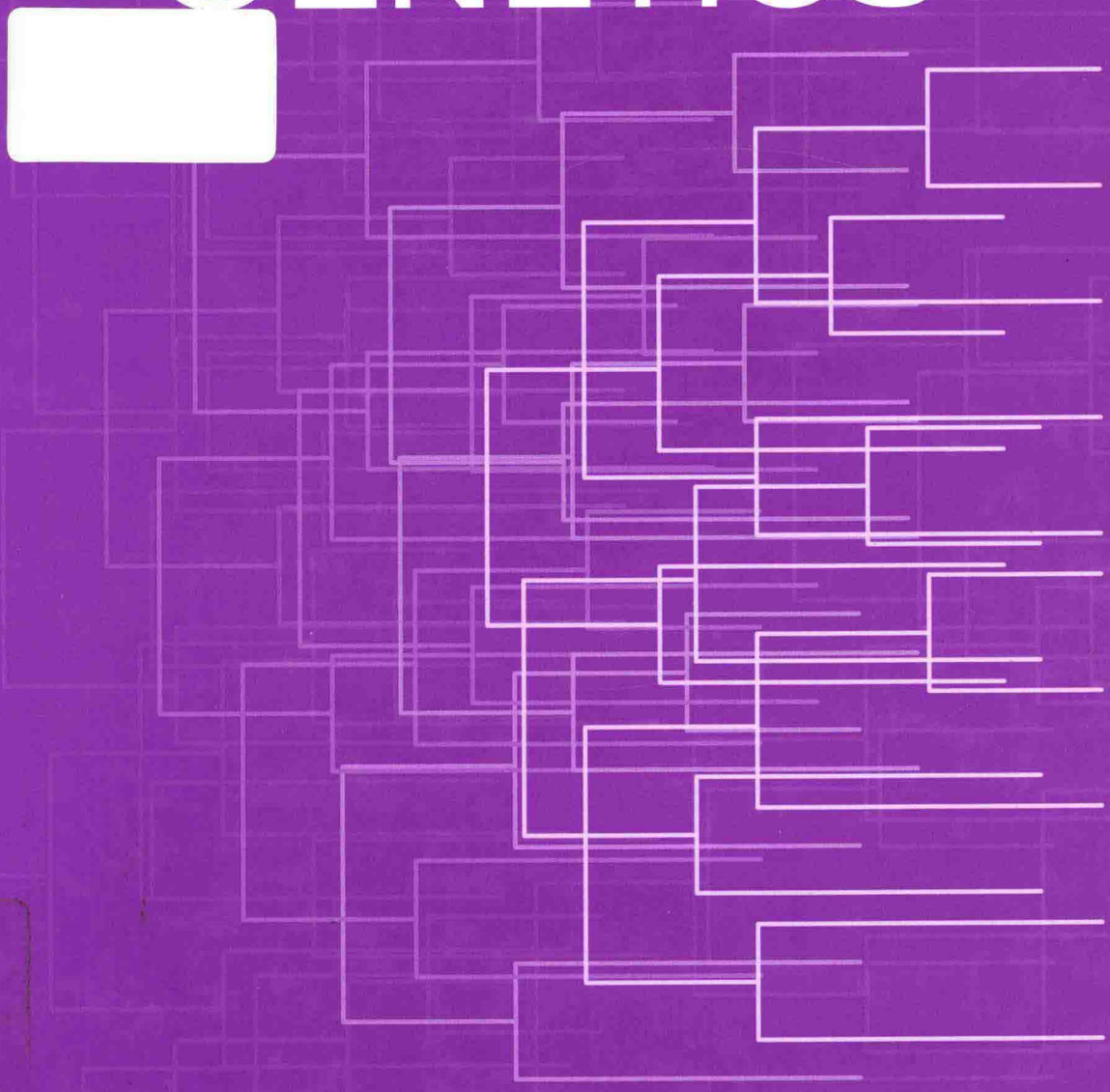
Microbial Population Genetics



Jianping Xu

Caister Academic Press

Caister Academic Press
Norfolk, UK

www.caister.com

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-904455-59-2

Cover image adapted from Fig. 3.2.

Printed and bound in Great Britain by Cromwell Press Group, Trowbridge, Wiltshire.

Microbial Population Genetics

Edited by

Jianping Xu

Department of Biology
McMaster University
Hamilton, ON
Canada

Caister Academic Press

Contributors

Fernando Gonzalez Candelas
Instituto Cavanilles de Biodiversidad y Biologia Evolutiva
University of Valencia
Valencia
Spain

fernando.gonzalez@uv.es

Yujun Cui
Beijing Institute of Microbiology and Epidemiology
Beijing
China

cuiyujun.lam@gmail.com

Bertrand D. Eardly
Department of Biology
Eberly College of Science
Pennsylvania State University
Reading, PA
USA

bde1@psu.edu

Beile Gao
Department of Biochemistry
McMaster University
Hamilton, ON
Canada

gaob@mcmaster.ca

G. Brian Golding
Department of Biology
McMaster University
Hamilton, ON
Canada

golding@mcmaster.ca

Radhey S. Gupta
Department of Biochemistry
McMaster University
Hamilton, ON
Canada

gupta@mcmaster.ca

Weilong Hao
Department of Biology
Indiana University
Bloomington, IN
USA

haow@indiana.edu

Deirdre A. Joy
Parasitology and International Programs Branch
Division of Microbiology and Infectious Diseases
NIAID/NIH/DHHS
Bethesda, MD
USA

djoy@mail.nih.gov

Yanjun Li
Beijing Institute of Microbiology and Epidemiology
Beijing
China

navylyj@163.com

Kui Lin
College of Life Sciences
Beijing Normal University
Beijing
China

linkui@bnu.edu.cn

Yingqin Luo
College of Life Sciences
Beijing Normal University
Beijing
China

yingqin.luo@asu.edu

Scott R. Miller
Division of Biological Sciences
The University of Montana
Missoula, MT
USA

