of the automated reconstruction of metabolic pathways in *Plasmodium*. *Trends in Parasitology*, 25:37–43, 2009.

[97] J. Gralnick and D. Downs. Protection from superoxide damage associated with an increased level of the YggX protein in *Salmonella enterica*. *Proceedings of the National Academy of Sciences*, 98(14):8030–8035, 2001.

[98] J. A. Gralnick and D. M. Downs. The YggX protein of *Salmonella enterica* is involved in Fe(II) trafficking and minimizes the DNA damage caused by hydroxyl radicals: residue Cys-7 is essential for YggX function. *Journal of Biological Chemistry*, 278(23):20708–20715, 2003.

[99] K. B. Gromadski, H. J. Wieden, and M. V. Rodnina. Kinetic mechanism of elongation factor Ts-catalyzed nucleotide exchange in elongation factor Tu. *Biochemistry*, 41(1):162–169, 2002.

[100] D. Gutnick, J. M. Calvo, T. Klopotowski, and B. N. Ames. Compounds which serve as the sole source of carbon or nitrogen for *Salmonella typhimurium* LT-2. *Journal of Bacteriology*, 100(1):215–219, 1969.

[101] E. Hajnsdorf and P. Regnier. *E. coli* RpsO mRNA decay: RNase E processing at the beginning of the coding sequence stimulates poly(A)-dependent degradation of the mRNA. *Journal of Molecular Biology*, 286(4):1033–1043, 1999.

[102] C. J. Harrison, M. Hayer-Hartl, M. Di Liberto, F. Hartl, and J. Kuriyan. Crystal structure of the nucleotide exchange factor GrpE bound to the ATPase domain of the molecular chaperone DnaK. *Science*, 276(5311):431–435, 1997.

[103] R. Harrison, B. Papp, C. Pál, S.G. Oliver, and D. Delneri. Plasticity of genetic interactions in metabolic networks of yeast. *Proceedings of the National Academy of Sciences*, 104(7):2307–2312, 2007.

[104] F. U. Hartl and M. Hayer-Hartl. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295(5561):1852–1858, 2002.

[105] V. Hatzimanikatis and K. H. Lee. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabolic Engineering*, 1(4), 1999.

[106] M. Heinemann, A. Kummel, R. Ruinatscha, and S. Panke. *in silico* genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. *Biotechnology and Bioengineering*, 92(7):850–864, 2005.

[107] D. Herbert, P. J. Phipps, and R. E. Strange. Chemical analysis of microbial cells. *Methods in Microbiology*, 5:209–344, 1971.

[108] M. J. Herrgard, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Bluthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novere, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasic, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttila, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26(10):1155–1160, 2008.

[109] C. D. Herring, A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. Ø. Palsson. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nature Genetics*, 38(12), 2006.

[110] T. Hesterkamp, S. Hauser, H. Lutcke, and B. Bukau. *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proceedings of the National Academy of Sciences*, 93(9):4437–4441, 1996.

[111] K. G. Hoff, J. R. Cupp-Vickery, and L. E. Vickery. Contributions of the lppvk motif of the iron-sulfur template protein IscU to interactions with the Hsc66-Hsc20 chaperone system. *Journal of Biological Chemistry*, 278(39):37582–37589, 2003.

[112] A. Hoffmann, F. Merz, A. Rutkowska, B. Zachmann-Brand, E. Deuerling, and B. Bukau. Trigger factor forms a protective shield for nascent polypeptides at the ribosome. *Journal of Biological Chemistry*, 281(10):6539–6545, 2006.

[113] C. Holden. Alliance launched to model *E. coli*. *Science*, 297(5586), 2002.

[114] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences*, 98(4):1693–168., 2001.

[115] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(15):8409–8414, 2000.

[116] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, 2003.

[117] S. I. Husnain, W. Meng, S. J. Busby, and M. S. Thomas. *Escherichia coli* can tolerate insertions of up to 16 amino acids in the RNA polymerase alpha subunit inter-domain linker. *Biochimica et Biophysica Acta*, 1678(1), 2004.

[118] R. U. Ibarra, J. S. Edwards, and B. Ø. Palsson. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*, 420(6912):186–189, 2002.

[119] Y. Ikeuchi, A. Soma, T. Ote, J. Kato, Y. Sekine, and T. Suzuki. Molecular mechanism of lysidine synthesis that determines tRNA identity and codon recognition. *Molecular Cell*, 19(2):235–246, 2005.

[120] J. Izard and R. J. Limberger. Rapid screening method for quantitation of bacterial cell lipids from whole cells. *Journal of Microbiological Methods*, 55:411–418, 2003.

[121] C. Jain. Degradation of mRNA in *Escherichia coli*. *IUBMB Life*, 54(6):315–321, 2002.

[122] N. Jamshidi and B. Ø. Palsson. Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology*, 1:26–26, 2007.

