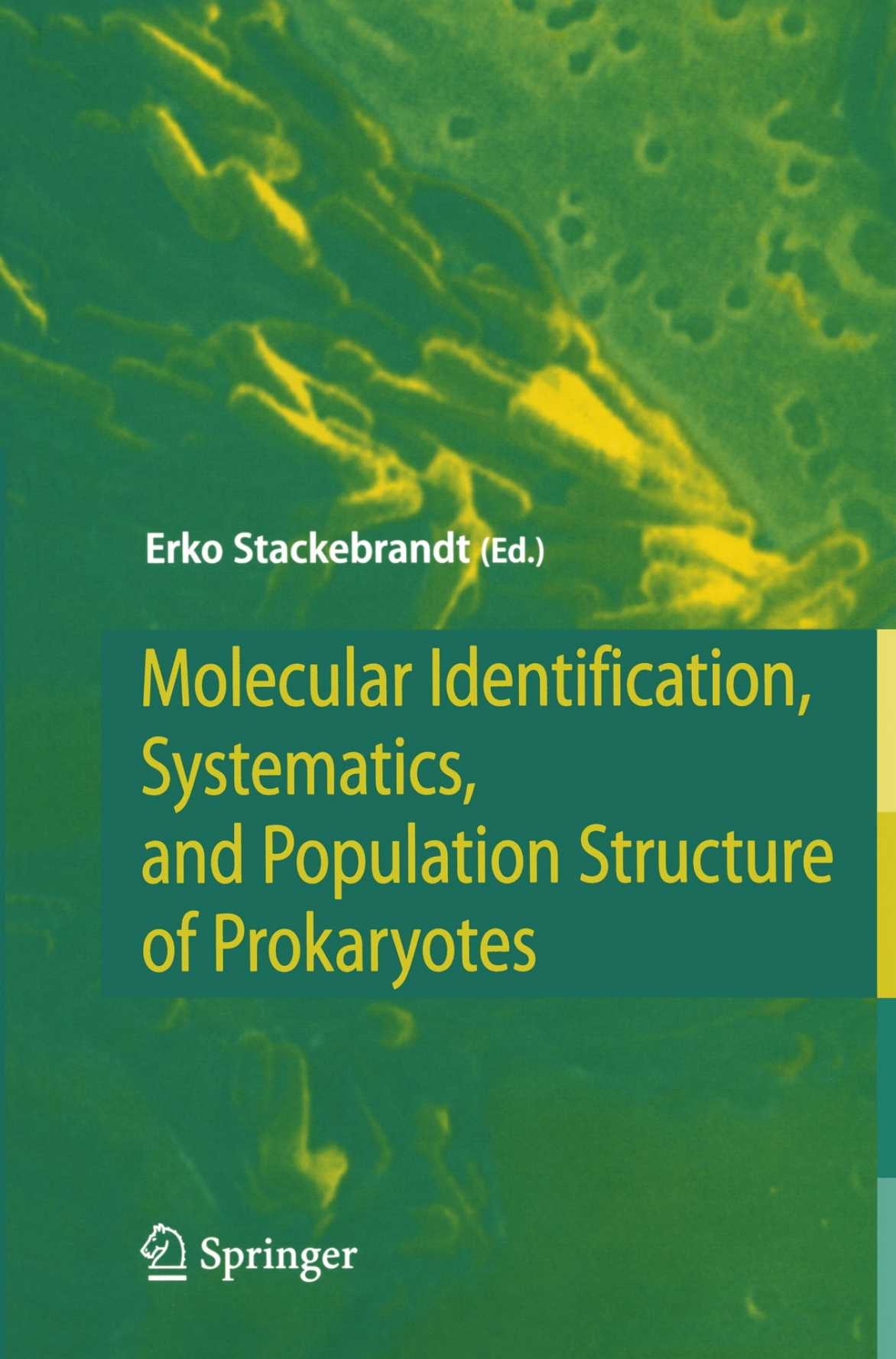

Erko Stackebrandt (Ed.)
Molecular Identification, Systematics, and Population Structure of Prokaryotes
Springer

Molecular Identification, Systematics, and Population Structure of Prokaryotes

Erko Stackebrandt (Ed.)

Erko Stackebrandt (Ed.)

Molecular Identification, Systematics, and Population Structure of Prokaryotes

With 56 Figures and 11 Tables

Springer

Professor Dr. Erko Stackebrandt
DSMZ GmbH
Mascheroder Weg 1b
38124 Braunschweig
Germany
E-mail: erko@dsmz.de

Library of Congress Control Number: 2005932380

ISBN-10 3-540-23155-2 Springer-Verlag Berlin Heidelberg New York
ISBN-13 978-3-540-23155-4 Springer-Verlag Berlin Heidelberg New York

Springer is a part of Springer Science + Business Media
springer.com

Printed in Germany

Cover design: *design&production*, Heidelberg, Germany
Typesetting and production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig, Germany
31/3150-YL - 5 4 3 2 1 0 - Printed on acid-free paper

Preface

One of the most exciting developments in biology is the change within bacterial systematics that has transformed a discipline of doubtful importance to most scientists into a respected discipline that now provides a phylogenetic framework for other areas in microbiology. Since the first description of a bacterial species 130 years ago, it has been the goal of systematists to work with a uniform classification system in which the genealogy of organisms provides the underlying basis of classification. Today, the 16S rRNA gene sequences of type strains of the vast majority of described species have been determined and have laid the foundations of a single hierarchic system. However, systematic biology has a far wider application than merely the provision of a reliable classification scheme for new strains. Within the framework of the stabilizing hierarchic system, genomes, non-coding regions, genes and their products can now be evaluated in an evolutionary context. Modifications in the tempo and mode of the evolution of individual markers may in turn modify the hierarchic system. Systematics, having left its ivory tower, is a dynamic process which is constantly progressing. The time has passed in which taxonomic schemes were outlined by a few individuals in a few countries: systematics today is a global, multi-disciplinary research field which has caught the attention of scientists who had never believed that they would follow with interest the development of systematics. Microbial ecologists, searching the vast and largely undetected microbial diversity by using the same molecular methods also in use by systematists, developed an interest in identification and classification. Population geneticists, searching for the spread of the causative agents of diseases by molecular methods can reflect on the delineation of the taxon 'species' and will unavoidably influence the thoughts on concepts and definitions. Clinical and environmental microbiologists develop and use DNA micro-arrays for rapid bacterial identification, biosafety and biosecurity issues and also for diversity assessment. Their technologies will soon be applied by and further enhance the power of systematics.

This book will summarize some of the recent developments in the molecular characterization of cultured and as yet uncultured Archaea and Bacteria, emphasising the strengths and weaknesses of individual approaches. All these techniques provide masses of molecular data which are retriev-

able from public databases. In contrast, the vast majority of phenotypic information on microorganisms is not electronically accessible. These data are needed for organisms already described and for species to be described. Systematics has not yet reached a state in which prokaryotes are defined solely on a molecular basis. The *Code of Nomenclature* requires the deposition of the type strain of each new species in public collections or resource centers, making it impossible to describe a handful of genes or gene products as a new species. Without any doubt, this information is extremely useful as it points towards the existence of new species, possibly providing information about their isolation. Most importantly, individual disciplines have learned to communicate on a common platform, leading to a newly emerging integrated discipline named 'systems biology'. In this environment, organisms will be embedded in a landscape[1] of information, in which organisms emerge as peaks on a topographic map. The denser the information net of genetic and epigenetic information, the clearer the view on the path of evolution of individual properties. Whether these peaks will receive taxon status will be open to thorough scrutiny.

The number of phyla defined by as yet uncultured strains exceeds those containing type strains of cultured species two-fold[2]. The neglible increase of novel species can be explained by the demanding mode with which species need to be circumscribed, the high costs involved in a proper description, the lack of trained taxonomists, and the directives of funding bodies not to collect and store diversity on a large scale. The chapters of this book are compiled to stimulate students to enter the field of bacterial diversity, to spread before them the patchwork carpet of fascinating multi-faceted disciplines which open the field to ecosystem functioning, communication within communities, symbiosis, life in extreme environments, astrobiology, and more.

Braunschweig, October 2005 *Erko Stackebrandt*

[1] The term 'landscape' in this context was first introduced to me by Jean Swings (Gent, Belgium).

[2] Fox JL (2005) Ribosomal gene milestones met, already left in dust. ASM News 71:6–7

Contents

Contributors

Markus Antwerpen
Robert Koch Institut, Burgstrasse 37, 38855 Wernigerode, Germany

Frank Oliver Glöckner
Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

Shigeaki Harayama
Department of Biotechnology, National Institute of Technology and Evaluation, 2-5-8 Kazusa-Kamatari, Kisarazu-shi,Chiba 292–0818, Japan

Hiroaki Kasai
Marine Biotechnology Institute, 3-75-1 Heita, Kamaishi, Iwate 026–0001, Japan

Anke Meyerdierks
Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

William W. Mohn
Department of Microbiology and Immunology, University of British Columbia, 300–6174 University Boulevard, Vancouver, British Columbia, V6T 1Z3, Canada

Josh D. Neufeld
Department of Biological Sciences, University of Warwick, Coventry, UK CV4 7AL

Ulrich Nübel
Robert Koch Institut, Burgstr. 37, 38855 Wernigerode

Christine Pourcel
GPMS laboratory, Institute of Genetics and Microbiology, University Paris XI, 91405 Orsay cedex, France

Rüdiger Pukall
DSMZ – Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Mascheroder Weg 1b, 38124 Braunschweig, Germany

Ramon Rosselló-Mora
Grup d'Oceanografia Interdisciplinar, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), C/Miquel Marqués 21, 07190 Esporles, Illes Balears, Spain

Erko Stackebrandt
DSMZ – Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Mascheroder Weg 1b, 38124 Braunschweig, Germany

Birgit Strommenger
Robert Koch Institut, Burgstrasse 37, 38855 Wernigerode, Germany

Luc Vauterin
Applied Maths BVBA, Keistraat 120, 9830 Sint-Martens-Latem, Belgium

Paul Vauterin
Applied Maths BVBA, Keistraat 120, 9830 Sint-Martens-Latem, Belgium

Gilles Vergnaud
Division of Analytical Microbiology, Centre d'Etudes du Bouchet, B.P. 3, 91710 Vert le Petit, France

Wolfgang Witte
Robert Koch Institut, Burgstrasse 37, 38855 Wernigerode, Germany

1 Exciting Times: The Challenge to be a Bacterial Systematist

Erko Stackebrandt

A comparison of the molar proportions reveals certain striking, but perhaps meaningless, regularities. *Vischer, Zamenhof and Chargaff (1949)*

Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organisation the largest amount of its own history. *Zuckerkandl and Pauling (1965)*

The species is man-made, and since it cannot be defined, the creation of taxa of higher categories based on species makes an absurd situation. *Cowan (1951)*

1.1 Introduction

In his overview "*Anaerobic life – a centennial view*" Ralf Wolfe (1999), referring to the dawn of complete genome sequencing of prokaryotes, states that "there has never been a more exciting time for the study of phylogeny and evolution". This citation complements the one by Hugenholtz and Pace (1996), referring to the encouraging development in microbial ecology, which is quoted by Neufeld and Mohn at the beginning of Chap. 7 in this book. These summaries are certainly more than personal opinions and highlight the enthusiasm that accompanies and drives microbiologists at unprecedented rates to new shores of understanding the biology of microorganisms. Can the history of microbiology be viewed as a series of isolated periods in which microbiologists considered themselves working in an exciting time? Is not the history of microbiology from the mid-nineteenth century a continuum of scientific achievements, in which scientists of any generation found it rewarding to contribute? When one considers not the short time periods, but the average generation time of 30–40 years as the productive years of a microbiologist, then this statement is correct (I am aware that the productive period of some microbiologists is significantly longer).

Looking backwards, there were times in which microbiologists must have been similarly impressed about developments in their own disciplines

Erko Stackebrandt: DSMZ – Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Mascheroder Weg 1b, 38124 Braunschweig, Germany, E-mail: erko@dsmz.de

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

as we are today. In retrospect, these events are named *milestones*, mainly single events which most probably are the crystallization of a much longer preceding period. For the discipline of bacterial systematics, I could think of a few such milestones or milestone eras while a non-taxonomist will certainly define others although, in many cases, milestones cover more than a single discipline. Persons mentioned in the following chapter are recognized for their achievements in microbiology but do not comprise an exhaustive list: this is not a comprehensive chapter on the history of bacterial systematics but rather a short introduction to a subject which has caught the attention of microbiologists from its very early beginnings. We are fortunate to work in a time in which bacterial systematics has been elevated to a scientific multidisciplinary field. For me, the exciting time spanned from 1970 until today, but I fully agree with Ralph Wolfe that this period will be significantly extended with new emerging directions and techniques, several of which are summarized in this book. The two main achievements that influenced my perspective of modern bacterial systematics were: first, the introduction of DNA–DNA reassociation studies in the early 1970s and, second, 16S rRNA oligonucleotide cataloguing in the late 1970s. The following years witnessed the application of reverse transcriptase and PCR-mediated sequence analysis of 16S rRNA genes and the analysis of genes coding for proteins. The combination of molecular, chemotaxonomic, physiological and other cellular traits led to first insights into the relatedness among prokaryotic species, changing each textbook chapter on microbial systematics. This development also fertilized ecological studies, leading to the recognition of as-yet uncultured organisms and the linkage of function to structure. It revolutionized the scale on which to look at prokaryotic diversity (Venter et al. 2004) and it revived the discussion on the concept and definition of the taxon 'species', sharpening the awareness that species are populations rather than genomically coherent entities (Coenye et al. 2005).

Advancements achieved during the period of an exciting time are the basis for the exciting times to come and are a fundamental driving force of visions that still motivate young people to dedicate themselves to science. The knowledge that we are only passengers in the 'train of science', which we enter at a certain station and alight at another as the train continues down the tracks, puts the achievement of scientists into perspective: we use the scientific platform provided by our predecessors and we broaden the basis, modify and sometimes radically change existing developments. Occasionally, we even may break with existing dogmas. The accumulated knowledge will be passed on to our successors who will continue the process, starting from a much higher and broader knowledge platform than the preceding generation. The following paragraphs will briefly summarize four milestone eras that have influenced the direction of microbial system-

atics. The past 130 years have been shaped by developments originating in various other disciplines and, still today, microbial taxonomists are often the users rather than the architects of concepts.

As milestones and highly productive eras should be recognized as such through their merits, the following subdivision is somewhat artificial, a personal view influenced by teachers, literature and my own experience. In no way can a contribution such as an introduction to a series of recent achievements and developments be sufficiently comprehensive to fully acknowledge the contributions and the influence of key scientists on the development of their own and on neighbouring scientific fields. The reader is referred to their original literature and to monographs in order to pay full tribute to their achievements.

1.2 The Early Heroes (1860–1900)

Even though the beginning of bacterial systematics can be placed with the description of the first bacterial species in 1872 by Ferdinand Cohn, his conclusions, mainly based on his own observations, were also influenced by the concepts, accurate observations and misinterpretations of scientists working in the early decades of the nineteenth century. Several developments ran in parallel. Above all, the morphology of micro-organisms was observed by light microscopy in combination with the application of specific staining procedures. Although stains were introduced as early as 1770 in the study of the structure of wood, it was not until 1839 that Christian Gottfried Ehrenberg (1795–1876) used stains to study microbes. At that time, the isolation of micro-organisms in pure culture had not been achieved. Although Louis Pasteur (1822–1895) and other scientists from that era described micro-organisms which fermented and caused diseases of sheep, cattle and other farm animals, as well as human illnesses, it was Robert Koch (1843–1910) who developed the technique of growing pure bacterial cultures. Most of the cultivation [on potato, gelatine, agar medium; later done in glass dishes introduced by Richard J. Petri (1852–1921)] and staining techniques were developed in the mid- to late 1800s by Robert Koch, Paul Ehrlich (1854–1915) and Hans Christian Gram (1853–1938). These various fundamental procedures were necessary to turn bacteriology into a respected science; and at that time the improvement of the health of livestock and man had absolute priority.

Pasteur established the view that microbes could be classified into fixed and unchangeable species and genera. Each species was believed to cause a specific disease. In contrast, Antoin Bechamp (1816–1908) declared that all animal and plant cells contained minuscule granules (granulations

moléculaires) that did not die when the organism died. These granules were believed to be the source of fermentation; and micro-organisms could arise from them as well. Several respected scientists believed that the morphological diversity of micro-organisms was due to variations of one and the same organism, e.g. Zopf (1846–1909), Wilhelm von Naegeli (1817–1891), Theodor Billroth (1829–1894), "missing the point that different stages of development, types of multiplication, the variety of size and form, and specific metabolic properties were associated with distinct species types" (Drews 1999).

Organisms that were observable under the microscope and later as pure cultures were named without guidelines (not to speak about rules). As summarized by Drews (2000) in his essay on the roots of microbiology, almost every scientist who observed micro-organisms gave them a new name without noticing that the same organisms may have already been named differently by another taxonomist. Synonyms accumulated as culture-dependent changes erroneously mirrored the existence of novel organisms (more than 40,000 invalid names and synonyms were counted at the end of the 1970s).

Called the 'father of systematics', Ferdinand Cohn studied algae, lichens and bacteria in media composed of defined mineral solutions complemented with different organic carbon sources. He was the first to propose a relationship among these organisms (Cohn 1867) and, summarizing his observations on shape, cellular structures, pigmentation and metabolic activities, he presented the first classification system of bacteria (Cohn 1872, 1876). He concluded that bacteria can be divided into distinct species with typical characteristics, which are transmitted to the following generations when bacteria multiply. Cohn also proposed that varieties exist within species, a notion that today plays an important role in the recognition of a bacterial species as a population, guiding scientists towards a new definition of this taxon more than 130 years later (Palys et al. 1997, 2000; Stackebrandt et al. 2002; Gevers et al. 2005).

The lack of recognizable characters other than morphological properties explains the superimposition of the botanical classification system to bacteria by the botanist Cohn (1872, 1876). Cohn, using the binominal nomenclature, affiliated the Schizomyceae (bacteria) and Schizophyceae (Cyanophyceae or cyanobacteria) to the group of Schizophyta (fission plants), but considered these micro-organisms as a group on their own. Bacteria were defined as chlorophyll-less cells of characteristic shape that multiply by cross-division and live as single cells, filamentous cell chains, or cell aggregates. [The fact that some Bacteria (sensu Woese et al. 1990) still carry the ending 'mycetes' is a reminder of the now discarded hypothesis that bacteria are fission fungi (schizomycetes). Note that even some of the archaeal taxa carry the ending 'bacteria', although the bacteria and archaea are members of two different Domains, indicating that nomen-

clature does not necessarily reflect phylogeny.]. The Schizomycetes contained four groups: 'Sphaerobacteria' (sphere-shaped, e.g. *Micrococcus*), 'Microbacteria' (rod-like, e.g. *Bacterium*), 'Desmobacteria' (filamentous, e.g. *Bacillus*, *Vibrio*), and 'Spirobacteria' (screw-like bacteria, e.g. *Spirillum*, *Spirochaeta*). On the basis of specific properties which were considered taxonomically less significant than morphology, Cohn divided some of his proposed genera, e.g. *Micrococcus*, into chromogenic (pigmented), zymogenic (fermenting) and pathogenic (contagious) species; and he described the purple bacteria in terms of their shape, pigments, gas vacuoles and sulfur globules.

It has to be stressed that Cohn already commented on the limited phylogenetic significance of the taxa he included in the morphology-based system: he was aware that the genera and species of bacteria have other meanings than for higher organisms, which reproduce sexually. He clearly stated that the proposed 'form-genera' and 'form-species' needed to be tested to determine whether they were indeed related in terms of descent. This, however, could not be achieved prior to 1970 at the level of genera (De Ley et al. 1970, Palleroni and Duodoroff 1971; Palleroni et al. 1973) and prior to 1977 at the level of higher taxa (Woese and Fox 1977; Woese et al. 1990). As a phylogenetic framework is still missing at the intraspecific level, appropriate methods need to be developed before systematists will be in a position to develop concepts.

1.3
The Dawn of Microbial Ecology and the Continuing Struggle with Classification Systems (1900–1930)

At the beginning of the twentieth century, the morphological basis of bacterial systematics was considerably broadened by the addition of physiological traits to the list of taxonomically important properties. Based on comparative morphological analysis and the hitherto unrecognized diversity of end-products and relation to oxygen, Orla-Jensen (1909) defined the main lines of bacterial systematics on the basis of physiological characteristics. However, as the system remained artificial (only elements of it were later found to have a phylogenetically sound basis) and the degree of the polyphyletic origin was not determinable, neither morphology, physiology, motility, nor any other property selected as the basis for a taxonomic scheme gave a satisfactory answer to conflicting alternatives. Even today, some of these discrepancies still complicate taxonomy.

At the turn of the century, microbial ecology was emerging as a new field, when Beijerinck (1895) described the formation of hydrogen sulfide from sulfate by a species later reclassified as *Desulfovibrio desulfuricans* and when

Winogradsky (1890) discovered chemoautotrophy (also see Winogradsky 1998). He was able to cultivate iron bacteria, described earlier by Cohn (1872), using mineral substrates from which ferrous iron was oxidized to ferric iron, obtaining energy for CO_2 assimilation. Analogous to this finding was the isolation of ammonium- and nitrite-oxidizing lithotrophic bacteria. At this time, microbial ecology was promoted mainly by members of the Delf School, e.g. Martinus Beijerinck, Cornelius B. van Niel and Albert J. Kluyver (to name a few with the greatest influence). They introduced the methods of selected isolation, including baiting micro-organisms with the properties they wanted to know about, by selecting the appropriate culture medium. The detection of a new range of physiologies considerably broadened the spectrum of taxonomically meaningful properties.

It must be mentioned in the context of this brief historical summary that, based on his own observations which were later supported by the theory of mutation of De Vries (1901), Martinus Beijerinck (1899) initiated experiments on changing physiological properties through variation and mutation, claiming that bacteria and fungi were more suitable objects for studies on heredity than higher evolved organisms (Beijerinck et al. 1940). These studies, later continued by members of the Delft school, led to the development of the genetics of micro-organisms (Delbrück and Luria 1942).

Though confronted with a broad spectrum of observations, the underlying genetic basis of the phenotype was missing. As pointed out by Palleroni (2003), the scientific community accepted the simplicity of Cohn's morphological system over the physiology-based concept for decades to come. His system was modified by adding new 'form-genera' to the inventory (Lehmann and Neumann 1896; Migula 1900; Pringsheim 1923; Janke 1924; Prévot 1933). Morphology continued to play a dominating conceptual role, far beyond the first morphology-based description of Ferdinand Cohn.

