MICROBIAL FUNCTIONAL GENOMICS

JIZHONG ZHOU
DOROTHEA K. THOMPSON
YING XU
JAMES M. TIEDJE

Contents

Foreword

Preface

Acknowledgments

1. Genomics: Toward a Genome-level Understanding of the Structure, Functions, and Evolution of Biological Systems 1
Jizhong Zhou, Dorothea K. Thompson, and James M. Tiedje

1.1 Introduction / 1
1.2 Definitions and Classifications / 2
1.2.1 Classification based on system attributes / 3
1.2.2 Classification based on relationships to other scientific disciplines / 5
1.2.3 Classification based on types of organisms studied / 7
1.3 Historical Perspective of Genomics / 7
1.4 Challenges of Studying Functional Genomics / 10
1.4.1 Defining gene function / 10
1.4.2 Indentifying and characterizing the molecular machines of life / 11
1.4.3 Delineating gene regulatory networks / 11
1.4.4 System-level understanding of biological systems beyond individual cells / 12
1.4.5 Computational challenges / 13
1.4.6 Multidisciplinary collaborations / 13
1.5 Scope and General Approaches / 13
1.5.1 Structural Genomics / 13
1.5.2 Transcriptomics / 14

1.5.3 Proteomics / 16
1.6 Importance of Microbial Functional Genomics to the Study of Eukaryotes / 17
1.7 Summary / 18

2. Microbial Diversity and Genomics **21**
Konstantinos Konstantinidis and James M. Tiedje

2.1 Introduction / 21
2.2 Biochemical Diversity / 21
2.3 Genetic Diversity / 24
2.3.1 The unseen majority / 24
2.3.2 How many prokaryotic species are there? / 25
2.4 The Challenge of Describing Prokaryotic Diversity / 27
2.4.1 Methods to study microbial diversity / 27
2.4.2 Limitations of culture-independent methods /
2.4.3 Interesting findings from culture-independent approaches / 28
2.5 Diversity of Microbial Genomes and Whole-Genome Sequencing / 31
2.5.1 Genomic diversity within species / 35
2.5.2 Genome structure and its relation to the ecological niche / 36
2.5.3 General trends in genome functional content / 38
2.5.4 Biases in the collection of sequenced species: a limit to understanding / 38
2.6 Summary / 39

3. Computational Genome Annotation **41**
Ying Xu

3.1 Introduction / 41
3.2 Prediction of Protein-Coding Genes / 42
3.2.1 Evaluation of coding potential / 43
3.2.2 Identification of translation start / 46
3.2.3 Ab initio gene prediction through information fusion / 47
3.2.4 Gene identification through comparative analysis / 51
3.2.5 Interpretation of gene prediction / 53
3.3 Prediction of RNA-Coding Genes / 54
3.4 Identification of Promoters / 57
3.4.1 Promoter prediction through feature recognition / 57
3.5 Operon Identification / 59
3.6 Functional Categories of Genes / 60
3.7 Characterization of Other Features in a Genome / 62
3.8 Genome-Scale Gene Mapping / 63

3.9 Existing Genome Annotation Systems / 63
3.10 Summary / 66

4. Microbial Evolution from a Genomics Perspective **67**
Jizhong Zhou and Dorothea K. Thompson

4.1 Introduction / 67
4.2 Identification of Orthologous Genes / 68
4.3 Genome Perspectives on Molecular Clock / 70
4.3.1 Historical overview / 70
4.3.2 Current genomic view on molecular evolutionary clock / 72
4.3.3 Timing genome divergence / 73
4.4 Genome Perspectives on Horizontal Gene Transfer / 75
4.4.1 Historical overview of horizontal gene transfer / 75
4.4.2 Identification of HGT / 76
4.4.3 Mechanisms underlying HGT / 78
4.4.4 Types of genes subjected to HGT / 79
4.4.5 Classification and scope of HGT / 81
4.4.6 Evolutionary impact of HGT / 83
4.5 Genomic Perspectives on Gene Duplication, Gene Loss, and Other Evolutionary Processes / 85
4.5.1 Gene and genome duplication / 85
4.5.2 Genomic perspectives on gene loss / 88
4.5.3 Genomic perspectives on other evolutionary processes / 90
4.6 Universal Tree of Life / 91
4.6.1 Establishment of a universal tree of life / 91
4.6.2 Challenges and current view of the universal tree / 96
4.6.3 Genome-based phylogenetic analysis / 98
4.7 Minimal Genomes / 102
4.8 Genomic Insights into Lifestyle Evolution / 103
4.9 Genome Perspective of Mitochondrial Evolution / 105
4.10 Summary / 109