[123] N. Jamshidi and B. Ø. Palsson. Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology*, 4:171–171, 2008.

[124] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophysical Journal*, 95(3):1487–1499, 2008.

[125] V. Jarlier and H. Nikaido. Mycobacterial cell wall: structure and role in natural resistance to antibiotics. *FEMS microbiology letters*, 123(1-2):11–18, 1994.

[126] M. Jishage, A. Iwata, S. Ueda, and A. Ishihama. Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions. *Journal of Bacteriology*, 178(18):5447–5451, 1996.

[127] D. C. Johnson, D. R. Dean, A. D. Smith, and M. K. Johnson. Structure, function, and formation of biological iron-sulfur clusters. *Annual Review of Biochemistry*, 74:247–281, 2005.

[128] R. Joubert, P. Brignon, C. Lehmann, C. Monribot, F. Gendre, and H. Boucherie. Two-dimensional gel analysis of the proteome of lager brewing yeasts. *Yeast*, 16(6):511–522, 2000.

[129] A. R. Joyce, J. L. Reed, A. White, R Edwards, A. Osterman, T. Baba, H. Mori, S. A. Lesely, B. Ø. Palsson, and S. Agarwalla. Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *Journal of Bacteriology*, 188(23), 2006.

[130] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database ISSUE):D354–D357, 2006.

[131] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32 Database ISSUE:D277–D280, 2004.

[132] C. W. Kang and C. R. Cantor. Structure of ribosome-bound messenger RNA as revealed by enzymatic accessibility studies. *Journal of Molecular Biology*, 181(2):241–251, 1985.

[133] P. D. Karp, M. Arnaud, J. Collado-Vides, J. Ingraham, I. T. Paulsen, and M.H. Saier. The *E. coli* ecocyc database: No longer just a metabolic pathway database. *ASM News*, 70(1), 2004.

[134] P. D. Karp, I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham. Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Research*, 2007.

[135] P. D. Karp, S. Paley, and P. Romero. The pathway tools software. *Bioinformatics*, 18 Suppl 1:S225–S232, 2002.

[136] A. Katayama, A. Tsujii, A. Wada, T. Nishino, and A. Ishihama. Systematic search for zinc-binding proteins in *Escherichia coli*. *European Journal of Biochemistry*, 269(9), 2002.

[137] S. Kato, H. Mihara, T. Kurihara, Y. Takahashi, U. Tokumoto, T. Yoshimura, and N. Esaki. Cys-328 of IscS and Cys-63 of IscU are the sites of disulfide bridge formation in a covalently bound IscS/IscU complex: implications for the mechanism of iron-sulfur cluster assembly. *Proceedings of the National Academy of Sciences*, 99(9), 2002.

[138] T. Kawashima, C. Berthet-Colominas, M. Wulff, S. Cusack, and R. Leberman. The structure of the *Escherichia coli* EF-Tu.EF-Ts complex at 2.5 A resolution. *Nature*, 379(6565), 1996.

[139] K. O. Kelly and M. P. Deutscher. The presence of only one of five exoribonucleases is sufficient to support the growth of *Escherichia coli*. *Journal of Bacteriology*, 174(20):6682–6684, 1992.

[140] M. J. Kerner, D. J. Naylor, Y. Ishihama, T. Maier, H. C. Chang, A. P. Stines, C. Georgopoulos, D. Frishman, M. Hayer-Hartl, M. Mann, and F. U. Hartl. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, 122(2):209–220, 2005.

[141] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp. Ecocyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research*, 33(Database ISSUE):D334–D337, 2005.

[142] P. Kharchenko, L. Chen, Y. Freund, D. Vitkup, and G. M. Church. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, 7(177), 2006.

[143] V. Khemici, I. Toesca, L. Poljak, N. F. Vanzo, and A. J. Carpousis. The RNase E of *Escherichia coli* has at least two binding sites for DEAD-box RNA helicases: functional replacement of RhlB by RhlE. *Molecular Microbiology*, 54(5):1422–1430, 2004.

[144] M. C. Kiel, V. S. Raj, H. Kaji, and A. Kaji. Release of ribosome-bound ribosome recycling factor by elongation factor G. *Journal of Biological Chemistry*, 278(48):48041–48050, 2003.

[145] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1:2–2, 2007.

[146] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2), 2003.

[147] G. Kramer, T. Rauch, W. Rist, S. Vorderwulbecke, H. Patzelt, A. Schulze-Specking, N. Ban, E. Deuerling, and B. Bukau. L23 protein functions as a chaperone docking site on the ribosome. *Nature*, 419(6903):171–174, 2002.

[148] L. Kuepfer, U. Sauer, and L. M. Blank. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Research*, 15(10):1421–1430, 2005.

[149] S. V. Kumar, M. S. Dasika, and C. D. Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8:212–212, 2007.