While Europe was setting the pace in the early years of bacterial systematics, America adopted its own bacterial classification system (Buchanan 1918; Winslow et al. 1920) by publishing the first edition of Bergey's *Manual of determinative bacteriology* (Bergey et al. 1923). This standard textbook was updated about every decade until 1990, when the first edition of Bergey's *Manual of systematic bacteriology* (Krieg 1986) was released. As the release of the new edition overlapped with the recognition about the restricted taxonomic value of morphology, these four volumes were composed to cluster groups of organisms under headings reflecting superficial morphological and physiological properties. Nevertheless, the merits of Bergey's manual has been recognized and the accumulated, systematized and published taxonomic knowledge in a single coherent volume constituted the "the first formal co-operation in the history of bacterial taxonomy" (Kluyver and van Niel 1936).

Today, knowing the basic phylogenetic lineages of cultured organisms, we consider most morphological and many physiological traits as being polyphyletic. Only a few morphologically complex traits are so far considered monophyletic, e.g. those of myxobacteria and spirochetes, as well as the formation of endospores. Even the thickness of the peptidoglycan, the basis for the Gram-staining reaction used to classify bacteria into two main groups, is not a monophyletic trait, as seen in the presence of Gram-positive cell walls in Archaea and Bacteria and the placement of Firmicutes, Actinobacteria and deinococci in separate higher taxa. The notion that morphologically different organisms may produce the same set of fermentation products or react similarly towards the presence of oxygen and light was first elucidated by deciphering metabolic pathways and recently by molecular analysis. Though certain physiological properties are indeed monophyletic, this information was not available to workers in the pre-molecular era. Rather than criticizing them for something they could not possibly have detected, we should acknowledge their attempts and those of the many others that followed for developing a range of systems, each of them devised to better serve the community of users.

1.4 Encouragement and Frustration (The Era 1930–1950)

Several key scientists from the early twentieth century influenced the science of bacterial systematics. There were the above-mentioned members of the Delft School, Albert J. Kluyver and his student Cornelius B. van Niel, as well as Robert E. Hungate, a student of the latter, and Roger Stanier. All of them were either involved in the isolation of bacteria, shifting the emphasis from clinical to environmental strains, or they were influencing the concepts of taxonomy. Hungate, the pioneer of anaerobic microbial microbiology and ecology (Chung and Bryant 1997), provided the fundament for the discovery of a new spectrum of microbial diversity, including the archaeabacteria (archaea), described about 40 years later (Woese and Fox 1977). Kluyver and van Niel are also recognized for their criticism against the system(s) outlined in the successive editions of Bergey's *Manual of determinative bacteriology*. Above all, they were critical of the "utter disregard for mutual relationships between natural groups" (Kluyver and van Niel 1936) and the disregard of other voices in the field (e.g. Rahn 1929, 1937). They also detailed many errors that arose as a consequence of the arbitrary use of morphological, physiological, cultural and pathogenic properties in bacterial classification (Palleroni 2003). This author also highlights the European tradition of favouring morphology as the first and most reliable guide of taxonomic systems (Kluyver and van Niel 1936) and disregarding

the use of physiology unless physiological principles could be subordinated to morphology. In the system of Kluyver and van Niel, morphological characters included the shape and size of cells, type of motility, presence of flagella, their number and type of insertion, the mode of reproduction, occurrence of endospores and various structural peculiarities. Certain physiological properties were indeed recognized but the overall importance of reactions for the cell was not reflected by their importance on taxonomic ranks. Pathogenicity was considered of doubtful value and differentiation of genera and even species on its basis was objectionable as a taxonomic criterion. Considering the genetic instability of many pathogenicity factors this may be judged as a wise decision; but the decision of Kluyver and van Niel was certainly not guided by genetic principles. It was inevitable that the basis of a true natural classification of bacteria would remain unsteady "inasmuch as the course of phylogeny will always remain unknown" (Kluyver and van Niel 1936). A call for a more prudent consideration of taxonomic systems was proposed by White (1937) who phrased: "the present call is not for newer, more ingenious, more pretentious, systems of classification, but for patient and incisive investigation". Later, Stanier and van Niel (1941) and van Niel (1946) commented on the inflexibility of Bergey's classification system that was based on the arbitrary selection of properties that could not be changed without replacing the existing system. The main advantage of Bergey's system was its practicability, i. e. identification and classification, but only if the key characters were mutually exclusive. The 'indications of relationships' should better be replaced by 'means of identification' and a broad range of differentiation characters rather than a few key properties should guide classification. This history of this period has been covered more extensively by Palleroni (2003).

It was not until the mid-1940s that van Niel (1946) agreed to add physiology, pathogenicity, nutrition and other easily determinable properties, e. g. colour, to the morphological properties used to devise an empirical key for bacteria. Obviously, systems were mainly devised to facilitate the affiliation of strains to species. The problem was the early adoption of names of taxonomic ranks from botanical and zoological systems where (at least in the majority of taxa) a taxon within a hierarchic system should indeed indicate genomic coherence and common ancestry. In microbiology, the majority of taxa (including even the taxon 'species') constituted a collection of entities of vastly different phylogenetic origin. van Niel (1946) pointed out the inability of phenotype-based classification systems to deduce phylogenetic interrelationships, though evolutionary consideration should have their place in bacterial taxonomy. Considering the general disbelief towards the emerging phylogenetic framework in 1980, it must be assumed that most microbiologists will have believed that determination of phylogenies were inherently indeterminable, at least at the higher

taxonomic levels. Woese (1987) criticized Roger Stanier who considered speculations on microbial evolution as being metascientific, by stating that "microbiology had reduced evolutionary matters to the status of dalliance was indeed unfortunate, for much of what is important and interesting about evolution lay hidden in the microbial world".

Not foreseeable by scientists in the 1940s, it was another 20 years before the pioneering work of Zuckerkandl and Pauling (1962) provided the framework of a phylogeny-based classification system. Today, with the broad outline of the system increasingly stable, a situation similar to that in the 1940s is occurring with the discussion of the concept of bacterial 'species' and the change from an artificial and arbitrary species definition (Staley and Konopka 1985; Wayne et al. 1987; Vandamme et al. 1996; Dijkshoorn et al. 2000; Rosselló-Mora and Amann 2001) to a definition that recognizes and describes natural mechanisms of speciation (summarized by Gevers et al. 2005).

Though the older systems have nothing less than historical value, they are important to remember as milestones of systematist's hybris to attempt to circumscribe the 'true' nature of the path of evolution. The merits and the correct perspective of early classification systems are discussed comprehensively by Kluyver and van Niel (1936) and by van Niel (1946). Still today we squeeze populations of more or less genomically diverse organisms into the taxon 'species' and define borders for genera, families and higher taxa, comforting ourselves by acknowledging the arbitrary nature of our definitions. Today, 130 years after Cohn's first description of species, our knowledge about the make-up and expression of a cell is breathtaking, but still we struggle with the definition of certain ranks.

Parallel to the discussion on the inappropriateness of phenotypic properties in reflecting evolutionary relationships, a possible solution through the linking of systematics to genealogy was slowly emerging. Originating in the nineteenth century, the discipline of Biochemistry was established together with the basis for a deeper understanding of heredity. Nucleic acids were isolated (Miescher, 1811–1887), terms like 'gene' and 'macromolecule' were introduced, the extraction of the first enzyme was described (Buchner 1897) and biochemical reactions were linked to genetic phenomena. The advantage of working with micro-organisms was recognized, but it was not until the 1940s when Avery et al. (1944) identified DNA as the responsible agent for the transfer of genetic markers in bacterial cultures. *Neurospora crassa* (Beadle and Tatum 1941) and bacterial species (Luria and Delbrück 1943) were study objects on physiological changes due to mutations. The mechanisms of the transfer of genetic information was described in *Escherichia coli* (Chargaff et al. 1949) and the genomic world was open to new research avenues, following the elucidation of the macromolecular structure of proteins (Pauling and Corey 1951) and nucleic acids (Watson and Crick 1953).

1.5 Expanding the Range of Properties: The Genetic and Epigenetic Levels (1950–1980)

The criticism on Bergey's classification system published in the 1940s and 1950s was accepted in the last edition of the Manual in 1974. Studies on the base composition of DNA, DNA–DNA reassociation studies and comparative biochemical and physiological studies did indeed demonstrate the phylogenetic coherence of some morphologically defined genera. However, major discrepancies were already noticed at the level of families and orders. The foreword to the eighth edition stated the inability of the present data set to deduce a hierarchic system of bacteria, as the majority of the key properties may have been the result of convergent evolution. Thus, the presentation of a fully developed system was abolished and taxa were clustered in 17 groups, according to the morphology and physiology of their members. In a few cases only were genera arranged into orders and families, only a few of which have survived the close scrutiny of phylogenetic analyses in recent years.

With a considerable delay of several years, several other important milestones in microbial systematics were accomplished, with their technical origins arising from ideas expressed in other disciplines. The most outstanding was the discovery of DNA, the full importance of which was recognized when the structure became available (Watson and Crick 1953) and appropriate methods for its analysis and manipulation were introduced. A second milestone was the development of computers in the 1950s and their use in handling phenetic and molecular data. A third milestone with direct implications on the future of systematics remained unnoticed by microbial systematists, who were involved in the daily struggle of identification and species description. Moreover, at the time of publication, microbiologists were not in a position to fully acknowledge that the ideas of Zuckerkandl and Pauling (1962, 1965) could be applied to bacteria. These visionaries postulated that "the amount of history preserved will be the greater, the greater the complexity of the elements at that level and the smaller the parts of the elements that have to be effected to bring about a significant change." They not only defined sematophoric molecules, i. e. genes and their transcripts [DNA (primary), mRNA (secondary), proteins (tertiary semantides)], as 'sense-carrying' units, i. e. the blueprint of an organisms' evolutionary history, but they also predicted that parts of the phylogenetic tree could be defined in terms of episemantic molecules, i. e. molecules that are synthesized under the control of proteins. Due to methodological constraints, the tertiary semantides (i. e. proteins) were the first molecules to be analysed, either by direct sequence analysis (e. g. cytochrome C, fibrinopeptides,

ferrodoxins), or by immunological approaches such as immunodiffusion and microcomplement fixation. Though protein sequencing lost its significance with the introduction of rapid sequencing techniques for DNA, its results already pointed towards the discrepancies between the outline bacterial classification schemes and the natural relationships of bacteria (Schwartz et al. 1975; Dickerson 1980; Ambler et al. 1987). Analysis of DNA and RNA was delayed for more than a decade by the lack of routine sequencing methods. In order to obtain at least general insights into the nucleotide similarities of primary and secondary semanides, hybridization techniques were introduced. DNA–DNA reassociation studies were the first to cluster organisms according to phylogenetic relationships and they played a decisive role in the definition of the taxon 'species' (Brenner et al. 1969; Palleroni et al. 1971; Johnson 1973; Grimont 1981); and still today it is considered the 'gold standard' for the delineation of species (see Chap. 2). The recommendation to use a 70% or so DNA–DNA reassociation value for defining species originated mainly from the experience made with numerous strains of enterobacterial species (Steigerwalt et al. 1976). Transferring the situation defined for a phylogenetically very shallow group of mainly eukaryote-associated organisms to all prokaryotes – which are the recent manifestations of different modes and times at which organisms evolve – is a dramatic underestimation of their phylogenetic status. But then one has to remember that the taxon thus delineated is an artificial construct, helpful in structuring the bacterial world at the level of species in a coherent way. Nevertheless, in times of whole genome sequencing approaches, the laborious DNA–DNA hybridization methodology seems to be out of date. As the number, identity and degree of conservatism of genes involved in the hybridization process remain unknown (even today), the ancestral genotype of a species cannot be determined. The obvious disadvantages (Stackebrandt et al. 2002), are more than compensated by the involvement of the majority of genes in the reassociation process. More recent attempts, concentrating on only a single or a few molecular markers, are significantly more biased, as one can only speculate whether these genes represent the evolutionary status of the complete genome. The artificial threshold value of about 70% reassociation (reflecting > 96% genome similarity; Schleifer and Stackebrandt 1983) indeed correlate well with those phenotypic properties of strains which are of general taxonomic value for the description of a species. DNA–DNA reassociation experiments confirmed the notion that a bacterial 'species' is not a genomically coherent entity but represents a population of highly related strains.

The recognition that translation mechanisms are highly conserved between species has opened a superior method of bacterial systematics (Dubnau et al. 1965). When methodologies to sequence RNA were not initially available, hybridization regimes between the rRNA gene and the gene prod-

uct were applied to groups of organisms known to be taxonomic dumping grounds, e.g. pseudomonads (Palleroni et al. 1973; De Smedt and De Ley 1977; De Vos and de Ley 1983) and clostridia (Johnson and Francis 1975). These bacteria lacked the chemical diversity found in many Gram-positive bacteria, such as actinobacteria and lactic acid bacteria. Within a few years, microbiologists noticed the phylogenetic unrelatedness of groups of bacteria which, based on morphological and metabolic grounds, has constituted well established genera for more than 80 years. For the first time in the history of microbiology, the failure of superficial properties to circumscribe natural relatedness became obvious. Results of DNA–rRNA reassociation studies unravelled deeper phylogenetic relationships than those obtained by DNA–DNA reassociation. While this finding alone was extremely satisfying, the restrictions of rRNA hybridization methods became apparent with the publication of the first results of rRNA oligonucleotide catalogue comparisons (Woese and Fox 1977). Phylogenetic analyses of catalogues, though limited at that time because of the lack of methods to sequence complete genes, were able to include any strain into a single dendrogram of relationship, including archaebacteria, eubacteria and eukaryotes.

Comparative studies highlighted the usefulness of the accumulated database of episemantic markers used in chemosystematics (chemotaxonomy, chemical taxonomy). Chemotaxonomy evolved as the by-product of biochemical and chemical work and developed in parallel with the introduction of chromatographic and other analytical methods. Without the support of peptidoglycan structure (Weidel and Pelzer 1964; Schleifer and Kandler 1967), isoprenoid quinones (Collins et al. 1977) and the lipid and fatty acid composition of cells (Lechevalier and Lechevalier 1970; Langworthy 1977; Lechevalier et al. 1977; Kates 1978), the acceptance of the phylogenetic uniqueness of many archaeal and bacterial taxa would have been delayed considerably. The determination of chemical markers, introduced during the 1950s, not only circumscribe the present state of a cell's chemical composition but indeed provide valuable properties used to critically analyse the phylogenetic clustering of groups of organisms at the genus level. This facet of systematics has not lost any of its attraction and, without its discriminatory power, many phylogenetically closely related species groups would not have been described as genera. Types and variation of peptidoglycan isoprenoid quinones, fatty acids, base composition of DNA, polar lipids, polyamines, pigments or mycolic acids and more are routinely used within the polyphasic approach to systematics. While single markers are rarely indicative of the phylogenetic coherence of a higher taxon, novel combinations of two or more of these properties are often highly correlated with the phylogenetic uniqueness of the respective organisms (Stackebrandt and Schumann 2000).

This period also witnessed the development of a third mainstream in bacterial systematics, numerical phenetic taxonomy (NT), introduced in the 1950s. Lasting for about 25 years, its influence on the recognition of coherence and lack thereof should not be underestimated, even if this approach is hardly in use anymore. This method is tightly connected with the development of algorithms, computers and the taxonomic concept that the reliability of the description of a taxon is improved by the provision of a comprehensive set of phenetic characters. Electronic computerization of microbiological data was first introduced by Sneath (1957) in order to handle the enormous amount of phenetic data collected during a taxonomic study of the genus *Chromobacterium*. This development ran in parallel with the work of Sokal and Michener (1958), who used an electric device to generate a classification of a eukaryotic taxon. Sokal and Sneath (1963) joined forces to develop the "*Principles of numerical taxonomy*" and they were among the first to develop and apply clustering and probabilistic distance coefficients in numerical taxonomy, e. g. single and average-linkage clustering, Jaccard's coefficient, scaling of multistate characters, parallelism and convergence, and equal weighting. Many of these algorithms and their modifications are still in use today in cluster analysis of the electrophoretic patterns of DNA and RNA digests (Riboprint, ARDRA, DGGE, AFLP, RFLP, etc.), protein patterns, fatty acid methyl ester patterns and the evaluation of ecological parameters, to name a few. Numerical analysis pointed out many inconsistencies in the classification at that time, leading to many taxonomic rearrangements. However, in the absence of a phylogenetic background, the resolving power of numerical analyses was overestimated, as the significance of individual properties remained unknown. Superficial characters were treated the same way as properties which indeed reflected the genealogy of the study object. With the advent of chemotaxonomy and a revised species definition, the numerical analysis lost its influence and present-day studies mainly target intraspecific variations.

1.6 Yet Another Exciting Time: Unravelling the Genealogy(ies) of Cultured and As-Yet Uncultured Prokaryotes

Being trained as a bacterial systematist during the late 1960s, I applied some of the key techniques of that period (determination of metabolic pathways, peptidoglycan structure and base composition of DNA, DNA–DNA reassociation studies) and witnessed the emergence of the breathtaking and historical development of molecular systematics. This era began, almost unnoticed by taxonomists, with a paper by Uchida et al (1974). 16S rRNA

oligonucleotide cataloguing changed the perception with which systematics was going to be executed in the future.

Though only a few species were investigated by this time-demanding technique before the advent of reverse transcriptase sequencing and, a few years later, PCR-based cycle sequencing, accelerating the analyses, the new approach of aligning systematics to the emerging tree of conservative macromolecules must be considered a powerful kickstart (Woese et al. 1985). While the power of these methods for the determination of intraspecific relationships was certainly overemphasized in the 1980s (which somehow discredited this method for some systematists), ribosomal RNA/rRNA gene sequencing remained the key to affiliate novel organisms to genera and to infer their phylogenetic novelty. After this short period of hesitation and disbelief that sequencing analysis of macromolecules would indeed benefit bacterial systematics other than as the provision of just another fragment in the general description of species, it was accepted so rapidly that, 20 years after its introduction, it is considered a routine and long-established method. The broad outline of higher taxa (Gibbons and Murray 1978) was not corrected but replaced. In 2001, the new editors of Bergey's Manual fully adopted the new system (Garrity et al. 2001, 2002) and are now, together with a new generation of systematists, actively involved in shaping the hierarchic structure of prokaryotes (Stackebrandt et al. 1997). The acceptance of molecular sequences to guide systematics has been facilitated by the availability of an enormous amount of phenetic data accumulated over the past decades. When superimposed on the phylogenetic clusters, many chemotaxonomic data gained new taxonomic significance as they were often the main criteria to delineate higher taxa. The fear that species and genera were described chiefly on the basis of 16S rDNA gene sequences (Palleroni 2003) is unjustified.

There were voices that considered the introduction of gene sequence comparison unfortunate, as it appeared the only method upon which phylogenetic relationships were based. However, soon after the analyses of 16S ribosomal RNA sequences began to influence systematics, scientists began wondering whether changes in nucleotide sequence of this single molecule solely represents its own evolution, rather than the evolution of a large portion of the genome, reflecting the genealogy of the host. However similar sequence analysis of the genes coding for 23S rDNA, elongation factors, ATPase, chaperons and many others demonstrated that the majority of the so-called housekeeping genes or core genes provided tree topologies that by and large matched that of the 16S rDNA tree (Gupta 1998, 2000), thus confirming the description of kingdoms and phyla in the two prokaryotic domains [the interested reader is referred to the scientific debate between Mayr (1998) and Carl Woese (1998) about "differing views as to what biology is and will be"]. Today, public databases contain sequences of hundreds

of fully sequenced genomes, offering a rich playground for studies on the micro- and macroevolution of genes and, crucial for systematists, providing information on the extent of horizontal gene transfer (Lawrence 2002). Like in previous times when taxonomists tried to avoid the use of genetically instable and plasmid-coded phenetic properties, the taxonomist of today will be prudent not to derive a phylogenetic framework on the basis of genes subjected to lateral gene transfer among members of the taxon concerned.

The discussion of the nature of the taxon 'species' has been provoked by the application of molecular tools, especially at the level of the species concept, i. e. the hypothetical basis of speciation. As a result of intensive multilocus enzyme electrophoreses (Selander et al. 1994), RAPDs (Istock et al. 1996) and multilocus sequence typing of housekeeping genes (Maiden et al. 1998), new ideas about speciation mechanisms have been expressed and mechanisms identified that contribute to the evolution of the genome. Some organisms are subjected to reticulate events or panmixis (Maynart-Smith et al. 1993, Istock et al. 1996) in which clonal relationships, due to mutational events and vertically transmitted accessoric genetic elements, are perturbated by horizontal genetic transfer, e. g. conjugation, phage transduction DNA transformation (Achtman 1998). Others, mostly endosymbionts and obligate pathogenic organisms, are mainly clonal because horizontal gene transfer appears to be a rare event. In an attempt to come to a biological species definition for bacteria, it has been proposed (Dykhuizen and Green 1991) to consider the following observations: (1) phylogenetic trees from different genes from members of a single species should be different and (2) phylogenetic trees from different genes from members of different species should be the same. What had been a challenge at the time when this definition was proposed has now become possible through high-throughput sequencing automation, allowing the analysis of five genes with a total of about 3,500 base pairs for each of about 2,000 strains of a single species. The intraspecific diversity recognizes centres of evolution leading to recognizable entities, named ecotypes (Cohan 2001, 2002). Their possible role in a redefined species description has been discussed in detail (Palys et al. 2000; Gevers et al. 2005).

The following chapters will highlight some of the key approaches used in microbial systematics and molecular ecology. These microbiological areas are somewhat related, as they originally evolved from the analysis of the same molecule, the 16S rRNA. Both disciplines will mutually benefit from progress made in either field. One set of approaches is based on the finding of taxon-specific signature sequences in the rapidly increasing database of rRNA catalogues and complete sequences from the late 1980s on (Brosius et al. 1987). Molecular probes are used in clinical diagnostic and most impressively in in-situ hybridization studies in ecology. The

database of more than 120,000 16S rDNA gene sequences results from the recognition of the unexplored microbial diversity that reinforces earlier notions about the inability of cultured organisms to represent diversity. The listing of exiting new developments in systematics will however not be complete without a mention of rapid DNA profiling methods, used routinely not only in bacterial identification and in the description of new taxa, but also in the assessment of the molecular diversity of populations in their natural environment. The handling and identification of the relatively small number of only about 6,000 validly described species (with an annual increase of 230 – 300 species) is manageable, but the situation may soon get out of hand once novel and innovative isolation methods have been devised. A prerequisite for the handling of a substantial increase in species numbers is the design of dynamic automated identification systems that access curated databases of molecular and non-molecular data, combined with advanced computational strategies and knowledge management. The search for novel organisms should run in parallel with the investment in reproducible authentification methods with a high resolving power, such as those based on mass spectrometry (MS) and mainly in use for clinical isolates and select agents (e. g. matrix adsorbed laser deionization/ionization time-of-flight MS, Fourier-transformed infrared MS).

