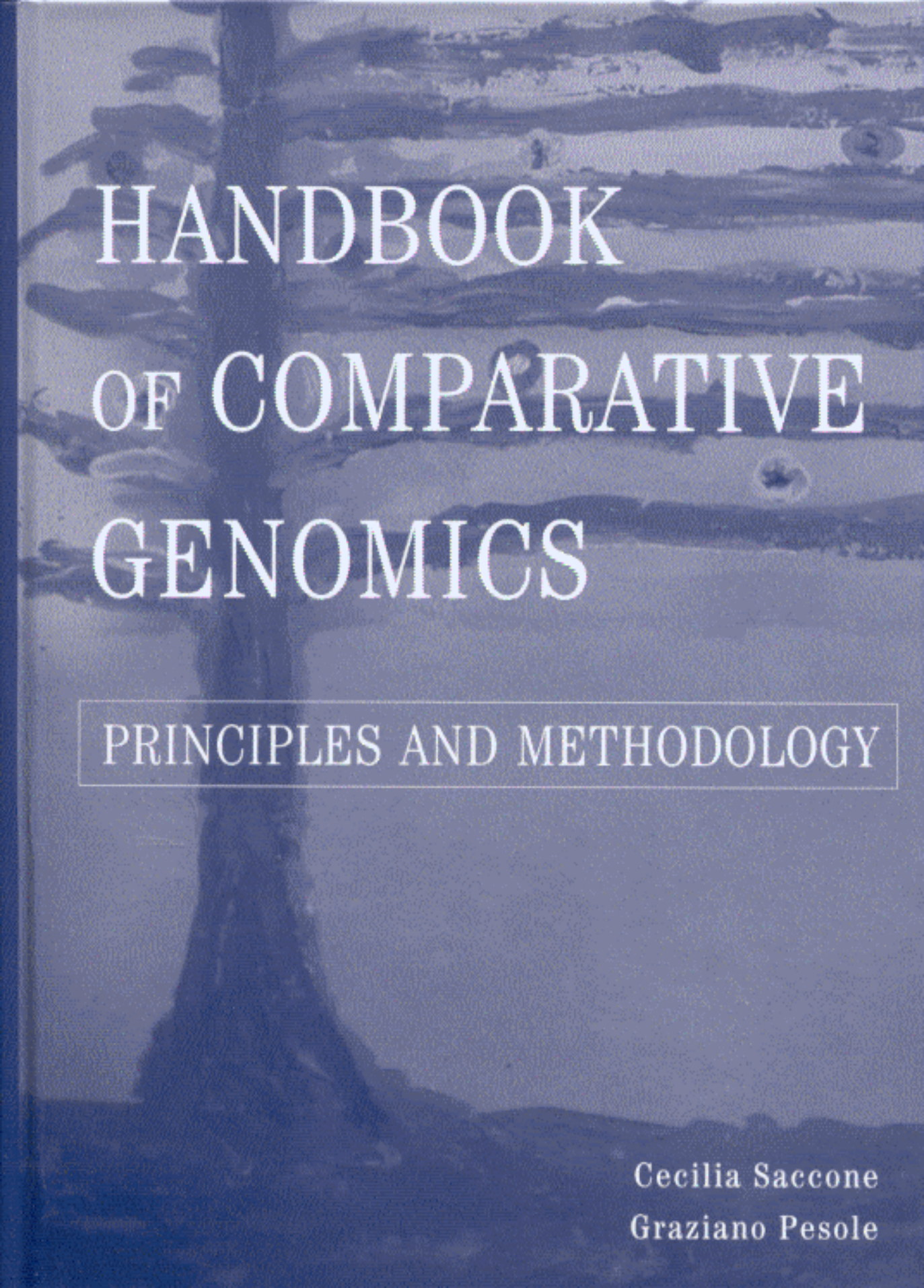




HANDBOOK OF COMPARATIVE GENOMICS

PRINCIPLES AND METHODOLOGY

Cecilia Saccone
Graziano Pesole

CONTENTS

PREFACE	xiii
I GENOME FEATURES	1
1 PROKARYOTES	3
1.1 Introduction / 3	
1.2 Morphology and Classification / 8	
1.3 Genome Shape and Size / 10	
1.4 Gene Content and Organization / 16	
1.5 Base Composition / 20	
1.6 Codon Use / 27	
1.7 Replication and Expression / 29	
2 EUKARYOTES	36
2.1 Introduction / 36	
2.2 Classification and Time Scale / 36	
2.3 Genome Shape and Size / 40	
2.4 Base Composition / 46	
2.5 Replication, Repair, and Recombination / 48	
2.6 Gene Expression / 53	
2.6.1 Transcription and Posttranscriptional Regulation / 53	
2.6.2 Genetic Code and Codon Use / 59	
2.6.3 Translation and Posttranslation Modifications / 61	

2.7	Completely Sequenced Eukaryotic Genomes / 64	
2.7.1	<i>Saccharomyces cerevisiae</i> Genome / 66	
2.7.2	<i>Schizosaccharomyces pombe</i> Genome / 69	
2.7.3	<i>Caenorhabditis elegans</i> Genome / 70	
2.7.4	<i>Drosophila melanogaster</i> Genome / 72	
2.7.5	<i>Arabidopsis thaliana</i> Genome / 74	
2.7.6	<i>Oryza sativa</i> Genome / 77	
2.7.7	<i>Homo sapiens</i> Genome / 78	
3	ORGANELLES	85
3.1	Mitochondria / 85	
3.1.1	General Structure and Function / 85	
3.1.2	DNA and Genetic System / 88	
3.1.3	Genome Features / 97	
3.2	Chloroplasts and Other Plastids / 125	
II	METHODOLOGIES	131
4	MOLECULAR BIOLOGY TECHNIQUES FOR GENOMICS	133
4.1	Genome DNA Sequencing / 133	
4.1.1	DNA-Sequencing Techniques / 133	
4.1.2	The Human Genome Project / 136	
4.2	Analysis of the Transcriptome / 137	
4.2.1	Analysis of Gene Expression / 137	
4.2.2	Expressed Sequence Tags / 138	
4.2.3	Serial Analysis of Gene Expression / 139	
4.2.4	Differential Display / 141	
4.2.5	Representational Difference Analysis / 142	
4.2.6	DNA Microarrays / 143	
4.3	Analysis of the Proteome / 149	
4.3.1	Two-Dimensional Gel Electrophoresis / 151	