[150] V. S. Kumar and C. D. Maranas. Growmatch: An automated method for reconciling *in silico/in vivo* growth predictions. *PLoS Computational Biology*, 5(3), 2009.

[151] T. Kurihara, H. Mihara, S. Kato, T. Yoshimura, and N. Esaki. Assembly of iron-sulfur clusters mediated by cysteine desulfurases, IscS, CsdB and CSD, from *Escherichia coli*. *Biochimica et Biophysica Acta*, 1647(1-2):303–309, 2003.

[152] L. Laffend and M. L. Shuler. Ribosomal protein limitations in *Escherichia coli* under conditions of high translational activity. *Biotechnology and Bioengineering*, 43(5), 1994.

[153] G. Layer, S. Ollagnier-de Choudens, Y. Sanakis, and M. Fontecave. Iron-sulfur cluster biosynthesis: characterization of *Escherichia coli* CyaY as an iron donor for the assembly of [2Fe-2S] clusters in the scaffold IscU. *Journal of Biological Chemistry*, 281(24):16256–16263, 2006.

[154] D. Y. Lee, H. Yun, S. Park, and S. Y. Lee. MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis. *Bioinformatics*, 19(16):2144–2146, 2003.

[155] J. M. Lee, E. P. Gianchandani, J. A. Eddy, and J. A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology*, 4(5):e1000086–e1000086, 2008.

[156] K. Lee, F. Berthiaume, G. N. Stephanopoulos, D. M. Yarmush, and M. L. Yarmush. Metabolic flux analysis of postburn hepatic hypermetabolism. *Metabolic Engineering*, 2(4):312–327, 2000.

[157] M. J. Lee, E. P. Gianchandani, J. A. Eddy, and J. A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology*, 4(5):e1000086–e1000086, 2008.

[158] T. T. Lee, S. Agarwalla, and R. M. Stroud. A unique RNA fold in the RumA-RNA-cofactor ternary complex contributes to substrate selectivity and enzymatic function. *Cell*, 120(5), 2005.

[159] B. Li, H. Wing, D. Lee, H. C. Wu, and S. Busby. Transcription activation by *Escherichia coli* FNR protein: similarities to, and differences from, the crp paradigm. *Nucleic Acids Research*, 26(9):2075–2081, 1998.

[160] D. S. Li, K. Ohshima, S. Jiralerspong, M. W. Bojanowski, and M. Pandolfo. Knockout of the cyaY gene in *Escherichia coli* does not affect cellular iron content and sensitivity to oxidants. *FEBS Letters*, 456(1):13–16, 1999.

[161] F. Li, I. Thiele, N. Jamshidi, and B. Ø. Palsson. Functional assessment of the TLR receptor network. *PLoS Computational Biology*, 5(2), 2009.

[162] Y. Li and S. Altman. A specific endoribonuclease, RNase P, affects gene expression of polycistronic operon mRNAs. *Proceedings of the National Academy of Sciences*, 100(23):13213–13218, 2003.

[163] Z. Li and M. P. Deutscher. The role of individual exoribonucleases in processing at the 3' end of *Escherichia coli* tRNA precursors. *Journal of Biological Chemistry*, 269(8):6064–6071, 1994.

[164] R. Lill, E. Crooke, B. Guthrie, and W. Wickner. The trigger factor cycle includes ribosomes, presecretory proteins, and the plasma membrane. *Cell*, 54(7):1013–1018, 1988.

[165] P. A. Limbach, P. F. Crain, and J. A. McCloskey. Summary: the modified nucleosides of RNA. *Nucleic Acids Research*, 22(12):2183–2196, 1994.

[166] K. Linke, T. Wolfram, J. Bussemer, and U. Jakob. The roles of the two zinc binding sites in DnaJ. *Journal of Biological Chemistry*, 278(45):44457–44466, 2003.

[167] J. Lu, A. Lal, B. Merriman, S. Nelson, and G. Riggins. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, 84(4):631–636, 2004.

[168] H. Maeda, N. Fujita, and A. Ishihama. Competition among seven *Escherichia coli* sigma subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Research*, 28(18), 2000.

[169] T. F. Mah, K. Kuznedelov, A. Mushegian, K. Severinov, and J. Greenblatt. The alpha subunit of *E. coli* RNA polymerase activates RNA binding by NusA. *Genes & Development*, 14(20):2664–2675, 2000.

[170] R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, 2003.

[171] R. A. Majewski and M. M. Domach. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering*, 35, 1990.

[172] A. Manichaikul, L. Ghamsari, E. F. Y. Hom, C. Lin, R.R Murray, R. L. Chang, T. Hao, Y. Shen, A. K. Chavali, I. Thiele, X. Yang, E. Mello, D. E. Hill, M. Vidal, K. Salehi-Ashtiani, and J. A. Papin. Metabolic network analysis integrated with genome-wide transcript verification. submitted(2009).

