Recent technological advances have enabled comprehensive determination of the molecular composition of living cells. The chemical interactions between many of these molecules are known, giving rise to genome-scale reconstructed biochemical reaction networks underlying cellular functions. Mathematical descriptions of the totality of these chemical interactions lead to genome-scale models that allow the computation of physiological functions.

Reflecting these recent developments, this textbook explains how such quantitative and computable genotype-phenotype relationships are built using a genome-wide basis of information about the gene portfolio of a target organism. It describes how biological knowledge is assembled to reconstruct biochemical reaction networks, the formulation of computational models of biological functions, and how these models can be used to address key biological questions and enable predictive biology.

Developed through extensive classroom use, the book is designed to provide students with a solid conceptual framework and an invaluable set of modeling tools and computational approaches.

Detailed lecture slides, along with MATLAB™ and Mathematica™ workbooks, are available for download at www.cambridge.org/sb.

**Bernhard O. Palsson** is the Galletti Professor of Bioengineering and Professor of Pediatrics at the University of California, San Diego. For almost 30 years, his research has focused on the development of large-scale models of biological functions and their use to solve basic and applied problems in the life sciences. He has authored three previous textbooks.
Systems Biology
Constraint-based Reconstruction and Analysis

BERNHARD O. PALSSON
Department of Bioengineering,
University of California at San Diego, USA
To SHIREEN and SIRUS
# Contents

*Preface* xv  
*List of abbreviations* xvii

1 **Introduction** 1  
1.1 The Genotype–Phenotype Relationship 1  
1.2 Some Concepts of Genome-scale Science 3  
1.3 The Emergence of Systems Biology 8  
1.4 Building Foundations 11  
1.5 About This Book 12  
1.6 Summary 13

## Part I Network Reconstruction 15

2 **Network Reconstruction: The Concept** 17  
2.1 Many Reactions and Their Stoichiometry 17  
2.2 Reconstructing a Pathway 18  
2.3 Module-by-module Reconstruction 22  
2.4 Proteins and Their Many States 25  
2.5 Central *E. coli* Energy Metabolism 29  
2.6 Genome-scale Networks 29  
2.7 Summary 32

3 **Network Reconstruction: The Process** 33  
3.1 Building Knowledge Bases 33  
3.2 Reconstruction is a Four-step Process 35  
3.3 Reconstruction is Iterative and Labor-intensive 43  
3.4 The Many Uses of Reconstructions 45  
3.5 Summary 49

4 **Metabolism in *Escherichia coli*** 50  
4.1 Some Basic Facts about *E. coli* 50  
4.2 History 51  
4.2.1 Pre-genome era reconstructions 53  
4.2.2 Genome era reconstructions 55  
4.3 Content of the jJO1366 Reconstruction 65  
4.4 From a Reconstruction to a Computational Model 67  
4.5 Validation of jJO1366 68  
4.6 Uses of the *E. coli* GEM 70  
4.7 Summary 73

5 **Prokaryotes** 75  
5.1 State of The Field 75
5.2 Metabolism in Pathogens 78
5.3 Metabolism in Blue-Green Algae 80
5.4 Metabolism in Microbial Communities 82
  5.4.1 Systems biology of communities 85
  5.4.2 Model-based analysis of microbial communities 86
5.5 An Environmentally Important Organism 88
  5.5.1 Geobacter sulfurreducens 88
  5.5.2 Genome-scale science for Geobacter 90
5.6 Summary 95

6 Eukaryotes 96
  6.1 Metabolism in Saccharomyces cerevisiae 96
    6.1.1 Reconstruction and its uses 96
    6.1.2 Community-based reconstruction 98
  6.2 Metabolism in Chlamydomonas reinhardtii 101
    6.2.1 Metabolic network reconstruction 101
    6.2.2 Description of photon usage 102
  6.3 Metabolism in Homo sapiens 103
    6.3.1 Recon 1 104
    6.3.2 Uses of Recon 1 106
    6.3.3 Building multi-cell and multi-tissue reconstructions 108
    6.3.4 Mapping Recon 1 onto other mammals 114
    6.3.5 Recon 2 114
  6.4 Summary 116

7 Biochemical Reaction Networks 117
  7.1 Protein Properties 117
  7.2 Structural Biology 119
  7.3 Transcription and Translation 123
  7.4 Integrating Network Reconstructions 128
  7.5 Signaling Networks 131
  7.6 Summary 133

8 Metastructures of Genomes 134
  8.1 The Concept of a Metastructure 134
  8.2 Transcriptional Regulatory Networks 138
  8.3 Refactoring DNA for Synthetic Biology 142
  8.4 The Challenge of Polyomic Data Integration 144
  8.5 Building Mathematical Descriptions 145
  8.6 Summary 148

