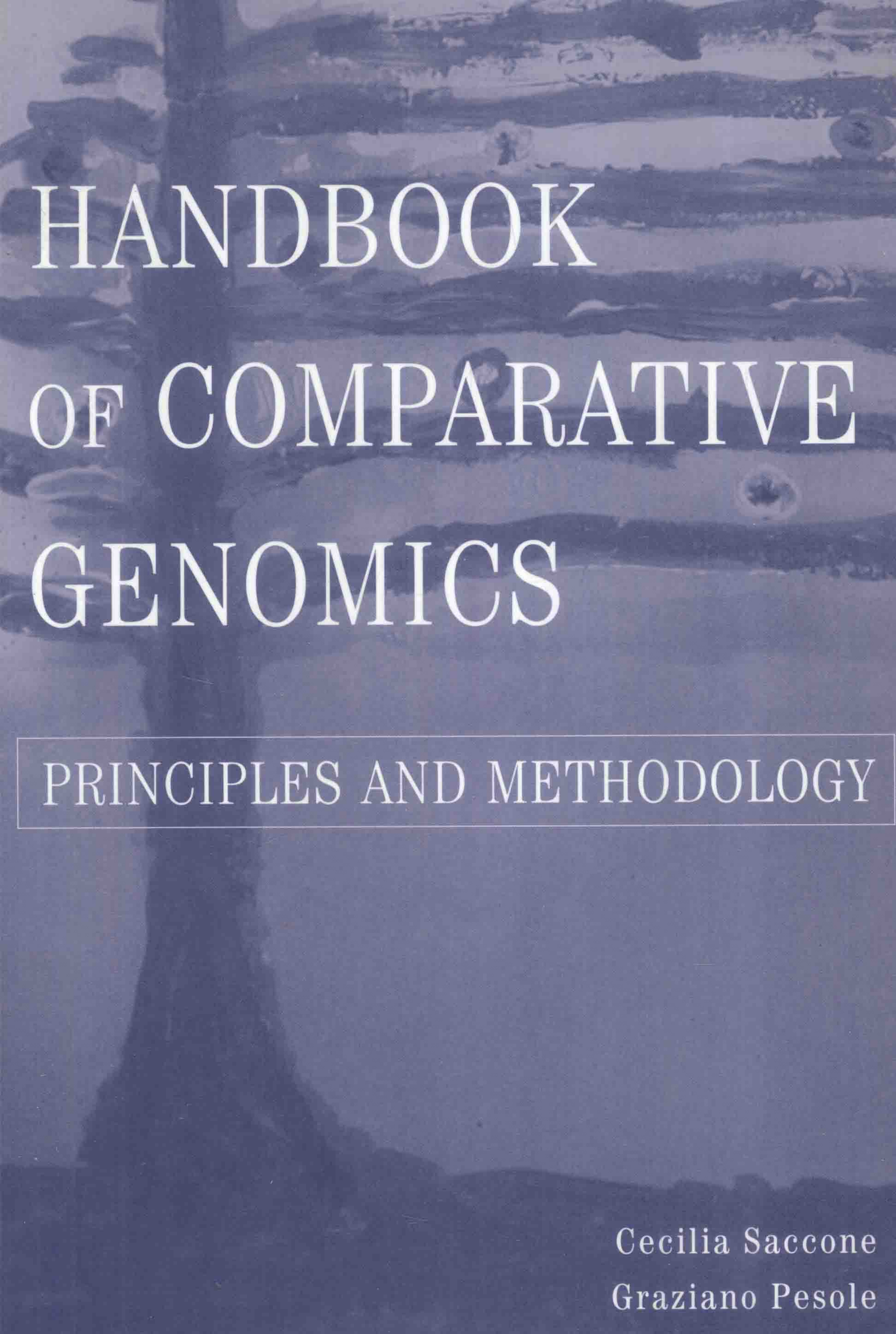

HANDBOOK OF COMPARATIVE GENOMICS

PRINCIPLES AND METHODOLOGY

Cecilia Saccone
Graziano Pesole

HANDBOOK OF COMPARATIVE GENOMICS
Principles and Methodology

CECILIA SACCONE
Department of Biochemistry and Molecular Biology
University of Bari
Italy