scott.miller@mso.umt.edu

Thomas G. Mitchell
Department of Molecular Genetics and Microbiology
Duke University Medical Center
Durham, NC
USA

tom.mitchell@duke.edu

Teresa E. Pawlowska
Department of Plant Pathology and Plant-Microbe Biology
Cornell University
Ithaca, NY
USA

tep8@cornell.edu

Jim Provan
School of Biological Sciences
Queen's University Belfast
Belfast
UK

j.provan@qub.ac.uk

Rafael Sanjuán
Instituto Cavanilles de Biodiversidad y Biologia Evolutiva
University of Valencia
Valencia
Spain

rafael.sanjuan@uv.es

Jianping Xu
Department of Biology
McMaster University
Hamilton, ON
Canada

jpxu@mcmaster.ca

Yanfeng Yan
Beijing Institute of Microbiology and Epidemiology
Beijing
China

yanyf1983@163.com

Ruifu Yang
Beijing Institute of Microbiology and Epidemiology
Beijing
China

ruifuyang@gmail.com

Preface

Population genetics investigates the spatial and temporal patterns of genetic variation among individuals and populations of organisms, including the mechanisms for such patterns. This field has a rich history, started with Charles Darwin in 1859 with his seminal publication *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. His revolutionary idea of Evolution by Natural Selection revealed the fundamental issues of population genetics: the existence of heritable variations among individuals, the contributions of these variations to the differential survival and reproduction among individuals, the accumulation of advantageous variants in a population over time, and the divergence among populations due to the accumulation of different variants. For most of the 19th and the early 20th century, there was very little agreement within the scientific community about the genetic basis of phenotypic variation among individuals in natural populations. However, this period laid some of the most important foundations of biology, including population genetics. The early pioneers in genetics and specifically population genetics include Francis Galton, Karl Pearson, Gregor Mendel, Thomas Morgan, Herman Nilsson-Ehle, Udny Yule, Ronald Fisher, J.B.S. Haldane, Sewall Wright, Sergei Chetverikov, and Theodore Dobzhansky.

Most early population geneticists used mathematical models to examine the factors that might help explain phenotypic differences among individuals and populations. When the principles of inheritance, the Mendelian laws, were rediscovered at the beginning of the 20th century, rapid advancements in population genetics soon followed, leading to the formulation of the New Synthesis of the theory of evolution. However, for the first half of the 20th century, most evolutionary biologists and population geneticists studied plants and animals. Few studies examined microbial populations.

The era of molecular population genetics began in 1966 when Lewontin and Hubby, and Harris, introduced the technique of protein electrophoresis to population genetics. Surprisingly abundant genetic variations were found in virtually all organisms examined, including some microorganisms. However, many of these variants had little phenotypic effect on the survival and reproduction of organisms. These findings have led to the emergence of the neutral theory of molecular evolution. The application of DNA sequencing technology into population genetics since the early 1980s further expanded our understanding of the extent of genetic variation among individuals and within and between populations. While most molecular population genetic studies from the 1960s to the early 1990s were still focused on macroorganisms such as plants and animals, the next phase of technical innovation, the development and application of high-throughput DNA sequencing and microarray technologies in the mid-1990s, were led by studies of microorganisms. These and other developments are now allowing unprecedented access to genetic materials from populations and communities

of micro- and macro-organisms in their natural habitats. Indeed, comparative genomics, phylogenomics, phylodynamics, evolutionary genomics, population genomics, metagenomics, and systems biology are now burgeoning fields of scientific investigations.

The objective of this book is to bring up-to-date research advances in broad areas of microbial population genetics and genomics. It reviews the application of various molecular tools in our understanding of the patterns and the potential mechanisms for genetic variation of microorganisms at a broad range of scales, from those within micro-niches to among populations from different continents. The book introduces both fundamental concepts and recent molecular population genetic tools and data from SNP surveys, whole-genome DNA sequences, and microarray hybridizations etc. It covers broad groups of microorganisms including viruses, bacteria, archaea, fungi, protozoa, and algae.

The 12 chapters in this book contain a wide range of topics. Chapter 1 introduces recent advances in microbial systematics, with a special focus on the unique utility of insertions/deletions in helping to resolve the phylogenetic relationships among bacterial and archaeal phyla and orders. Chapter 2 reviews comparative microbial genomics and the impact of whole genome sequences in helping us understand the large-scale population and evolutionary issues. Chapter 3 summarizes our current understanding of horizontal gene transfer in bacteria, including the patterns, rates, and fate of horizontally transferred genes. Chapter 4 describes the molecular diversity of a deadly group of human bacterial pathogens and discussed how such information could be used for source tracking and rapid identification of human pathogenic microbes in general. Chapters 5 to 10 review the population genetics of large representative groups of microorganisms that include the nitrogen-fixing bacteria (Chapter 5), the photosynthetic cyanobacteria (Chapter 6), the eukaryotic microalgae (Chapter 7), the fungal mutualists (Chapter 8), the human fungal pathogens (Chapter 9), the human malaria parasites (Chapter 10), and human pathogenic viruses (Chapter 11). Chapter 12 presents an overview of the broad field of metagenomics and its wide-ranging impact on our understanding of microbial diversity, function, inter-relationships, and population genetics, from organisms ranging from viruses to bacteria, archaea, and eukaryotic microbes.

Microbial population genetics is a rapidly advancing field of investigation with relevance to many other theoretical and applied areas of scientific investigations. The theoretical issues that many of the chapters touched upon include the origins and evolution of species, of sex and recombination, and of life (Chapters 1, 2, 3, 11, and 12). On the applied side, population genetics lays the foundations for tracking the origin and evolution of antibiotic resistance and deadly infectious pathogens (Chapters 3, 4, 7, 9, 10, and 11). Population genetics is also an essential factor for devising strategies for the conservation and better utilization of beneficial microbes (Chapters 5, 6, 7, 8, and 12).