5. Computational Methods for Functional Prediction of Genes **113**
Ying Xu

5.1 Introduction / 113
5.2 Methods for Gene Function Inference / 114
5.2.1 Gene functions at different levels / 114
5.2.2 Searching for clues to gene function / 114
5.3 From Gene Sequence to Function / 117
5.3.1 Hierarchies of protein families / 117
5.3.2 Searching family trees / 122

5.3.3 Orthologous vs. paralogous genes / 125
5.3.4 Genes with multiple domains / 126
5.4 Identification of Sequence Motifs / 128
5.5 Structure-Based Function Prediction / 129
5.5.1 Protein fold recognition through protein threading / 130
5.5.2 From structure to function / 132
5.5.3 Disordered vs. ordered regions in proteins / 134
5.6 Nonhomologous Approaches to Functional Inference / 135
5.7 Functional Inference at a Systems Level / 136
5.8 Summary / 139

6. DNA Microarray Technology **141**
Jizhong Zhou and Dorothea K. Thompson

6.1 Introduction / 141
6.2 Types of Microarrays and Advantages / 143
6.2.1 Concepts, principles, and history / 143
6.2.2 Microarray types and their advantages / 144
6.3 Microarray Fabrication / 146
6.3.1 Microarray fabrication substrates and modification / 146
6.3.2 Arraying technology / 151
6.3.3 Critical issues for microarray fabrication / 155
6.4 Microarray Hybridization and Detection / 158
6.4.1 Probe design, target preparation, and quality / 158
6.4.2 Hybridization / 163
6.4.3 Detection / 163
6.4.4 Critical issues in hybridization and detection / 164
6.5 Microarray Image Processing / 166
6.5.1 Data acquisition / 166
6.5.2 Assessment of spot quality and reliability, and background subtraction / 167
6.6 Using Microarrays to Monitor Gene Expression / 170
6.6.1 General approaches to revealing differences in gene expression / 170
6.6.2 Specificity, sensitivity, reproducibility, and quantitation of microarray-based detection for monitoring gene expression / 172
6.6.3 Microarray experimental design for monitoring gene expression / 173
6.7 Summary / 173

7. Microarray Gene Expression Data Analysis **177**
Ying Xu

7.1 Introduction / 177
7.2 Normalization of Microarray Gene Expression Data / 179

7.2.1 Sources of systematic errors / 179
7.2.2 Experimental design to minimize systematic variations / 180
7.2.3 Selection of reference points for data normalization / 181
7.2.4 Normalization methods / 182
7.3 Data Analysis / 185
7.3.1 Data transformation / 185
7.3.2 Principle component analysis / 185
7.4 Identification of Differentially Expressed Genes / 186
7.5 Identification of Coexpressed Genes / 189
7.5.1 Basics of gene expression data clustering / 191
7.5.2 Clustering of gene expression data / 193
7.5.3 Cluster identification from noisy background / 197
7.5.4 EXCAVATOR: a software for gene expression data analysis / 199
7.5.5 Discovering subtypes through data clustering / 203
7.6 Applications of Gene Expression Data Analysis for Pathway Inference / 203
7.6.1 Data-constrained pathway construction / 204
7.7 Summary / 205

8. Mutagenesis as a Genomic Tool for Studying Gene Function **207**
Alexander S. Beliaev

8.1 Introduction / 207
8.2 Transposon Mutagenesis / 205
8.2.1 Overview of transposition in bacteria / 208
8.2.2 Transposons as tools for mutagenesis / 211
8.2.3 Transposon-based approaches for identification of essential genes / 213
8.2.4 Signature-tagged mutagenesis for studying bacterial pathogenicity / 219
8.3 Targeted Mutagenesis Through Allelic Exchange / 224
8.3.1 Suicide vector systems for allelic exchange / 224
8.3.2 Strategies commonly utilized for targeted mutagenesis by allelic exchange / 225
8.3.3 Application of allele exchange approach in functional genomic studies for sequenced microorganisms / 231
8.4 Gene Silencing Using Antisense mRNA Molecules / 233
8.4.1 Antisense RNA regulation in vivo / 233
8.4.2 Antisense approach to large-scale functional genomic studies / 234
8.5 Summary / 238

