
Contents

Preface, xv

Authors, xix

Contributor, xxi

CHAPTER 1 ■ DNA Sequencing Strategies 1

1.1 THE EVOLUTION OF DNA SEQUENCING TECHNOLOGIES	1
1.2 DNA SEQUENCE ASSEMBLY STRATEGIES	2
1.3 NEXT-GENERATION SEQUENCING	5
1.4 SEQUENCING BIAS AND ERROR RATES	6
REFERENCES	7

CHAPTER 2 ■ Coding Sequence Prediction 9

2.1 INTRODUCTION	9
2.2 MAPPING MESSENGER RNA (mRNA)	9
2.3 STATISTICAL MODELS	13
2.3.1 5' Untranslated Region	14
2.3.2 -35 Signal	14
2.3.3 B Recognition Element	14
2.3.4 TATA Box	15
2.3.5 Ribosomal Binding Site	15
2.3.6 Start Codon	16
2.3.7 Protein Coding Sequence	16
2.3.8 Donor Splice Site	17

2.3.9	Intron Sequence	17
2.3.10	Acceptor Splice Site	18
2.3.11	Stop Codon	18
2.3.12	3' Untranslated Region	18
2.3.13	Terminator	18
2.4	CROSS-SPECIES METHODS	19
2.4.1	Nucleotide Homology	19
2.4.2	Protein Homology	20
2.4.3	Domain Homology	22
2.5	COMBINING GENE PREDICTIONS	23
2.6	SPLICE VARIANTS	24
	REFERENCES	27
CHAPTER 3 ■ Between the Genes		31
3.1	INTRODUCTION	31
3.2	TRANSCRIPTION FACTORS	31
3.2.1	Transcription Factor Binding Site (TFBS) Motifs	32
3.2.2	TFBS Location	34
3.2.3	TFBS Neighborhood	34
3.2.4	TFBS Conservation	35
3.3	RNA	36
3.3.1	Ribosomal RNA	37
3.3.2	Transfer RNA	37
3.3.3	Small Nucleolar RNA	39
3.3.4	MicroRNA	41
3.3.5	Other Types of RNAs	43
3.4	PSEUDOGENES	44
3.4.1	Transposable Elements	46
3.4.2	DNA Transposons	46
3.4.3	Retrotransposons	47
3.5	OTHER REPEATS	49
	REFERENCES	50

CHAPTER 4 ■ Genome-Associated Data	55
4.1 INTRODUCTION	55
4.2 OPERONS	55
4.3 METAGENOMICS	56
4.3.1 Population Statistics	56
4.3.2 Data Size	58
4.3.3 Phylogenetic Sorting	58
4.3.4 Assembly Quality	59
4.4 INDIVIDUAL GENOMES	60
4.4.1 Epigenetics	60
4.4.1.1 DNA Methylation	60
4.4.1.2 Histone Modification	60
4.4.1.3 Nucleosome Positioning	61
4.4.2 Single Nucleotide Polymorphisms	62
4.4.2.1 Nomenclature	62
4.4.2.2 Effects	62
4.4.3 Insertions and Deletions	63
4.4.3.1 Nomenclature	63
4.4.3.2 Effects	64
4.4.4 Copy Number Variation	64
REFERENCES	65
CHAPTER 5 ■ Characterization of Gene Function through Bioinformatics: The Early Days	69
5.1 OVERVIEW	69
5.2 STAND-ALONE TOOLS AND TOOLS FOR THE EARLY INTERNET	71
5.3 PACKAGES	73
5.3.1 IBI/Pustell	73
5.3.2 PC/GENE	73
5.3.3 GCG	74
5.3.4 From EGCG to EMBOSS	74
5.3.5 The Staden Package	75

5.3.6	GeneSkipper	77
5.3.7	Sequencher	77
5.4	FROM FASTA FILES TO ANNOTATED GENOMES	77
5.4.1	ACeDB	77
5.4.2	One Genome Project, the Beginning of Three Genome Annotation Systems	78
5.5	CONCLUSION	79
	REFERENCES	79
CHAPTER 6 ■ Visualization Techniques and Tools for Genomic Data		83
6.1	INTRODUCTION	83
6.2	VISUALIZATION OF SEQUENCING DATA	84
6.3	VISUALIZATION OF MULTIPLE SEQUENCE ALIGNMENTS	88
6.3.1	Pairwise Alignment Viewers	88
6.3.2	Multiple Alignment Viewers	90
6.4	VISUALIZATION OF HIERARCHICAL STRUCTURES	91
6.4.1	Tree Visualization Styles	92
6.4.2	Tree Visualization Tools	94
6.5	VISUALIZATION OF GENE EXPRESSION DATA	96
6.5.1	Expression Data Visualization Techniques	97
6.5.2	Visualization for Biological Interpretation	99
	REFERENCES	101
CHAPTER 7 ■ Functional Annotation		105
7.1	INTRODUCTION	105
7.2	BIOPHYSICAL AND BIOCHEMICAL FEATURE PREDICTION	105
7.2.1	Physical Chemistry Features	105
7.2.2	Sequence Motif Prediction	106
7.2.2.1	<i>Protein Modification</i>	106
7.2.2.2	<i>Protein Localization</i>	107