I want to thank Hugh Griffin at Horizon Scientific Press for his support and patience with this project. I am grateful to all the authors for their contributions. It has been a great privilege and honour to work with this group of scientists.

Jianping (JP) Xu, PhD
McMaster University
Hamilton, Ontario, Canada

Contents

Recent Advances in Understanding Microbial Systematics

1

Radhey S. Gupta and Beile Gao

Abstract

The higher taxonomic groups within Prokaryotes are presently distinguished mainly on the basis of their branching in phylogenetic trees. In most cases, no molecular, biochemical or physiological characteristics are known that are uniquely shared by species from these groups. Analyses of genome sequences are leading to discovery of novel molecular characteristics that are specific for different groups of *Bacteria* and *Archaea* and provide more precise means for identifying and circumscribing these groups of microbes in clear molecular terms and for understanding their evolution. These new approaches and their limited applications for clarifying microbial systematics are described here. Because of their taxa specificities, further studies on these newly discovered molecular characteristics should lead to discovery of novel biochemical and physiological characteristics that are unique to different groups of microbes.

Introduction

An understanding of the evolutionary history of life, which spans a period of more than 3.5 billion years (Ga), constitutes one of the most fascinating problems in life sciences (Schopf, 1978; Woese *et al.*, 1990; Margulis, 1993; Gould, 1994; Gupta, 1998a; Cavalier-Smith, 2002). The seminal work of Charles Darwin (Darwin, 1859) provided evidence that all living organisms shared a common ancestry and it also revealed key insights as to how the evolutionary process works to generate different life forms. Although Darwin's theory of evolution is applicable to all forms of life, his work primarily dealt with the macroscopic life forms, covering a period of < 1.0 Ga. There was almost no information available at that time about prokaryotic organisms, which based upon available evidence were the sole inhabitants of this planet for at least the first 1.5–2.0 Ga history of life (Schopf, 1978; Kasting, 1993). In the past 25–30 years, much has been learnt about the diversity of prokaryotic organisms (Woese, 1987; Olsen and Woese, 1993), but many critical issues remain unresolved (Gupta and Griffiths, 2002; Gupta, 2005a). However, with the dawn of genomic era, the prospects of gaining an understanding of these critical issues regarding the evolutionary relationships among prokaryotic organisms are beginning to emerge (Doolittle, 1999; Gupta, 2002; Ciccarelli *et al.*, 2006). This chapter describes some critical unresolved issues in prokaryotic phylogeny and new approaches that are proving helpful in understanding of microbial systematics and phylogeny.

Evolutionary relationships among prokaryotes: critical issues that need to be understood

The prokaryotic organisms are presently divided into two main domains, *Bacteria* and *Archaea* (Woese *et al.*, 1990; Ludwig and Klenk, 2005). Of these, *Bacteria* constitute the vast majority (>98%) of known prokaryotic organisms; hence, an understanding of the relationship among them constitutes a major part of prokaryotic phylogeny. Our current understanding of the evolutionary relationships among prokaryotes is mainly based on 16S rRNA sequences (Woese *et al.*, 1990;

Ludwig and Klenk, 2005). Based on branching in the 16S rRNA trees, the cultivable *Bacteria* are presently divided into about 24 main groups or phyla. These groups include *Thermotogae, Aquificae,* green non-sulphur bacteria or *Chloroflexi, Deinococcus-Thermus, Cyanobacteria,* low G+C Gram-positive (*Firmicutes*), high G+C Gram-positive (*Actinobacteria*), *spirochaetes,* green sulphur bacteria (*Chlorobi*), *Bacteroidetes, Chlamydiae, Planctomycetes, Proteobacteria, Thermodesulfobacteria, Thermomicrobia, Chrysiogenetes, Deferribacteres, Dictyoglomi, Fusobacteria, Acidobacteria, Fibrobacteres, Nitrospira, Flexistipes* and *Verrucomicrobia* (Ludwig and Klenk, 2005). Some of these phyla, namely *Thermodesulfobacteria, Thermomicrobia, Chrysiogenetes, Deferribacteres, Dictyoglomi, Fusobacteria, Acidobacteria, Fibrobacteres, Nitrospira* and *Flexistipes,* consist of only a few species, whereas other phyla such as *Proteobacteria, Cyanobacteria, Actinobacteria, Firmicutes* and *Bacteroidetes* contain thousands of species accounting for more than 90–95% of all known bacteria (Ludwig and Klenk, 2005).

The division of *Bacteria* into these 24 or so main groups is quite arbitrary and it currently has little evolutionary or taxonomic significance (Stackebrandt, 2006; Ludwig and Klenk, 2005). The main problem in this regard is that currently there are no objective criteria as to what constitute a phylum or other higher taxonomic levels such as Class, Order or Family (Stackebrandt, 2006; Ludwig and Klenk, 2005). The arbitrariness of the present bacterial classification is illustrated by the example of the *Proteobacteria* phylum. Proteobacteria comprise the largest group within prokaryotes accounting for nearly 50% of all cultured bacteria (Maidak *et al.*, 2001; Ludwig and Klenk, 2005; Kersters *et al.*, 2006). Based on their branching in the 16S rRNA trees, they have been further divided into five classes or divisions, named α, β, γ, δ, and ε (Maidak *et al.*, 2001; Ludwig and Klenk, 2005; Kersters *et al.*, 2006). Of these, α-, β- and γ-divisions harbour approximately 12%, 8% and 26% of all cultured bacteria (Maidak *et al.*, 2001). The species from these subgroups can be clearly distinguished from each other and from all other bacteria based on large numbers of molecular characteristics (Gupta, 2000b, 2005b, 2006; Kersters *et al.*, 2006; Ciccarelli *et al.*, 2006; Gupta and Sneath, 2007; Gupta and Mok, 2007; Gao *et al.*, 2009). However, despite their phylogenetic and molecular distinctness, these large groups of *Bacteria* are presently not recognized as distinct phyla, whereas numerous other poorly studied bacteria consisting of only a few species are recognized as separate phyla of bacteria.

