

Contents

1	Introduction	1
	Henrik Christensen	
1.1	Basics and History of Bioinformatics	2
1.1.1	Bioinformatics Is Integrating Many Scientific Fields	2
1.1.2	Homology	3
1.1.3	Evolution Is Related to Sequence Diversity Caused by Mutations	4
1.2	The Aim, Structure, and Outline of the Book	5
1.3	Computers and Operating Systems Required for Bioinformatics	6
1.4	Computer Programs and Pipelines	7
1.5	Activity	8
	Further Reading	8
2	DNA Sequence Assembly and Annotation of Genes	9
	Henrik Christensen and Arshnee Moodley	
2.1	DNA Sequencing	10
2.1.1	Sanger Sequencing	12
2.1.2	Massive Parallel, Short-Read Sequencing	13
2.1.3	DNA Sequencing in Metagenomic and for Single Cell Sequencing	15
2.1.4	Real-Time, Single Molecule Sequencing	15
2.2	DNA Sequence Assembly	16
2.2.1	Base Calling and Trimming	16
2.2.2	Assembly of DNA Sequences	17
2.3	Closing of Genomes	21
2.4	DNA Sequence Formats	22
2.5	Annotation	23
2.5.1	Elements from Comparative Genomics Aiming Annotation	24
2.6	Activities	24
2.6.1	Annotation at RAST	24
	Further Reading	25

3	Databases and Protein Structures	29
	Henrik Christensen and Lisbeth de Vries	
3.1	Introduction to Bioinformatics Databases	30
3.1.1	Data Formats Used with Bioinformatics Databases	33
3.2	Organization of Databases and Bioinformatics Institutions	33
3.3	Major Bioinformatics Databases	35
3.3.1	GenBank	35
3.3.2	The European Nucleotide Archive (ENA)	35
3.3.3	Swiss-Prot and UniProt	36
3.3.4	Genomics Databases	38
3.3.5	Raw Sequence Read Datasets	38
3.3.6	Other Databases	39
3.3.7	Primary and Secondary Bioinformatics Databases	40
3.3.8	Data Formats in Bioinformatics Databases	41
3.4	Accession Numbers	42
3.5	Protein Structure Databases and Predictions	43
3.5.1	Primary and Secondary Structures	44
3.5.2	Domain Prediction and Databases	45
3.5.3	Protein 3D Structure	47
3.6	Overview of Proteomics Databases and Servers	49
3.7	Help to Databases	51
3.8	Activities	52
3.8.1	Download a Sequence from NCBI	52
3.8.2	Download a Genome from NCBI	52
3.8.3	Deposition of Sequence with GenBank	53
3.8.4	Protein Structure Prediction with Swiss Model and SPDBV	53
3.9	Example Illustrating Variable Information and Redundancy of a Primary Genomic Database and Comparison to Some Other Information in the Book	54
	References	55
4	Pairwise Alignment, Multiple Alignment, and BLAST	59
	Henrik Christensen and John Elmerdahl Olsen	
4.1	The Pairwise Alignment Problem	60
4.1.1	Global or Local Pairwise Alignments?	61
4.1.2	Substitution Matrices	62
4.1.3	Gaps	65
4.1.4	Dynamic Programming	66
4.2	Multiple Alignment	74
4.2.1	Clustal	74
4.2.2	Other Multiple Alignment Programs	78

