

Systems Biology

Constraint-based Reconstruction and Analysis

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



CAMBRIDGE
UNIVERSITY PRESS



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107038851

© B. O. Palsson 2015

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

First published 2015

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data

Palsson, Bernhard, author.

Systems biology : constraint-based reconstruction and analysis / Bernhard O. Palsson.

p. ; cm.

Includes bibliographical references.

ISBN 978-1-107-03885-1 (Hardback)

1. Title.

[DNLM: 1. Models, Biological. 2. Systems Biology. 3. Metabolic Networks and
Pathways—physiology. QU 26.5]

QH508

571.7—dc23 2014031793

ISBN 978-1-107-03885-1 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

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 *iJO1366* Reconstruction 65
4.4 From a Reconstruction to a Computational Model 67
4.5 Validation of *iJO1366* 68
4.6 Uses of the *E. coli* GEM 70
4.7 Summary 73

5 Prokaryotes 75
5.1 State of The Field 75

viii CONTENTS

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	<i>Geobacter sulfurreducens</i>	88
5.5.2	Genome-scale science for <i>Geobacter</i>	90
5.6	Summary	95
6	Eukaryotes	96
6.1	Metabolism in <i>Saccharomyces cerevisiae</i>	96
6.1.1	Reconstruction and its uses	96
6.1.2	Community-based reconstruction	98
6.2	Metabolism in <i>Chlamydomonas reinhardtii</i>	101
6.2.1	Metabolic network reconstruction	101
6.2.2	Description of photon usage	102
6.3	Metabolism in <i>Homo sapiens</i>	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

9.2.1	Elementary biochemical reactions	154
9.2.2	Basic chemistry	155
9.2.3	Example: glycolysis	158
9.3	Network Structure: \mathbf{S} as a Connectivity Matrix	161
9.3.1	The maps of \mathbf{S}	161
9.3.2	Biological quantities displayed on maps	161
9.3.3	Linearity of maps	165
9.4	Mathematics: \mathbf{S} as a Linear Transformation	165
9.4.1	Mapping fluxes onto concentration time derivatives	165
9.4.2	The four fundamental subspaces	166
9.4.3	Looking into the four fundamental subspaces	167
9.5	Systems Science: \mathbf{S} and Network Models	168
9.6	Summary	171
10	Simple Topological Network Properties	172
10.1	The Binary Form of \mathbf{S}	172
10.2	Participation and Connectivity	173
10.2.1	Rearranging the stoichiometric matrix	174
10.2.2	Connectivities in genome-scale matrices	175
10.3	Linked Participation and Connectivities	179
10.3.1	The adjacency matrices of \mathbf{S}	179
10.3.2	Computation of the adjacency matrices	180
10.4	Summary	182
11	Fundamental Network Properties	184
11.1	Singular Value Decomposition	184
11.1.1	Decomposition into three matrices	184
11.1.2	The content of \mathbf{U} , $\mathbf{\Sigma}$, and \mathbf{V}	185
11.1.3	Key properties of the SVD	188
11.2	SVD and Properties of Reaction Networks	189
11.3	Studying Elementary Reactions using SVD	191
11.3.1	The linear reversible reaction	191
11.3.2	The bi-linear association reaction	193
11.4	Studying Network Structure Using SVD	196
11.5	Drivers and Directions	199
11.5.1	Directions: the column space	199
11.5.2	Drivers: the row space	201
11.5.3	The fundamental subspaces are of a finite size	201
11.6	Summary	202
12	Pathways	204
12.1	Network-based Pathway Definitions	204
12.2	Choice of a Basis	205
12.3	Confining the Steady-state Flux Vector	208
12.3.1	Finite or closed spaces	208
12.3.2	Importance of constraints	210

x CONTENTS

12.4 Pathways as Basis Vectors	212
12.4.1 Some perspective	212
12.4.2 Extreme pathways	214
12.4.3 Classifying extreme pathways	215
12.4.4 The simplest set of linearly independent basis vectors	217
12.4.5 Examples of pathway computation	217
12.5 Summary	220

13 Use of Pathway Vectors 221

13.1 The Matrix of Pathway Vectors	221
13.2 Pathway Length and Flux Maps	222
13.3 Reaction Participation and Correlated Subsets	224
13.4 Input–output Relationships and Crosstalk	228
13.5 Regulation Eliminates Active Pathways	230
13.6 Summary	232

14 Randomized Sampling 233

14.1 The Basics	233
14.2 Sampling Low-dimensional Spaces	234
14.3 Sampling High-dimensional Spaces	237
14.4 Sampling Network States in Human Metabolism	241
14.5 Summary	247

Part III Determining the Phenotypic Potential of Reconstructed Networks 249

15 Dual Causality 251

15.1 Causation in Physics and Biology	251
15.2 Building Quantitative Models	255
15.2.1 The physical sciences	255
15.2.2 The life sciences	255
15.2.3 Genome-scale models	256
15.3 Constraints in Biology	260
15.4 Summary	263

16 Functional States 264

16.1 Components vs. Systems	264
16.2 Properties of Links	266
16.3 Links to Networks to Biological Functions	267
16.4 Constraining Allowable Functional States	271
16.5 Biological Consequences of Constraints	272
16.6 Summary	276