[173] S. S. Mansy and J. A. Cowan. Iron-sulfur cluster biosynthesis: toward an understanding of cellular machinery and molecular mechanism. *Acc Chem Res*, 37(9), 2004.

[174] J. Martin, T. Langer, R. Boteva, A. Schramel, A. L. Horwich, and F. U. Hartl. Chaperonin-mediated protein folding at the surface of groEL through a 'molten globule'-like intermediate. *Nature*, 352(6330), 1991.

[175] J. A. McCloskey and J. Rozenski. The small subunit rRNA modification database. *Nucleic Acids Research*, 33(Database ISSUE), 2005.

[176] A. Mehra and V. Hatzimanikatis. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophysical Journal*, 90(4):1136–1146, 2006.

[177] A. Mehra, K. H. Lee, and V. Hatzimanikatis. Insights into the relation between mRNA and protein expression patterns: I. theoretical considerations. *Biotechnology and Bioengineering*, 84(7):822–833, 2003.

[178] T. Mizuno. Compilation of all genes encoding two-component phosphotransfer signal transducers in the genome of *Escherichia coli*. *DNA Research*, 4(2):161–168, 1997.

[179] M. L. Mo, B. Ø. Palsson, and M. J. Herrgård. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Systems Biology*, 3(37), 2009.

[180] Todd K. Moon and Wynn C. Stirling. *Mathematical methods and algorithms for signal processing.* Prentice Hall, Upper Saddle River, NJ, 2000.

[181] Nomenclature Committee of the International Union of Biochemistry NC-IUBMB and Molecular Biology. *Enzyme Nomenclature.* Academic Press, San Diego, California, 6th edition, 1992.

[182] F. C. Neidhardt. The regulation RNA synthesis in bacteria. *Prog Nucleic Acid Res Mol Biol*, 3:145–181, 1964.

[183] F. C. Neidhardt, editor. *Chemical Composition of* Escherichia coli, volume 2. ASM Press, Washington, D.C., 2nd edition, 1996.

[184] F. C. Neidhardt, editor. *Chemical Composition of* Escherichia coli, volume 1. ASM Press, Washington, D.C., 2nd edition, 1996.

[185] F. C. Neidhardt, J. L. Ingraham, and M. Schaechter. *Physiology of the bacterial cell: a molecular approach.* Sinauer Associates, Sunderland, Mass., 1990.

[186] B. E. Nickels, S. J. Garrity, V. Mekler, L. Minakhin, K. Severinov, R. H. Ebright, and A. Hochschild. The interaction between sigma70 and the beta-flap of *Escherichia coli* RNA polymerase inhibits extension of nascent RNA during early elongation. *Proceedings of the National Academy of Sciences*, 102(12):4488–4493, 2005.

[187] J. Nogales, B. Ø. Palsson, and I. Thiele. A genome-scale metabolic reconstruction for *Pseudomonas putida* KT2440: *i*JN746 as cell factory. *BMC Systems Biology*, 2(1):79, 2008.

[188] M. Nomura. Regulation of ribosome biosynthesis in *Escherichia coli* and *Saccharomyces cerevisiae*: diversity and common principles. *Journal of Bacteriology*, 181(22):6857–6864, 1999.

[189] M. Nomura, R. Gourse, and G. Baughman. Regulation of the synthesis of ribosomes and ribosomal components. *Annual Review of Biochemistry*, 53, 1984.

[190] M. Nomura and E. A. Morgan. Genetics of bacterial ribosomes. *Annual Review of Genetics*, 11, 1977.

[191] I. Nookaew, M.C. Jewett, A. Meechai, C. Thammarongtham, K. Laoteng, S. Cheevadhanarak, J. Nielsen, S. Bhumiratana, C.R. Yang, O. Dror, et al. The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC Systems Biology*, 2(1):71, 2008.

[192] S. Normark, S. Bergstrom, T. Edlund, T. Grundstrom, B. Jaurin, F. P. Lindberg, and O. Olsson. Overlapping genes. *Annual Review of Genetics*, 17, 1983.

[193] R. A. Notebaart, F. H. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, 7(1):296–296, 2006.

[194] M. Nuth, T. Yoon, and J. A. Cowan. Iron-sulfur cluster biosynthesis: characterization of iron nucleation sites for assembly of the [2Fe-2S]2+ cluster core in IscU proteins. *Journal of the American Chemical Society*, 124(30):8774–8775, 2002.

[195] M. A. Oberhardt, J. Puchalka, K. E. Fryer, V. A. Martins dos Santos, and J. A. Papin. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *Journal of Bacteriology*, 190(8):2790–2803, 2008.

[196] Y. K. Oh, B. Ø. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *Journal of Biological Chemistry*, 2007.