It is important to point out that when these main groups of *Bacteria* were first described, only a limited number of sequences were available and these groups could be clearly distinguished based on long internal branches that separated them in the 16S rRNA trees (Woese *et al.*, 1985; Woese, 1987). However, with the enormous increase in the number of sequences, boundaries between these groups have become blurred making it difficult to clearly demarcate these groups in phylogenetic terms (Ludwig and Klenk, 2005). Further, except for their branching pattern in phylogenetic trees, for most bacterial groups, no molecular, biochemical or physiological characteristics are known that are unique to them. Hence, a central issue of fundamental importance to microbiology that remains to be understood and resolved is: 'In what aspects do different main groups of bacteria differ from each other and do species from these groups share any unique molecular, biochemical, structural or physiological characteristics that are distinctive of each group?' Another central issue in bacterial phylogeny is to understand how different main groups within *Bacteria* are related to each other and evolved from a common ancestor. Phylogenetic trees based on rRNA and other gene/protein sequences have not been able to resolve these relationships leading to notion that this important problem is insolvable (Doolittle, 1999; Ludwig and Klenk, 2005).

Based on this brief overview, it should be evident that in order to develop a reliable understanding of microbial systematics and phylogeny it is necessary at first to develop *new well-defined* (*molecular or biochemical*) *criteria* for identifying all of the main groups or divisions within *Bacteria* in a precise and definitive manner. These new criteria or properties should be such that they should enable identification and circumscription of all of the major taxa (at various taxonomic levels) in clear molecular and/or biochemical terms (Gupta and Griffiths, 2002). Further, it is also of central importance to understand how different groups of *Bacteria* are related to each other

and have branched off from a common ancestor (Gupta, 2001).

New molecular markers for systematic and evolutionary studies

The availability of genome sequences from large numbers of microbes in recent years has opened up new windows of opportunities for discovering novel molecular characteristics that are unique for different groups of bacteria and can be used for their identification as well as for biochemical and functional studies (Nelson *et al.*, 2001; Korbel *et al.*, 2005). Comparative genomics provides the primary means for mining information from genomic sequences. These studies are leading to identification of different kinds of molecular markers that are proving of great value for understanding microbial systematics and phylogeny (Gupta, 1998a; Lerat *et al.*, 2005; Gupta and Griffiths, 2006). The ideal markers for such studies should have the following characteristics: '*These markers should be homologous apomorphic characters that evolved only once (synapomorphy) but not by convergence*' (Stackebrandt, 2006). Such markers also should not be affected by factors such as multiple changes at a given site, long-branch attraction effect, differences in evolutionary rates between and among species, lateral gene transfers, etc., which confound the inferences from phylogenetic trees (Delsuc *et al.*, 2005). Our recent work in this area describes two different types of molecular markers or rare genetic changes that generally satisfy these characteristics.

The first of these newly discovered molecular markers consist of conserved insertions and deletions (indels) in gene/protein sequences. The indels that provide useful phylogenetic markers are generally of defined size and they are flanked on both sides by conserved regions to ensure that they are reliable characteristics (Gupta, 1998a, 2000b, 2004, 2005b; Gupta and Griffiths, 2002; Gupta *et al.*, 2003; Gao and Gupta, 2005). Because of the highly specific nature of genetic changes that give rise to a given conserved indel, such changes are less likely to arise independently in different groups or taxa by either convergent or parallel evolution (i.e. homoplasy) (Gupta, 1998a; Rokas and Holland, 2000). Hence, when a conserved signature indel (CSI) of defined size is uniquely found in a phylogenetically defined group(s) of species, the simplest explanation for this observation is that the genetic change responsible for this CSI occurred once in a common ancestor of this group of species and then passed on to various descendents. Because the presence or absence of a given CSI in different species is not affected by factors such as differences in evolutionary rates, CSIs, which are restricted to particular clade(s), have generally provided good phylogenetic markers of common evolutionary descent. In addition, genetic changes leading to CSIs could be introduced at various stages during evolution, it is possible to identify CSIs in gene/protein sequences at different phylogenetic depths corresponding to various high taxonomic groupings (e.g. phylum, order, family or genus). Such CSIs, in turn, can provide well-defined markers for identifying different taxonomic groups of bacteria in molecular terms. Indeed, identified CSIs that are commonly shared by species from a number of different phyla have provided valuable information regarding branching order and interrelationships among different main groups of bacteria (Gupta, 2001, 2003; Gupta and Griffiths, 2002; Griffiths and Gupta, 2004b).

The second kind of molecular markers that have proven very useful for systematic and phylogenetic studies are whole proteins that are uniquely found in particular groups or subgroups of bacteria (Gupta and Lorenzini, 2007; Gupta and Mok, 2007). Comparative analyses of genomic sequences have indicated that such conserved signature proteins (CSPs), which are also referred to as ORFans (i.e. ORFs that have no known homologues), are also present at different phylogenetic depths (Siew and Fischer, 2003; Daubin and Ochman, 2004; Lerat *et al.*, 2005; Dutilh *et al.*, 2008). Recent studies show that many of these CSPs are uniquely present in all species from particular groups (Gao and Gupta, 2007; Gupta and Lorenzini, 2007; Gupta and Mok, 2007); hence it is likely that genes for these proteins evolved once in a common ancestor of these groups and then retained by all of its descendants. Because of their taxa specificity, these CSPs, again provide valuable molecular markers for identifying different groups of species in molecular terms. Further, similar to the CSIs, based

upon species distribution patterns of these CSPs, it is again possible to draw robust phylogenetic inferences regarding interrelationships among various bacterial groups.