7.3	PROTEIN DOMAINS	108
7.4	SIMILARITY SEARCHES	112
7.4.1	Paralogs	112
7.4.2	Orthologs	113
7.4.3	Xenologs	115
7.4.4	Analogs	116
7.5	PAIRWISE ALIGNMENT METHODS	116
7.5.1	Canonical Methods	116
7.5.2	Heuristic Methods	117
7.5.3	Scoring Matrices	118
7.6	CONCLUSION	119
	REFERENCES	119
CHAPTER 8 ■ Automated Annotation Systems		125
8.1	INTRODUCTION	125
8.2	MAGPIE	126
8.2.1	Analysis Management	126
8.2.2	Structural Annotation	128
8.2.3	Functional Annotation	130
8.2.4	User Interface	133
8.3	GENERIC MODEL ORGANISM DATABASE	135
8.3.1	Analysis Management	135
8.3.2	Structural Annotation	135
8.3.3	Functional Annotation	136
8.3.4	User Interface	136
8.4	AGeS	137
8.4.1	Analysis Management	137
8.4.2	Structural Annotation	138
8.4.3	Functional Annotation	138
8.4.4	User Interface	138
8.5	ENSEMBL	139
8.5.1	Analysis Management	140

8.5.2	Structural Annotation	140
8.5.3	Functional Annotation	141
8.6	SUMMARY	142
	REFERENCES	143

CHAPTER 9 ■ Dynamic Annotation Systems: End-User-Driven Annotation and Visualization 145

9.1	INTRODUCTION	145
9.2	WEB-BASED GENOME ANNOTATION BROWSERS	146
9.2.1	University of California, Santa Cruz (UCSC) Genome Browser	147
9.2.2	Ensembl Genome Browser	149
9.2.3	NCBI Map Viewer	151
9.2.4	Generic Genome Browser	154
9.3	STAND-ALONE GENOME ANNOTATION BROWSERS	155
9.3.1	Bluejay: An XML-Based Genome Visualization and Data Integration System	155
9.3.2	NCBI Genome Workbench	158
9.3.3	Integrated Genome Browser	160
9.3.4	Apollo	161
9.4	COMPARATIVE VISUALIZATION OF GENOMES	162
9.4.1	Dot Plots	162
9.4.2	Linear Representation	163
9.4.3	Circular Representation	165
	REFERENCES	169

CHAPTER 10 ■ Web-Based Workflows 173

10.1	INTRODUCTION	173
10.2	PRINCIPLES OF WEB-BASED WORKFLOWS	173
10.2.1	Motivation	173
10.2.2	Early Workflow Environments	174
10.3	GALAXY	175
10.3.1	Interactive Analysis	175

10.3.2	Workflows	177
10.3.3	Component Repository	178
10.4	TAVERNA	179
10.4.1	The Design Interface	180
10.4.2	The Results Interface	181
10.4.3	Workflow Repository	183
10.5	SEAHAWK	183
10.5.1	Demonstration-to-Workflow	183
10.5.2	The Search Widget	186
10.5.3	Data Filters and Labels	189
10.5.4	Taverna Enactment of Seahawk-Generated Workflows	190
10.6	CONCLUSION	192
	REFERENCES	192
CHAPTER 11	■ Analysis Pipelines for Next-Generation Sequencing Data	195
11.1	INTRODUCTION	195
11.2	GENOME SEQUENCE RECONSTRUCTION	196
11.2.1	Alignment to the Reference Genome	197
11.2.2	<i>De Novo</i> Assembly	198
11.3	ANALYSIS PIPELINES: CASE STUDIES	200
11.3.1	16S rRNA Analysis	200
11.3.2	Targeted EST Assembly	202
11.3.3	Gene Prediction	204
11.4	NEXT-GENERATION GENOME BROWSING	205
11.4.1	Integration of Different Types of Genomic Data	205
11.4.2	Decentralization	209
	REFERENCES	210
INDEX, 215		