[197] N. Opalka, M. Chlenov, P. Chacon, W. J. Rice, W. Wriggers, and S. A. Darst. Structure and function of the transcription elongation factor GreB bound to bacterial RNA polymerase. *Cell*, 114(3):335–345, 2003.

[198] D. S. Oppenheim and C. Yanofsky. Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics*, 95(4):785–795, 1980.

[199] M. J. Osborne, N. Siddiqui, D. Landgraf, P. J. Pomposiello, and K. Gehring. The solution structure of the oxidative stress-related protein YggX from *Escherichia coli*. *Protein Science*, 14(6):1673–1678, 2005.

[200] K. A. Ost and M. P. Deutscher. RNase PH catalyzes a synthetic reaction, the addition of nucleotides to the 3' end of RNA. *Biochimie*, 72(11):813–818, 1990.

[201] Y. Otsuka and T. Yonesaki. A novel endoribonuclease, RNase LS, in *Escherichia coli*. *Genetics*, 169(1), 2005.

[202] R. Overbeek, D. Bartels, V. Vonstein, and F. Meyer. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chemical Reviews*, 107:3431–3447, 2007.

[203] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17):5691–5702, 2005.

[204] B. Ø. Palsson. *in silico* biotechnology: Era of reconstruction and interrogation. *Current Opinion in Biotechnology*, 15(1):50–51, 2004.

[205] B. Ø. Palsson. Two-dimensional annotation of genomes. *Nature Biotechnology*, 22(10):1218–1219, 2004.

[206] B. Ø. Palsson. *Systems biology: properties of reconstructed networks*. Cambridge University Press, New York, 2006.

[207] J. A. Papin, T. Hunter, B. Ø. Palsson, and S. Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2), 2005.

[208] J. A. Papin and B. Ø. Palsson. Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *Journal of Theoretical Biology*, 227(2), 2004.

[209] J. A. Papin, N. D. Price, and B. Ø. Palsson. *in silico* cells: studying genotype-phenotype relationships with constraints-based models. In B. Kholodenko and H. V. Westerhoff, editors, *Metabolic Engineering in the Post-Genomic Era*. Horizon Bioscience, 2004.

[210] J. H. Park, K. H. Lee, T. Y. Kim, and S. Y. Lee. Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proceedings of the National Academy of Sciences*, 104(19):7797–7802, 2007.

[211] J. S. Park, M. T. Marr, and J. W. Roberts. *E. coli* transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell*, 109(6):757–767, 2002.

[212] A. M. Patel and S. D. Dunn. RNase E-dependent cleavages in the 5' and 3' regions of the *Escherichia coli* unc mRNA. *Journal of Bacteriology*, 174(11):3541–3548, 1992.

[213] S. W. Peretti and J.E. Bailey. Mechanistically detailed model of cellular metabolism for glucose-limited growth of *Escherichia coli* B/r-A. *Biotechnology and Bioengineering*, 28(11), 1986.

[214] E. Perez-Rueda and J. Collado-Vides. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Research*, 28(8):1838–1847, 2000.

[215] J. W. Pinney, M. W. Shirley, G. A. McConkey, and D. R. Westhead. metashark: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Research*, 33(4):1399–1409, 2005.

[216] J. Pramanik and J. D. Keasling. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, 56(4), 1997.

[217] N. D. Price, I. Famili, D. A. Beard, and B. Ø. Palsson. Extreme pathways and Kirchhoff's second law. *Biophysical Journal*, 83(5):2879–2882, 2002.

[218] N. D. Price, J. A. Papin, C. H. Schilling, and B. Palsson. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends in Biotechnology*, 21(4), 2003.

[219] N. D. Price, J. L. Reed, and B. Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11), 2004.

[220] N. D. Price, I. Thiele, and B. Ø. Palsson. Candidate states of *Helicobacter pylori*'s genome-scale metabolic network upon application of loop law thermodynamic constraints. *Biophysical Journal*, 90:3919–3928, 2006.

[221] J. Puchalka, M. A. Oberhardt, M. Godinho, A. Bielecka, D. Regenhardt, K. N. Timmis, J. A. Papin, and V. A. Martins dos Santos. Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. *PLoS Computational Biology*, 4(10):e1000210–e1000210, 2008.

[222] H. Putzer and S. Laalami. Regulation of the expression of aminoacyl-tRNA synthetases and translation factors. In L. Lapointe, J. & Brakier-Gingras, editor, *Translation Mechanisms*. Landes Bioscience, Georgetown, Texas, USA., 2003.

[223] L. M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, H. V. Westerhoff, K. van Dam, and S. G. Oliver. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19(1), 2001.

[224] A. Raghunathan, J.L. Reed, S. Shin, B. Ø. Palsson, and S. Daefler. Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. submitted (2008).

[225] R. Ramakrishna, J. S. Edwards, A. McCulloch, and B. Ø. Palsson. Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 280(3):R695–R704, 2001.