Our recent work in this area has led to identification of large numbers of CSIs and CSPs that are distinctive characteristics of various main groups within *Bacteria*. Based on these characteristics all of the main groups within *Bacteria* including α-, β-, γ-, δ and ε-*Proteobacteria* (Gupta, 2000b; Gupta, 2005b; Gupta, 2006; Gupta and Sneath, 2007; Gupta and Mok, 2007; Gao *et al.*, 2009), *Cyanobacteria* (Gupta *et al.*, 2003; Gupta, 2009), *Deinococcus-Thermus* (Griffiths and Gupta, 2004a; Griffiths and Gupta, 2007a), *Chlamydiae–Verrucomicrobia* (Griffiths *et al.*, 2005; Griffiths *et al.*, 2006; Gupta and Griffiths, 2006; Griffiths and Gupta, 2007b), *Fibrobacter--Chlorobi–Bacteroidetes* (Gupta, 2004; Gupta and Lorenzini, 2007), *Actinobacteria* (Gao and Gupta, 2005; Gao *et al.*, 2006), *Aquificales* (Griffiths and Gupta, 2004b; Griffiths and Gupta, 2006) and *Firmicutes* (Gupta and Gao, 2009), as well as *Archaea* (Gao and Gupta, 2007), can now be described and circumscribed in molecular terms. Information for some of the CSIs and CSPs for some of these microbial groups is provided in Table 1.1. These newly discovered molecular markers also provide powerful means for the discovery of novel biochemical and physiological characteristics that are unique and distinguishing characteristics of different groups of bacteria. To illustrate the usefulness of these new approaches for understanding microbial phylogeny and systematics, some of the work that has been done in this regard on *Archaea* and *Actinobacteria* is reviewed here.

Molecular markers for Archaea and its main groups

Archaea are widely regarded as one of the three domains of life (Woese, 1987, 1998; Woese *et al.*, 1990; Doolittle, 1999; Ludwig and Klenk, 2005), although a number of observations indicate this group of prokaryotes exhibits a close relationship to the Gram-positive bacteria (Mayr, 1998; Gupta, 1998a,b, 2000a; Koch, 2003; Skophammer *et al.*, 2007). Archaeal species were earlier believed to inhabit only extreme environments such as extremely hot, extremely saline, or very acidic or alkaline conditions (Woese, 1987; Woese *et al.*, 1990). However, recent studies provide evidence that they are widespread in different environments (Pace, 1997). Presently, very few molecular characteristics are known that are uniquely shared by either all archaea or different main groups within archaea (Woese, 1987; Woese *et al.*, 1990). The phylogenetic analyses of cultivable *Archaea* have led to their division into two major groups or phyla designated as *Crenarchaeota* and *Euryarchaeota* (Woese *et al.*, 1990; Ludwig and Klenk, 2005; Gribaldo and Brochier-Armanet, 2006). The species from both these groups, particularly *Euryarchaeota*, are highly diverse in terms of their metabolism and physiology. Based on their metabolic and physiological characteristics and other unique features, five functionally distinct groups within *Euryarchaeota* are currently recognized: methanogens, sulphate reducers, extreme halophiles, cell wall-less archaea, and extremely thermophilic sulphur-metabolizing archaea (Ludwig and Klenk, 2005; Gribaldo and Brochier-Armanet, 2006). The methanogens form the largest group within the *Euryarchaeota* and they are distinguished from all other prokaryotes by their ability to obtain all or most of their energy via the reduction of CO_2 to methane, the process of methanogenesis. However, methanogens are polyphyletic in different phylogenetic trees (Brochier *et al.*, 2005; Bapteste *et al.*, 2005; Gribaldo and Brochier-Armanet, 2006).

We have carried out comprehensive analyses on all available sequenced archaeal genomes to search for CSPs that are unique to either all archaea or its main subgroups (Gao and Gupta, 2007). These studies have identified >1400 proteins that are distinctive characteristics of *Archaea* and its various subgroups and whose homologues are not found in any other organisms. Six of these identified proteins are unique to all *Archaea*, 11 proteins are specific for *Crenarchaeota* and seven proteins are only found in various *Euryarchaeota* (Gao and Gupta, 2007). Additionally, many other proteins are specific for various subgroups within the *Archaea* (e.g. *Sulfolobales*, *Halobacteriales*, *Thermococci*, *Thermoplasmata*, all methanogenic archaea or particular groups of methanogens). Based upon the species distributions of these proteins, the evolutionary stages where the genes for these proteins have probably evolved are shown in Fig. 1.1. These proteins provide novel

Table 1.1 Summary of molecular markers (CSIs and CSPs) for some of the major bacterial phyla