These times are so rich in new techniques, new technical support, new insights and fresh ideas that not only students find it difficult to maintain an overview about advances in the field of microbial systematics and diversity. Most obviously, it is a good time to be part of this exciting avenue. I am confident that the next generation of microbiologists will benefit from the scientific progress achieved at the turn of the twenty-first century. It is the hope of the authors of this book that newcomers to the field of microbial diversity may have the enthusiasm to equip themselves with a sufficiently qualified background and experience to carry on the exploration of the microbial world. To quote somebody who knew what it is all about: "The best way to have a good idea is to have a lot of ideas" (Linus Pauling, "*The nature of the chemical bond*")

References

Achtman M (1998) Microevolution during epidemic spread of *Neisseria meningitidis*. Electrophoresis 19:593–596

Ambler RP, Daniel M, McLellan L, Meyer TE, Cusanovich MA, Kamen MD (1987) Amino acid sequences of cytochrome c-554(548) and cytochrome c' from a halophilic denitrifying bacterium of the genus *Paracoccus*. Biochem J 248:365–371

Avery OT, Macleod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type 111. J Exp Med 79:137–158

Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in *Neurospora*. Proc Natl Acad Sci USA 27:499–506

Beijerinck MW (1895) Ueber *Spirillum desulfuricans* als Ursache von Sulfat-reduction. Centralbl Bakteriol Parasitkd Infekt Abt II 1:49–59

Beijerinck MW (1899) Über ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabakblätter. Centralbl Bakteriol Parasitkd Infekt Abt II 5:27–33

Beijerinck IGV, Dooren de Jong LE den, Kluyver AJ (1940) Martinus Willem Beijerinck, his life and work. W13, The Hague

Bergey DH, Harrison FC, Breed RS, Hammer BW, Huntoon FM (1923) Bergey's manual of determinative bacteriology, 1st edn. Williams and Wilkins, Baltimore

Brenner DJ, Fanning GR, Johnson KE, Citrella RV, Falkow S (1969) Polynucleotide sequence relationships among members of the *Enterobacteriaceae*. J Bacteriol 98:637–650

Brosius J, Palmer ML, Kennedy PJ, Noller HF (1987) Complete nucleotide sequence of the 16S ribosomal RNA gene from *Escherichia coli*. Proc Natl Acad Sci USA 75:4801–4805

Buchanan RE (1918) Studies in the nomenclature and classification of the bacteria. V. Subgroups and genera of the *Bacteriaceae*. J Bacteriol 3:27–61

Buchner E (1897) Alkoholische Gährung ohne Hefezellen. Ber Dtsch Chem Ges 30:117–124

Chargaff E, Vischer E, Doniger R, Green C, Misani, F (1949) The composition of the desoxypentose nucleic acids of thymus and spleen. J Biol Chem 177:405–416

Chung KT, Bryant MP (1997) Robert E. Hungate: pioneer of anaerobic microbial ecology. Anaerobe 3:213–217

Coenye T, Gevers D, Van de Peer Y, Vandamme P, Swings J (2005) Reevaluating prokaryotic species. FEMS Microbiol Rev 29:147–167

Cohan FM (2001) Bacterial species and speciation. Syst Biol 50:513–524

Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56:457–487

Cohn F (1867) Beiträge zur Physiologie der Phycochromaceen and Florideen. Arch Mikrosk Anat Entwicklungsmech 3:1–60

Cohn F (1872) Untersuchungen über Bakterien II. Beitr Biol Pflanz 1:127–224

Cohn F (1876) Untersuchungen über Bakterien IV. Beiträge zur Biologie der Bacillen. Beitr Biol Pflanz 2:249–276

Collins MD, Pirouz T, Goodfellow M, Minnikin DE (1977) Distribution of menaquinones in actinomycetes and corynebacteria. J Gen Microbiol 100:221–230

Cowan ST (1951) Sense and nonsense in taxonomy. J Gen Microbiol 67:1–8

De Ley J (1970) Reexamination of the association between melting point, buoyant density, and chemical base composition of deoxyribonucleic acid. J Bacteriol 101:738–754

De Smedt J, De Ley J (1977) Intra- and intergeneric similarities of *Agrobacterium* ribosomal ribonucleic acid cistrons. Int J Syst Bacteriol 27:222–240

De Vos P, De Ley J (1983) Intra- and intergeneric similarities of *Pseudomonas* and *Xanthomonas* ribosomal ribonucleic acid cistrons. Int J Syst Bacteriol 33:487–509

De Vries H (1901) Die Mutationstheorie. Veit, Leipzig

Delbrück M, Luria SE (1942) Interference between bacterial viruses. I. Interference between two bacterial viruses acting upon the same host, and the mechanism of virus growth. Arch Biochem 1:111–141

Dickerson RE (1980) Cytochrome c and the evolution of energy metabolism. Sci Am 242:136–153

Dijkshoorn L, Ursing BM, Ursing JB (2000) Strain, clone and species: comments on three basic concepts of bacteriology. J Med Microbiol 49:397–401

Drews G (1999) Ferdinand Cohn: a promoter of modern microbiology. Nova Acta Leopold 80

Drews G (2000) The roots of microbiology and the influence of Ferdinand Cohn on Microbiology of the 19th century. FEMS Microbiol Rev 24:225–249

Dubnau D, Smith I, Porell P, Marmur J (1965) Genetic conservation in *Bacillus* species and nucleic acid homologies. Proc Natl Acad Sci USA 54:491–498

Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268

Garrity GM, Boone DR, Castenholz RW (2001) The Archaea and the deeply branching and phototrophic bacteria. In: Garrity GM, Boone DR, Castenholz RW (eds) Bergey's manual of systematic bacteriology vol 1, 2nd edn. Springer, Berlin Heidelberg New York

Garrity GM, Johnson KL, Bell J, Searles DB (2002) Taxonomic outline of the procaryotes, rel 3.0, http://dx.doi.org/10.1007/bergeysoutline

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL Swings J (2005) Reevaluating prokaryotic species. Opinion paper. Nat Rev Microbiol 3:733–739

Gibbons NE, Murray RGE (1978) Proposals concerning the higher taxa of bacteria. Int J Syst Bacteriol 28:1–6

Grimont PAD (1981) Use of DNA reassociation in bacterial classification. Can J Microbiol 34:541–546

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. Microbiol Mol Biol Rev 62:1435–1491

Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24:367–402

Hugenholtz P, Pace NR (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol 14:190–197

Istock CA, Bell JA, Ferguson N, Istock NL (1996) Bacterial species and evolution: theoretical and practical perspectives. J Ind Microbiol 17:137–150

Janke A (1924) Allgemeine Technische Mikrobiologie, I Teil: Die Mikroorganismen. Steinkopf, Dresden

Johnson JL (1973) The use of nucleic acid homologies in the taxonomy of anaerobic bacteria. Int J Syst Bacteriol 23:308–315

Johnson JL, Francis BS (1975) Taxonomy of the clostridia: ribosomal ribonucleic acid homologies among the species. J Gen Microbiol 88:229–244

Kates M (1978) The phytanyl ether-linked polar lipids and isoprenoid neutral lipids of extremely halophilic bacteria. Prog Chem Fats Other Lipids 15:301–342

Kluyver AJ, van Niel CB (1936) Prospects for a natural system of classification of bacteria. Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt II 94:369–403

Krieg NR (ed) (1986) Bergey's manual of systematic bacteriology, vol 1. Williams and Wilkins, Baltimore

Langworthy TA (1977) Long-chain diglycerol tetraethers from *Thermoplasma acidophilum*. Biochim Biophys Acta 487:37–50

Lawrence JG (2002) Gene transfer in bacteria: speciation wihtout species. Theor Popul Biol 61:449–460

Lechevalier MP, Lechevalier H (1970) Chemical composition as a criterion in the classification of aerobic actinomycetes. Int J Syst Bacteriol 20:435–443

Lechevalier MP, Bièvre C de, Lechevalier HA (1977) Chemotaxonomy of aerobic actinomycetes: phospholipid composition. Biochem Syst Ecol 5:249–260

Lehmann KB, Neumann RO (1896) Atlas und Grundriss der Bakteriologie und Lehrbuch der Speciellen Bakteriologischen Diagnostik, 1st edn. Lehmann, Munich

Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491–511

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russel JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic organisms. Proc Natl Acad Sci USA 95:3140–3145

Maynart-Smith J, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388

Mayr E (1998) Two empires or three? Proc Natl Acad Sci USA 95:9720–9723

Migula W (1900) Specielle Systematik der Bakterien. Fischer, Jena

van Niel CB (1946) The classification and natural relationships of bacteria. Cold Spring Harbor Symp Quant Biol 11:285–301

Orla-Jensen S (1909) Die Hauptlinien der natürlichen Bakteriensystems. Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt II 22:305–346

Palleroni NJ (2003) Prokaryote taxonomy of the 20th century and the impact of studies on the genus *Pseudomonas*: a personal view. Microbiology 149:1–7

Palleroni NJ, Doudoroff M (1971) Phenotypic characterization and deoxyribonucleic acid homologies of *Pseudomonas solanacearum*. J Bacteriol 107:690–696

Palleroni NJ, Kunisawa R, Doudoroff M (1973) Nucleic acid homologies in the genus *Pseudomonas*. Int J Syst Bacteriol 23:333–339

Palys T, Nakamura LK, Cohan FM (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol 47:1145–1156

Palys T, Berger E, Mitrica I, Nakamura LK, Cohan FM (2000) Protein-coding genes as molecular markers for ecologically distinct populations: the case of two *Bacillus* species. Int J Syst Evol Microbiol 50:1021–1028

Pauling L, Corey RB (1951)The structure of synthetic polypeptides. Proc Natl Acad Sci USA 37:241–250

Prévot AR (1933) Études de systématique bactérienne. I. Lois générales. II. Cocci anaérobies. Ann Sci Nat Bot Biol Veg 15:23–260

Pringsheim EG (1923) Zur Kritik der Bakteriensystematik. Lotos 71:357–377

Rahn O (1929) Contributions to the classification of bacteria, V–X. Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt II 79:321–343

Rahn O (1937) New principles for the classification of bacteria. Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt II 96:273–286

Rosselló-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67

Schleifer KH, Kandler O (1967) On the chemical composition of the cell wall of streptococci. I. The amino acid sequence of the murein of *Str. thermophilus* and *Str. faecalis*. Arch Mikrobiol 57:335–64

Schleifer KH, Stackebrandt E (1983) Molecular systematics of prokaryotes. Annu Rev Microbiol 37:143–187

Schwartz RM, Barker WC, Dayhoff MO (1975) Early events in the emergence of eukaryotes and prokaryotes inferred from RNA and protein sequences. In: Second college park colloquium on chemical evolution. University of Maryland, Baltimore

Selander RK, Li J, Boyd F, Wang F-S, Nelson K (1994) DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*. In: Priest FG, Ramos-Cormenzana A, Tindall B (eds) Bacterial diversity and systematics. Plenum, New York, pp 17–50

Sneath PHA (1957), The application of computers to taxonomy. J Gen Microbiol 17:201–226

Sokal R, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438

Sokal R, Sneath PHA (1963) Principles of numerical taxonomy. San Francisco

Stackebrandt E, Schumann P (2000) Introduction to the taxonomy of the class Actinobacteria. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds) The prokaryotes, 3rd edn. Springer, Berlin Heidelberg New York

Stackebrandt E, Rainey FA, Ward-Rainey NL (1997) Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. Int J Syst Bacteriol 47:479–491

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1052

Staley JT, Konopka A (1985) Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346

Stanier RY, van Niel CB (1941) The main outlines of bacterial classification. J Bacteriol 42:437–466

Steigerwalt AG, Fanning GR, Fife-Asbury MA, Brenner DJ (1976) DNA relatedness among species of *Enterobacter* and *Serratia*. Can J Microbiol 22:121–137

Uchida T, Bonen L, Schaup HW, Lewis BJ, Zablen L, Woese C (1974) The use of ribonuclease U2 in RNA sequence determination. Some corrections in the catalog of oligomers produced by ribonuclease T1 digestion of *Escherichia coli* 16S ribosomal RNA. J Mol Evol 28:63–77

Vandamme P, Pot B, Gillis M, Vos P de, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Vischer E, Zamenhof S, Chargaff E (1949) Microbial nucleic acids: the desoxypentose nucleic acids of avian tubercle bacilli and yeast. J Biol Chem 177:429–438

Watson JD, Crick FH (1953) Molecular structure of nucleic acids. Nature 171:737–738

Wayne L, Brenner DJ, Colwell RR, Grimont, PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Trüper HG (1987) International committee on systematic bacteriology: report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Weidel W, Pelzer H (1964) Bagshaped macromolecules – a new outlook on bacterial cell walls. Adv Enzymol Relat Areas Mol Biol 26:193–232

White PB (1937) Remarks on bacterial taxonomy. Zentralbl Bakteriol Parasitenkd Infektionskr. Hyg. Abt II 96:145–149

Winogradsky S (1890) Recherches sur les organismes de la nitrification. Compts Rendu 110:1013–1016

Winogradsky S (1998) Research on nitrifying organisms (1890: Compts Rendu 110:1013–1016). In: Brock TD (ed) Milestones in microbiology: 1556 to 1940. ASM, Washington, D.C., pp 231–233

Winslow CEA, Broadhurst J, Buchanan RE, Krumwiede C Jr, Rogers LA, Smith GH (1920) The families and genera of bacteria. Final report of the Committee of the Society of American Bacteriologists on characterization and classification of bacterial types. J Bacteriol 5:191–229

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR (1998) Default taxonomy: Ernst Mayr's view of the microbial world. Proc Natl Acad Sci USA 95:11043–11046

Woese CR, Stackebrandt E, Macke T, Fox GE (1985) A phylogenetic definition of the major eubacterial taxa. System Appl Microbiol 6:143–151
Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucaryas. Proc Natl Acad Sci USA 87:4576–4579
Woese G, Fox E (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090
Wolfe RS (1999) Anaerobic life – a centennial view. J Bacteriol 181:3317–3320
Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic, New York, pp 189–225
Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

2 DNA–DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation

Ramon Rosselló-Mora

2.1 Introduction

DNA–DNA reassociation techniques are used for many purposes, but in the field of microbial systematics they are in most cases linked to the circumscription of prokaryotic species. Actually, as we will see, the use of whole genome hybridizations in the definition of prokaryotic species has had an enormous influence since the origin of the polythetic classification system (Rosselló-Mora and Kämpfer 2004). The importance of morphology in the middle of the eighteenth century was substituted for that of biochemical properties at the beginning of the nineteenth century; and subsequently the emerging "modern spectrum" techniques emphasized the importance of genetic measurements, such as DNA–DNA reassociation experiments. However, after almost 50 years of the application of these techniques to circumscribe species, there is increasing reluctance to use them because of the intrinsic pitfalls in the methods (e.g. Stackebrandt 2003; Stackebrandt et al. 2002). Consequently, the question that arises is: if DNA reassociation techniques are to be substituted, what will take their place? However, in my opinion, it is still too soon to substitute these techniques because of several reasons: (a) the use of such parameters in the definition of species has been of paramount influence and has actually determined the size and shape of what we call 'species', (b) there are almost 5,000 species described (Garrity et al. 2004), many of them based on reassociation experiments, and the legitimacy of new circumscription methods should be validated and (c) the alternatives proposed are not yet standardized and tested sufficiently enough to offer a reliable, pragmatic and easy to use circumscription tool. Any new technique with the potential to act as a substitute for DNA–DNA reassociation experiments should demonstrate that: (a) it is more reliable, workable and pragmatic, (b) it does not radically change the present classification system and (c) it leads to results that fit into a genomically based perspective without losing sight of the organisms themselves. Any

Ramon Rosselló-Mora: Grup d'Oceanografia Interdisciplinar, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), C/Miquel Marqués 21, 07190 Esporles, Illes Balears, Spain, E-mail: rossello-mora@uib.es

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

intended substitution of a technique that has implications for the circumscription parameters that have served as a basis for the establishment of the current taxonomic system should also take into account the purpose of taxonomy. The end result itself is to provide a system that is operative and predictive; and the information behind a name should be more than a mere set of genes with no meaning. What has hitherto been constructed is a classification system based on the circumscription of taxa when the overall information collected indicated that such circumscription would be enough to recognize them as unique and identifiable. Behind a species name there is more than a binomial, there is a collection of data that allows identification from several independent sources that gives a prediction of how an organism may be and might behave. Our system is perhaps not perfect and deserves improvement, but as already noted "it is the envy of those who wish to implement similar systems in botany or zoology" (Euzéby and Tindall 2004).

DNA–DNA reassociation techniques, also known as DNA–DNA hybridization techniques, are based on an attempt to make raw comparisons of whole genomes between different organisms in order to calculate their overall genetic similarities. Just after the discovery of the intrinsic properties of DNA (i. e. information content and secondary structure resilience), a good number of techniques were developed and applied to microbial taxonomy in order to circumscribe its basic unit, the species. At that time, it was believed that such genetic comparisons would render more stable classifications than those simply based on phenotypic similarities (Krieg 1988). There is no doubt that the first attempt to elucidate taxonomic relationships based on single-stranded DNA reassociation conducted by Schildkraut et al. (1961) was a breakthrough for microbial systematics and for the construction of the current microbial classification system. They demonstrated that duplex formation between the denatured DNA of one organism and that of another organism would only occur if the overall DNA base compositions were similar and if the organisms from which the DNA was extracted were genetically related. At the time when a monothetic classification was abandoned in favour of a polythetic (or phenetic; Rosselló-Mora and Amann 2001) classification, these developments in DNA techniques led to microbial taxonomists extending the definition of the species by using reassociation results and by determining the GC mole percentage of each individual genome. The great practical advantage seen in DNA–DNA hybridization experiments was that the results did not show the continua often observed between groups defined by phenotypic characteristics, but instead the genomes appeared clustered in discrete groups, whether organisms tended to be closely related or not (Krieg 1988). Since then, such techniques have routinely been applied in most of the new species characterizations, especially those that involved new taxa in already existing

genera and/or those where more than a single isolate was used to circumscribe the taxon. The application of these techniques to circumscribe species was reinforced by a recommendation from an ad hoc committee on systematics (Wayne et al. 1987). In fact, the committee (using ΔT_m to indicate melting temperature increment) stated that "the phylogenetic definition of a species generally would include strains with approximately 70% or greater DNA-DNA relatedness and with 5 °C or less ΔT_m. Both values must be considered. Phenotypic characteristics should agree with this definition and would be allowed to override the phylogenetic concept of species only in a few exceptional cases". In addition, they reinforced that "it is recommended that a distinct genospecies that cannot be differentiated from another genospecies on the basis of any known phenotypic property not be named until they can be differentiated by some phenotypic property". That recommendation had two main effects. On the one hand, it forced descriptions based on both genomic and phenotypic properties but, on the other hand, it unwittingly created the belief that a rigid boundary of 70% genome similarity would be sufficient for the recognition of species. Both aspects have had an enormous influence on prokaryotic taxonomy.

Emerging techniques at the end of the twentieth century, such as rRNA gene sequencing and phylogenetic reconstructions, were expected to help in the replacement of DNA-DNA reassociation experiments. However, it was soon realized that, due to the length and information of the molecule, the resolution power needed to discriminate different species within a genus was not always adequate (e.g. Amann et al. 1992; Fox et al. 1992; Martínez-Murcia et al. 1992). For these reasons, it was accepted at that time that no other methodology could replace genome similarity analysis (Stackebrandt and Goebel 1994). It has always been clear that the best way to understand similarities would be to truly compare whole genome sequences (e.g. Owen and Pitcher 1985), a fact that has nowadays almost become possible. The increasing number of completely sequenced genomes allows such comparisons and the first speculations on how species can be circumscribed by this newly emerging information (Konstantinidis and Tiedje 2005; Santos and Ochman 2004; Stackebrandt et al. 2002; Zeigler 2003). However, all these new circumscription attempts should be previously validated by contrasting them with the criteria used to construct the current taxonomic schema.

DNA-DNA reassociation experiments have often been criticized due to their high experimental error and their failure at generating cumulative databases (e.g. Sneath 1989; Stackebrandt 2003). However, their use has never been abandoned because no other alternative has been either found or tested. In order to illustrate how often DNA-DNA reassociation experiments are still used to circumscribe species, a survey on all the publications that appeared in 'Int. J. Syst. Evol. Microbiol.' during 2004 has been under-

Table 2.1. 'Int. J. Syst. Evol. Microbiol.' survey: absolute numbers and percentages of articles or new descriptions that were published in the six issues of vol 54 of the journal during 2004

Articles with new descriptions	305	
Articles with reassociation experiments	199	65%[a]
Articles without reassociation experiments	106	35%[a]
Spectrophotometric reassociation experiments	67	34%[b]
Non-radioactive microtitre-plate hybridizations	96	48%[b]
Non-radioactive filter methods (chemiluminescence)	9	5%[b]
Radioactive filter, S1, or hydroxyapatite methods	27	14%[b]
New species	351	
New species with a single isolate	191	54%[c]
New genera	65	
New 'candidatus'	17	

[a] percentages refer to the 305 articles with new descriptions

[b] percentages refer to the 199 articles where reassociation experiments were performed

[c] percentages refer to the total number of 351 new species classifications

taken (Table 2.1). In that year, around 305 articles appeared that compiled the description of about 351 new species, 65 new genera, and 17 new 'candidatus'. Among all these new species descriptions, about 65% of them used DNA–DNA reassociation experiments. From the 35% of the remaining descriptions where no reassociation was used, more than 75% were based on a single isolate and more than half corresponded to new genera. In such cases, the rationale for taxa descriptions were mainly based on 16S rDNA sequence dissimilarities. However, it is also worth noting that among all the descriptions where DNA–DNA reassociation was used, nearly 60% of them were also based on a single isolate. In these cases, the use of hybridizations was to show enough dissimilarity to their closest relative species.

There is a desire to replace DNA–DNA reassociation for other more accurate techniques (Stackebrandt et al. 2002) but its use still cannot be avoided. Consequently, this is a timely review concerning existing techniques, their pitfalls and the meaning of their results. In addition, the possibility to replace them will also be discussed.