Part II Mathematical Properties of Reconstructed Networks 149

9 The Stoichiometric Matrix 151
  9.1 The Many Attributes of S 151
  9.2 Chemistry: S as a Data Matrix 153
### CONTENTS

12.4 Pathways as Basis Vectors  
12.4.1 Some perspective  
12.4.2 Extreme pathways  
12.4.3 Classifying extreme pathways  
12.4.4 The simplest set of linearly independent basis vectors  
12.4.5 Examples of pathway computation

12.5 Summary  

13 Use of Pathway Vectors  
13.1 The Matrix of Pathway Vectors  
13.2 Pathway Length and Flux Maps  
13.3 Reaction Participation and Correlated Subsets  
13.4 Input–output Relationships and Crosstalk  
13.5 Regulation Eliminates Active Pathways  
13.6 Summary

14 Randomized Sampling  
14.1 The Basics  
14.2 Sampling Low-dimensional Spaces  
14.3 Sampling High-dimensional Spaces  
14.4 Sampling Network States in Human Metabolism  
14.5 Summary

#### Part III Determining the Phenotypic Potential of Reconstructed Networks

15 Dual Causality  
15.1 Causation in Physics and Biology  
15.2 Building Quantitative Models  
15.2.1 The physical sciences  
15.2.2 The life sciences  
15.2.3 Genome-scale models  
15.3 Constraints in Biology  
15.4 Summary

16 Functional States  
16.1 Components vs. Systems  
16.2 Properties of Links  
16.3 Links to Networks to Biological Functions  
16.4 Constraining Allowable Functional States  
16.5 Biological Consequences of Constraints  
16.6 Summary

17 Constraints  
17.1 Genome-scale Viewpoints  
17.2 Stating and Imposing Constraints
21.2 Types of Objective Functions 343
21.3 Producing Biomass 344
21.4 Formulating The Biomass Objective Function 350
21.5 Studying the Objective Function 353
21.6 Objective Functions in Practice 353
21.7 Summary 355

Part IV Basic and Applied Uses 357

22 Environmental Parameters 359
22.1 Varying a Single Parameter 359
   22.1.1 Robustness analysis 359
   22.1.2 The effects of oxygen on ATP production 359
   22.1.3 The effects of oxygen uptake rate on growth rate 362
   22.1.4 Sensitivity with respect to key processes 364
   22.1.5 Uses of robustness analysis 366
22.2 Varying Two Parameters 368
   22.2.1 Phenotypic phase planes 368
   22.2.2 Using the PhPP at a small scale 371
   22.2.3 Using the PhPP at the genome-scale 371
22.3 Summary 377

23 Genetic Parameters 378
23.1 Single Gene Knock-outs 378
   23.1.1 Concept 378
   23.1.2 Core \textit{E. coli} metabolic network 379
   23.1.3 Genome-scale studies of essential genes 383
   23.1.4 Studying non-lethal gene KOs 386
23.2 Double Gene Knock-outs 390
   23.2.1 Core \textit{E. coli} metabolic network 391
   23.2.2 Genome-scale studies 392
23.3 Gene Dosage and Sequence Variation 395
23.4 Summary 397

24 Analysis of Omic Data 398
24.1 Context for Content 398
24.2 Omics Data-mapping and Network Topology 402
24.3 Omics Data as Constraints 403
24.4 Omics Data and Validation of GEM Predictions 405
24.5 Summary 406

25 Model-Driven Discovery 407
25.1 Models Can Drive Discovery 407
25.2 Predicting Gap-filling Reactions 412
25.3 Predicting Metabolic Gene Functions 416
25.4 Summary 420
26 Adaptive Laboratory Evolution 422
   26.1 A New Line of Biological Inquiry 422
   26.2 Determining the Genetic Basis 424
   26.3 Interpretation of Outcomes 427
   26.4 A Specific Example of Nutrient Adaptation 431
   26.5 General Uses of ALE 433
   26.6 Complex Examples of Adaptive Evolution 433
   26.7 Summary 437

27 Model-driven Design 438
   27.1 Historical Background 438
   27.2 GEMs and Design Algorithms 444
   27.3 GEMs and Cell Factory Design 446
   27.4 Summary 450

Part V Conceptual Foundations 451

28 Teaching Systems Biology 453
   28.1 The Core Paradigm 453
   28.2 High-throughput Technologies 455
   28.3 Network Reconstruction 455
   28.4 Computing Functional States of Networks 458
      28.4.1 Conversion to a computational model 458
      28.4.2 Topological properties 459
      28.4.3 Determining the capabilities of networks 459
      28.4.4 Dynamic states 462
   28.5 Prospective Experimentation 462
   28.6 Building a Curriculum 464
   28.7 Summary 466