17 Constraints 277

17.1 Genome-scale Viewpoints	277
17.2 Stating and Imposing Constraints	279

17.3	Capacity Constraints	282
17.4	Constraints from Chemistry	285
17.4.1	Mass conservation	286
17.4.2	Thermodynamics	287
17.4.3	Fluxomics	288
17.5	Regulatory Constraints	289
17.6	Coupling Constraints	291
17.7	Simultaneous Satisfaction of All Constraints	296
17.8	Summary	297
18	Optimization	298
18.1	Overview of Constraint-based Methods	298
18.2	Finding Functional States	300
18.3	Linear Programming: the basics	302
18.4	Genome-scale Models	306
18.5	Summary	311
19	Determining Capabilities	312
19.1	Optimal Network Performance	312
19.1.1	Co-factors	312
19.1.2	Biosynthetic Precursors	313
19.2	Production of ATP	315
19.2.1	Producing ATP aerobically from glucose	315
19.2.2	Producing ATP anaerobically from glucose	319
19.2.3	Optimal ATP production from other substrates	320
19.3	Production of Redox Potential	320
19.3.1	Aerobic production of NADH from glucose	320
19.3.2	Anaerobic production of NADH	324
19.4	Capabilities of Genome-scale Models	325
19.5	Summary	325
20	Equivalent States	327
20.1	Equivalent Ways to Reach a Network Objective	327
20.2	Flux Variability Analysis	330
20.2.1	The concept	330
20.2.2	Flux variability in the core <i>E. coli</i> model	332
20.2.3	Genome-scale results	335
20.3	Extreme Pathways and Optimal States	336
20.3.1	The concept	336
20.3.2	Extreme pathways in the core <i>E. coli</i> metabolic network	337
20.3.3	Genome-scale results	337
20.4	Enumerating Alternative Optima	339
20.5	Summary	341
21	Distal Causation	342
21.1	The Objective Function	342

xii CONTENTS

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 <i>E. coli</i> 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 <i>E. coli</i> 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
	<i>References</i>	481
	<i>Index</i>	510

Preface

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 L^AT_EX 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:

Ramy Aziz, Aarash Bordbar, Roger Chang, Addiel U. de Alba Solis, Andreas Drger, Juan Nogales Enrique, Gabriela Guzman, Hooman Hefzi, Daniel Hyduke, Neema Jamshidi, Ryan LaCroix, Haythem Latif, Josh Lerman, Douglas McCloskey, Jonathan Monk, Harish Nagarajan, Jeff Orth, Troy Sandberg, Nikolaus Sonnenschein, Alex Thomas, and Daniel Zielinski.

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

Abbreviations

ALE	adaptive laboratory evolution
AOS	alternative optimal solutions
BiGG	biochemical genetic and genomic
BOF	biomass objective function
CDS	coding sequence
COBRA	constraint-based reconstruction and analysis
CoSy	community systems
DIET	direct interspecies electron transfer
DIP	di- <i>myo</i> -inositol 1,1'-phosphate
DMMM	dynamic multi-species metabolic modeling
EnMe	endo-metabolome
ETS	electron-transport system
ExME	exo-metabolome
FA	fraction of agreement
FBA	flux balance
FCF	flux coupling finder
FIG	Fellowship for Interpretation of Genomes
FVA	flux variability analysis
GAM	growth-associated maintenance
GDLS	genetic design through local search
GEM	genome-scale model
GENRE	genome-scale reconstruction
GOF	gain of function
GPR	gene-to-protein-to-reaction
GUR	glucose uptake rate
HGP	human genome project
HMDB	Human Metabolome Database
HT	high-throughput
I/O	input/output
IDV	isotopomer distribution vector
IEM	inborn error of metabolism
IOFA	input-output feasibility array
k-base	knowledge base
KI	knock-in
KO	knock-out
LIMS	laboratory information management system
LO	line of optimality
LOF	loss of function
LPR	ligand to protein to reaction
LPS	lipid polysaccharide

xviii LIST OF ABBREVIATIONS

MDV	mass distribution vector
MILP	mixed-integer linear programming
MOMA	minimization of metabolic adjustment
MS	mass spectrometry
MU	modular unit
NGAM	non-growth-associated maintenance
NMR	nuclear magnetic resonance
NTP	nucleotide triphosphate
ORF	open reading frame
PCA	principal component analysis
PDB	Protein Data Bank
PFL	pyruvate formate lyase
PhPP	phenotypic phase plane
POR	pyruvate oxidoreductase
PPS	pentose phosphate shunt
PVT	pressure volume temperature
QA	quality-assured
QC	quality-controlled
RBR	RNA polymerase binding region
RBS	ribosome binding site
rFBA	regulated flux balance
ROOM	regulation off/on modification
RTS	RNAP-guided transcript segment
SKI	species knowledge index
SNP	single nucleotide polymorphism
SOP	standard operating procedure
SVD	singular value decomposition
TCA	tricarboxylic acid
TF	transcription factor
Tr/Tr	transcription/translation
TRN	transcriptional regulatory network
TSS	transcription start site
TU	transcription unit