[226] N. A. Ranson, D. K. Clare, G. W. Farr, D. Houldershaw, A. L. Horwich, and H. R. Saibil. Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes. *Nature Structural & Molecular Biology*, 13(2), 2006.

[227] R.L. Rardin. *Optimization in operations research.* Prentice Hall, 1998.

[228] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–155., 2002.

[229] J. L. Reed, I. Famili, I. Thiele, and B. Ø. Palsson. Towards multidimensional genome annotation. *Nature Reviews Genetics*, 7(2):130–141, 2006.

[230] J. L. Reed and B. Ø. Palsson. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states. *Genome Research*, 14(9):1797–1805, 2004.

[231] J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong, and B. Ø. Palsson. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences*, 103(46):17480–17484, 2006.

[232] J.L. Reed, T.D. Vo, C. H. Schilling, and B. Ø. Palsson. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biologyogy*, 4(9), 2003.

[233] P. Regnier and M. Grunberg-Manago. Cleavage by RNase III in the transcripts of the metY-nus-A-infB operon of *Escherichia coli* releases the tRNA and initiates the decay of the downstream mRNA. *Journal of Molecular Biology*, 210(2), 1989.

[234] Q. Ren, K. Chen, and I. T. Paulsen. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research*, 35(Database ISSUE):D274–D279, 2007.

[235] N. B. Reuven and M. P. Deutscher. Multiple exoribonucleases are required for the 3' processing of *Escherichia coli* tRNA precursors *in vivo*. *FASEB Journal*, 7(1):143–148, 1993.

[236] J. P. Richardson and J. Greenblatt. Control of RNA chain elongation and termination. In F.C. Neidhardt, editor, Escherichia coli *and* Salmonella typhimurium, pages 822848.–822848. American Society for Microbiology, Washington, DC, 2nd edition, 1996.

[237] M. Riley, T. Abe, M. B. Arnaud, M. K. Berlyn, F. R. Blattner, R. R. Chaudhuri, J. D. Glasner, T. Horiuchi, I. M. Keseler, T. Kosuge, H. Mori, N. T. Perna, 3rd Plunkett, G., K. E. Rudd, M. H. Serres, G. H. Thomas, N. R. Thomson, D. Wishart, and B. L. Wanner. *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. *Nucleic Acids Research*, 34(1), 2006.

[238] B. B. Roa, D. M. Connolly, and M. E. Winkler. Overlap between PdxA and KsgA in the complex PdxA-KsgA-ApaG-ApaH operon of *Escherichia coli* K-12. *Journal of Bacteriology*, 171(9):4767–4777, 1989.

[239] J. Roberts and J. S. Park. Mfd, the bacterial transcription repair coupling factor: translocation, repair and termination. *Current Opinion in Microbiology*, 7(2):120–125, 2004.

[240] P. Ross-Macdonald, P. S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heidtman, F. K. Nelson, H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402(6760):413–48., 1999.

[241] J. Rozenski, P. F. Crain, and J. A. McCloskey. The RNA modification database: 1999 update. *Nucleic Acids Research*, 27(1):196–197, 1999.

[242] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. Regulondb (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Research*, 34(Database ISSUE):D394–D397, 2006.

[243] M. Santillan and M. C. Mackey. Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data. *Proceedings of the National Academy of Sciences*, 98(4):1364–1369, 2001.

[244] N. Sarkar. Polyadenylation of mRNA in prokaryotes. *Annual Review of Biochemistry*, 66:173–197, 1997.

[245] J. M. Savinell and B. Ø. Palsson. Network analysis of intermediary metabolism using linear optimization. II. interpretation of hybridoma cell metabolism. *Journal of Theoretical Biology*, 154(4):455–473, 1992.

[246] M. Schaechter, O. Maaløe, and N. 0. Kjeldgaard. Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonella typhimurium*. *Journal of General Microbiology*, 19:592–606, 1958.

[247] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø Palsson. BiGG: A biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinfomratics*, submitted(2009).

[248] C.H. Schilling, M.W. Covert, I. Famili, G.M. Church, J.S. Edwards, and B. Ø. Palsson. Genome-scale metabolic model of *Helicobacter pylori* 26695. *Journal of Bacteriology*, 184(16), 2002.

[249] C.H. Schilling, J.S. Edwards, and B. Ø. Palsson. Towards metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnology Progress*, 15(3):288–295, 1999.

[250] H. J. Schonfeld, D. Schmidt, H. Schroder, and B. Bukau. The DnaK chaperone system of *Escherichia coli*: quaternary structures and interactions of the DnaK and GrpE components. *Journal of Biological Chemistry*, 270(5):2183–2189, 1995.

[251] R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology*, 3(119), 2007.

[252] D. Segre, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.