29 Epilogue 467
   29.1 The Brief History of COBRA 467
   29.2 Common Misunderstandings 469
   29.3 Questions in Biology and in Systems Biology 470
   29.4 Why Build Mathematical Models? 472
   29.5 What Lies Ahead? 473

References 481
Index 510
The genesis of the bottom-up approach to systems biology was the availability of the first full genome sequences. In principle, these sequences had information about all the genetic elements that underlie the function of the sequenced organism. Enough information was available about the function of subsets of these genes – namely the genes encoding metabolic functions – that an organized assembly of all the biochemical, genetic, and genomic information was achievable. Such an organized assembly is de facto a knowledge base, or a k-base, that gives rise to a network reconstruction at the genome-scale. Since such reconstructions are represented with accurate chemical equations, they can be mathematically described. A mathematical description can be used to compute functional states of a network that correspond to observable phenotypes and biological functions. With these elements in place, a new genome-scale science was born that focused on mechanistic genotype–phenotype relationships.

The first genome-scale models of metabolism appeared in 1999 and 2000. In the next half-decade or so, an enthusiastic group of investigators developed many fundamental concepts, in silico methods, and algorithms to analyze their properties. At times, and to many, these initial efforts seemed mostly exploratory. Fortunately, in the mid-2000s an abundance of data sets and data types became available to validate and demonstrate the utility of genome-scale models for research and discovery. At the end of the decade several highly curated models were available for model organisms. These models gained predictive power and over the next 5 years or so, a number of prospective uses of genome-scale models appeared. In other words, predictive genotype–phenotype relationships had appeared. These predictions were somewhat limited in scope, but proved useful for a series of applications. Currently the range of possible predictions of biological properties and functions is growing rapidly and it appears that this approach to genome-scale science is in its early stages of development, with a bright future ahead of it.

For most of this fifteen-year history, the focus of genome-scale models has been metabolism. After initial successes with metabolic genome-scale models, it became clear that the same approach that led to their genesis could be applied to any other cellular process reconstructed in biochemically accurate detail. Thus, a vision was laid out in 2003 that the path to whole-cell models was conceptually possible and that such models could be used as a context for mechanistically integrating disparate omic data types. Ten years later, this vision started to be realized and a rapidly growing number of cellular functions are being reconstructed and addressed computationally. Given the fact that the genotype–phenotype relationship is fundamental to biology, this development has a broad transformative potential for the life sciences.

Writing this book was hard. It represents an attempt to summarize the concepts that have been developing over the past 15 years or so, that underlie what has become a true genome-scale science. Looking at the history of the field after the writing process, it is quite remarkable to see its rapid emergence, development, and maturation. Furthermore, looking forward, it appears that numerous areas of microbiology, cell
biology, and developmental biology will be influenced by the approach and methods described in this book.

To master this field one needs familiarity with an unusual range of disciplines. One needs to understand the basics of life sciences: biochemistry, molecular biology, genetics, microbiology, and cell biology. High-throughput measurements call for an understanding of basic technological characteristics, such as multiplexing, miniaturization, and automation. The large data sets generated call for proficiency in bioinformatics and a comfort level with big data. Mathematically modeling such data sets from a fundamental standpoint requires familiarity with the mathematical language of linear algebra and logistical relationships. Simulations require the use of constraint-based optimization and an understanding of the evolutionary principles of generation of diversity and selection. Bottom-up systems biology is thus a field with a broad conceptual basis. This book attempts to bring all these concepts from the expert level to the general senior or first-year graduate student level in bioengineering, bioinformatics, and life sciences.

As with all major undertakings, this project could not have been completed without the help of several individuals.

Marc Abrams managed all aspects of the preparation of the manuscript. He tirelessly helped me with preparing the text and the illustrations, assembling the references, correcting LATEX scripts, and interacting with the publisher. Without him this book would not have been completed.

Nathan Lewis and Adam Feist were responsible for the challenging task of managing the original illustrations in the book. Their contribution was immense, making the concepts in the text and the material as a whole more accessible.

The following people were generous with their time and expertise, improving the manuscript with their contributions to the text, figures, or proofreading of the final manuscript. I am very grateful to these individuals:


The conceptual framework that this book describes has been under development since the birth of my two children, to whom it is dedicated.