GRAZIANO PESOLE
Department of Physiology and General Biochemistry
University of Milan
Italy

WILEY-LISS
A JOHN WILEY & SONS PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Saccone, Cecilia.
Handbook of comparative genomics: Principles and methodology / Cecilia Saccone, Graziano Pesole.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-39128-X (cloth : alk. paper)
1. Genomics—Handbooks, manuals, etc. 2. Evolutionary genetics—Handbooks, manuals, etc. I. Pesole, Graziano. II. Title.
[DNLM: 1. Genomics. 2. Computational Biology. 3. Evolution, Molecular. 4. Sequence Analysis, DNA. QU 58.5 S119h 2003]
QH447.S23 2003
572.8—dc21 2002011158

Printed in the United States of America

10 9 8 7 6 5 4 3 2

HANDBOOK OF COMPARATIVE GENOMICS

To
Ernesto Quagliariello

PREFACE

No sooner had the genomics era begun when a post-genomics era was declared. Since the beginning of large-scale sequencing, there has been a need for new approaches and tools for the study of living matter at the molecular level, and particularly of new hypotheses for progress in biology. There are tremendous expectations about future benefits of sequencing complete genomes; realistically, however, much still remains to be done. Indeed, the exploitation of genome information is still in its infancy and new methodologies may be required to make the most of them.

Some fundamental questions are: To what extent has large-scale sequencing increased our knowledge of the properties and functions of species and, in general, of the structure and function of the genome? What value-adding notions have emerged since sequencing? How has our knowledge of the molecular basis of biological processes improved, if at all? What is our capability to perform comparative genomic studies? What general rules can we derive?

It is well known that advances in comparative biology are based largely on the concepts of analogy and homology. For this reason, comparative analyses of complete genomes rather than single genes, gene families, or specific genome regions must be performed. The sequencing of complete genomes from prokaryotes, eukaryotes, and organelles has aided research on the structure and evolution of the genome as a unit, as opposed to the previous focus on its components. Hence, the new discipline of evolutionary genomics is emerging and will make a revolutionary impact in terms of helping to disclose the linear structure of nucleic acids and proteins, their three-dimensional folding, cytogenetics, gene expression, and regulatory pathways.

We are persuaded that comparative evolutionary genomics is the key to unraveling the hidden messages of living matter. Our use of *genomics* includes its expression and the regulatory mechanisms that are at the basis of all biological processes. In other words, genomics also includes transcriptomics, proteomics, and other "omics," concepts that interconnect and overlap. "Omics" research needs broad databases and dynamic technologies capable of facilitating collaborative efforts across many disciplines, such as biology, chemistry, informatics, mathematics, and physics.

The ambitious goal of this book is to offer a tool for trained biologists who want to tackle the new genomic dimension of modern biology. It will also be useful for technology developers, managers, industries, and funding agencies—in essence anyone interested in the exciting new applications in this new dimension. We are aware that the book is a personal vision and because of its extent and complexity, cannot cover all the literature. Nonetheless, we have tried to highlight milestones, emerging principles, key methods used, and the most urgent needs of this new field. Beginning with a description of complete genomes sequenced from living organisms, this handbook pinpoints new concepts emerging from available data. At the same time we describe the leading methods used to study complete genomes and their evolution.

In summary, we have written this book as a guide for students and researchers not necessarily specialized in genomics. It is written for anyone who approaches, with theoretical and practical aims, this intriguing new chapter of biology—perhaps already a new discipline. We have organized the book in three parts.