[253] C. P. Selby and A. Sancar. Molecular mechanism of transcription-repair coupling. *Science*, 260(5104):53–58, 1993.

[254] R. S. Senger and E. T. Papoutsakis. Genome-scale model for *Clostridium acetobutylicum*: Part I. metabolic network resolution and analysis. *Biotechnology and Bioengineering*, 101(5), 2008.

[255] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, 2002.

[256] J. J. Silberg, T. L. Tapley, K. G. Hoff, and L. E. Vickery. Regulation of the HscA ATPase reaction cycle by the co-chaperone HscB and the iron-sulfur cluster assembly protein IscU. *Journal of Biological Chemistry*, 279(52):53924–53931, 2004.

[257] S. Sinha. Theoretical study of tryptophan operon application in microbial technology. *Biotechnology and Bioengineering*, 31(2), 1988.

[258] E. Skordalakes and J. M. Berger. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell*, 114(1):135–146, 2003.

[259] A. D. Smith, J. N. Agar, K. A. Johnson, J. Frazzon, I. J. Amster, D. R. Dean, and M. K. Johnson. Sulfur transfer from IscS to IscU: the first step in iron-sulfur cluster biosynthesis. *Journal of the American Chemical Society*, 123(44):11103–11104, 2001.

[260] M. E. Smulson and R. J. Suhadolnik. The biosynthesis of the 7-deazaadenine ribonucleoside, tubercidin, by *Streptomyces tubercidicus*. *Journal of Biological Chemistry*, 242(12):2872–2876, 1967.

[261] A. Soma, Y. Ikeuchi, S. Kanemasa, K. Kobayashi, N. Ogasawara, T. Ote, J. Kato, K. Watanabe, Y. Sekine, and T. Suzuki. An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Molecular Cell*, 12(3):689–698, 2003.

[262] M. Sprinzl and K. S. Vassilenko. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 33(Database ISSUE):D139–D140, 2005.

[263] L. Stein. Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2(7), 2001.

[264] B. S. Stevenson and T. M. Schmidt. Life history implications of rRNA gene copy number in *Escherichia coli*. *Applied and Environmental Microbiology*, 70(11):6670–6677, 2004.

[265] B. L. Stitt. *Escherichia coli* transcription termination protein Rho has three hydrolytic sites for ATP. *Journal of Biological Chemistry*, 263(23):11130–11137, 1988.

[266] B. L. Stitt. *Escherichia coli* transcription termination factor Rho binds and hydrolyzes ATP using a single class of three sites. *Biochemistry*, 40(7):2276–2281, 2001.

[267] P. R. Subbarayan and M. P. Deutscher. *Escherichia coli* RNase M is a multiply altered form of RNase i. *RNA*, 7(12):1702–1707, 2001.

[268] R. J. Suhadolnik and T. Uematsu. Biosynthesis of the pyrrolopyrimidine nucleoside antibiotic, toyocamycin. vII. origin of the pyrrole carbons and the cyano carbon. *Journal of Biological Chemistry*, 245(17):4365–4371, 1970.

[269] S. Sundararaj, A. Guo, B. Habibi-Nazhad, M. Rouani, P. Stothard, M. Ellison, and D. S. Wishart. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*. *Nucleic Acids Research*, 32(Database ISSUE):D293–D295, 2004.

[270] P. F. Suthers, A. P. Burgard, M. S. Dasika, F. Nowroozi, S. Van Dien, J. D. Keasling, and C. D. Maranas. Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metabolic Engineering*, 9(5-6), 2007.

[271] M. F. Symmons, M. G. Williams, B. F. Luisi, G. H. Jones, and A. J. Carpousis. Running rings around RNA: a superfamily of phosphate-dependent RNases. *Trends in Biochemical Sciences*, 27(1):11–18, 2002.

[272] A. D. Tadmor and T. Tlusty. A coarse-grained biophysical model of *E. coli* and its application to perturbation of the rRNA operon copy number. *PLoS Computational Biology*, 4(4):e1000038–e1000038, 2008.

[273] Y. Takahashi and M. Nakamura. Functional assignment of the ORF2-IscS-IscU-IscA-hscB-hscA-fdx-ORF3 gene cluster involved in the assembly of Fe-S clusters in *Escherichia coli*. *Journal of Biochemistry (Tokyo)*, 126(5):917–926, 1999.

[274] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41–41, 2003.

[275] S. A. Teter, W. A. Houry, D. Ang, T. Tradler, D. Rockabrand, G. Fischer, P. Blum, C. Georgopoulos, and F. U. Hartl. Polypeptide flux through bacterial hsp70: DnaK cooperates with trigger factor in chaperoning nascent chains. *Cell*, 97(6):755–765, 1999.

[276] I. Thiele, R. M. T. Fleming, A. Bordbar, and B. Ø. Palsson. Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional and translational machinery. *Biophysical Journal*, submitted(2009).