Bernhard Palsson
On the Oracle, August 2014
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALE</td>
<td>adaptive laboratory evolution</td>
</tr>
<tr>
<td>AOS</td>
<td>alternative optimal solutions</td>
</tr>
<tr>
<td>BiGG</td>
<td>biochemical genetic and genomic</td>
</tr>
<tr>
<td>BOF</td>
<td>biomass objective function</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>COBRA</td>
<td>constraint-based reconstruction and analysis</td>
</tr>
<tr>
<td>CoSy</td>
<td>community systems</td>
</tr>
<tr>
<td>DIET</td>
<td>direct interspecies electron transfer</td>
</tr>
<tr>
<td>DIP</td>
<td>di-myoinositol 1,1’-phosphate</td>
</tr>
<tr>
<td>DMMM</td>
<td>dynamic multi-species metabolic modeling</td>
</tr>
<tr>
<td>EnMe</td>
<td>endo-metabolome</td>
</tr>
<tr>
<td>ETS</td>
<td>electron-transport system</td>
</tr>
<tr>
<td>ExMe</td>
<td>exo-metabolome</td>
</tr>
<tr>
<td>FA</td>
<td>fraction of agreement</td>
</tr>
<tr>
<td>FBA</td>
<td>flux balance</td>
</tr>
<tr>
<td>FCF</td>
<td>flux coupling finder</td>
</tr>
<tr>
<td>FIG</td>
<td>Fellowship for Interpretation of Genomes</td>
</tr>
<tr>
<td>FVA</td>
<td>flux variability analysis</td>
</tr>
<tr>
<td>GAM</td>
<td>growth-associated maintenance</td>
</tr>
<tr>
<td>GDS</td>
<td>genetic design through local search</td>
</tr>
<tr>
<td>GEM</td>
<td>genome-scale model</td>
</tr>
<tr>
<td>GENRE</td>
<td>genome-scale reconstruction</td>
</tr>
<tr>
<td>GOF</td>
<td>gain of function</td>
</tr>
<tr>
<td>GPR</td>
<td>gene–protein–to–reaction</td>
</tr>
<tr>
<td>GUR</td>
<td>glucose uptake rate</td>
</tr>
<tr>
<td>HGP</td>
<td>human genome project</td>
</tr>
<tr>
<td>HMDB</td>
<td>Human Metabolome Database</td>
</tr>
<tr>
<td>HT</td>
<td>high-throughput</td>
</tr>
<tr>
<td>I/O</td>
<td>input/output</td>
</tr>
<tr>
<td>IDV</td>
<td>isotopomer distribution vector</td>
</tr>
<tr>
<td>IEM</td>
<td>inborn error of metabolism</td>
</tr>
<tr>
<td>IOFA</td>
<td>input-output feasibility array</td>
</tr>
<tr>
<td>k-base</td>
<td>knowledge base</td>
</tr>
<tr>
<td>Ki</td>
<td>knock-in</td>
</tr>
<tr>
<td>KO</td>
<td>knock-out</td>
</tr>
<tr>
<td>LIMS</td>
<td>laboratory information management system</td>
</tr>
<tr>
<td>LO</td>
<td>line of optimality</td>
</tr>
<tr>
<td>LOF</td>
<td>loss of function</td>
</tr>
<tr>
<td>LPR</td>
<td>ligand to protein to reaction</td>
</tr>
<tr>
<td>LPS</td>
<td>lipid polysaccharide</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>MDV</td>
<td>mass distribution vector</td>
</tr>
<tr>
<td>MILP</td>
<td>mixed-integer linear programming</td>
</tr>
<tr>
<td>MOMA</td>
<td>minimization of metabolic adjustment</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MU</td>
<td>modular unit</td>
</tr>
<tr>
<td>NGAM</td>
<td>non-growth-associated maintenance</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>NTP</td>
<td>nucleotide triphosphate</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PFL</td>
<td>pyruvate formate lyase</td>
</tr>
<tr>
<td>PhPP</td>
<td>phenotypic phase plane</td>
</tr>
<tr>
<td>POR</td>
<td>pyruvate oxidoreductase</td>
</tr>
<tr>
<td>PPS</td>
<td>pentose phosphate shunt</td>
</tr>
<tr>
<td>PVT</td>
<td>pressure volume temperature</td>
</tr>
<tr>
<td>QA</td>
<td>quality-assured</td>
</tr>
<tr>
<td>QC</td>
<td>quality-controlled</td>
</tr>
<tr>
<td>RBR</td>
<td>RNA polymerase binding region</td>
</tr>
<tr>
<td>RBS</td>
<td>ribosome binding site</td>
</tr>
<tr>
<td>rFBA</td>
<td>regulated flux balance</td>
</tr>
<tr>
<td>ROOM</td>
<td>regulation off/on modification</td>
</tr>
<tr>
<td>RTS</td>
<td>RNAP-guided transcript segment</td>
</tr>
<tr>
<td>SKI</td>
<td>species knowledge index</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOP</td>
<td>standard operating procedure</td>
</tr>
<tr>
<td>SVD</td>
<td>singular value decomposition</td>
</tr>
<tr>
<td>TCA</td>
<td>tricarboxylic acid</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>Tr/Tr</td>
<td>transcription/translation</td>
</tr>
<tr>
<td>TRN</td>
<td>transcriptional regulatory network</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>TU</td>
<td>transcription unit</td>
</tr>
</tbody>
</table>