4.3	BLAST.....	80
4.3.1	NCBI BLAST	81
4.3.2	Ortholog Detection.	83
4.3.3	BLAST2 sequences	83
4.3.4	Statistics.....	83
4.3.5	Variants of BLAST	84
4.4	Activities	86
4.4.1	Pairwise Alignment	86
4.4.2	Multiple Alignment with ClustalX	86
4.4.3	BLAST.....	87
	References.....	87
5	Primer Design	89
	Henrik Christensen and John Elmerdahl Olsen	
5.1	Background for Oligonucleotide Design.....	90
5.1.1	Practical Approach to Oligonucleotide Design Whether of Exploratory Nature or for Diagnostic Purpose	90
5.2	General Rules for Design of Oligonucleotides	93
5.2.1	Lengths of PCR Primers and Products.....	95
5.2.2	Lengths of Oligonucleotide Hybridization Probes	95
5.3	Sequence Comparison	96
5.3.1	String Comparison by Score	96
5.3.2	Nearest Neighbor Comparisons of Duplex Stability	96
5.3.3	Design of Primers for PCR and “Kwok’s Rules”	97
5.3.4	Design of Probes for Hybridization.....	98
5.4	<i>T_m</i> Calculations	99
5.4.1	Estimation of <i>T_m</i> by Formula	100
5.4.2	Formamide Considerations	100
5.4.3	Estimation of <i>T_m</i> by Nearest Neighbor Prediction	100
5.5	Special Applications.....	101
5.5.1	Exploratory Applications.....	101
5.5.2	Diagnostic Applications.....	103
5.6	Data Formats	104
5.7	Programs	105
5.8	Activities	105
5.8.1	Exploratory Primers with Primer3 for Recognition of Single DNA Sequences	105
5.8.2	Diagnostic Primers with PrimerBLAST	105
	Further Reading	107

6	Introduction to Phylogenetic Analysis of Molecular Sequence Data	111
	Henrico Christensen and John Elmer Dahl Olsen	
6.1	Background	112
6.2	Understanding the Phylogenetic Tree	113
6.3	Assumptions About Data in Order to Perform Phylogenetic Analysis	117
6.4	Phylogenetic Model Parameters.	118
6.4.1	The Tree Structure	118
6.4.2	Substitution Matrix and Evolutionary Models.	118
6.4.3	Weighting of Characters	119
6.5	Phylogenetic Methods	119
6.5.1	Maximum Parsimony.	120
6.5.2	Distance Matrix/Neighbor Joining.	121
6.5.3	Maximum Likelihood	121
6.5.4	Bayesian (Mr. Bayes) Inference of Phylogeny	123
6.6	Comparison of Phylogenetic Methods.	123
6.6.1	Bootstrap	124
6.7	Whole Genome Phylogeny	125
6.7.1	Core Phylogeny	125
6.7.2	Genome Distance Phylogeny.	126
6.8	Data Formats	126
6.9	Phylogenetic Program Packages	127
6.10	Activities	128
6.10.1	Neighbor Joining Phylogeny	128
	Further Reading	129
7	Sequence-Based Classification and Identification	131
	Henrik Christensen and John Elmerdahl Olsen	
7.1	Introduction	132
7.2	Classification of Prokaryotes	134
7.2.1	Classification Based on 16S rRNA Gene Sequence Comparison	134
7.2.2	From DNA–DNA Hybridization of Total DNA to WGS for Classification	135
7.2.3	Classification Based on Core Genome Analysis and Multilocus Sequence Analysis (MLSA)	138
7.3	Classification of the Taxonomic Hierarchy	139
7.3.1	Classification of Species	139
7.3.2	Classification of Genera.	139
7.3.3	Classification of Families, Orders, Classes, and Phyla	139
7.4	Rules for the Naming of a New Prokaryote.	140
7.4.1	Bacterial Species Names Are Linked to the Type Strain	141
7.5	The Benefits of Sequence-Based Identification.	142
7.5.1	16S rRNA Sequence-Based Identification, Step-by-Step	142

7.5.2	16S rRNA-Based Identification Without Culture	143
7.5.3	Sequence-Based Identification of Fungi	143
7.5.4	Sequence-Based Identification of Protists	143
7.6	Strain Identification by WGS Analysis	143
7.6.1	SNP	144
7.6.2	OGRI for Strain Level Conformation and Identification	145
7.6.3	OGRI for Fungi	146
7.6.4	Identification of <i>E. coli</i> K12	146
7.7	Activity	146
7.7.1	16S rRNA Gene Sequence-Based Identification	146
	References	147

8 16S rRNA Amplicon Sequencing 153

Henrik Christensen, Jasmine Andersson, Steffen Lyng Jørgensen, and
Josef Korbinian Vogt