2.2 Semantic Considerations

Prokaryotic taxonomy, like eukaryotic taxonomy, is filled with semantic misuses. There are several examples that in some respect are responsible for the so-called 'species problem': (a) the use of homology as a synonym of similarity, (b) the persistent homonymy of the term species and (c) the

synonymy between concept and definition. Although these issues will be thoroughly discussed elsewhere, it is worth providing some clarifications at this point:

1. Homology vs similarity: since the early days of the interpretation of DNA–DNA reassociation results, homology and similarity have been used as synonyms. However, it was soon noted that the use of the term homology would not be appropriate for interpreting hybridization results, because there was no certainty that bound stretches of DNA from different organisms would contain identical nucleotide sequences and the use of terms such as relatedness or DNA binding would be more accurate (Brenner and Cowie 1968; De Ley et al. 1970). However, these recommendations were not taken into account and for decades the term homology has been used to express DNA–DNA reassociation results. Later, there was again the temptation to abandon the term homology (Stackebrandt and Liesack 1993) by arguing that the values observed were not linearly correlated with sequence identity. Homology is not a measurable parameter: either two characters (in this case sequences or DNA fragments) are homologous or not, which means that either they have the same evolutionary origin or not (Fitch 2000; Mindell and Meyer 2001; Tindall 2002). Homology basically has an evolutionary meaning and thus cannot be applied either as a synonym for sequence identity or to express DNA–DNA reassociation results. The term similarity is perhaps the best choice because it does not imply any evolutionary nor phylogenetic meaning. Despite the reiterated recommendations, there are still quite a few publications that wrongly use the term homology.

2. Homonymy of the term species: perhaps the most important cause of the 'species problem' is the persistent homonymy (Reydon 2004). This means that different scientific disciplines adopt different concepts to embrace their devised units, but the same term 'species' is given to all of them. This has always been regarded as a clear case of pluralism (Brigandt 2002; Ereshefsky 1998; Mishler and Donoghue 1982; Reydon 2004). For some, it would be better to eliminate the term species and each scientific discipline should instead adopt a unique and specially tailored basic unit, such as 'biospecies', 'ecospecies' or 'phylospecies' (Ereshefsky 1998). However, for others, pluralism is still an adequate choice, with the term 'species' being kept for general-purpose classification, which should retain binomials as a property of the taxonomic system (Brigandt 2002). These problems, which have been thoroughly discussed in eukaryotic taxonomies, are well represented when classifying prokaryotes. Actually, what taxonomists mean by a species does not satisfy, for instance, microbial ecologists or population geneticists, although it would probably not be possible for these groups to come to any mutual agreement on terminology. It is also important to note that, for example "evolution was inferred from the classification, not vice

versa" (Sneath 1988) and thus the ultimate concept of 'species' is a property of taxonomy. These disagreements are the basis for most of the discussions on the adequacy of the current species concept in use (Rosselló-Mora and Kämpfer 2004) and, therefore, most probably it would be recommendable to adopt a clear pluralistic approach. Taking into account that the term and idea of 'species' is the basal taxonomic unit originally devised to support a universal hierarchic system (Ereshefsky 1994), the main arguments expressed here are within the framework of taxonomy and refer to the species concept currently applied to the classification of prokaryotes. Perhaps the most updated version of the prokaryotic species concept is "a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions" (Stackebrandt et al. 2002). The whole critical viewpoint here revolves around the adequacy of DNA–DNA reassociation experiments to circumscribe genomically coherent groups.

3. Concept and definition: another exponent example of semantic misunderstanding is the confusion between concept and definition. Both terms are often used as synonyms, but it is important to take into account that distinguishing them may very much help in clarifying our prokaryotic species 'problem'. The species concept is the idea that explains and circumscribes the patterns of recurrence observed in nature. It is the essence of what we think is the basic unit for constructing an operative and predictive classification. Within the concept, we should find the reasons for including or excluding naturally occurring individuals within a category. However, the species definition is the way we recognize that individuals belong to a category. The definition provides a set of parameters that are sufficient to recognize that a certain group of individuals belong to a recurrent pattern in nature. Actually, this responds in the most pragmatic way to identify what we think is a unit. Our reductionistic approach to understanding nature allows us to formulate the simplest way to recognize units (Rosselló-Mora 2003). For example, in this chapter, 'genomic coherency' applies to the concept, whereas the relaxed (or not) results or values of DNA–DNA reassociation experiments would apply to the definition. For example, changing the method and parameters to recognize coherent genomic groups, such as substituting DNA–DNA reassociation experiments (e. g. MLST), would result in a change in how we define species but not how we conceive them. The concept remains the same.

2.3 DNA–DNA Reassociation Measurement, Parameters and Methods

During the almost 50 years of use of whole genome hybridization studies for microbial taxonomy, quite a few techniques have been developed (Table 2.2). All such techniques have in common the measurement of the extent and/or stability of the hybrid double-stranded DNA resulting from a denatured mixture of DNAs incubated under stringent conditions that allow only renaturation of complementary sequences. Actually, the use of different techniques, and their comparisons have been extensively discussed (e.g. Brenner 1978; De Ley and Tijtgat 1970; Goris et al. 1998; Grimont 1988; Grimont et al. 1980; Johnson 1985, 1991; Owen and Pitcher 1985; Stackebrandt and Liesack 1993; Tjernberg et al. 1989). As will be clarified later and despite any apparent diversity, all methods rely on a few common properties with the differences between them being basically variations in the DNA labelling type and/or the measurement technique. It seems that with time, the multiple techniques published have been developed following the need to simplify the manipulation procedures, and allow a larger number of simultaneous measurements.

There are two main strategies for performing reassociation experiments: those where the hybridization reaction is carried out in free solution and those that imply previous fixation of the test DNA onto a solid surface. Among the free-solution methods, the most ancestral required one of the test DNAs to be labelled with heavy isotopes; and the separation of homologous renatured strands from the hybrids was carried out under buoyant density ultracentrifugation procedures (Schildkraut et al. 1961). However, better accuracy in the measurement of hybrid molecules was achieved by the use of radiolabels. A labelled DNA, commonly sheared into small single-stranded polynucleotide molecules, is hybridized against an excess of unlabelled high-molecular-weight target DNA. Double-stranded DNA is then separated from single-stranded unhybridized DNA either by the use of a selective binding to hydroxyapatite (Brenner et al. 1969b), or by the selective digestion of single-stranded DNA with nuclease S1 (Crosa et al. 1973; Popoff and Coynault 1980). Both strategies gauge the measurement of the extent of labelled DNA that has hybridized against an unlabelled target and its comparisons against homologous reassociations. Due to methodological and health concerns, the use of radiolabels is not easily implemented in laboratories, promoting the development and establishment of non-radioactive methods. For example, there is a non-radioactive and miniaturized method equivalent to the original hydroxyapatite method where DNA is double-labelled with biotin- and digoxigenin-modified nu-

Table 2.2. Comparison of methods used for DNA–DNA reassociation experiments, description of their methodological basis and their advantages and disadvantages. *M* Method, *L* labelling, *Ms* measured parameter

Method	Short description	Remarks
Free-solution methods		
M: buoyant density L: heavy isotopes Ms: RBR Schildkraut et al. (1961)	One of the two genomes to test is labelled with a heavy isotope. Renatured strands are separated on a caesium chloride gradient by ultracentrifugation. Quantification is made on the relative amounts of mixed DNA of intermediate molecular weight.	The method was rapidly abandoned due to the advantages of the newly developed methods.
M: hydroxyapatite L: radioactive isotope Ms: RBR; ΔT_m Brenner et al. (1969b) Lind and Ursing (1986)	Radioactively labelled DNA is hybridized in solution under stringent conditions against unlabelled DNA. Single DNA strands are separated from renatured double DNA strands on the basis of physical affinity of a hydroxyapatite matrix. Quantifications are made upon the relative amounts of hybrid DNAs formed in respect to the homoduplex renatured DNA. Likewise, temperature denaturation profiles can be performed to quantify the thermal stability of the hybrids in respect to the homoduplex (ΔT_m). Reference DNAs have to be labelled with radioactive isotopes by either nick translation or iodination. Quantifications are done as specific radioactivity (cpm).	Advantages: multiple simultaneous hybridizations can be performed. The method allows a short protocol for RBR measurements and a longer one for ΔT_m measurements. Hybridizations can be done under stringent conditions and without adding extra components (e.g. formamide, blocking reagents, etc.). The hybridization temperature is not restricted and can be as high as needed. Disadvantages: reference DNAs should be labelled with radioactive isotopes.

Table 2.2. (continued)

Method	Short description	Remarks
M: hydroxyapatite/microtitre plate L: digoxigenin–biotin Ms: RBR Ziemke et al. (1998)	The method is basically a modification of the Brenner et al. (1969b). Double-labelled DNA (biotin/digoxigenin) is hybridized against unlabelled DNA. Single- and double-strand DNAs are separated by the use of hydroxyapatite. Quantifications of both fractions are done on streptavidin-coated microtitre plates, using an alkaline phosphatase-based colorimetric enzymatic bioassay. Reference DNAs have to be double-labelled by using nick translation and modified nucleotides bound to digoxigenin or biotin. Quantifications are done by reading the absorbance at 405 nm in a microtitre plate reader.	Advantages: the same as the radioactive method, but no ΔT_m can be performed due to the lower sensitivity of the method. Labelling is performed with non-radioactive labels. Disadvantages: double-labelling should be done with nick-translation and only RBR can be determined.
M: spectrophotometry L: none Ms: RBR; ΔT_m De Ley et al. (1970) Huß et al. (1983)	Renaturation rates of equimolar mixtures of two DNAs are measured by the resulting slope recorded by the decrease in absorbance at 260 nm, and compared to those of the respective homoduplexes.	Advantages: there is no labelling step in the method and DNAs can be used almost without any manipulation. Disadvantages: Special spectrophotometers are needed. The amounts and quality (length and purity) of the two DNAs in the hybridization should be identical, only pairwise results can be achieved. Perhaps the most time-consuming and DNA quantity- and quality-demanding technique.

Table 2.2. (continued)

Method	Short description	Remarks
M: Fluorimetric L: none Ms: ΔT_m González and Sáiz-Jiménez (2004)	Method comparable to that of spectrophotometry. Unlabelled DNAs are hybridized under stringent conditions. Hybrid denaturation rates are measured by following the decrease of fluorescence of double-strand DNA specific fluorescent dye (e. g. SYBR green I). Quantification is made upon the difference between the thermal mid point of the homoduplex and that of the hybrids (ΔT_m).	Advantages: relatively low amounts of DNAs are needed, there is no labelling step and multiple simultaneous experiments can be performed. Disadvantages: The technique needs the use of real-time fluorescent detectors, such as real-time PCR thermocyclers. As for the spectrophotometric method, DNAs should have identical quality and quantity conditions. The method is new and needs to be validated before being compared with those in routine use.
M: endonuclease L: radioactive isotopes Ms: RBR Crosa et al. (1973): S1-TCA Popoff and Coynault (1980): S1-DE81	Labelled and unlabelled DNAs are hybridized in solution under stringent conditions and further digested with a single-strand-specific endonuclease. Double-strand DNAs are then separated from free radioactive nucleotides either by precipitation (S1-TCA) or filtration (S1-DE81). Amounts of hybrid DNAs are referred to renatured homoduplex DNA. Labelling and quantification are done as in the hydroxylapatite method.	Advantages: the same as for the hydroxyapatite method. Disadvantages: reference DNAs should be labelled with radioactive isotopes.
Bound-DNA methods		
M: agar embedded L: radioactive isotopes Ms: RBR Bolton and McCarthy (1962) Brenner et al. (1969a)	Denatured DNA is embedded in agar, left to solidify and then disaggregated into small fragments by pressing through a mesh. The small fragments of agar containing unlabelled DNA are hybridized against radioactively labelled DNA in solution. Once hybridization is done, the agar is washed and the radioactivity quantified.	Advantages: accuracy of the radioactivity measurements. Disadvantages: accessibility of labelled DNA to the embedded DNA, difficulties in incubating the samples at high temperatures.

Table 2.2. (continued)

Method	Short description	Remarks
M: membrane filters L: radioactive isotopes Ms: RBR; ΔT_m Johnson (1981) Owen and Pitcher (1985) Tjernberg et al. (1989)	Unlabelled and denatured DNAs are covalently bound to a membrane filter and then hybridized against free radioactively labelled DNA. Hybridization is done under stringent conditions where, for example, competition with unlabelled DNA can be performed. Hybridization solution includes blocking reagents to avoid unspecific binding to the membrane filter. RBR measurements are done in comparison to the homoduplex reaction. Likewise, temperature denaturation profiles can be performed to quantify the thermal stability of the hybrids with respect to the homoduplex (ΔT_m). Labelling and quantification are done as in the hydroxyapatite method. Additionally, sample-specific activity can also be measured by the use of autoradiography accompanied by densitometric analysis.	Advantages: multiple simultaneous hybridizations with a single-labelled DNA can be performed. These methods calculating thermal stability of the hybrids are independent of the quality of the bound DNA [e.g. Tjernberg et al. (1989) use lysed cell cultures instead of purified DNA]. Disadvantages: the use of radiolabels and, for RBR calculations, the amount of bound DNA may not be equivalent for the different samples and may saturate the dot. Additionally, DNA can be released from the membrane during the incubation/washing steps, therefore biasing the results.
M: membrane filters L: non-radioactive labels Ms: RBR Jahnke (1994) Cardinali et al. (2000) Gade et al. (2004)	Hybridization conditions are similar to those of the radioactive method, only the detection steps are a bit more tedious. Quantifications are done by using colorimetric enzymatic bioassays and differences are mainly due to the type of detection. There are several quantification possibilities, e.g. soluble enzymatic product (Jahnke 1994), chemiluminescence, film development and densitometric analysis (Cardinali et al. 2000), insoluble precipitated enzymatic product and densitometric analysis (Gade et al. 2004), among others.	Advantages: the same as the radioactive methods, but using non-radioactive labels instead. Disadvantages: the same as the radioactive methods, but with more incubation/washing steps. Due to the sensitivity of the detection only RBR can be determined. Densitometric measurements might not be accurate enough. Being an enzymatic reaction, the development should be done during the period of time when the reaction is linear.

Table 2.2. (continued)

Method	Short description	Remarks
M: microtitre plate-bound DNA L: photobiotin Ms: RBR Ezaki et al. (1989) Adnan et al. (1993) Kaznowski (1995) Christensen et al. (2000)	Unlabelled and denatured DNAs are adsorbed or covalently bound to the wells of a microtitre plate where the hybridization will take place. Reference DNAs are labelled with photobiotin and hybridized against bound DNA. To reduce hybridization temperatures and in order to maintain stringency, formamide is used. The amount of hybridized DNA is revealed after a fluorometric or colorimetric enzymatic bioassay.	Advantages: depending on the protocol, hybridizations are reduced to a few hours. Multiple simultaneous measurements can be performed. Labelling is performed with non-radioactive labels. Disadvantages: as in any assay where DNA is bound, and due to the large number of steps and incubation times, DNA can be released therefore biasing the results. Only RBR can be calculated while no temperature profiles can be used. Hybridization has to include the use of denaturing agents such as formamide to avoid incubations at high temperatures.
M: microtitre plate-bound DNA L: digoxigenin Ms: ΔT_m Mehlen et al. (2004)	Reference and test DNAs are digested with *Sau*3A restriction enzyme and then ligated to oligonucleotide linkers that serve as a target for amplification purposes. Genomic DNAs are then amplified, and the reference DNA is labelled with digoxigenin modified dUTP. Non-labelled DNAs are covalently bound to microtiter plates and hybridized with free-labelled reference DNA. Hybridization is followed by different washes with buffer of ion concentrations of increasing stringency. Detection of bound DNA is revealed after a colorimetric enzymatic bioassay.	Advantages: in principle, there is no need for high quantities of genomic DNAs. Labelling is performed with non-radioactive labels, and detection can be simply done with a standard microtitre reader. The calculations of melting temperatures minimize the effects of unspecific binding, and are unaffected by different amounts of unlabelled microtitre-bound genomic DNA. Disadvantages: any problem derived from the amplification procedure. No RBR can be calculated. Hybridization has to include the use of denaturing agents such as formamide to avoid incubations at high temperatures.

cleotides; and the detection is simply undertaken as a bioassay in microtitre plates (Ziemke et al. 1998). As an alternative to labelling DNA, a spectrophotometric method was developed by De Ley et al. (1970) where a mixture of two unlabelled DNAs of identical quality and concentration are denatured, and their renaturation is optically followed under stringent conditions with a special spectrophotometer. The measurement of reassociation is made by the decrease in absorbance that single-stranded DNA shows when it renatures as a double strand. The extent of hybrid molecules is extrapolated from the comparisons of the differences in the reassociation rates of homologous and heterologous DNAs. Recently, a new fluorometric method that uses a real-time PCR thermocycler has been developed with a similar basis as the spectrophotometric method (González and Sáiz-Jiménez 2004). This method is based on measuring the thermal stability of the hybrid molecules with the use of SYBR green I. Although this method is still to be validated by evaluating the results with other techniques, preliminary comparisons indicate its adequacy (Jurado et al. 2005).

All methods implying fixed DNA rely on the same principle, where the denatured target DNA is bound to a solid surface and then hybridized against a labelled reference DNA in free solution. Labelled DNA is dissolved in a solution with an ionic strength that provides enough stringency to allow only renaturation of complementary strands at a given temperature. Additionally, the hybridization buffer includes several coating compounds that hamper unspecific binding of labelled DNA to the DNA-free solid surface. The first experiments were performed with agar as the solid surface for binding DNA (Bolton and McCarthy 1962). However, such a supporting matrix was rapidly abandoned in favour of the use of macroporous supports such as nitrocellulose or Nylon filters which provided covalent surface binding of the DNA, and thus a minimization of the loss of the target DNA from the support. There are quite a few published procedures using membrane filters, with the main differences between them being basically the type of label for the reference DNA and thus the quantification measurement procedures. DNA can be radiolabelled and the hybridization extent can be either quantified by scintillation (e.g. De Ley and Tijtgat 1970), or by the densitometric measurement of the spot generated through autoradiography (e.g. Amann et al. 1992). However, similar methods have been developed by the use of non-radioactive labels, such as digoxigenin- or biotin-modified nucleotides; and the measurement is carried out after densitometric quantification of the spots generated, for instance, from chemiluminescence on X-ray films (e.g. Cardinali et al. 2000), or directly onto the membrane with a precipitated product (e.g. Gade et al. 2004). A colorimetric measurement with the combined use of microtitre plates has even been used (Jahnke 1994). More modern attempts to combine genomics technology with classic species circumscription have been undertaken by

the use of micro- or macroarrays (Cho and Tiedje 2001; Ramisse et al. 2003; Watanabe et al. 2004). However, most probably if the classic technologies are considered difficult to implement and only a few laboratories use them (Cho and Tiedje 2001; Stackebrandt 2003), the use of genomics technology might be even more restricted.

Finally, one of the most currently applied methods that implies immobilization of DNA onto a solid surface is the one that uses microtitre plates instead of macroporous membranes. The success of these methods relies on the possibility of performing fast and radioactivity-free assays, all in the same container. There are several published methods, but the most known and used is that of Ezaki et al. (1989) which binds the target DNA in the wells of a microtitre plate and the test DNA is labelled with biotin. First, measurements were undertaken by the use of fluorogenic substrates, but later these were substituted by a chemiluminescent substrate and by covalent binding onto the microtitre plate surface (Adnan et al. 1993). However, similar methods have been developed that use colorimetric reactions for the detection (Kaznowski 1995) which, importantly, reduce the cost of the equipment used. Lately, more sophisticated and reliable methods have been developed which allow experimentation with fastidious organisms whose DNA is difficult to recover (Mehlen et al. 2004) and, in this case, genomic DNA is previously amplified before being bound to the microtitre well. Then, digoxygenin-labelled reference DNA is used to perform the hybridization and the stringency is accomplished by washing with decreasing ion strength buffers, which allows a determination of melting profiles for hybrid molecules. Detection is achieved colorimetrically.

Depending on the method used, there are two main parameters that can be determined: the relative binding ratio (RBR) and the increment of melting temperature (ΔT_m). Sometimes the same procedure can provide both parameters, but most of the techniques just provide one or the other (Table 2.2). It is important to note that RBR values especially depend on the stringency of the method used. At a given ionic strength, hybridizations may be carried out under what are considered to be optimal conditions (25–30 °C below the melting point of the reference native DNA, i.e. T_m), under stringent or exacting conditions (10–15 °C below T_m), or under relaxed, non-exacting conditions (30–50 °C below T_m), although most results correspond to optimal-condition experiments (Schleifer and Stackebrandt 1983).

The RBR is the measurement of the extent of double-stranded hybrid DNA for a given pair of genomes relative to that measured for the reference DNA performed under identical renaturation conditions. RBR is expressed as a percentage, considering that the reference genome hybridizes 100% with itself. For those methods that use labelled DNA, large amounts of labelled DNA may still remain as single-stranded DNA after the hybridiza-

tion experiment; and then the binding ratio (BR) is calculated as the extent of double-stranded hybrid DNA in relation to the total labelled DNA added in each single experiment. RBR is then determined by comparing the percent reassociation of each heterologous reaction to that of the homologous reaction, which is considered to be 100%. Spectrophotometric methods calculate the extent of hybrid DNA by basically comparing the reassociation kinetics with those of homologous DNA. The RBR is the most used parameter in the circumscription of species.

A more reliable parameter to determine is the ΔT_m, simply because it is independent from the quantity and quality of the DNAs used for the experiment (Tjernberg et al. 1989). However, ΔT_m requires more time-consuming methods and is generally only achievable using radioactive labels. This parameter is a reflection of the thermal stability of the DNA duplexes. ΔT_m is actually the difference between the melting temperature of a given homologous DNA and that of a hybrid DNA. At a given ionic strength, the melting temperature of a DNA (or thermal denaturation midpoint, T_m; where 50% of DNA strands appear denatured) is directly related to its GC content (Schildkraut and Lifson 1965; Turner 1996). Hybrid DNAs tend to melt earlier. The less related a pair of DNAs, the higher the difference between their melting points (in degrees Celsius), in comparison with their corresponding homologues. This is because a lower base pairing will render a less thermally stable base complementation. When the measurements are carried out with a labelled reference DNA, the melting temperatures are solely related to the extent of base pairing and remain independent from the quality and quantity of each of the DNAs used for the hybridization. Consequently, the results of analysing melting profiles are very reproducible and less subject to experimental error than RBR. However, because of the technical difficulties, RBR is much more popular when trying to calculate raw genome similarities. In principle, the two parameters do not need to be related: RBR reflects the extent of double-stranded DNA with a base complementarity of less than 15% base mispairing (Stackebrandt and Goebel 1994; Ullmann and McCarthy 1973) and ΔT_m reflects the extent of sequence identity. However, it has been demonstrated empirically that there is indeed a linear correlation between them (e. g. Grimont 1988; Johnson 1989; Rosselló-Mora and Amann 2001; Tjernberg and Ursing 1989); and generally values of RBR above 50% correlate with a ΔT_m value below 4–5 °C.