Phylum	CSPs/CSIs	Reference
Actinobacteria	Phylum-specific: 5CSPs/CSIs (CoxI; CTPS; GluRS; 23S rRNA) Subgroups: *CMN subgroup*: 13CSPs/CSI (CarA) *Micrococcineae*: 8CSPs/CSI (S3)	Gao and Gupta (2005), Gao and Gupta (2006)
Deinococcus-Thermus	Phylum-specific: 65CSPs/CSIs (σ70, RpoC, ThrRS, L1, UvrA, Ffh, SerRS) Subgroups: *Deinococcoci*: 206CSPs	Griffiths and Gupta (2004), Griffiths and Gupta (2007)
Cyanobacteria	Phylum-specific: CSIs (UvrD, SecA, EF-Tu, S1, Pol I, IMPDH, FtsH, GlgC, PSY, σ70) Subgroups: *Cyano Clade A*: CSI (EF-G) *Other Cyano expcept Clade A*: CSIs (Pol I, DnaX, TrpRS, TSB) *Synechococcophycidae*: CSIs (Pol I, RpoB, KgsA, TyrRS, RpoC)	Gupta (2003, 2009), Gupta *et al.* (2003)
Bacteroidetes, Chlorobi and Fibrobacter	Phylum-specific: 1CSPs/CSIs (RpoC; SHMT) Subgroups: *Bacteroidetes*: 27CSPs/CSIs (GyrB; SecA) *Chlorobi*: 51CSPs/CSIs (DnaE; AlaRS) *Bacteroidetes and Chlorobi*: 5CSPs/CSIs (ATPsyn; FtsK; UvrB)	Gupta and Lorenzini (2007), Gupta (2004), Griffiths and Gupta (2001)
Chlamydiae	Phylum-specific: 59CSPs/CSIs (RpoA; EF-Tu; GyrB; EF-P; LysRS; MgtE; MurA; TrmD) Subgroups: *Chlamydiaceae*: 79CSPs *Chlamydophila*: 20CSPs *Chlamydia*: 20CSPs	Griffiths *et al.* (2005, 2006), Gupta and Griffiths (2006)
Aquificae	Phylum-specific: 10CSPs/CSIs (GidA; RpoC; PolA; EF-Tu; SecA)	Griffiths and Gupta (2004, 2006)
α-Proteobacteria	Phylum-specific: 6CSPs/CSIs (Ctag; PurC; SAICAR Synthetase; DnaB; ATP1, etc. >13) Subgroups: α-*Proteobacteria except Rickettsiales*: 10CSPs/CSIs (Cox I; AlaRS; MutS) *Rickettsiales*: 3CSPs/CSIs (XerD integrase; Lap) *Rickettsiaceae*: 4CSPs/CSIs (Mfd; L19; FtsZ; σ70; ExoVII) *Anaplasmataaceae*: 5CSPs/CSIs (RP-314; Tgt) *Rhodobacterales, Caulobacter and Rhizobiales*: CSIs (DnaA; RP-057) *Rhodobacterales and Caulobacter*: CSI (AsnB) *Rhizobiales*: 6CSPs/CSI (TrpRS) *Rhizobiaceae, Brucellaceae and Phyllobacteriaceae*: CSIs (KGD; LytB; Lep A; SucC; GyrA) *Bradyrhizobiaceae*: 62CSPs/CSIs (SerS; LIG1)	Gupta and Mok (2007), Kainth and Gupta (2005), Guputa (2000, 2005)
γ-Proteobacteria	Phylum-specific: 4CSPs/CSI (PurH) Subgroups: *Enterobacteriales,Pasteurellales, Vibrionales, Aeromonadales and Alteromonadales*: 20CSPs/CSIs (RpoB; L16)	Gao and Gupta (2009), Gupta (2000)
ε-Proteobacteria	Phylum-specific: 49CSPs/CSIs (UvrB; PheRS; RecA; FtsH; RpoC) Subgroups: *Wolinella and Helicobacter:* 11CSPs/CSI (RpoB/RpoC) *Campylobacter:* 18CSPs/CSI (RpoC)	Gupta (2000, 2006)

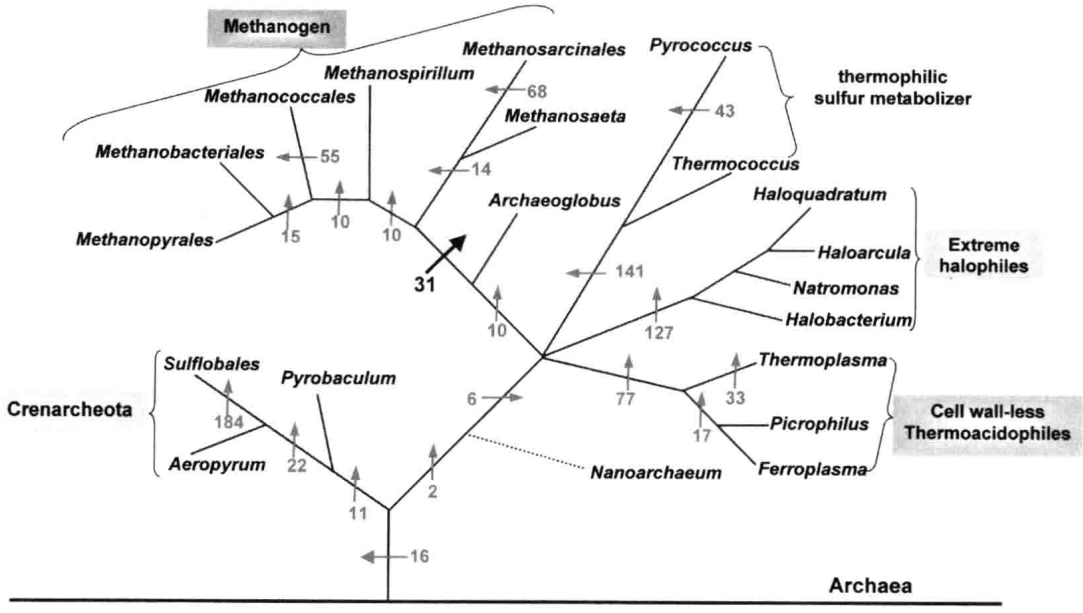


Figure 1.1 A summary diagram showing the species distribution patterns of various *Archaea*-specific proteins. The arrows mark the suggested evolutionary stages where proteins that are uniquely shared by the indicated groups (numbers indicated correspond to the CSPs that are specific for these groups) were introduced. The branching pattern shown here is based upon the species distribution patterns of these proteins and it is unrooted. The dotted line for *Nanoarchaeum* indicates that its placement within *Euryarchaeota* is uncertain. Modified from Gao and Gupta (2007).

molecular markers or signature proteins (CSPs) that are distinctive characteristics of *Archaea* and all of its major subgroups. Most of these proteins are of unknown function (Gao and Gupta, 2007) and further studies on them should lead to discovery of novel biochemical and physiological characteristics that are unique to these groups.