- In Part I we describe the state-of-the-art in molecular knowledge of the main biological processes achieved with modern biotechnological tools. We introduce recent insights brought about by genome sequencing. In particular, we summarize the major features of the genomes completely sequenced in prokaryotes, eukaryotes, and organelles.
- In Part II we illustrate the most recent methodologies used in genomics. We describe the available experimental and bioinformatics tools with particular emphasis on molecular biology techniques, biological databases, and computational methods for the analysis of sequence data.
- Part III contains data derived from comparative studies. We discuss fundamental, cutting-edge topics such as the evolution of genome size, base compositional constraints, and the structure and origin of organisms at the molecular level. We conclude by addressing recent advances in molecular phylogenetics.

ACKNOWLEDGMENTS

We thank our dearly missed friend and colleague Giuliano Preparata, a distinguished theoretical physicist who introduced us to the multidisciplinary approach in the study of molecular evolution. Warm thanks also to several colleagues and friends who offered advice and suggestions on several topics covered in the book; they include Marcella Attimonelli, Giorgio Bernardi, Rita Casadio, Nicla Cataldo, Victor De Lorenzo, Annamaria D'Erchia, Ilenia D'Errico, Carmela Gissi, Alessandro Minelli, Aurelio Reyes, Teresa M. R. Regina, Elisabetta Sbisà, and Apollonia Tullo. Finally, special thanks to Alessandra Larizza, for recovering the bibliography and organizing the material, and to Marilina Lonigro, for assistance in English.

CONTENTS

PART I

GENOME FEATURES

CHAPTER 1

PROKARYOTES

1.1 INTRODUCTION

While this book is being written, complete sequences of bacterial genomes are being produced at a rate of about two genomes per month, and the National Center for Biotechnology Information (NCBI) Web site (see the URL in Table 5.1) reports about 60 completely sequenced prokaryotic genomes. Data reported in this chapter refer to the status of completely sequenced genomes, summarized in Table 1.1. Obviously, by the time you read this book, many more will have been sequenced and perhaps some of the aspects dealt with could be viewed differently, although we do not expect dramatic changes in our knowledge unless technology speeds its pace considerably.

Table 1.1 reports the prokaryotic genomes completely sequenced up to now and includes such features as species name, EMBL data library accession number, size, shape, presence of extrachromosomal elements, and bibliographic references. From a look at this list, one can gain an appreciation of the diverse reasons for promoting the sequencing of one species rather than another. Bacterial species are sequenced according to their research interest in basic or applied science: their importance for phylogenetic investigations, to shed light into the metabolic machinery (mainly Archaea) as well as for their importance as human and/or animal pathogens, and for their role as a source of industrial enzymes. In other words, priority has been given to species already well known or species presenting attractive opportunities in applied fields; thus from a phylogenetic point of view, the choice turns out to be very random.

We know we are at the infancy of the genomic era; despite the fact that completely sequenced organisms are still tiny in number, they have already turned out to be full of surprises. In this chapter we summarize the principal sequencing achievements that have improved our knowledge of the prokaryotic genomes and have contributed to outlining methods and approaches to be used in such studies.