8.1	Use of 16S rRNA Amplicon Sequencing, Generation of Data, and Bioinformatics Pipelines	154
8.1.1	Use of 16S rRNA Amplicon Sequencing and Generation of Raw Data	154
8.1.2	Bioinformatical Pipelines to Analyze Data	157
8.2	Data Analysis	158
8.2.1	Quality Trim by Sequence	158
8.2.2	Pairing of Reads	159
8.3	Removal of Chimeras and DNA from Other Domains or Life	159
8.4	Grouping of Reads into OTUs	159
8.5	Alignment of OTUs and Association of OTUs with Taxonomic Units	159
8.6	α -(Within Group) and β -Diversity (Between Groups) Comparison	160
8.6.1	Rarefaction Analysis	161
8.7	Taxonomic Comparisons	163
8.7.1	Generation and Interpretation of Heatmaps and Boxplots	163
8.8	Principal Coordinates Analysis (PCoA)	166
8.9	Prediction of Function	167
8.10	Activity QIIME2	167
8.10.1	Installation	168
8.10.2	Running QIIME2	168
8.10.3	Download Data	169
8.10.4	Change Data Format to QIIME2 Artifacts	170
8.10.5	Demultiplexing Sequences	170
8.10.6	Denoising	170
8.10.7	Visualization Summaries of the Data	171
8.10.8	Phylogenetic Tree	172
8.10.9	α - and β -Diversity Analyses	173

8.10.10	Alpha Rarefaction Plotting	173
8.10.11	Taxonomic Analysis	174
8.10.12	Exporting Data	177
8.10.13	Filtering Data	178
8.10.14	Good to Know When Working with QIIME2	178
	References	179
9	Full Shotgun DNA Metagenomics	183
	Henrik Christensen and John Elmerdahl Olsen	
9.1	Background	184
9.2	Sequencing Strategies and Data Types	186
9.3	Analysis of Full DNA Shotgun Sequence Data	188
9.4	MG-RAST	190
9.5	Upload and Analyze on MG-RAST	193
9.6	Activities	197
9.6.1	Full DNA Metagenome–Shotgun DNA Metagenomics	197
	References	198
10	Transcriptomics	201
	Rikke Heidemann Olsen and Henrik Christensen	
10.1	Introduction to Transcriptomics	202
10.2	Experimental Design	204
10.3	Preparing a RNA-seq Library	205
10.3.1	Step 1: Isolation of RNA	205
10.3.2	Step 2: Depletion or Removal of rRNA	207
10.3.3	Step 3: Conversion of RNA into Complementary DNA (cDNA)	207
10.3.4	Step 4: Addition of Sequencing Adaptors	207
10.3.5	Step 5: PCR Amplification (Enrichment)	207
10.3.6	Step 6: Quality Control of the Library	208
10.4	Sequencing	208
10.5	Data Management (Sequence Reads)	208
10.5.1	Raw Data	209
10.5.2	Alignments of Sequence Reads	209
10.5.3	Normalization of Data	210
10.6	Differential Gene Expression	210
10.7	Conclusion	212
	References	213
11	Sequenced-Based Typing and Prediction of Function	215
	Henrik Christensen and John Elmerdahl Olsen	
11.1	Background of Prokaryotic Populations and Population Genetics	216
11.1.1	Mutation	218
11.1.2	Selection	220

11.1.3	Genetic Drift.....	221
11.1.4	Migration	221
11.1.5	The Biological Consequences of Population Genetics of Prokaryotes.....	221
11.2	Multilocus Sequence Typing (MLST)	223
11.2.1	MLST	223
11.2.2	Multilocus Sequence Analysis	223
11.3	Whole Genome-Based Typing.....	223
11.3.1	Whole Genomic Multilocus Sequence Typing (wgMLST)	224
11.3.2	Single Nucleotide Polymorphisms (SNP)	225
11.3.3	Typing of Virulence, Antibiotic Resistance, and Serotype Based on the Whole Genomic Sequence	226
11.4	Organisms Specific Platforms for Whole Genome Sequence-Based Typing	228
11.5	Association of Genetic Traits with Metadata and Phylogeny.....	228
11.6	Activities	229
11.6.1	MLST Typing of <i>Pasteurella multocida</i>	229
11.6.2	Graphics	229
	Further Reading	231
	Appendix	235
	Index	239