4.3.2	Protein Identification / 151	
4.3.3	Study of Protein–DNA and Protein–Protein Interactions / 155	
4.3.4	Proteome Analysis Using Biochips / 157	
5	BIOLOGICAL DATABASES IN THE GENOMIC ERA	159
5.1	Introduction / 159	
5.2	Primary and Specialized Databases / 159	
5.3	Database Structures / 162	
5.4	Linked Databases and Database Interoperability / 163	

5.5	Database Annotation / 166	
5.6	Retrieval Systems / 169	
5.6.1	SRS / 169	
5.6.2	Entrez / 169	
5.6.3	Other Retrieval Systems / 170	
5.7	Nucleotide Databases / 170	
5.8	Protein Databases / 171	
5.9	Other Protein Databases / 171	
5.10	Genomic Databases and Resources / 174	
5.11	Gene Databases and Resources / 179	
5.12	Transcriptome Databases / 180	
5.13	Metabolism Databases / 181	
5.14	Mutation Databases / 182	
5.15	Mitochondrial Databases and Resources / 184	
6	COMPUTATIONAL METHODS FOR THE ANALYSIS OF GENOME SEQUENCE DATA	187
6.1	Introduction / 187	
6.2	Dot-Plot Matrix / 188	
6.3	Sequence Pairwise Alignment / 189	
6.3.1	Needleman–Wunsch Global Alignment Algorithm / 191	
6.3.2	Smith–Waterman Algorithm for the Identification of Common Molecular Subsequences / 193	
6.3.3	Alignment of cDNA and Genomic DNA Sequences / 196	
6.3.4	Genome Alignment / 198	
6.3.5	Cleanup of Sequence Databases from Redundancy / 201	
6.3.6	Measure of the Similarity Degree between Homologous Sequences / 203	
6.4	Database Searching / 208	
6.4.1	FASTA / 209	
6.4.2	BLAST / 211	
6.4.3	BLAST and FASTA Family of Programs / 215	
6.4.4	Filtering Matches to Unwanted Sequences / 216	
6.4.5	Filtering Matches to Repetitive Sequences / 221	
6.4.6	Statistical Significance of Alignment Scores / 223	
6.5	Multiple Alignment / 226	
6.6	Alignment Profiles to Recognize Distantly Related Protein or Protein Modules / 230	
6.7	Methods for Sequence Assembly / 234	
6.7.1	Sequence Cleaning / 236	
6.7.2	Sequence Clustering / 237	
6.7.3	Construction of Alignment Consensus / 238	