9. Mass Spectrometry **241**
Nathan VerBerkmoes, Joshua Sharp, and Robert Hettich

9.1 Introduction / 241

9.2 Fundamentals of Mass Spectrometry / 242
9.2.1 Basic components of any mass spectrometer / 242
9.2.2 Ionization methods / 243
9.2.3 Mass analyzers / 244
9.2.4 Coupling separation methods with mass spectrometry / 246
9.2.5 Ion structural characterization / 247
9.3 Fundamentals of Protein and Peptide Mass Spectrometry / 248
9.3.1 Protein measurements / 248
9.3.2 Peptide measurements / 250
9.4 Mass Spectrometry for Protein and Proteome Characterization / 253
9.4.1 Overview of mass spectrometry approaches for protein studies / 253
9.4.2 Bottom-up mass spectrometry proteomics / 259
9.4.3 Top-down mass spectrometry proteomics / 275
9.4.4 Relating mass spectrometry proteomic data to biological information / 279
9.5 Summary / 282

10. Identification of Protein–Ligand Interactions **285**
Timothy Palzkill

10.1 Introduction / 285
10.2 High-Throughput Cloning of Open Reading Frames / 286
10.2.1 Bacteriophage λ att recombination-based cloning / 286
10.2.2 Topoisomerase-based cloning / 288
10.2.3 In vivo recombination-based cloning in yeast / 289
10.2.4 Advantages and disadvantages of recombinational cloning systems / 290
10.3 Yeast Two-Hybrid Selection System / 291
10.3.1 Analysis of genome-wide protein–protein interactions in yeast / 292
10.3.2 Genome-wide yeast two-hybrid analysis of other organisms / 297
10.4 Use of Phage Display to Detect Protein–Ligand Interactions / 298
10.4.1 Display of proteins on M13 filamentous phage / 298
10.4.2 Display of proteins on the T7 bacteriophage / 300
10.4.3 Combining yeast two-hybrid and phage display data / 301
10.5 Detecting Interactions with Protein Fragment Complementation Assays / 301
10.5.1 Overview / 301
10.5.2 Protein fragment complementation using dihydrofolate reductase / 302
10.5.3 Monitoring protein interactions by intracistronic β-galactosidase complementation / 303
10.6 Use of Mass Spectrometry for Protein–Protein Interaction Mapping / 304
10.6.1 Overview / 304

10.6.2 Identification of substrates for *E. coli* GroEL / 304
10.6.3 Identification of protein complexes in *Saccharomyces cerevisiae* / 306
10.7 Protein Assays for Protein Expression Profiling and Interactions / 306
10.7.1 Antibody arrays for protein expression profiling / 307
10.7.2 Functional analysis using peptide, protein, and small-molecule arrays / 307
10.8 Surface Plasmon Resonance Biosensor Analysis / 319
10.8.1 Measuring interactions of biomolecules with SPR / 320
10.8.2 Integration of SPR biosensors with mass spectrometry / 321
10.9 Summary / 322

11. The Functional Genomics of Model Organisms: Addressing Old Questions from a New Perspective **325**
Dorothea K. Thompson and Jizhong Zhou

11.1 Introduction / 325
11.2 *Escherichia coli*: A Model Eubacterium / 326
11.2.1 *E. coli* genome / 327
11.2.2 *E. coli* transcriptomics / 327
11.2.3 *E. coli* proteomics / 332
11.2.4 Modeling *E. coli* metabolism: *in silico* metabolomics / 334
11.3 *Bacillus subtilis*: A Paradigm for Gram-Positive Bacteria / 336
11.3.1 *B. subtilis* genome / 335
11.3.2 *B. subtilis* transcriptomics / 339
11.3.3 *B. subtilis* proteomics / 346
11.4 *Saccharomyces cerevisiae*: A Model for Higher Eukaryotes / 348
11.4.1 Yeast genome / 349
11.4.2 Yeast transcriptomics / 351
11.4.3 Yeast proteomics / 360
11.4.4 Yeast interactome: mapping protein–protein interactions / 363
11.5 Comparative Genomics of Model Eukaryotic Organisms / 370
11.6 Summary / 373

12. Functional Genomic Analysis of Bacterial Pathogens and Environmentally Significant Microorganisms **377**
Dorothea K. Thompson and Jizhong Zhou

12.1 Introduction / 377
12.2 Advancing Knowledge of Bacterial Pathogenesis through Genome Sequence and Function Annotation / 379
12.2.1 Predicting virulence genes from sequence homology / 379
12.2.2 Repeated DNA elements indicate potential virulence factors / 380
12.2.3 Evolution of bacterial pathogens: gene acquisition and loss / 382