To calculate ΔT_m, multiple-step washing profiles have to be carried out. However, a parameter named $\%DR_7$ was developed to simplify the washing profiles without losing accuracy in the measurements (Tjernberg et al. 1989). $\%DR_7$ is calculated after two steps of washing the hybridized molecules: the first wash is undertaken at 7 °C below the melting temperature of the reference DNA and a second wash is performed at 100 °C in order to achieve complete denaturation. $\%DR_7$ is the amount of DNA released in

the first step as a percentage of the total amount of eluted DNA. Thus, for a given pair of DNAs, the higher the $\%DR_7$, the less they are related. However, although this parameter could have been a good compromise between the accuracy of ΔT_m measurements and the simplicity of RBR calculations, it has never been applied to any great extent.

It is not easy to recommend a method, or a parameter, for circumscribing species when using DNA–DNA reassociation experiments. It is a question of the equipment that one possesses and the accuracy of the measurements that one wants to achieve. The sensitivity of radioactive measurements means these are the ones that provide the most accurate and reproducible data. Actually, such methods generally allow the measurement of both parameters, RBR and ΔT_m; and an additional advantage of using radioactive labels is that, when measuring melting temperatures, the results are independent of the quality and quantity of the DNA. It is even possible to use cell extracts directly and dot-blot them onto filters instead of previously having to isolate high-quality DNA (Rosselló et al. 1991; Tjernberg et al. 1989). The non-radioactive methods are currently the methods of choice, simply because of the security advantages of not using radiolabels. However, it has to be understood that the accuracy may be less because of the larger standard deviations of the experiments. Spectrophotometric methods, like real-time PCR measurements, require the determination of the exact amounts of the DNAs to be used; and for hybridization purposes both should have very similar conditions of quality. Additionally, they can only be undertaken as pair-wise assays, especially spectrophotometric methods; and for multiple determinations the experiment is quite time-consuming. Despite this, such experiments are currently some of the most popular for use in bacterial taxonomy (Table 2.1). The most used methods for determining genome similarities are those that imply attachment of the nucleic acids onto a solid surface, either on a filter or in microtitre plates (Table 2.1). All of them imply either adsorption or covalent attachment of the DNA onto a surface, with the expectation that: (a) identical test DNA amounts are attached per spot/well and (b) the loss of attached bound DNA due to washes and incubations is negligible. Despite this, these methods and especially those using microtitre plates (e.g. Christensen et al. 2000; Ezaki et al. 1989) are the most used (Table 2.1). Microtitre plate methods that use colorimetric bioassays, such as for instance modifications of the Ezaki method (Kaznowski 1995), or those that adapt radioactive methods to miniaturized non-radioactive procedures (Ziemke et al. 1998), may also be chosen because of the lower costs of the equipment used (i.e. regular microtitre plate readers are less expensive than special spectrophotometers, fluorometers or phosphor-imagers, among others).

Most of the methods have been thoroughly compared in order to validate their results (e.g. Christensen et al. 2000; De Ley and Tijtgat 1970; De Ley

et al. 1970; Ezaki et al. 1989; Goris et al. 1998; Grimont et al. 1980; Jahnke 1994; Mehlen et al. 2004; Tjernberg et al. 1989; Ziemke et al. 1998). From the comparisons, it can be deduced that the level of agreement is quite good, especially for those hybridizations of closely related strains; and generally values are above 50%. However, the level of agreement might decrease when the genome similarities are lower, just because the background of the techniques might be different. Additionally, it is important to take into account that the standard deviations are relatively high, especially for those techniques that are non-radioactive, and values might be as high as 8% (Christensen et al. 2000; Johnson 1991; Sneath 1989). Nevertheless, as will be argued later, the evaluation of the hybridization results may be better read as if evaluating, for instance, chemotaxonomic markers, where the patterns shown by the relative amounts of the components are of higher importance than those of each absolute value.

Finally, there is a belief that hybridization methods are difficult to implement in a regular laboratory because of the laborious procedures involved; and they are also of high cost because of the equipment required (e. g. Gillis et al. 2001; Stackebrandt 2003; Stackebrandt et al. 2002; Young 1998). However, I would argue here that this may be true only for such methods that require radiolabels, expensive spectrophotometers, fluorimeters, real-time thermocyclers, or X-ray film exposure and development. The methods adapted to colorimetric measurements (e. g. Kaznowski 1995; Mehlen et al. 2004; Ziemke et al. 1998), in contrast, require nothing more than the regular apparatus found in any microbiology laboratory, such as microtitre readers for visible light (which can be substituted by regular spectrophotometers), water baths, microfuges and even a low-cost thermocycler. The protocols developed are no more laborious than others dealing with molecular techniques; and, once DNA is isolated, the procedures can take one or at most two days.

2.4 Interpretation of Results and the Boundaries for Species Circumscription

The importance of the results generated by DNA–DNA hybridization techniques have been empirically emphasized after years of using such techniques. The original experiments were designed simply to understand raw genome similarities. However, soon the empirical observation that genomically coherent groups (later named genospecies; Ravin 1963) did frequently match phenotypically well defined species (taxospecies) gave paramount importance to hybridization results. Additionally, the occasionally found continua between phenotypically defined groups were usually resolved,

since organisms tended to be either closely related or not (Goodfellow et al. 1997). It is important to note here that DNA–DNA reassociation results are rough estimations of the average genetic relationship of two highly related organisms and that the actual sequence similarity of the compared DNA strands may be significantly higher. The interpretation of DNA–DNA hybridization results acquired predominance in the development of a species concept for prokaryotes; and their use over a period of decades has had an influence that cannot be underestimated. Nowadays, the idea of placing a group of organisms within a single group named 'species' is unavoidably linked to genomic coherency. However, there is a need to substitute such methods by others that give better scientific assistance (Stackebrandt et al. 2002), but such substitution in taxonomy could only be done if the new information retrieved confirms that of the standardized methods.

The genomic size of a species had been empirically circumscribed after the observation of how taxospecies fitted to genospecies. For some, cut-off values above 60% similarity (<7 °C of ΔT_m) would embrace coherent species (Johnson 1973). However, others might find more robustness by setting the boundaries as high as 80% similarity (<5 °C of ΔT_m; Grimont 1988). All such observations made an ad hoc committee recommend that a robust species definition could be circumscribed by the inclusion of organisms sharing more than 70% DNA similarity, or less than 5 °C ΔT_m (Wayne et al. 1987). However, such values were only a recommendation, since it had also been empirically observed that there was a transitional range of values (between 50 – 80% similarity, or 5–7 °C ΔT_m) where subgrouping could sometimes be complicated because different taxospecies could appear within a single genospecies and vice versa (Grimont 1988; Johnson 1989). Despite this, many scientists took the value of 70% as a rigid boundary for species circumscription, thereby unnecessarily forcing their descriptions (Rosselló-Mora 2003). Re-evaluations of the species definition have led to recommendations of more relaxed boundaries without rigid genomic boundaries for species circumscriptions but, in addition, the sound re-evaluation of such results, using additional taxonomic parameters (e.g. Stackebrandt et al. 2002; Ursing et al. 1995). It is clear that the original recommendations were produced after empirical observations were made with easily cultured organisms, such as enterobacteria (Grimont 1988; Stackebrandt 2003), anaerobic low-GC Gram-positive or Gram-negative organisms (Johnson 1973), or pseudomonads (Palleroni 2003). However, the use has undoubtedly been extended to a much wider range of organisms, as can be seen in the many new classifications. Given the vast diversity expected in the prokaryotic world (Whitman et al. 1998), it is clear that the parameters used to circumscribe the basic unit of diversity may not equally fit all organisms. Trying to evaluate the whole of microbial diversity with a single measuring stick is a reductionistic approach that

cannot be sound, especially if the parameters used in circumscriptions are taken as being rigid and immutable (Rosselló-Mora 2003).

The taxonomic schema should follow a pragmatic approach in order to provide the scientific community with an operative system (Rosselló-Mora and Kämpfer 2004; Young 2001). In this regard, it is accepted that the circumscription of the basic unit of prokaryotic classification should be based on the simultaneous evaluation of multiple parameters that cover both genomic properties and phenotype and that no single parameter is given undue prominence (Stackebrandt et al. 2002; Vandamme et al. 1996). DNA-DNA reassociation may not be regarded as the 'gold standard' for circumscribing species; but it has to be evaluated within the framework of a collection of parameters showing coherency in both genomic and phenotypic terms. For pragmatic reasons, it is recommended not to classify new species if one or either premise fails (Stackebrandt et al. 2002). For example, a clear-cut genomic group based on reassociation experiments that cannot be phenotypically distinguished from its related organisms may be regarded as a genomovar of a single species (Ursing et al. 1995). In a similar way, a clear-cut phenotypic group that cannot be genomically distinguished from its closest relatives should be considered as a biovar (Sneath 1992). Circumscription of a species within the framework of taxonomy must not simply rely on DNA-DNA reassociation results, although these are of paramount help to understand if one is dealing with a coherent group of strains that can be discriminated from their closest relatives.

Finally, there are some anecdotal examples where the relevance of DNA-DNA hybridization results has been disregarded when circumscribing prokaryotic species. Cases such as maintaining *Neisseria gohorreae* and *N. meningitidis* in two different species although genomically they should be one, or separating two genera such as *Shigella* and *Escherichia*, as well as many other examples for genera like *Yersinia*, *Bacillus*, *Brucella*, etc., respond to pragmatic reasons for their identification, often because of their medical implications. This was clearly stated by an ad hoc committee (Wayne et al. 1987) as: "phenotypic characteristics should agree with this definition and would be allowed to override the phylogenetic concept of species only in a few exceptional cases". This statement has also been ignored by many readers and such incongruities have been interpreted as unwarrantable pitfalls of the taxonomic principles (e.g. Palys et al. 1997; Sneath 1989; Stackebrandt 2003). It is worth emphasizing at this point that taxonomy pursues the construction of an operative, predictive and generally applicable classification schema. If the operationality of the system leads towards an impracticable but exhaustive classification, then the aim of taxonomy has failed. For pragmatic reasons, taxonomists are tolerant to the pitfalls of the measurements.

2.5 The Impact of DNA–DNA Hybridizations on the Conception of a Species and Changes in the Concept and/or the Definition

It is important to note here that the species is an artificial construct of the human mind basically addressed to classify the patterns of recurrence that can be observed in nature (Hey et al. 2003). The understanding of the prokaryotic world improved in parallel to technological developments, but some of these improvements have simultaneously fastened certain criteria in scientific belief, which over time have become tenets. One finds clear examples in prokaryotic taxonomy. The discrete units circumscribed by DNA–DNA reassociation which mostly agreed with a phenotypic framework were taken to represent those recurrence patterns understandable as species. That principle permitted the establishment of a rather stable and operative classification system for prokaryotes (Stackebrandt et al. 2002). However, there are criticisms of current circumscription because it is too conservative and because, by using the DNA–DNA reassociation circumscription criteria, no comparisons with higher eukaryote taxonomies can be carried out (see Rosselló-Mora and Amann 2001; Staley 2004). As is thoroughly discussed in eukaryotic taxonomy, the patterns of recurrence may be necessarily different for different kinds of organisms that exhibit distinct levels of morphological and/or physiological complexity (Hey 2001); and, thus, the parameters used to circumscribe species may be different for different taxonomies. Additionally, for given kinds of organism, one can view them from a variety of perspectives and, since each perspective is legitimate (Hull 1997), it is a question of accepting that pluralism in taxonomy may solve the so called 'species problem' (Ereshefsky 1998; Rosselló-Mora 2003; Young 2001). Taking such premises into account, a universal species concept may be impossible to achieve; and the basic essence of the prokaryotic species may not be comparable to any other species originating from other taxonomies. However, this is perhaps the most pragmatic position.

The principle of genomic coherency based on DNA–DNA reassociation results has had an influence on prokaryotic taxonomy comparable to that of 'breeding true' in the animal and plant species concept. The finding of a parameter that seems to unify criteria towards the recognition of recurrence patterns soon materializes as a tenet. For example, it is clear now that the 'breeding true' concept, which is the basis for the biological species concept (Mayr 1942), can no longer be taken as a universal parameter to embrace all eukaryotic species; and this has brought decades of heated debates (for reviews, see e. g. Hull 1997; Mayden 1997). Actually, the history of microbial taxonomy repeats that of eukaryotes and, in parallel to the

understanding of the extent of the organism's diversity, the validity of the circumscription parameters tends to be relative. Once, DNA–DNA reassociation was considered to be 'the gold standard' for many taxonomists for circumscribing species, such experiments were mostly used in the new descriptions and its use was even more reinforced after the recommendation of an ad hoc committee (Wayne et al. 1987). To date, taxonomists have succeeded in formulating a classification system of about 5,000 species, many of them circumscribed after DNA–DNA hybridization experiments were made available. Any change in the definition of the species should take that fact into account.

As has been discussed, the methods providing raw genomic similarities are submitted to a relatively large experimental error, in addition to impractical properties such as the impossible construction of an interactive and cumulative database (Sneath 1989; Stackebrandt 2003). These are indeed important pitfalls of the method that can lead to its use towards the emerging technologies being questioned (Stackebrandt et al. 2002). Actually, an ad hoc committee for the evaluation of the current definition of species (Stackebrandt et al. 2002) has recommended the search for new methods to replace the use of DNA–DNA reassociation experiments. Special emphasis is being placed on the evaluation of methods such as: (a) sequencing protein-coding genes, an extension of MLST, or (b) DNA profiling, such as AFLP, ribotyping, REP-PCR, or PCR-RFLP. However, any method that is to be used as a substitute for DNA–DNA reassociation should be previously validated. The reluctance to use sets of genes for their phylogenetic evaluation is mainly due to the difficulties in selecting them and designing proper amplification primers and, as criticized for the 16S rDNA analysis, also corresponds to the insignificant portion of the genome that they represent. Indeed, there have been some attempts to design universal primers for some of the reduced sets of universally present genes, but only with about 60% amplification success (Santos and Ochman 2004). Primer redesign or improvement can only be carried out if the genome of closely related organisms is available. However, this approach becomes very impractical when the new isolates belong to unknown phyla. Yet, it seems that there might be a correlation between some single gene sequence identities and genomic similarities (Zeigler 2003), especially the *recN* and *dnaX* genes that have been selected as being discriminative between species. However, as the author also claims, it is too soon to place strong emphasis on this because the data set used was very limited, and all genomes analysed belonged to pathogenic or human saprophytic microbiota.

In principle, reassociation experiments represent raw data on whole genome comparisons, which is an advantage for those techniques that analyse a reduced portion of the genome (Mallet and Willmott 2003; Young 1998). Of course, the best substitute for reassociation experiments in tax-

onomy would be pure genome comparisons after undertaking complete sequencing programs, but this is still utopian because of the relatively high costs of sequencing. Despite the technical difficulties in achieving complete genomes, the first insights into their comparisons and the concordances with classic taxonomic circumscriptions are ongoing, and encouraging (Zeigler 2003). For example, Konstantinidis and Tiedje (2005) carried out an exhaustive comparative survey of about 70 closely related and completely sequenced genomes and their corresponding hybridization values. The best parameter found for taxonomic purposes was the average nucleotide identity (ANI) of shared genes. The values obtained correlated with both 16S rRNA gene sequence identity and DNA–DNA similarity values with pairwise comparisons. Nevertheless, it is still too soon to be able to use this parameter, since there are many comparisons still to carry out before it can be validated. However, the final goal of such techniques in taxonomy should be to undertake the comparisons using the understanding of the information behind the genes or genomes that are under study. Ignoring this fact and treating genes or genome information as mere quantitative data would mean that the substitution would not result in an improvement to the use of DNA–DNA reassociation experiments.

2.6 Epilogue

The species concept for prokaryotes has been especially devised by taxonomists to create an operative and predictive classification system. The first formulation of what a species could be was made by Aristotle about 2,400 years ago and the idea of species was understood as being the basis for a hierarchic classification schema. Since then, the concept of 'species' should be regarded as a property of taxonomy; and its formulation has been improved by taxonomists in parallel to conceptual and methodological scientific developments. Other uses of the term to name essentially different units has led to heated debates, but as Sneath (1988) remarked, taxonomy has been the primary basis for conceptual developments in evolution (and I would say also ecology). The species concept for prokaryotes is well consolidated in microbial taxonomy, but of course it can be improved. DNA–DNA reassociation results gave, for the first time, a measurable way to circumscribe units and therefore the use of the method was established as a priority when classifying new species. This gave the concept a 'genomic coherency' dimension that cannot be misinterpreted and which may equally apply to the 'phenotypic coherency' and 'monophyly' dimensions provided by established taxonomic approaches. Consequently, I am reasonably confident that most taxono-

mists would agree that it is the best concept that we can achieve at this point in time.

DNA–DNA reassociation experiments applied to taxonomy should be taken as a method that allows raw genomic coherency to be understood. This means that, when analysing a group of strains that appear monophyletic and are genetically and phenotypically related, the hybridization results will help to show if they belong to the same genomic circumscription or not. Rigid boundaries, such as 70%, are not to be taken dogmatically, but one has to understand that the classification of new species should follow pragmatic and logical premises. In some cases, a defined phenotypic and genetic group will be circumscribed by cut-off values of 60% or even 50%, but they could still be considered as a single species. In other cases, if the phenotypic and genetic information supports them, two different species may even be distinguished by cut-off values of 80%. The most important point here is that when describing new species, no single value can be given undue prominence, and, altogether, the information retrieved should show enough consistency for the classification. Classifying new species when they cannot be differentiated from their closest relatives hinders the operationality of the system. The aim of a taxonomist is not the classification of everything as a means to an end, but to provide a system that can be used by the rest of the scientific community who find it easy, useful and workable.

DNA–DNA reassociation experiments have been predominantly taken as the measuring basis for circumscribing species for nearly half a century. Most of the current taxonomic schema have been constructed with them and they have been of paramount importance in the way we understand prokaryotic classification. Nevertheless, such techniques suffer from important disadvantages, especially when compared with the newly emerging molecular approaches. Sooner or later, DNA–DNA reassociation will be replaced by analyses that provide more accurate measurements and cumulative databases. However, given the influence that genomic similarities have had on the circumscription of most of the species during the construction of the current classification schema, new methodologies may have to reproduce similar observations. Whole genome sequence comparisons are surely the choice for replacement, and parameters such as ANI could be of enormous help in understanding genomic coherency. This will be true though only if these new species definitions render units that are comparable to the hitherto classified species and that represent the basic structure of our current, indeed defective, but operative and predictive taxonomic classification system for prokaryotes. However, for the time being and until whole genome sequencing is as routine as single gene sequencing is now, DNA–DNA reassociation experiments will have to be used to circumscribe species.

Acknowledgements. I want to thank Rudolf Amann, Hans-Jürgen Busse, Peter Kämpfer, Kostas Konstantinidis, Jorge Lalucat, Karl-Heinz Schleifer, Johannes Sikorsky, Jim Staley, Maria Valens, and Brian Tindall for reviewing and improving the manuscript by giving helpful recommendations; and I thank Chris Rodgers for his excellent corrections of the manuscript. R.R.M.'s research is supported by grants BOS-2003-05198-C02-01 and -02 from the Spanish Ministry of Science and Technology.