TABLE 1.1. Prokaryotic Genomes Completely Sequenced

Species	Main Chromosome		Extrachromosomal Elements		References
	Accession Number	Size (bp)	Accession Number	Size (bp)	
			Archaea		
Aeropyrum pernix	BA000002	1,669,695			Kawarabayasi, Hino et al. (1999)
Archaeoglobus fulgidus	AE000782	2,178,400			Klenk, Clayton et al. (1997)
Halobacterium sp. NRC-1 (3 chromosomes)	AE004437	2,014,239	AF016485	191,346	Ng, Kennedy et al. (2000)
	AE004438	365,425	AE004438	365,425	
	AF016485	191,346			
Methanobacterium thermoautotrophicum	AE000666	1,751,377			Smith, Doucette-Stamm et al. (1997)
Methanococcus jannaschii	L77117	1,664,970	L77118	58,407	Bult, White et al. (1996)
			L77119	16,550	
Methanococcus kandleri AV19	AE009439	1,694,969			Slesarev, Mezhevaya et al. (2002)
Methanosarcina acetivorans str. C2A	AE010299	5,751,492			Galagan, Nusbaum et al. (2002)
Methanosarcina mazei Goe1	AE008384	4,096,345			Deppenmeier, Johann et al. (2002)
Pyrobaculum aerophilum	AE009441	2,222,430			Fitz-Gibbon, Ladner et al. (2002)
Pyrococcus abyssi	AL096836	1,765,118			Lecompte, Ripp et al. (2001)
Pyrococcus furiosus DSM 3638	AE009950	1,908,256			Robb, Maeder et al. (2001)
Pyrococcus horikoshii	AP000001–AP000007	1,738,505			Kawarabayasi, Sawada et al. (1998)
Sulfolobus solfataricus	AE006641	2,992,245			She, Singh et al. (2001)
Sulfolobus tokodaii	BA000023	2,694,765	AJ010405	41,229	Kawarabayasi, Hino et al. (2001)
Thermoplasma acidophilum	AL445063–AL445067	1,564,905			Ruepp, Graml et al. (2000)
Thermoplasma volcanium	AP000991–AP000996	1,584,799			Kawashima, Amano et al. (2000)
			Bacteria		
Agrobacterium tumefaciens str.C58(Cereon)	AE007869	2,841,581			Goodner, Hinkle et al. (2001)
Agrobacterium tumefaciens str.C58(U.Washington)	AE008688	2,841,490	AE008687	542,780	Wood, Setubal et al. (2001)
			AE008690	214,234	

Aquifex aeolicus	AE000657	1,551,335	AE000667	39,456	Deckert, Warren et al. (1998)
Bacillus halodurans	BA000004	4,202,353			Takami, Nakasone et al. (2000)
Bacillus subtilis	AL009126	4,214,814			Kunst, Ogasawara et al. (1997)
Borrelia burgdorferi[a]	AE000783	910,725	AE000791	9,386	Fraser, Casjens et al. (1997);
			AE000792	26,498	Casjens, Palmer et al. (2000)
			AE001575	30,750	
			AE001576	30,223	
			AE001577	30,299	
			AE001578	29,838	
			AE001579	30,800	
			AE001580	30,885	
			AE001581	30,651	
			AE001583[a]	5,228	
			AE000793[a]	16,823	
			AE001582[a]	18,753	
			AE000785[a]	24,177	
			AE000794[a]	26,921	
			AE000786[a]	29,766	
			AE000784[a]	28,601	
			AE000789[a]	27,323	
			AE000788[a]	36,849	
			AE000787[a]	38,829	
			AE000790[a]	53,561	
			AE001584[a]	52,971	
Brucella melitensis	AE008917	2,117,144			DelVecchio, Kapatral et al. (2002)
Buchnera sp. APS	AP000398	640,681	AP001070	7,258	Shigenobu, Watanabe et al. (2000)
			AP001071	7,786	
Campylobacter jejuni	AL111168	1,641,481			Parkhill, Wren et al. (2000)
Caulobacter crescentus	AE005673	4,016,947			Nierman, Feldblyum et al. (2001)
Chlamydia pneumoniae AR39	AE002161	1,229,853		4,524	Read, Brunham et al. (2000)
Chlamydia pneumoniae CWL029	AE001363	1,230,230			Kalman, Mitchell et al. (1999)
Chlamydia trachomatis MoPn	AE002160	1,069,412	AE002162	7,501	Read, Brunham et al. (2000)
Chlamydia trachomatis serovar D	AE001273	1,042,519			Stephens, Kalman et al. (1998)
Chlamydophila pneumoniae J138	BA000008	1,226,565			Shirai, Hirakawa et al. (2000)
Clostridium acetobutylicum	AE001437	3,940,880	NC_001988	192,000	Nolling, Breton et al. (2001)

(*Continued*)