12.3 Comparative Genomics: Clues to Bacterial Pathogenicity / 388
12.3.1 The genomics of *Mycobacterium tuberculosis*: virulence gene identification and genome plasticity / 389
12.3.2 Microarray-based comparative genomics of *Helicobacter pylori* / 391
12.3.3 Comparative analysis of the *Borrelia burgdorferi* and *Treponema pallidum* genomes / 394
12.3.4 Sequence comparison of pathogenic and nonpathogenic species of *Listeria* / 396
12.3.5 Comparative genomics of *Chlamydia pneumoniae* and *Chlamydia trachomatis*: two closely related obligate intracellular pathogens / 398
12.4 Discovery of Novel Infection-Related Genes Using Signature-Tagged Mutagenesis / 399
12.4.1 *Vibrio cholerae* genes critical for colonization / 400
12.4.2 Virulence genes of *Staphylococcus aureus* infection / 401
12.4.3 *Escherichia coli* K1: identification of invasion genes / 401
12.4.4 Diverse genes implicated in *Streptococcus pneumoniae* virulence / 402
12.5 Application of Microarrays to Delineating Gene Function and Interaction / 403
12.5.1 Exploring the transcriptome of bacterial pathogens / 404
12.5.2 Elucidating the molecular intricacies of host–pathogen interactions / 406
12.5.3 Identification of antimicrobial drug targets / 409
12.6 The Proteomics of Bacterial Pathogenesis / 410
12.6.1 Comparative proteomics / 411
12.6.2 Defining the proteome of individual bacterial pathogens / 412
12.6.3 Proteomic approach to host–pathogen interactions / 413
12.7 Genome Sequence and Functional Analysis of Environmentally Important Microorganisms / 414
12.7.1 Dissimilatory metal ion-reducing bacterium *Shewanella oneidensis* / 414
12.7.2 Extreme radiation-resistant bacterium *Deinococcus radiodurans* / 416
12.7.3 Hyperthermophilic archaeon *Pyrococcus furiosus* / 417
12.8 Summary / 420

13. The Impact of Genomics on Antimicrobial Drug Discovery and Toxicology 423
Dorothea K. Thompson and Jizhong Zhou

13.1 Introduction / 423
13.2 Antibacterial Drug Discovery: A Historical Perspective / 424
13.3 Challenges of New Drug Discovery / 427

13.3.1 Resistance to antimicrobial agents and the need for new antibiotic discovery / 427
13.3.2 Desirable properties of antimicrobial targets / 430
13.4 Microbial Genomics and Drug Target Selection / 431
13.4.1 Mining genomes for antimicrobial drug targets / 431
13.4.2 Comparative genomics: assessing target spectrum and selectivity / 434
13.4.3 Genetic strategies: verifying the essentiality or expression of gene targets / 435
13.4.4 Microarray analysis: establishing functionality for novel drug targets / 436
13.5 Determining Therapeutic Utility: Drug Target Screening and Validation / 438
13.5.1 Target-based drug screening / 439
13.5.2 Microarrays and drug target validation / 442
13.6 Genomics and Toxicology: The Emergence of Toxicogenomics / 446
13.6.1 Microarrays in mechanistic toxicology / 446
13.6.2 Microarrays in predictive toxicology / 449
13.7 Summary / 449

14. Application of Microarray-based Genomic Technology to Mutation Analysis and Microbial Detection **451**
Jizhong Zhou and Dorothea K. Thompson

14.1 Introduction / 451
14.2 Oligonucleotide Microarrays for Mutation Analysis / 452
14.2.1 Microarray-based hybridization assay with allele-specific oligonucleotides / 452
14.2.2 Microarray-based single-base extension for genotyping / 458
14.2.3 Microarray-based ligation detection reaction for genotyping / 461
14.3 Microarrays for Microbial Detection in Natural Environments / 461
14.3.1 Limitations of conventional molecular methods for microbial detection / 461
14.3.2 Advantages and challenges of microbial detection in natural environments / 462
14.3.3 Functional gene arrays / 463
14.3.4 Phylogenetic oligonucleotide arrays / 469
14.3.5 Community genome arrays / 471
14.3.6 Whole-genome open reading frame arrays for revealing genome differences and relatedness / 473
14.3.7 Other types of microarrays for microbial detection and characterization / 474
14.4 Summary / 475

15. Future Perspectives: Genomics Beyond Single Cells 477
James M. Tiedje and Jizhong Zhou

15.1 Introduction / 477
15.2 The Informational Base of Microbial Biology: Genome Sequences / 478
15.2.1 Determination of the Genetic Content of Both Cultured and Uncultured Microorganisms / 478
15.2.2 Community Genomics or Metagenomics / 480
15.3 Gene Functions and Regulatory Networks / 482
15.4 Ecology and Evolution / 483
15.5 System-level Understanding of Microbial Community Dynamics / 483
15.6 Summary / 485

Glossary 487

References 499

Index 575