References

Adnan S, Li N, Miura H, Hashimoto Y, Yamamoto H, Ezaki T (1993) Covalently immobilized DNA plate for luminometric DNA–DNA hybridization to identify viridans streptococci in under 2 hours. FEMS Microbiol Lett 106:139–142

Amann RI, Lin C, Key R, Montgomery L, Stahl D (1992) Diversity among *Fibrobacter* isolates: Towards a phylogenetic classification. System Appl Microbiol 15:23–31

Bolton ET, McCarthy BJ (1962) A general method for the isolation of RNA complementary to DNA. Proc Nat Acad Sci USA 48:1390–1397

Brenner DJ (1978) Characterization and clinical identification of *Enterobacteriaceae* by DNA hybridization. Prog Clin Pathol 7:71–17

Brenner DJ, Cowie DB (1968) Thermal stability of *Escherichia coli–Salmonella typhymurium* deoxyribonucleic acid duplexes. J Bacteriol 95:2258–2262

Brenner DJ, Fanning GR, Johnson KE, Citarella RV, Falkow S (1969a) Polynucleotide sequence relationships among members of Enterobacteriaceae. J Bacteriol 98:637–650

Brenner DJ, Fanning GR, Rake AV, Johnson KE (1969b) Batch procedure for thermal elution of DNA from hydroxyapatite. Anal Biochem 28:447–459

Brigandt I (2002) Species pluralism does not imply species eliminativism. Phyl Sci 70:1305–1316

Cardinali G, Liti G, Martini A (2000) Non-radioactive dot-blot DNA reassociation for unequivocal yeast identification. Int J Syst Evol Microbiol 50:931–936

Cho J-C, Tiedje JM (2001) Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol 67:3677–3682

Christensen H, Angen Ø, Mutters R, Olsen JE, Bisgaard M (2000) DNA–DNA hybridization determined in micro-wells using covalent attachment of DNA. Int J Syst Evol Microbiol 50:1095–1102

Crosa JH, Brenner DJ, Falkow S (1973) Use of a single-strand specific nuclease for analysis of bacterial and plasmid deoxyribonucleic acid homo- and hetero-duplexes. J Bacteriol 115:904–911

De Ley J, Cattoir H, Reynaerts A (1970) The quantitative measurement of DNA hybridization from renaturation rates. Eur J Biochem 12:133–142

De Ley J, Tijtgat R (1970) Evaluation of membrane filter methods for DNA–DNA hybridization. Antonie Van Leeuwenhoek 36:461–474

Ereshefsky M (1994) Some problems with the Linnaean hierarchy. Phyl Sci 61:186–205

Ereshefsky M (1998) Species pluralism and anti-realism. Phyl Sci 65:103–120

Euzéby JP, Tindall BJ (2004) Valid publication of new names or new combinations: making use of the validation lists. ASM News 70:258–259

Ezaki T, Hashimoto Y, Yabuuchi E (1989) Fluorometric deoxyribonucleic acid–deoxyribonucleic acid hybridization in microdilution wells as an alternative to membrane filter hybridization in which radioisotopes are used to determine genetic relatedness among bacterial strains. Int J Syst Bacteriol 39:224–229

Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16:227–231

Fox GE, Wisotzkey JD, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170

Gade D, Schlesner H, Glöckner FO, Amann R, Pfeiffer S, Thomm M (2004) Identification of Planctomycetes with order-, genus-, and strain-specific 16S rRNA-targeted probes. Microbiol Ecol 47:243–251

Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the Prokaryotes. In: Boone DR, Castenholz RW, Garrity GM (eds) Bergey's manual of systematic bacteriology, 2nd edn, rel 5.0. Springer, Berlin Heidelberg New York, DOI:10.1007/bergeysoutline200405

Gillis M, Vandamme P, De Vos P, Swings J, Kersters K (2001) Polyphasic taxonomy. In: Boone DR, Castenholz RW, Garrity GM (eds) Bergey's manual of systematic bacteriology, 2nd edn. Springer, Berlin Heidelberg New York, pp 43–48

González JM, Sáiz-Jiménez C (2004) A simple fluorimetric method for the estimation of DNA–DNA relatedness between closely related microorganisms by thermal denaturation temperatures. Extremophiles 9:75–79

Goodfellow M, Manfio GP, Chun J (1997) Towards a practical species concept for cultivable bacteria. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: the units of biodiversity. Chapman & Hall, London, pp 25–59

Goris J, Suzuki K-I, De Vos P, Nakase T, Kersters K (1998) Evaluation of a microplate DNA–DNA hybridization method compared with the initial renaturation method. Can J Microbiol 44:1148–1153

Grimont PAD (1988) Use of DNA reassociation in bacterial classification. Can J Microbiol 34:541–546

Grimont PAD, Popoff MY, Grimont F. Coynault C, Lemelin M (1980) Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. Curr Microbiol 4:325–330

Hey J (2001) The mind of the species problem. Trends Ecol Evol 16:326–329

Hey J, Waples RS, Arnold ML, Butlin RK, Harrison RG (2003) Understanding and confronting species uncertainty in biology and conservation. Trends Ecol Evol 18:597–603

Hull DL (1997) The ideal species concept – and why we can't get it. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: the units of biodiversity. Chapman & Hall, London, pp 357–380

Huß VAR, Festl H, Schleifer KH (1983) Studies on the spectrometric determination of DNA hybridization from renaturation rates. Syst Appl Microbiol 4:184–192

Jahnke K-D (1994) A modified method of quantitative colorimetric DNA–DNA hybridization on membrane filters for bacterial identification. J Microbiol Methods 20:237–288

Johnson JL (1973) Use of nucleic acid homologies in the taxonomy of anaerobic bacteria. Int J Syst Bacteriol 23:308–315

Johnson JL (1981) Genetic characterization. In: Gerhardt P, Murray RGE, Costilow RN, Nester EW, Wood WA, Krieg NR, Philips GB (eds) Manual of methods for general microbiology. ASM, Washington, D.C., pp 450–472

Johnson JL (1985) DNA reassociation and RNA hybridisation of bacterial nucleic acids. Methods Microbiol 18:33–74

Johnson JL (1989) Nucleic acids in bacterial classification. In: Williams ST, Sharpe ME, Holt JG (eds) Bergey's manual of systematic bacteriology, vol 4. Williams and Wilkins, Baltimore, pp 2306–2309

Johnson JL (1991) DNA reassociation experiments. In: Stackebrandt E, Goodfellow M (eds) Nucleic acid techniques in bacterial systematics. Wiley, pp 21–44

Jurado V, Láiz L, González JM, Hernández-Marine M, Valens M, Sáiz-Jimenez C (2005) *Phyllobacterium catacumbae*, sp. nov., a member of the Rhizobiales isolated from roman catacombs. Int J Syst Evol Microbiol (in press)

Kaznowski A (1995) A method of colorimetric DNA–DNA hybridization in microplates with covalently immobilized DNA for identification of *Aeromonas* spp. Med Microbiol Lett 4:362–369

Krieg N (1988) Bacterial classification: an overview. Can J Microbiol 34:536–540

Konstantinidis KT, Tiedje JM (2005) Genomic insights to advance the species definition for prokaryotic species. Proc Natl Acad Sci USA, DOI: 10.1073/pnas.0409727102

Lind E, Ursing J (1986) Clinical strains of *Enterobacter agglomerans* (synonyms: *Erwinia herbicola, Erwinia mellitiae*) identified by DNA–DNA hybridization. Acta Pathol Microbiol Immunol Scand B 94:250–231

Mallet J, Willmott K (2003) Taxonomy: renaissance of tower of Babel? Trends Ecol Evol 18:57–59

Martínez-Murcia AJ, Benlloch S, Collins MD (1992) Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16 s ribosomal DNA sequencing: lack of congruence with results of DNA–DNA hybridizations. Int J Syst Bacteriol 42:412–421

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF, Dawah HA, Wilson MR (eds.) Species: the units of biodiversity. Chapman & Hall, London, pp 381–421

Mayr E (1942) Systematics and the origin of species from the view point of a zoologist. Columbia University, New York

Mehlen A, Goeldner M, Ried S, Stindl S, Ludwig W, Schleifer K-H (2004) Development of a fast DNA–DNA hybridization method based on melting profiles in microplates. System Appl Microbiol 27:689–695

Mindell DP, Meyer A (2001) Homology evolving. Trends Ecol Evol 16:434–439

Mishler BD, Donoghue MJ (1982) Species concepts: a case for pluralism. Syst Zool 31:491–503

Owen RJ, Pitcher D (1985) Current methods for estimating DNA base composition and levels of DNA–DNA hybridization. In: Goodfellow M, Minikin E (eds) Chemical methods in bacterial systematics. Academic, London, pp 67–93

Palleroni NJ (2003) Prokaryote taxonomy of the 20th century and the impact of studies on genus *Pseudomonas*: a personal view. Microbiology 149:1–7

Palys T, Nakamura LK, Cohan FM (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol 47:1145–1156

Popoff M, Coynault C (1980) Use of DEAE-cellulose filters in the S1 nuclease method for bacterial deoxyribonucleic acid hybridization. Ann Microbiol 131A:151–155

Ramisse V, Balandreau J, Thibault F, Vidal D, Vergnaud G, Normand P (2003) DNA–DNA hybridization study of *Burkholderia* species using genomic DNA macro-array analysis coupled to reverse genome probing. Int J Syst Evol Microbiol 53:739–746

Ravin AW (1963) Experimental approaches to the study of bacterial phylogeny. Am Nat 97:307–318

Reydon TAC (2004) Why does the species problem still persist? Bioessays 26:300–305

Rosselló R, Garcia-Valdés E, Lalucat J, Ursing J (1991) Genotypic and phenotypic diversity of *Pseudomonas stutzeri*. System Appl Microbiol 14:150–157

Rosselló-Mora R (2003) Opinion: the species problem, can we achieve a universal concept? System Appl Microbiol 26:323–326

Rosselló-Mora R, Amann, R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67

Rosselló-Mora R, Kämpfer P (2004) Defining microbial diversity – the species concept for prokaryotic and eukaryotic microorganisms. In: Bull AT (ed) Microbial diversity and bioprospecting. ASM, Washington, D.C., pp 29–39

Santos SR, Ochman H (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environ Microbiol 6:754–759

Schildkraut C, Lifson S (1965) Dependence of the melting temperature of DNA on salt concentration. Biopolymers 3:195–208

Schildkraut CL, Marmur J, Doty P (1961) The formation of hybrid DNA molecules and their use in studies of DNA homologies. J Mol Biol 3:595–617

Schleifer K-H, Stackebrandt E (1983) Molecular systematics of prokaryotes. Annu Rev Microbiol 37:143–187

Sneath PHA (1988) The phenetic and cladistic approaches. In: Hawksworth DL (ed) Prospects in systematics. Systematics Association/Clarendon, Oxford, pp 252–273

Sneath PHA (1989) Analysis and interpretation of sequence data for bacterial systematics: the view of a numerical taxonomist. System Appl Microbiol 12:15–31

Sneath PHA (1992) International code of nomenclature of bacteria, 1990 revision. American Society for Microbiology, Washington, D.C.

Stackebrandt E (2003) The richness of prokaryotic diversity: there must be a species somewhere. Food Technol Biotechnol 41:17–22

Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44:846–849

Stackebrandt E, Liesack W (1993) Nucleic acids and classification. In: Goodfellow M, O'Donnell AG (eds) Handbook of new bacterial systematics. Academic, London, pp 151–194

Stackebrandt E, Frederiksen W, Garrity G, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047

Staley JT (2004) Speciation and bacterial phylospecies. In: Bull AT (ed) Microbial diversity and bioprospecting. ASM, Washington, D.C., pp 40–48

Tindall BJ (2002) Prokaryotic systematics: a theoretical overview. Encycl Life Sci 15:244–251

Tjernberg I, Ursing J (1989) Clinical strains of *Acinetobacter* classified by DNA–DNA hybridization. APMIS 97:595–605

Tjernberg I, Lindth E, Ursing J (1989) A quantitative bacterial dot method for DNA–DNA hybridization and its correlation to the hydroxyapatite method. Curr Microbiol 18:77–81

Turner, DJ (1996) Thermodynamics of base pairing. Curr Opin Struct Biol 6:299–304

Ullman SJ, McCarthy BJ (1973) The relationship between mismatched base pairs and the thermal stability of DNA duplexes. Biochim Biophys Acta 294:416–424

Ursing JB, Rosselló-Mora RA, Garcia-Valdes E, Lalucat J (1995) Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. Int J Syst Bacteriol 45:604

Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Watanabe T, Murata Y, Oka S, Iwahashi H (2004) A new approach to species determination for yeast strains: DNA microarray-based comparative genomic hybridization using a yeast DNA microarray with 6000 genes. Yeast 21:351–365

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Trüper HG (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583

Young JPW (1998) Bacterial evolution and the nature of species. In: Carvalho GR (ed) Advances in molecular ecology. IOS, Amsterdam, pp 119–131

Young JM (2001) Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy. Int J Syst Evol Microbiol 51:945–953

Zeigler DR (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol 53:1893–1900

Ziemke F, Höfle MG, Lalucat J, Rosselló-Mora R (1998) Reclassification of *Shewanella putrefaciens* Owen's genomic group II as *Shewanella baltica* sp. nov. Int J Syst Bacteriol 48:179–186

3 DNA Fingerprinting Techniques Applied to the Identification, Taxonomy and Community Analysis of Prokaryotes

Rüdiger Pukall

3.1 Introduction

The characterization of bacteria or microbial communities at the genotypic level is of crucial importance to medical, industrial and environmental microbiology, as well as microbial ecology and taxonomy. Compared to phenotypic testing, molecular methods based on the investigation of total DNA or segments of DNA are superior because the analysis is independent of possible variations in cultures due to growth and media conditions (e. g. temperature, pH, composition of media). Furthermore, other methods such as serotyping have been shown to be an excellent method for typing certain strains, e. g. *Salmonella*. However, the discriminatory power of serotyoing may be low for other groups of strains. For example, most strains of *Staphylococcus aureus* (Karakawa et al. 1985) express a single serotype only. During recent years, a broad spectrum of DNA-fingerprinting techniques have been developed, covering all ranks between phylum and strains, including those taxa that have not yet been cultured in the laboratory. At the beginning of the molecular era, pulsed-field gel electrophoresis was applied to the discrimination of yeast strains (Schwartz and Cantor 1984). In the following decade, the resolution power of genes coding ribosomal RNA for the identification and taxonomy of species (Pace et al. 1986; Woese 1987) was investigated. Researchers can now choose from a wide spectrum of techniques, spanning applications such as the identification and authentification of strains, phylogenetic analysis and the elucidation of microbial epidemiology and population structures.

Today, the most detailed form of typing is full-genome sequencing. To date, 210 microbial genomes have been completed and the sequences deposited in public databases; and a list of these genomes is given at www.ncbi.nlm.nih.gov/genomes/proks.cgi. In addition, 536 ongoing prokaryotic genome projects are listed in the genomes online database (GOLD),

Rüdiger Pukall: DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Mascheroder Weg 1b, 38124 Braunschweig, Germany, E-mail: rpu@dsmz.de

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

available at www.genomesonline.org. Information from different gene loci led to the development of PCR primers used in phylogenetic analysis (see Chap. 5), comparative DNA typing (multi-locus sequence typing (MLST, see Chap. 6), multi-locus variable number tandem repeat analysis (MLVA, see Chap. 4) and the selection of targets in the DNA microarray technology (see Chap. 9). Full genomic sequences are still rare today for multiple isolates of a single microbial species. However, the need to provide a broader basis of sequence information is obvious in order to rapidly identify and discriminate at the molecular level in clinical environments, to search for the mechanisms of speciation in nature and to look into the function of microorganisms in their habitat (see Chap. 8). Nevertheless, even with the lack of sufficient molecular information, sequences of genes and gene clusters are regularly evaluated for their use in designing appropriate methodologies to discriminate at the inter- and intraspecific level (see Chap. 7). These methods often follow a "trial and error" strategy, such as those targeting whole genomes (PFGE, RAPD-PCR, REP-PCR, AFLP), gene clusters (ribotyping of rrn operons), parts of gene clusters (intergenetic 16S–23S rDNA spacer regions) or individual genes (ARDRA of 16S rDNA, T-RFLP, SSCP, DDGE, TGGE).

For some of these approaches, automation and standardization procedures have been utilized to maximize reproducibility among laboratories. Also, computer programs have been developed for translating typing data into coherent genetic profiles and to calculate inter-strain relatedness in a combinatorial manner (e.g. BioNumerics, Applied Math).

Most methods listed above have been extensively used in the past decade for the characterization of pro- and eukaryotic pathogenic microorganisms, epidemiological typing, classification and sub-typing of producing strains and contaminations in food microbiology and biotechnology. They were also indispensable in polyphasic taxonomy and are responsible for the breakthrough in the elucidation of microbial diversity in different environments. The choice of the DNA typing method used for a given application depends strongly on the needs and resources available in a laboratory. In general, the different molecular typing methods described below vary within the following criteria: discriminatory power, reproducibility, ease of performance, processing time in which data can be obtained, interpretation of data (e.g. equipment, computerized analysis), standardization for intra- and interlaboratory use, expense and, finally, experience.

3.2 DNA Typing Methods

3.2.1 DNA Typing Methods Targeting the Whole Genome of a Bacterial Strain

Pulsed Field Gel Electrophoresis for Macro-restriction of Total DNA

Conventional agarose gel electrophoresis is restricted to the separation of DNA fragments < 50 kb. To overcome this limitation, pulsed field gel electrophoresis (PFGE) was developed, which allows clear separation of large DNA fragments of up to 1,000 kb in size. For the extraction of high-molecular-weight genomic DNA, bacterial cells are harvested from a freshly grown culture. Cells are resuspended in saline/EDTA buffer and aliquots are embedded in a low-melting-point agarose in order to prevent DNA release and non-specific fragmentation. The most critical point is the cell concentration of the suspension embedded in the agarose plug. Cell suspensions of up to 10^9 cells/ml (equivalent to McFarland standard 3) are used, but it is recommended to pour several plugs with different cell densities to obtain at least one well lysed cell batch. The agarose blocks containing the cells are then treated with lysozyme, proteinase K and Pefabloc SC solution for irreversible inactivation of enzymes. After washing the plugs in TE buffer, the released chromosomal DNA is digested with a restriction endonuclease that recognizes only a small number of fragments (rare cutter). The resulting fragments can be separated in an agarose gel which is subjected to alternate multidirectional fields. Normally, up to 20 fragments ranging over 10–800 kb in length are generated from macro-restriction (Olive and Bean 1999). After electrophoresis, the gel is stained with ethidium bromide for detection of the DNA profile. Recognition sites for restriction endonucleases are highly specific, but may be changed by single-base substitution or major changes such as insertions, deletions or transpositions. Variation in the pattern of fragments obtained is called restriction fragment length polymorphism (RFLP). PFGE represents one type of RFLP, others are discussed below. PFGE is a reliable tool for the determination of the genome size of a bacterium, including the analysis of its genome organization. PFGE was used by Pradella et al. (2002, 2004) to discriminate between *Desulfurella* strains and has also been used in the characterization of lactic acid bacteria, including probiotic strains and other biotechnological relevant bacteria (Tynkkynen et al. 1999; Ventura and Zink 2002b; Ventura et al. 2003; Yeung et al. 2004).

Compared to other typing methods, PFGE is laborious, expensive and only recommended for the characterization of small sets of isolates. Due to

its excellent reproducibility and discriminatory power, PFGE is widely used as a "gold standard" for epidemiological typing of bacterial pathogens, including multi-resistant *Staphylococcus aureus* strains. Furthermore, guidelines for the interpretation of PFGE-derived fingerprints of genetically related strains (clones) were published by Tenover et al. (1995), including interpretation of differences in the number and length of fragments caused by genetic events. Guidelines for the recognition of clones of *Streptococcus pneumoniae* using molecular typing methods like BOX-PCR (see below), PFGE and MLST were published by McGee et al. (2001). PFGE provides a high level of discrimination which may be useful for local epidemiological studies of, e.g. *S. pneumoniae* clones, whereas MLST provides a more rigorous way of assigning isolates to individual clonal clusters. In McGee's study, isolates whose allelic profile differed at three or more of the seven loci analysed were regarded as distinct clones (see MLST, Chap. 6).

Standardized protocols have been developed in networked projects in order to increase the intra- and inter-laboratory reproducibility for typing specific organisms. Protocols for the characterization of strains affiliated to the genera *Salmonella*, *Escherichia*, *Campylobacter*, *Shigella* and *Listeria* were published by the molecular subtyping network for food-borne bacterial disease surveillance "PulseNet" (Graves and Swaminathan 2001) and the German part of the network (summarized at www.foodborne-net.de). Harmonization of PFGE protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus* was accomplished within HARMONY, a European Union-funded project (Murchan et al. 2003). Recommended procedures are given at the homepage (www.harmony-microbe.net/microtyping.htm). Standardization of DNA preparation and digestion was not considered necessary for reproducibility. Rather, standardization of running conditions and electrophoretic parameters was found to be an absolute requirement (concentration of agarose, standard gel volume, DNA concentration within the plug, ionic strength and volume of running buffer, running temperature, voltage, switching times).

Random Amplified Polymorphic DNA Assay

Random amplified polymorphic DNA (RAPD) analysis was first described by Williams et al. (1990) and Welsh and McClelland (1990) when they fingerprinted bacterial genomes using arbitrary primers for the amplification of DNA polymorphisms in a PCR-dependent approach. Therefore, this assay is often referred to as "arbitrarily primed PCR" (AP-PCR). In the experiment published by Welsh and McClelland (1990), M13-Primer was used as a single primer for the amplification of DNA stretches in order to detect polymorphisms in different *Staphylococcus* strains. Normally, short random primers are used, 9–10 bases in length, which hybridize with suf-

ficient affinity to multiple loci on the chromosomal DNA at low annealing temperatures. The number and location of these random primer sites may vary in the genome of different strains of a species. Therefore, the selection of primer and conditions which generate the best pattern for differentiation of strains must be evaluated empirically. DNA fragments obtained by PCR can be separated by size, using conventional agarose gel electrophoresis. This method is easy to perform and may be used for processing a large number of strains. RADP analysis has been used for the analysis of microbial diversity and in a number of studies focussed on strain differentiation, including closely related strains which could not be differentiated by the 16S rDNA sequencing approach (van Reenen and Dicks 1996). The molecular identity of RAPD-generated fragments was investigated by partial DNA sequence analysis (van Leeuwen et al. 1999). RAPD amplicons were sequenced after cloning and the sequence information subsequently used for generating probes able to detect sequence variations between genomes for binary typing of *S. aureus* strains. Although the discriminatory power of RAPD-PCR was described as high (Olive and Bean 1999), ambiguous findings were reported to be the result of the low stringency used for primer annealing and the lack of standardization. A low annealing temperature may result in imperfect hybridization of the primers, resulting in the formation of faint, non-specific bands and decreased reproducibility. In order to reduce the number of non-specific bands, stringency of primer annealing can be increased after the first PCR cycles (Blixt et al. 2003). Cusick and O'Sullivan (2000) used three different annealing temperatures simultaneously in a triplicate arbitrarily primed PCR approach for molecular fingerprinting of lactic acid bacteria (TAP-PCR). In addition to the annealing temperature, the concentration of primers and the brand of Taq polymerase used influence the quality of the fingerprint. Adjusting the primer/template ratio considerably decreases the intensity of background smearing visible after staining the gel with ethidium bromide (del Tufo and Tingey 1994; Tyler et al. 1997). Several reports point towards numerous additional factors which have an effect on the stability and reproducibility of this method. PCR reagents (template, primer, $MgCl_2$ concentration), PCR conditions and the type of thermal cycler have all been reported to make inter-laboratory reproducibility difficult (MacPherson et al. 1993; Penner et al. 1993; Grundmann et al. 1997). Due to its inability to discriminate between non-specific variation and true polymorphism, RAPD-PCR was evaluated by Tyler et al. (1997) as not being suitable for unravelling evolutionary relationships, tracking epidemiological relatedness or surveying genetic variations within natural populations. To overcome these limitations, multi-center studies were undertaken to develop standardized RAPD-PCR protocols for typing strains of *S. aureus* (van Belkum et al. 1995), *Acinetobacter* (Grundmann et al. 1997) and *Yersinia enterocolitica*

(Blixt et al. 2003). All three studies clearly showed that PCR derived fingerprint patterns can achieve a higher degree of reproducibility if standardized concentrations of template DNA and standardized PCR reagents (including optimized concentrations of primer and $MgCl_2$) are used in addition to standardized amplification conditions. Standardization of fragment separation and data analysis can be achieved by automation. DNA typing by RAPD-PCR can be combined with automated online laser fluorescence analysis by using a fluorescently labeled primer and DNA fragment analysis, based upon an automated DNA sequencer (Webster et al. 1996; Webster and Towner 2000; see also Sect. 3.2.3: ARDRA, T-RFLP).

Amplified Fragment Length Polymorphism Analysis

Similar to RAPD-PCR, amplified fragment length polymorphism analysis (AFLP) permits the simultaneous sampling of multiple loci distributed throughout a bacterial genome. In contrast to RAPD-PCR, restriction site and adaptor-specific primers for PCR amplification are used under highly stringent conditions. For AFLP analysis, two restriction enzymes (a rare cutter and a more frequently cutting enzyme) are selected to digest purified cellular DNA. Double-stranded oligonucleotide adaptors, specific to one or the other restriction site, are ligated to the termini of the DNA fragments and serve as primer-binding sites. The adaptors are designed in such a way that the original restriction site is not restored after ligation of the adaptor to a restriction fragment (Janssen et al. 1996). The amplification primers contain sequence stretches homologous to those of the adaptor and the restriction site. Applying primers with an extension of one to three nucleotides at the 3′ end beyond the sequence complementary to the restriction site (Janssen and Dijkshoorn 1996; Janssen et al. 1996; Arnold et al. 1999a, b) ensures only a subset of DNA fragments is generated. Stringent PCR conditions guarantee that only perfectly matching adaptor/primer hybrids are elongated, resulting in specific amplification products. Using either radioactively or fluorescently labeled primers, these fragments can be analyzed by polyacrylamide gel electrophoresis, followed by automatic banding pattern recognition systems. Fluorescently labeled primer systems and adaptors are commercially available and are optimized for specific DNA-sequencing equipment. The separation of fragments by agarose gel electrophoresis and visualization of patterns following ethidium bromide staining (Clerc et al. 1998) represents a more simplified AFLP procedure.