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

[278] I. Thiele and B. Ø. Palsson. Bringing genomes to life: The use of genome-scale *in silico* models. In S. Choi, editor, *Introduction to Systems Biology*, volume 1. Humana Press, 2007.

[279] I. Thiele and B. Ø. Palsson. 2D genome annotation jamborees: A community effort in systems biology. *Nature Biotechnology*, submitted(2009).

[280] I. Thiele, N. D. Price, T. D. Vo, and B. Ø. Palsson. Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet. *Journal of Biological Chemistry*, 280(12):11683–11695, 2005.

[281] I. Thiele, T. D. Vo, N. D. Price, and B. Palsson. An expanded metabolic reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants. *Journal of Bacteriology*, 187(16):5818–5830, 2005.

[282] M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. A. III Hutchison. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1), 1999.

[283] S. Tong, A. Porco, T. Isturiz, and T. Conway. Cloning and molecular genetic characterization of the *Escherichia coli* gntR, gntK, and gntU genes of GntI, the main system for gluconate metabolism. *Journal of Bacteriology*, 178(11):3260–3269, 1996.

[284] M. Torres, C. Condon, J. M. Balada, C. Squires, and C. L. Squires. Ribosomal protein S4 is a transcription factor with properties remarkably similar to NusA, a protein involved in both non-ribosomal and ribosomal RNA antitermination. *EMBO Journal*, 20(14):3811–3820, 2001.

[285] H. D. Urbina, J. J. Silberg, K. G. Hoff, and L. E. Vickery. Transfer of sulfur from IscS to IscU during Fe/S cluster assembly. *Journal of Biological Chemistry*, 276(48):44521–44526, 2001.

[286] N. F. Vanzo, Y. S. Li, B. Py, E. Blum, C. F. Higgins, L. C. Raynal, H. M. Krisch, and A. J. Carpousis. Ribonuclease E organizes the protein interactions in the *Escherichia coli* RNA degradosome. *Genes & Development*, 12(17):2770–2781, 1998.

[287] A. Varma, B. W. Boesch, and B. Ø. Palsson. Biochemical production capabilities of *Escherichia coli*. *Biotechnology and Bioengineering*, 42(1), 1993.

[288] A. Varma, B. W. Boesch, and B. Ø. Palsson. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology*, 59(8):2465–2473, 1993.

[289] A. Varma and B. Ø. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Nature Biotechnology*, 12, 1994.

[290] A. Varma and B. Ø. Palsson. Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. *Biotechnology and Bioengineering*, 45(1), 1995.

[291] E. Vivas, E. Skovran, and D. M. Downs. *Salmonella enterica* strains lacking the frataxin homolog CyaY show defects in Fe-S cluster metabolism *in vivo*. *Journal of Bacteriology*, 188(3):1175–1179, 2006.

[292] T. D. Vo, H. J. Greenberg, and B. Ø. Palsson. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *Journal of Biological Chemistry*, 279(38):39532–39540, 2004.

[293] J. Voulgaris, S. French, R. L. Gourse, C. Squires, and C. L. Squires. Increased rrn gene dosage causes intermittent transcription of rRNA in *Escherichia coli*. *Journal of Bacteriology*, 181(14):4170–4175, 1999.

[294] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 1988.

[295] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R.

Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 36(Database ISSUE):D13–D21, 2008.

[296] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 35(Database ISSUE), 2007.

[297] H. Willenbrock, C. Friis, A. Juncker, and D. Ussery. An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biology*, 7(12):R114, 2006.

[298] A. J. Williams. Internet-based tools for communication and collaboration in chemistry. *Drug Discovery Today*, 13(11-12):502–506, 2008.

[299] C. J. Wilson, D. Apiyo, and P. Wittung-Stafshede. Role of cofactors in metalloprotein folding. *Quarterly Reviews of Biophysics*, 37(3-4), 2004.

[300] P. Wong, S. Gladney, and J. D. Keasling. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnology Progress*, 13(2):132–143, 1997.

[301] J. Yang, J. P. Bitoun, and H. Ding. Interplay of IscA and IscU in biogenesis of iron-sulfur clusters. *Journal of Biological Chemistry*, 2006.

[302] B. Zeeberg. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome research*, 12(6):944–955, 2002.

[303] L. Zheng, V. L. Cash, D. H. Flint, and D. R. Dean. Assembly of iron-sulfur clusters. identification of an iscSUA-hscBA-fdx gene cluster from *Azotobacter vinelandii*. *Journal of Biological Chemistry*, 273(21):13264–13272, 1998.

[304] H. Zouridis and V. Hatzimanikatis. A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. *Biophysical Journal*, 92(3):717–730, 2007.

[305] Y. Zuo and M. P. Deutscher. The DNase activity of RNase T and its application to DNA cloning. *Nucleic Acids Research*, 27(20):4077–4082, 1999.