6.7.4	Sequence Mapping by Electronic PCR /	240
6.7.5	Sequence Assembly in Genome and EST Projects /	240
6.7.6	Sequence Assembly for Gene Index Construction /	243
6.8	Linguistic Analysis of Biosequences /	245
6.8.1	Biosequences as Markov Chains /	248
6.8.2	Linguistic Complexity of Biosequences /	249
6.8.3	Identification of Repeats in Genomic Sequences /	253
6.8.4	Pattern Searching in Biosequences /	254
6.8.5	Identification of Promoter Regions in Chromosomal Sequences /	261
6.8.6	Pattern Discovery for the Identification of Gene Regulatory Elements and of Protein Motif Models /	262
6.8.7	Gene Prediction /	265
6.8.8	Identification of CpG Islands in Genomic Sequences /	269
6.8.9	Analysis of Codon Use Strategy /	271
6.9	Prediction of RNA Secondary Structures /	273
6.10	Protein Sequence Analysis /	280
6.10.1	Analysis of Protein Primary Sequences /	281
6.10.2	Prediction of Transmembrane Protein Helices /	285
6.10.3	Identification of Protein Signal Peptides and Prediction of Their Subcellular Location /	290
6.10.4	Prediction of Protein Secondary Structure /	293
6.10.5	Prediction of Coiled-Coil and Helix-Turn-Helix Structures /	298
6.10.6	Prediction of Protein Tertiary Structure /	300
6.10.7	Protein Fold Recognition and Classification /	302
6.10.8	Comparative Evolutionary Genomic Tools for Predicting Protein Function /	303
6.11	Evolutionary and Phylogenetic Analysis /	306
6.11.1	Estimating Genetic Distances between Homologous Sequences /	306
6.11.2	Molecular Phylogeny /	309

III COMPARATIVE GENOMICS **325**

7 MOLECULAR EVOLUTION **327**

7.1	Introduction /	327
7.2	Evolution of Genome Size /	328
7.3	Role of Base Composition in Evolution /	332
7.4	Evolution of the Prokaryotic Genome /	337
7.5	From Prokaryotes to Eukaryotes /	338
7.5.1	Origin of the Eukaryotic Cell /	338

7.5.2	Evolution of the Mitochondrial Genome /	340
7.5.3	Origin and Evolution of Plastids /	346
7.6	From Unicellular to Multicellular State /	349
7.7	Evolution of the Nuclear Genome /	351
7.7.1	Introns /	351
7.7.2	Gene and Protein Number /	352
7.7.3	Noncoding Elements /	353
7.7.4	Expansion of Gene Families /	353
7.7.5	Genome Duplication /	360
7.7.6	Conclusion /	361
8	MOLECULAR PHYLOGENY	362
8.1	Introduction /	362
8.2	Molecular Clock /	363
8.3	Similarity Measure: Orthology Versus Paralogy /	364
8.4	Molecular Phylogeny in the Genomics Era /	367
8.5	Interrelationships between Distant Taxa: The Tree of Life /	369
8.6	Phylogeny of Metazoans /	370
8.6.1	Organelar versus Nuclear Taxonomy /	370
8.6.2	Phylogeny of Mammals /	371
	APPENDIX: URLs Cited in the Text	375
	REFERENCES	377
	INDEX	423