AFLP studies have demonstrated the robustness and reliability of this technique, which displays a higher discriminatory power when compared to RAPD-PCR, REP-PCR (Augustynowicz et al. 2003; Jonas et al. 2003, 2004) and ribotyping (Arias et al. 1997). AFLP has been applied successfully not

only to bacterial taxonomy, epidemiology and diversity studies, but also to the analysis of genetic variation in yeast, plant and animal genetics (Savelkoul et al. 1999). As AFLP, like other methods targeting restriction length polymorphism, detects changes in the nucleotide composition of restriction sites, discrimination between strains is influenced by the evolution of the genome, i. e. nucleotide variations caused by insertion and deletion events across the chromosome. It is obvious that the discriminatory power strongly depends on the selected restriction enzymes, because of the limitation imposed by the randomness of the restriction site for these enzymes. The choice of a suitable enzyme can be modeled by surveying complete genome sequences available today. As a consequence, AFLP is described as an excellent tool for the assessment of genetic polymorphism in clonal strains of *Mycobacterium tuberculosis* (van den Braak et al. 2004) when using optimal enzyme and primer combinations. A high-throughput AFLP procedure was described by Melles et al. (2004), analyzing the natural population dynamics of more than 1,000 *Staphylococcus aureus* strains. The pattern clusters were compared to sequence types and clonal complexes obtained by multi locus sequence typing (MLST). Both methods were similar in the composition of strain clusters emerging, though one method (MLST) focussed on sequences of a few selected housekeeping genes, while the other (AFLP) mirrored the relatedness of genomes.

As compared to RAPD-PCR, fluorescent AFLP is easier to standardize and is more specific because amplification is accomplished under high stringency conditions, brought about through the use of longer primer sequences and higher annealing temperatures. Inter-laboratory reproducibility of AFLP was tested by Jones et al. (1997) and intra-gel-specific correlation was found to be high (95.0 – 98.5%; Huys et al. 1996). However, it must be pointed out that the reproducibility of the AFLP analysis can decrease if one of the enzyme-dependent reactions (digestion of genomic DNA, ligation of adaptors, PCR) is incomplete (Witte 2002).

Amplification of Repetitive Elements Dispersed through the Whole Genome

The repetitive extragenic palindromic (REP) sequences, which are randomly distributed in bacterial genomes, are the targets of this PCR-based method. In contrast to RAPD-PCR, repetitive elements contain conserved regions and a primer designed to anneal to enterobacterial repetitive intergenic consensus sequences (ERIC) and BOX motifs can be hybridized under more stringent conditions. REP sequences were first detected in *Escherichia coli* and *Salmonella typhimurium* by Stern et al. (1984) but are also found, like ERIC motifs, in Gram-positive bacteria. The sequence structure, a palindrome, consists of a conserved consensus sequence, 38 nt in length, which can form a stable stem loop structure with a 5 bp variable central

region. A second type of repetitive element is presented by ERIC sequences, alternatively designated as intergenic repeat units (IRU). ERIC sequences, are 126 bp in length, contain a highly conserved central inverted repeat and are located in non-coding extragenic regions of the bacterial chromosome (Sharples and Lloyd 1990; Hulton et al. 1991). The function of these short interspersed repetitive DNA sequences was discussed by Lupski and Weinstock (1992), while Versalovic et al. (1991, 1994) described the distribution of repetitive elements in eubacteria and broadened their application for fingerprinting bacterial genomes through the development of oligonucleotide primers which were designed from each half of the conserved stem of the palindrome.

Interspersed repetitive elements (BOX), characterized by a modular structure, were first detected in the Gram-positive *Streptococcus pneumoniae* (Martin et al. 1992). BOX elements are also located within intergenic regions. These mosaic repetitive elements are composed of three subunits referred to as *box*A, *box*B and *box*C. The Box elements have no sequence relationship to either REP or ERIC motifs. In contrast to REP- and ERIC-PCR amplification, a single primer is used to amplify BOX-like elements. Amplicons represent genomic segments that are positioned between the conserved repetitive sequences. Initially thought to be unique to strains of the genus *Streptococcus*, BOX elements were subsequently found in various bacterial species. A PCR primer specific for the *box*A subunit can generally be used for the genomic fingerprinting of Gram-positive and Gram-negative bacteria. Amongst others, Box-PCR was successfully applied for differentiating between strains of *S. pneumoniae* (van Belkum et al. 1996; Overweg et al. 1999), *Bacillus anthracis* and *B. cereus* (Kim et al. 2002), *Bifidobacterium* species (Masco et al. 2003) and members of *Streptomyces* (Lanoot et al. 2004).

Several published reports have reviewed the strengths of REP genomic fingerprinting methods, including: (1) detailed protocols for the amplification, separation and detection of repetitive elements revealed from whole cells or purified DNA as target, (2) computer-assisted pattern analysis of REP fingerprints and (3) protocols for fluorophore-enhanced REP-PCR, electrophoresis and pattern detection using an automated DNA sequencer (Versalovic et al. 1991; Rademaker et al. 1998, 1999). Today, a high number of references are cited in public databases for applying REP-PCR, ERIC-PCR and BOX-PCR methods. Although the genomic fingerprinting of bacteria by amplification of repetitive elements has been used widely for the characterization and differentiation of strains in various fields of microbiology, taxon-dependent typability and discrimination power may differ within REP-, ERIC- and BOX-PCR techniques. Preliminary experiments should be performed in order to examine the optimum primer set and amplification conditions for a given application. For example, Szczuka

and Kasnowski (2004) found that, out of a collection of 120 *Aeromonas* strains, 25 isolates were not typable with REP-PCR, whereas ERIC-PCR worked well and resulted in excellent correlation with data obtained by RAPD-PCR. ERIC-PCR was also successfully applied to the characterization of *Staphylococcus epidermidis* (Wieser and Busse 2000), *Bifidobacterium* species (Ventura and Zink 2002a, 2003; Ventura et al. 2003) and *Listeria monocytogenes* (Harvey et al. 2004). Sequences of ERIC elements obtained from *Sinorhizobium meliloti* resulted in valuable information for the design of an oligonucleotide probe used for the rapid identification of *S. meliloti* strains (Niemann et al. 1999). The stability of ERIC profiles generated for five bacterial species (*P. aeruginosa*, *E. coli*, *Acinetobacter baumannii*, *Staphylococcus aureus*, *S. epidermidis*) after 24, 48, 72 h of incubation was analyzed by Kang and Dunne (2003). In addition, the same species were subcultured daily, representing up to 15 generational divisions. ERIC-PCR analysis from both experiments demonstrated that PCR fingerprints obtained for a single species were identical. In contrast, Reboli et al. (1994) reported that ERIC was not able to distinguish between different strains of *A. baumanii*, whereas REP-PCR was useful for determining the intraspecific relationships of these organisms. Discriminatory power could be increased significantly by combining BOX, ERIC and REP fingerprints (BER) in a single study (Rademaker et al. 2000).

Box and ERIC-PCR methods were also successfully applied for resolving the diversity of fluorescent pseudomonads (Dawson et al. 2002), while the combination of Box and REP-PCR approaches was able to differentiate *E. coli* 0157 serotypes from other *E. coli* strains (Hahm et al. 2003). As compared to RAPD-PCR, REP fingerprints result in a better resolution of *Helicobacter pylori* strains. Other authors have pointed out that fingerprinting with amplified repetitive elements is not suitable for species identification, as the resolution power is restricted to the strain level (Alam et al. 1999; Wieser and Busse 2000). However, clusters emerging as a result of these typing methods correlated well with those obtained by DNA:DNA hybridization experiments, i. e. they were suitable for delineating species, as demonstrated on strains of *Rhizobia* and xanthomonads (Nick et al. 1999; Rademaker et al. 2000). Some authors recommend REP fingerprinting for preliminary clustering of isolates in screening programs, for monitoring strain colonization or as a molecular tool in the polyphasic approach to taxonomy (Jersek et al. 1999; Antonio and Hillier 2003; Meacham et al. 2003).

As already observed with other DNA typing methods, lack of intra- and inter-laboratory reproducibility is the main obstacle for their general application. There are numerous examples where method-related artefacts obscured the outcome of the study. e. g. the study by Meacham et al. (2003) which genotyped a large number of *E. coli* isolates using ERIC-PCR. Al-

though higher annealing temperatures were used in this study for the amplification of ERIC elements, the reproducibility of patterns was decreased by inconsistencies in the presence and absence of bands within a single isolate tested in different PCR reactions. Several other studies have confirmed that REP-RCR and ERIC-PCR performed under less stringent conditions with low annealing temperatures (below 40 °C) can be considered a variant of RAPD-PCR; and therefore standardization is recommended in order to increase reproducibility (Snelling et al. 1996; Deplano et al. 2000), as already described for RAPD-PCR.

New strategies have recently been described which should circumvent some of the problems mentioned above. In order to avoid lack of standardization with respect to low intensity bands and to circumvent problems regarding background smearing following electrophoresis, Kingsley et al. (2002) used a non-gel-based technique which focussed on nucleic acid microarray technology for fingerprinting closely related *Xanthomonas* pathovars. Recently, a commercial REP-PCR fingerprinting kit (Bacterial Barcodes, Houston, Tex.) became available, enabling the standardized use of PCR reagents; and this kit was included in the genomic fingerprinting of *Clostridium difficile* (Spigaglia and Mastrantonio 2003) and *S. pneumoniae* clones (Gonzales et al. 2004). Automated pattern analysis of fluorescently labeled fragments and their separation by an automated DNA sequencer was part of a multicenter evaluation for epidemiological typing of methicillin-resistant *S. aureus* strains (Deplano et al. 2000). Furthermore, a commercial system has been described (DiversiLab) that electrophoretically separates REP-PCR amplicons on microfluid chips, combined with computer analysis of results (Healy et al. 2005). High-throughput REP fingerprinting on microfluidic chips has also been evaluated for *Mycobacterium* strains (Cangelosi et al. 2004).

3.2.2 DNA Typing Methods Targeting Gene Clusters (Operons)

RFLP Analysis with Southern Blotting and Probe Hybridization (Ribotyping)

During the early 1980s, the method of "rRNA gene restriction pattern" was developed for the characterization of bacteria (Grimont and Grimont 1986). After digesting extracted chromosomal DNA with a single or several restriction enzymes, fragments were size-separated by agarose gel electrophoresis, transferred to a Nylon membrane by Southern blotting and probed with γ-^{32}P ATP-labeled rRNA. Subsequently, radioactive probe labeling was replaced by chemical labeling (Kessler 1992) and rRNA by amplified rDNA. A chemically labeled oligonucleotide probe mixture, designed from 16S and 23S rRNAs of the *rrn* operon from *E. coli* was used

by Regnault et al. (1997) for universal ribotyping of bacteria. Using rare cutting restriction enzymes, fragments containing the complete *rrn* operon were obtained, used to determined the number of *rrn* operons and to investigate whether 16S rRNA and 23S rRNA genes are disconnected (Menke et al. 1991). A recombinant plasmid containing the entire *rrn*B operon from *E. coli* was constructed by Brosius et al. (1981). A plasmid containing this rRNA operon (*rrn*B) was linearized with *Eco*RI and used as a DNA probe (Webster et al. 1996). Although this technology has been used widely in the characterization of bacteria, manual ribotyping (also called traditional ribotyping) is more laborious and often lacks inter-laboratory reproducibility as compared to standardized, automated ribotyping with the Qualicon RiboPrinter system. The RiboPrinter system combines molecular processing steps in a stand-alone, automated robot, including cell lysis, digestion of chromosomal DNA with restriction enzymes (kits for *Eco*RI, *Pst*I and *Pvu*II are available, but use of other enzymes is possible), separation of fragments by electrophoresis, transfer of DNA fragments to a Nylon membrane, hybridization to a *E. coli rrn*B probe (a mixture of labeled fragments, similar to the DNA probe described by Webster et al. 1996), chemiluminescent detection of the probe to the fragments containing *rrn* operon sequences, image detection and computerized analysis and storage of RiboPrint patterns. The system processes a batch of eight isolates within 8 h. However, new batches can be started every 2 h, therefore enabling the potential characterization of 32 isolates per day. By including the highly variable parts of the spacer, lying between the conserved rRNA genes (16S rRNA–23S rRNA–5S rRNA), subtyping below the species level is possible. The resulting DNA fingerprint patterns are digitally stored and automatically aligned against an identification database, consisting of more than 6,000 profiles from more than 200 species, which is included in the accompanying software. During analysis, the RiboPrint profile of samples are grouped based on similarity scores. A ribogroup describes the genetic relationship of samples. Similarity scores of greater than 94% assign a new strain to the same ribogroup. Ribogroups are dynamic. This means that each time a new sample is processed, the ribogroups are reorganized and a reference pattern for each ribogroup is defined. After the RiboPrinter system has characterized a profile, it will be assigned to a species if the similarity coefficient obtained reaches a value higher than 85%.

Although slight deviations from previously obtained profiles may occur when strains are regularly subcultured and analyzed (personal observation), the highly standardized format and the single source of consumables, including even distilled water, guarantee high reproducibility. The system has been successfully applied in quality control and source-tracking of contaminations in the food and feed industry (e.g. lactic acid bacteria), epidemiology (e.g. *Campylobacter*, *Helicobacter*, *P. aeruginosa*,

enterohemorhagic *E. coli*, different *E. coli* and *Salmonella* serovars, *Klebsiella pneumoniae*, *Burkholderia cepacia* complex), differentiation between closely related species (e.g. members of the *B. cereus*/*B. anthracis* group) and description of a large number of species. The reader is referred to the "*International Journal of Systematic and Evolutionary Microbiology*" for references. Although reproducibility of automated ribotyping is very good, discrimination power was described to be low when compared to macrorestriction pattern analysis (PFGE) and traditional ribotyping (Dalsgaard et al. 1999). In general, the number of bands visualized depends on the restriction enzyme selected for analysis and on the degree of polymorphism that exists within and around the *rrn* operons. The signal intensity of bands is also influenced by the number of *rrn* operons, which may vary from one to 12 (Klappenbach et al. 2001). For further details see the ribosomal rRNA copy number database (http://rrndb.cme.msu.edu/rrndb). Recently, the standard *rrn* probe of the RiboPrinter was replaced by a peptide synthetase probe, suitable for detection of this enzyme in several *Streptomyces* soil strains (Ritacco et al. 2003).

Analysis of the 16S–23S Ribosomal Intergenic Spacer Region

This technique focusses on the characterization of the PCR-amplified 16S–23S spacer region of *rrn* operons, also described as the internal transcribed spacer (ITS) region, which evolves faster than the conserved *rrn* genes. The region contains elements that, still at the level of mRNA, form structures with 5′ and 3′ regions of the 16S rRNA and 23S rRNA genes, respectively, and are crucial in the maturation of rRNA species. Some taxa, e.g. *E. coli*, have one of two types of spacer separating the 16S and 23S coding regions. The spacers of four operons encode tRNA(Glu) and the other three encode both tRNA(Ile) and tRNA(Ala). Rapid identification of bacteria is based on the characterization of PCR-amplified ribosomal DNA of the internal spacer. The region can be amplified easily by the use of primers designed from the conserved regions of the 3′ terminus of the 16S rRNA and the 5′ terminus of the 23S rRNA. As the size of the spacer may vary considerably for different species and even among the different *rrn* operons of one genome, the discriminatory power is high. Differences in size can be determined by agarose gel electrophoresis or, for fluorescently labeled PCR fragments, by DNA sequencer as published by Hain et al. (1997) and Fisher and Triplett (1999). Differences in sequence may be investigated by restriction enzyme digestion of ITS amplicons or by direct sequencing.

Spacer polymorphism analysis was first described by Jensen et al. (1993) and applied to over 300 bacterial strains belonging to eight different genera of Gram-negative and Gram-positive bacteria. Gürtler et al. (1993) applied this method to typing clinical *Clostridium difficile* strains. Among others,

restriction enzyme digestion of the 16S–23S ITS region has been used for the phylogenetic analysis of lactic acid bacteria (Chenoll et al. 2003; Ventura and Zink 2003), *Acinetobacter* sp. (Dolzani et al. 1995), *Bacillus subtilis* (Shaver et al. 2002) and Gram-positive anaerobic cocci (Hill et al. 2002). Identification and analysis of population diversity within the staphylococci was investigated by several groups. 16S–23S rDNA intergenic spacer polymorphism analysis was found to be a reliable tool for the identification of 31 *Staphylococcus* species, although discrimination of subspecies was not possible (Mendoza et al. 1998; Bes et al. 2002). A review of the use of the 16S–23S ribosomal gene spacer region in studies of prokaryotic diversity was published by Garcia-Martinez et al. (1999).

Restriction enzyme analysis of ITS regions resulting in the occurrence of more than one band points towards the presence of polymorphisms within the different *rrn* operons. Because two electrophoretically separated fragments of the same size may differ in their sequence, sequencing of the ITS amplicon provides the most useful information. As shown by Boyer et al. (2001), the 16S–23S ITS composition mirrors higher phylogenetic grouping, reflected by the t-RNA type present or absent. Among others, sequencing of the ITS region was successfully applied for the differentiation of *Bifidobacteria* (Leblond-Bourget et al. 1996), *Streptococci* (Hassan et al. 2003; Mora et al. 2003) *Mycobacterium* sp. (Hamid et al. 2002), *Lactobacillus* (Song et al. 2000), *Pseudomonas* (Milyutina et al. 2004), *Gluconobacter* (Yukphan et al. 2004), *Micrococcus luteus* (Haga et al. 2003), *Bradyrhizobium* (Willems et al. 2003), *Bacillus* sp. (Xu and Cote 2003), *Legionella pneumophilia* (Perez-Luz et al. 2002), *Roseobacter*-related species (Söller et al. 2000), cyanobacteria (Boyer et al. 2001), the analysis of *Ralstonia solanacearum* pathovars (Pastrik et al. 2002) and also the characterization of anaerobes, e.g. *Fusobacterium* (Conrads et al. 2004) and *Porphyromonas gingivalis* (Rumpf et al. 2000). ITS sequences are compiled in the ribosomal internal spacer sequence collection (http://ulises.umh.es/RISSC/; Garcia-Martinez et al. 2001). ITS spacer sequences useful in the identification of *Mycobacterium* species are available from the ribosomal differentiation of medical microorganisms (RIDOM) database (http://www.ridom.de/; Harmsen et al. 2003). Sequence information is also useful for the design of taxon-specific probes or PCR primers for strain identification. Recently, an oligonucleotide microarray was developed for the identification of *Bacillus anthracis*, based on intergenic transcribed spacers in ribosomal DNA (Nübel et al. 2004).

3.2.3
DNA Typing Methods Targeting the 16S rRNA Gene

Amplified Ribosomal rDNA Restriction Analysis

With the establishment of molecular-based methods at the beginning of the 1980s, Carl Woese and Norman Pace revolutionized microbial taxonomy and microbial ecology by demonstrating that *rrn* sequences are the most useful chronometers for deciphering the phylogeny and evolution of organisms and the characterization of natural microbial populations. The secondary structure of the rRNA consists of a mixture of conserved regions involved in helix formation and variable regions linking the conserved regions, e. g. in stem loops. In total, nine highly variable regions are present, consisting of sequences that may differ to varying extents even for species of the same genus. Therefore, the discrimination of taxa depends upon the varying location of restriction sites of amplified 16S rRNA.

The amplified ribosomal rDNA restriction analysis (ARDRA) technique was first applied to the identification of medically important strains, but has since been reported as a reliable and valuable tool for phylogenetic and taxonomic studies of large sets of cultured or uncultured organisms from different habitats (Vaneechoutte 1992, 1993; Martinez-Murcia et al. 1995). As compared to methods based on the detection of RFLPs, the discriminatory power of the ARDRA method depends on the restriction enzyme(s) selected for digestion of PCR-amplified 16S rDNA. Restricted DNA fragments are separated by size via agarose electrophoresis or, when using fluorescently labeled primers or amplicons, by polyacrylamide electrophoresis. A detailed protocol for bacterial fingerprinting of amplified 16S rDNA, including amplified 16S–23S spacer region, is given by Massol-Deya et al. (1995). The discriminatory power can be increased by the simultaneous use of three different restriction enzymes. The choice of restriction enzyme applied in the analysis may be tested in silico using computer programs, e. g. Nebcutter (http://tools.neb.com/NEBcutter2/index.php/; New England Biolabs). 16S rDNA sequences of interest, available from public databases like EMBL or GenBank, can be imported directly into the computer program and restricted with a set of different restriction enzymes. For example, the restriction enzymes *Hae*III, *Hha*I (isoschizomer *Cfo*I) and *Bst*UI are suitable for many bacterial groups (Pukall et al. 1998). In addition, the enzymes *Alu*I, *Rsa*I and *Msp*I (isoschizomer *Sau*3AI) are frequently used for the characterization of isolates and bacterial communities. For standardization, protocols defining the enzymes with the most discriminatory power must be developed, in addition to the primer system used for amplification of 16S rDNA. Automated fragment length analysis of fluorescently labeled 16S rDNA following digestion with a four-base cutting restriction enzyme was recently described (Pukall et al. 1998). In this approach, fluorescently

labeled dUTPs were incorporated directly into the amplified DNA during PCR cycling. As compared to fluorescent end-labeling of primers, this technique has the advantage of a higher incorporation rate of fluorescent molecules into the DNA, resulting in an increased detection sensitivity.

Terminal Restriction Fragment Length Polymorphism Analysis

The procedure of terminal restriction fragment length polymorphism (T-RFLP) is similar to that described for the automated ARDRA method, but T-RFLP is mostly used as a tool for analyzing microbial communities. Similar to denaturing gradient gel electrophoresis (DGGE) or TGGE (see below), DNA from as-yet non-cultured bacteria is included in the T-RFLP analysis, replacing the cloning and sequencing approach for determination of population structures. In the first step, DNA is isolated from the community (see Chap. 7) and the 16S rDNA is PCR-amplified, consisting of a mixture of 16S rDNAs derived from different members of the community. Primers can be designed to be non-discriminative, amplifying the 16S rDNA of most members of a community, or more selective, targeting specific groups only (Liu et al. 1997, Marsh et al. 2000). The 5′ primer is fluorescently labeled to tag the products. The amplicon is then digested with a four-base cutter and the terminal fragments containing the fluorescent label are separated by size on an automated DNA sequencer or capillary electrophoresis unit, combined with commercial gene fragment analysis software. Terminal fragments of size > 550 bp resolve poorly under denaturing conditions, whereas non-denaturing conditions (Long Ranger matrix, BioRad) or capillary systems offer longer sequence reads.