Among the archaea-specific proteins (Gao and Gupta, 2007), of particular significance is the observation that 31 proteins are uniquely present in virtually all methanogens including *Methanopyrus kandleri* (Fig. 1.1). As indicated above, in phylogenetic trees based on 16S rRNA and various proteins sequences, the methanogenic archaea form at least two distinct clusters, with *M. kandleri* branching distinctly from both these clusters (Brochier *et al.*, 2005; Bapteste *et al.*, 2005; Gribaldo and Brochier-Armanet, 2006). The methanogenic archaea in these trees are also interspersed by other groups of non-methanogenic archaea such as *Halobacteriales*, *Archaeoglobus*, *Thermoplasmatales* and *Thermococcales* (Brochier *et al.*, 2005; Bapteste *et al.*, 2005; Gribaldo and Brochier-Armanet, 2006). This has led to important questions concerning the origin of methanogenesis. To account for these results, it has been suggested that methanogenesis evolved once in a common ancestor of different methanogenic archaea, *Halobacteriales*, *Archaeoglobus*, *Thermoplasmatales* and also possibly *Thermococcales*; this was followed by loss of various genes involved in methanogenesis from various other archaeal groups except the methanogens (Brochier *et al.*, 2005; Bapteste *et al.*, 2005; Gribaldo and Brochier-Armanet, 2006). According to this scenario, a methanogenic archaea was the common ancestor of different physiologically and metabolically distinct groups within *Euryarchaeota* and this capability was subsequently independently lost in all other lineages.

In contrast to this proposal, our results showing the presence of 31 proteins that are uniquely found in all methanogens strongly suggest that this group of archaea form a monophyletic lineage exclusive of other archaea (Gao and Gupta, 2007). Importantly, our analyses have also identified 10 additional proteins that are uniquely shared by various methanogens and *Archaeoglobus fulgidus* (Fig. 1.1). In contrast to *A. fulgidus*, no protein was identified that was uniquely shared

by various methanogenic archaea and any of the *Halobacterales* or *Thermoplasmatales*. These observations are highly significant because they strongly suggest that *Archaeoglobus* and all of the methanogens shared a common ancestor exclusive of various other archaea (Gao and Gupta, 2007). In other words, the ancestral lineage that led to the origin of methanogenesis very probably evolved from the *Archaeoglobus* lineage (Fig. 1.1). It is also significant that among the proteins that are uniquely shared by *Archaeoglobus* and methanogens, several are parts of the complexes that are important for nitrogen assimilation and methanogenesis. These results support the view that these characteristics have their origin within the *Archaeoglobus* lineage (Gao and Gupta, 2007).

Molecular markers for the Actinobacteria phylum and its subgroups

Gram-positive bacteria with high G+C DNA content are currently recognized as a distinct phylum, *Actinobacteria*, on the basis of their branching in 16S rRNA trees (Embley and Stackebrandt, 1993; Stackebrandt and Schumann, 2000; Ludwig and Klenk, 2005). This phylum constitutes one of the largest groups among *Bacteria*, comprising 130 genera (Garrity *et al.*, 2005). Actinobacterial species exhibit high levels of diversity in terms of their morphology and physiology. They also play important roles in medicine, industry and environment; some genera such as *Streptomyces* are major antibiotic producers while many others (e.g. *Mycobacterium*, *Corynebacterium*, *Nocardia*, *Leifsonia*, *Tropheryma*, etc.) cause serious human, animal and plant diseases (Embley and Stackebrandt, 1993; Stackebrandt and Schumann, 2000; Ludwig and Klenk, 2005; Ventura *et al.*, 2007). However, except for their distinct branching in phylogenetic trees, until recently no other biochemical or molecular characteristics were known that could distinguish species of this group from all other bacteria (Stackebrandt and Schumann, 2000; Ludwig and Klenk, 2005).

Sequence alignments of various proteins from different bacterial species have identified a number of CSIs that are specific for either all actinobacteria or certain subgroups within them (Gao and Gupta, 2005). An example of a CSI that is specific for the entire *Actinobacteria* phylum is shown in Fig. 1.2. In the partial sequence alignment of cytochrome c oxidase subunit 1 (CoxI) that is presented in this figure, a 2-aa indel is present in a conserved region that is unique to various actinobacterial species, but not seen in any other bacteria (Gao and Gupta, 2005). The shared presence of this 2-aa indel in all actinobacteria strongly indicates that the genetic change leading to this occurred only once in a common ancestor of actinobacteria and passed on to all descendant species. Besides this signature indel, several other CSIs including a 4-aa indel in CTP synthetase, a 5-aa indel in glutamyl-tRNA synthetase (GluRS), and a large insert in the 23S rRNA that are also specific for actinobacteria have been identified (Roller C *et al.*, 1992; Gao and Gupta, 2005). The actinobacteria-specificity of several of these CSIs (viz. CoxI, GluRS and CTP synthase) has been examined by sequencing fragments of these genes from 23 actinobacterial species, covering many different families. All of these gene fragments, except two in GluRS, were found to contain these CSIs providing strong evidence that they are distinctive characteristics of the entire phylum (Gao and Gupta, 2005). In view of their actinobacteria-specificity, these CSIs provide good molecular markers for circumscribing the *Actinobacteria* phylum and distinguishing species of this group from all other bacteria.

In addition to these CSIs, which are specific for all *Actinobacteria*, we have also identified many other CSIs that are specific for certain subgroups within this large phylum. Some of these CSIs also provide information regarding the interrelationships among different subgroups. As an example, in the ribosomal protein S3, we have identified a 5-aa indel that is uniquely shared by various actinobacterial species belonging to the suborders *Micrococcineae* and *Bifidobacterineae* (Fig. 1.3). The species from these two suborders are quite diverse in terms of their phenotypic and physiological characteristics. However, the uniquely shared presence of this indel by various *Micrococcineae* and *Bifidobacterineae* species provide evidence that these two suborders are phylogenetically close and they shared a common ancestor exclusive of other actinobacteria. The CSIs in a number of other proteins, as well as the clustering of these two subgroups in phylogenetic trees based on combined sequences for multiple

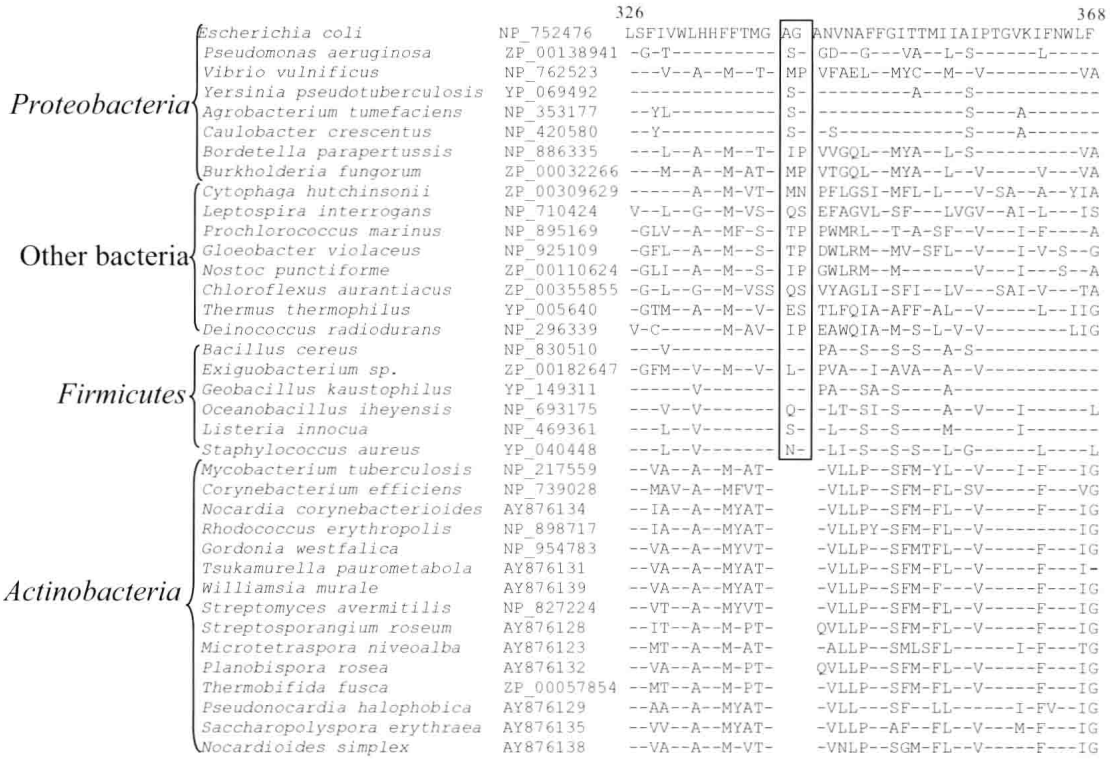


Figure 1.2 Partial alignment of cytochrome c oxidase subunit 1 sequences showing a 2-aa indel (boxed) that is specific for various actinobacterial species. Dashes in all sequence alignments indicate identity with the amino acid on the top line. Modified from Gao and Gupta (2005).

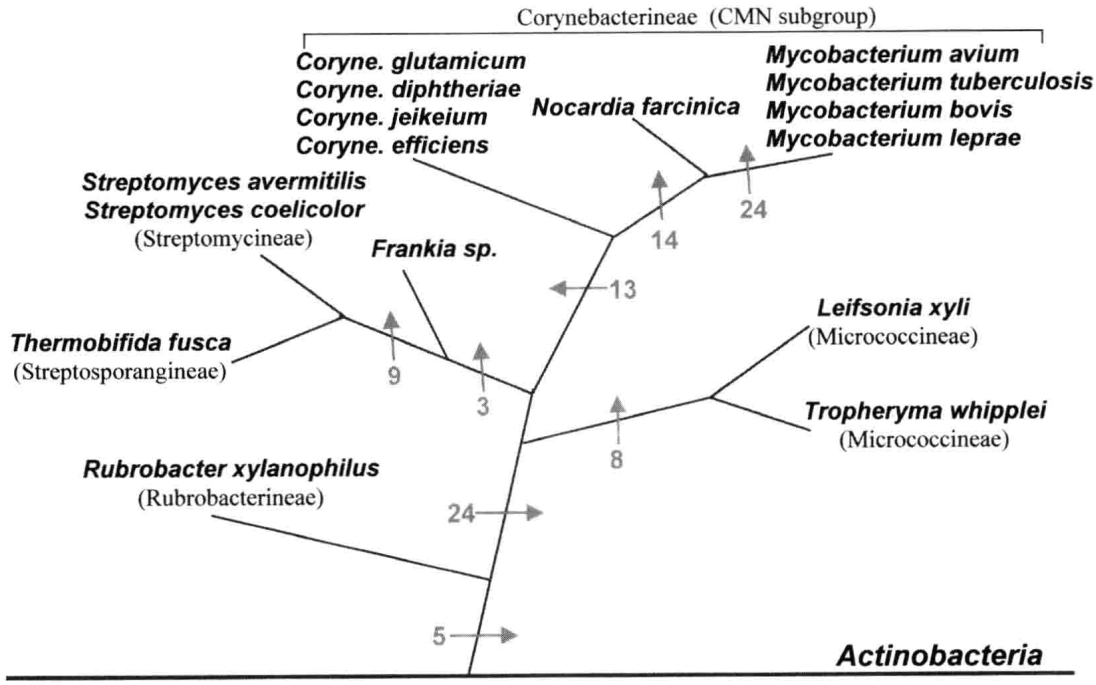


Figure 1.3 Partial alignment of ribosomal protein S3 sequences showing a 5-aa indel (boxed), which is specific for *Micrococcineae* and *Bifidobacterineae* (Gao and Gupta, unpublished results).