Similar to the ARDRA method, the use of two to three different restriction enzymes is recommended in order to increase the discriminatory power of the analysis. It must be kept in mind that fragments of the same size may nevertheless contain information from different organisms. Furthermore, phylogenetically highly related species or subgroups may not be distinguished even with an analysis based on three digests (Marsh et al. 2000). Using capillary electrophoresis and laser-induced fluorescence detection, a protocol for the optimized separation and detection of fragments between 20 bp and 1,632 bp in size was developed by Moeseneder et al. (1999). Using a complex bacterioplankton community, the authors demonstrated this technique to be more sensitive than DGGE (see below). A protocol specifically developed for separation by capillary electrophoresis is publicly available on the website of the ribosomal database project (RDP II; http://rdp.cme.msu.edu/). A web-based program was established and is included in the online analysis programs of the RDP II project (http://rdp8.cme.msu.edu/html/analyses.html). The analysis function permits the user to perform in silico restriction digestions of the entire 16S

sequence database of the RDP II and derive terminal restriction fragment sizes from the 5′ terminus of the user-specified primer to the 3′ terminus of the restriction endonuclease target site. The output can be sorted and viewed either phylogenetically or by size (Marsh et al. 2000). The program also allows for the testing of the discriminatory activity of enzymes and enzyme combinations. Individual researchers' terminal fragment data may be submitted for subsequent analysis. Unfortunately, the databases are no longer fully maintained at the new release (ver. 9) of the RDP II website, but are still available at the old RDP site.

Lukow et al. (2000) demonstrated that T-RFLP fingerprinting enables the detection of both spatial and temporal heterogeneities in the structural composition of highly diverse communities. T-RFLP was reported to be a useful analysis tool for the assessment of diversity and rapid comparison of community structures of *Bifidobacteria* (Sakamoto et al. 2003) and for the analysis of genes other than rRNA (Horz et al. 2000). Other authors have pointed out that T-RFLP may be used as a rapid tool for the analysis of replicate samples, but less so for providing reliable information on phylotype richness and evenness or consistency (Dunbar et al. 2000; a phylotype is defined by a 16S rDNA sequence, usually showing less than 99.0% similarity to its nearest neighbors). Lüdemann et al. (2000) showed that, in contrast to DGGE, T-RFLP prints derived from DNA and RNA revealed the most similar patterns. Lueders and Friedrich (2000), investigating an archaeal community, reported a relatively constant population structure analyzed at daily intervals on the basis of T-RFLP fingerprints, while significant shifts were observed when clone libraries were screened. Most criticism of T-RFLP focusses on sample preparation, efficient extraction of community DNA and pitfalls associated with PCR amplification-based gene analysis. Dunbar et al. (2000) suggested that as little as 0.1% and 1.0% of the populations comprising a bacterial community could be detected in T-RFLP profiles. A review describing the pitfalls of PCR-based rRNA analysis for determination of microbial diversity in environmental samples was produced by von Wintzingerode et al. (1997). Purity of community DNA, total amount of species-specific DNA, G+C content of DNA and 16S rRNA sequence variations due to operon heterogeneity are import parameters to be noted when analyzing T-RFLP, but they also influence all other PCR-based methods used for community analysis.

Denaturing Gradient Electrophoresis and Temperature Gradient Gel Electrophoresis

Denaturing gradient electrophoresis (DGGE), first described by Fischer and Lerman (1979), and applied in community analysis by Muyzer et al. (1993), and the "sister" technique temperature gradient gel electrophoresis (TGGE;

Rosenbaum and Riesner 1987; Zoetendal et al. 1998) are employed to separate DNA fragments of the same length, but of different sequence composition. Separation is based on the electrophoretic mobility of partially melted double-stranded DNA molecules. A GC-rich sequence (GC-clamp) attached to the 5′ end of the forward primer may act as a high melting domain to prevent the double-stranded DNA fragment from complete dissociation into single strands. Separation takes place in a polyacrylamide gel containing a linear denaturing gradient, generated either chemically using a mixture of urea and formamide (DGGE) or by a temperature gradient (TGGE).

Molecules with different sequences have different melting behaviors and therefore finish migrating at different positions in the gel. The sequence differences might be as small as a single nucleotide. In this regard, gradient gel electrophoresis was originally developed for use in medical applications, in order to detect point mutations. In general, single-stranded RNA forming specific secondary structures, double-stranded DNA and proteins can be analyzed. Analysis of short amplified DNA fragments is a widely used method in molecular ecological studies. A DNA fragment migrating in the gel matrix remains double-stranded until it reaches the conditions that cause the melting of the lower-temperature melting domains. Partial separation of double-stranded DNA decreases the mobility of the fragment until denaturation is complete (except for the GC clamp) and the fragment stops migrating. The temperature gradient applied for optimal separation of bands should be determined by running a perpendicular gel and/or running a time-course in a parallel gel using different gradients, as described by Heuer and Smalla (1997). DNA bands can be visualized after electrophoresis either by ethidium bromide, SYBR green staining or by silver staining. Whereas background staining can be reduced by the use of SYBR green, silver staining is more sensitive and also detects faint bands and single-stranded DNA. For community analysis, extracted and purified DNA is used as the target for PCR amplification of a molecular marker, e.g. a specific stretch of the rDNA or a housekeeping gene. DGGE and TGGE have become popular techniques in molecular microbial ecology because these methods are inexpensive and rapid, thereby allowing the simultaneous analysis of multiple samples. Commercial equipment, e.g. DGGE (BioRad) and TGGE (Biometra), are widely used for the investigation of bacterial community structures and for the determination of genetic diversity and population dynamics (Ferris et al. 1996; Murray et al. 1996; Brinkhoff and Muyzer 1997; Buchholz-Cleven et al. 1997; Kowalchuk et al. 1997; Øvreas et al. 1997; Vallaeys et al. 1997). This technique was also used to investigate gene expression in mixed populations (Wawer et al. 1997), monitor enrichment and isolation of bacteria (Rölleke et al. 1996; Teske et al. 1996; Jackson et al. 1998; Heuer et al. 1999), study phylogenetic

relationships (Muyzer et al. 1995; Heuer et al. 1997; Felske et al. 1999) and microheterogeneity in rRNA-encoding genes (Nübel et al. 1996) and to screen clone libraries. Several reviews have been published on the application of DGGE/TGGE in microbial ecology (Heuer and Smalla 1997; Muyzer and Smalla 1998; Muyzer 1999). The use of DGGE or TGGE goes beyond mere typing, as these approaches also offer the possibility of phylogenetic assessment of community members. In this case, selected bands are excised, DNA-purified and either sequenced directly or cloned and sequenced. Gels stained with ethidium bromide may also be blotted to membranes and used in oligonucleotide hybridization experiments.

Several factors may influence the assessment of diversity, such as methods used for harvesting cells, DNA extraction protocols, quality of extracted DNA (Stackebrandt et al. 2004), genome size and *rrn* operon numbers (Farelly et al. 1995), as well as biases caused by PCR (von Winzingerode et al. 1997). Whereas the 16S rDNA primer sets 341F/534r, 41f/927r and 1055f/1406r were successfully applied in studies for the characterization of bacterial communities from aquatic sites, primer set 968f/1401r was recommended for characterization of soil communities (Muyzer et al. 1995; Ferris et al. 1996; Heuer et al. 1997). Based on intrinsic properties of denaturants used in the polyacrylamide gel matrix, only DNA stretches up to 500 bp in length may be successfully analyzed with TGGE and DGGE. Although analysis of a 500-bp stretch will contain sufficient information for mutation analysis, it may be too short for phylogenetic inferences. Analysis of different regions of the 16S rRNA molecule and varying electrophoretic conditions will result in the formation of different fingerprints of a community. When amplification of PCR products is performed with primers derived from conserved regions of the 16S rRNA molecule (universal primer), only predominant members of the population may emerge, resulting in the suppression of minority members. Differential amplification of rRNA genes by PCR has been described by Reysenbach et al. (1992). Resolution of DGGE/TGGE fingerprints can be increased by the use of more group-specific primers or the fractionation of DNA according to its G+C content prior to PCR, as described by Heuer and Smalla (1997), though only selected members of the community will be enriched. Furthermore, a single DGGE band does not always represent a single PCR fragment (Sekiguchi et al. 2001), as fragments differing in sequence may nevertheless migrate to the same location because of similar melting properties. In addition, the formation of heteroduplex molecules formed by re-annealing of denatured PCR products within a PCR reaction may lead to misinterpretation of community complexity. However, homoduplex–heteroduplex polymorphism can be used for the characterization of cultivatable isolates (Daffonchio et al. 2000).

When the hypervariable 16S–23S intergenic spacer regions are amplified from the conserved adjacent sequences, homoduplex double-stranded DNA and heteroduplex structures may be formed which contain substantial regions of single-stranded DNA, depending on the PCR conditions used (Jensen and Straus 1993). Homoduplex–heteroduplex polymorphism (HHP) formed during PCR between amplicons from different ribosomal operons, or stretches within or outside of t-RNA genes, allowed the discrimination of *Salmonella* serovars (Jensen and Hubner 1996) and *Bacillus* and related genera (Daffonchio et al. 2000, 2003). For detection of ITS-HHP polymorphisms, electrophoresis is carried out in polyacrylamide in which, compared to homoduplex DNA fragments, the mobility of heteroduplex structures is reduced, depending on the secondary structure formed within the single-stranded regions.

A modified DGGE technique was recently described by Gürtler et al. (2001) for the analysis of mutations in the VS2 region of the 16S–23S spacer from *Staphylococcus aureus*. Amplicons of this region were separated by double-gradient denaturating gel electrophoresis (DG-DGGE), using denaturing conditions within a polyacrylamide matrix that itself contained a concentration gradient of 6 – 12%. This analysis allowed the detection of different genotypes of *S. aureus* isolates, characterized exclusively by homoduplex bands or a combination of homo- and heteroduplex bands. The authors were able to associate a single mutation to methicillin-resistant isolates from different geographic locations. A combined approach of 16S–23S ITS analysis and TGGE was described by Yasuda and Shiaris (2005), using non-GC-clamped PCR primers for the differentiation of diverse bacterial species. The high GC content site at the t-rRNA coding region of 16S–23S rDNA served as an internal self GC-clamp for TGGE. Gürtler et al. (2002) evaluated DGGE-mediated multi-locus sequence typing (MLST; see Chap. 6) for the characterization of *S. aureus* isolates. DGGE was used for the differentiation of amplicons obtained from seven housekeeping genes, thus avoiding the sequencing step. The authors pointed out that the DGGE-MLST method is a rapid, accurate and less expensive alternative to DNA sequencing.

PCR-based Single-stranded Confirmation Polymorphism

Similar to DGGE, PCR-based single-stranded confirmation polymorphism (SSCP) analysis was originally designed for the detection of polymorphisms and mutations in human genes (Orita et al. 1989). In this technique, DNA fragments derived from PCR amplification are denaturated to obtain single-stranded DNA that is subjected to electrophoresis on a non-denaturating gel. Under these conditions, single-stranded DNA has a folded conformation, which influences the electropheretic mobility. Therefore, as also

shown for DGGE and TGGE, DNA fragments of the same size but in different sequences, are separated on the basis of differences in structure. Silver-stained bands of ssDNA localized at different positions in a polyacrylamide gel indicate different sequences. Automated SSCP analysis with capillary electrophoresis and fluorescently labeled primers was described by Ghozzi et al. (1999) and King et al. (2005). In mutation analysis, single-base substitutions are detectable by analyzing small fragments of up to 200 bp (Hayashi 1992); and analysis of larger fragments follows digestion with restriction enzymes. SSCP was successfully applied for detecting differences in *rec*A operon fragments of *Burkholderia cepacia* (Moore et al. 2001), detecting *rpo*B gene mutations in *Mycobacterium tuberculosis* (Bobadilla-de-Valle et al. 2001) and detecting *gyr*A or *fla* gene polymorphism in *Campylobacter jejunii* (Hakanen et al. 2002; Hein et al. 2003). SSCP has also been performed for the analysis of housekeeping genes like *gro*EL and for the confirmation of an epidemic clonal complex of *Vibrio cholerae* serogroups 01 and 0139 (O'Shea et al. 2004). Widjojoatmodjo et al. (1994) were the first to evaluate SSCP for the rapid identification of bacteria, using different primer sets derived from specific regions of the 16S rRNA molecule for the discrimination of strains at genus and species level. Lee et al. (1996) introduced SSCP to study genetic profiles of natural bacterial communities; and Schwieger and Tebbe (1998) used a modified SSCP protocol for 16S rRNA gene-based microbial community analysis of rhizosphere and soil habitats. To overcome re-annealing of DNA strands and heteroduplex formation during electrophoresis and to reduce the number of bands per organism, the authors used one phosphorylated primer in the PCR reaction followed by specific digestion of the phosphorylated strands with a lambda exonuclease. SSCP was also applied to the characterization of pyrite-oxidizing bacterial populations (Battaglia-Brunet et al. 2002), to follow alterations in intestinal microbiota in fecal samples during storage (Ott et al. 2004), and to analyze biofilm compositions formed on different dental implant surfaces exposed in the oral cavity of humans (Groessner-Schreiber et al. 2004). Furthermore, the SSCP method was combined with ITS analysis for the identification of streptococci (Mora et al. 2003).

Similar to DGGE, bands can be excised and sequenced for further analysis and SSCP profiles can be hybridized against specific probes. In fact, gene probing may be a useful control to investigate the community profiles obtained, because the profile may consist of more sequences than are detectable by staining (Schmalenberger and Tebbe 2003). Although SSCP analysis is simple and requires neither a GC clamp nor the construction of gradients, most limitations of the SSCP technique are the same as those discussed for DGGE when applied to microbial communities. As a result of potential intraspecies operon heterogeneities of rRNA genes, more than one band per organism may be detectable. Identical ssDNA sequences can form

more than one stable formation and PCR-based community analysis is affected by the selection of primers derived from 16S rRNA (Schmalenberger et al. 2001) and electrophoretic conditions (e.g. gel matrix, temperature, addition of glycerol; Widjojoatmodjo et al. 1994).

References

Alam S, Brailsford SR, Whiley RA, Beighton D (1999) PCR-based methods for genotyping viridans group streptococci. J Clin Microbiol 37:2772–2776

Antonio MA, Hillier SL (2003) DNA fingerprinting of *Lactobacillus crispatus* strain CTV-05 by repetitive element sequence-based PCR analysis in a pilot study of vaginal colonization. J Clin Microbiol 41:1881–1887

Arias, CR, Verdonck L, Swings, J, Garay E, Aznar R (1997) Intraspecific differentiation of *Vibrio vulnificus* biotypes ba amplified fragment length polymorphism and ribotyping. Appl Environ Microbiol 63:2600–2606

Arnold C, Metherell L, Willshaw G, Maggs A, Stanley J (1999a) Predictive fluorescent amplified-fragment length polymorphism analysis of *Escherichia coli*: high-resolution typing method with phylogenetic significance. J Clin Microbiol 37:1274–1279

Arnold C, Metherell L, Clewley JP, Stanley J (1999b) Predictive modelling of fluorescent AFLP: a new approach to the molecular epidemiology of *E. coli*. Res Microbiol 150:33–44

Augustynowicz E, Gzyl A, Szenborn L, Banys D, Gniadek G, Slusarczyk J (2003) Comparison of usefulness of randomly amplified polymorphic DNA and amplified-fragment length polymorphism techniques in epidemiological studies on nasopharyngeal carriage of non-typable *Haemophilus influenzae*. J Med Microbiol 52:1005–1014

Battaglia-Brunet F, Clarens M, D'Hugues P, Godon JJ, Foucher S, Morin D (2002) Monitoring of a pyrite-oxidising bacterial population using DNA single-strand conformation polymorphism and microscopic techniques. Appl Microbiol Biotechnol 60:206–211

van Belkum A, Kluytmans J, van Leeuwen W, Bax R, Quint W, Peters E, Fluit A, Vandenbroucke-Grauls C, van den Koeleman H (1995) Multicenter evaluation of arbitrarily primed PCR for typing of *Staphylococcus aureus* strains. J Clin Microbiol 33:1537–1547

van Belkum A, Sluijuter M, de Groot R, Verbrugh H, Hermans PW (1996) Novel BOX repeat PCR assay for high-resolution typing of *Streptococcus pneumoniae* strains. J Clin Microbiol 34:1176–1179

Bes M, Saidi SL, Becharnia F, Meugnier H, Vandenesch F, Etienne J, Freney J (2002) Population diversity of *Staphylococcus intermedius* isolates from various host species: typing by 16S–23S intergenic ribosomal DNA spacer polymorphism analysis. J Clin Microbiol 40:2275–2277

Blixt Y, Knutsson R, Borch E, Radstrom P (2003) Interlaboratory random amplified polymorphic DNA typing of *Yersinia enterocolitica* and *Y. enterocolitica*-like bacteria. Int J Food Microbiol 83:15–26

Bobadilla-del-Valle M, Ponce-de-Leon A, Arenas-Huertero C, Vargas-Alarcon G, Kato-Maeda M, Small PM, Couary P, Ruiz-Palacios GM, Sifuentes-Osornio J (2001) *rpoB* gene mutations in rifampin-resistant *Mycobacterium tuberculosis* identified by polymerase chain reaction single-stranded conformational polymorphism. Emerg Infect Dis 7:1010–1013

Boyer SL, Flechtner VR, Johansen JR (2001) Is the 16S–23S rRNA internal transcribed spacer region a good tool for use in molecular systematics and population genetics? A case study in cyanobacteria. Mol Biol Evol 18:1057–1069

van den Braak N, Simons G, Gorkink R, Reijans M, Eadie K, Kremers K, van Soolingen D, Savelkoul P, Verbrugh H, van Belkum A (2004) A new high-throughput AFLP approach for identification of new genetic polymorphism in the genome of the clonal microorganism *Mycobacterium tuberculosis*. J Microbiol Methods 56:49–62

Brinkhoff T, Muyzer G (1997) Increased species diversity and extended habitat range of sulfur-oxidizing *Thiomicrospira* spp. Appl Environ Microbiol 63:3789–3796

Brosius J, Ullrich A, Raker MA, Gray A, Dull TJ, Gutell RR, Noller HF (1981) Construction and fine mapping of recombinant plasmids containing the *rrn*B ribosomal RNA operon from *E. coli*. Plasmid 6:112–118

Buchholz-Cleven, BEE, Rattunde B, Straub, KL (1997) Screening for genetic diversity of isolates of anaerobic Fe(II)-oxidizing bacteria using DGGE and whole-cell hybridization. System Appl Microbiol 20:301–309

Cangelosi GA, Freeman RJ, Lewis KN, Livingston-Rosanoff D, Shah KS, Milan SJ, Goldberg SV (2004) Evaluation of a high-throughput repetitive-sequence-based PCR system for DNA fingerprinting of *Mycobacterium tuberculosis* and *Mycobacterium avium* complex strains. J Clin Microbiol 42:2685–2693

Chenoll E, Macian MC, Aznar R (2003) Identification of *Carnobacterium*, *Lactobacillus*, *Leuconostoc* and *Pediococcus* by rDNA-based techniques. Syst Appl Microbiol 26:546–556

Clerc A, Manveau C, Nesme X (1998) Comparison of randomly amplified polymorphic DNA with amplified fragment length polymorphism to asssess genetic diversity and genetic relatedness with genospecies III of *Pseudomonas syringae*. Appl Environ Microbiol 64:1180–1187

Conrads G, Citron DM, Mutters R, Jang S, Goldstein EJ (2004) *Fusobacterium canifelinum* sp. nov., from the oral cavity of cats and dogs. Syst Appl Microbiol 27:407–413

Cusick SM, O'Sullivan DJ (2000) Use of a single, triplicate arbitrarily primed-PCR procedure for molecular fingerprinting of lactic acid bacteria. Appl Environ Microbiol 66:2227–2231

Daffonchio D, Cherif A, Borin S (2000) Homoduplex and heteroduplex polymorphisms of the amplified ribosomal 16S–23S internal transcribed spacers describe genetic relationships in the "*Bacillus cereus* group". Appl Environ Microbiol 66:5460–5468

Daffonchio D, Cherif A, Brusetti L, Rizzi A, Mora D, Boudabous A, Borin S (2003) Nature of polymorphisms in 16S–23S rRNA gene intergenic transcribed spacer fingerprinting of *Bacillus* and related genera. Appl Environ Microbiol 69:5128–5137

Dalsgaard A, Forslund A, Fussing V (1999) Traditional ribotyping shows a higher discrimination than the automated RiboPrinter system in typing *Vibrio cholerae* O1. Lett Appl Microbiol 28:327–333

Dawson SL, Fry JC, Dancer BN (2002) A comparative evaluation of five typing techniques for determining the diversity of fluorescent pseudomonads. J Microbiol Methods 50:9–22

Deplano A, Schuermans A, Van Eldere J, Witte W, Meugnier H, Etienne J, Grundmann H, Jonas D, Noordhoek GT, Dijkstra J, van Belkum A, van Leeuwen W, Tassios PT, Legakis NJ, van der Bergmans A, Blanc DS, Tenover FC, Cookson BC, O'Neil G, Struelens MJ (2000) Multicenter evaluation of epidemiological typing of methicillin-resistant *Staphylococcus aureus* strains by repetitive-element PCR analysis. (The European study group on epidemiological markers of the ESCMID.) J Clin Microbiol 38:3527–3533

Dolzani L, Tonin E, Lagatolla C, Prandin L, Monti-Bragadin C (1995) Identification of *Acinetobacter* isolates in the *A. calcoaceticus–A. baumannii* complex by restriction analysis of the 16S–23S rRNA intergenic-spacer sequences. J Clin Microbiol 33:1108–1113

Dunbar J, Ticknor LO, Kuske CR (2000) Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis. Appl Environ Microbiol 66:2943–2950

Farelly V, Rainey FA, Stackebrandt E (1995) Effect of genomic size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl Environ Microbiol 61:2798–2801

Felske A, Vancanneyt M, Kersters K, Akkermans AD (1999) Application of temperature-gradient gel electrophoresis in taxonomy of coryneform bacteria. Int J Syst Bacteriol 49:113–121

Ferris MJ, Muyzer G, Ward DM (1996) Denaturing gradient gel electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat community. Appl Environ Microbiol 62:340–346

Fischer SG, Lerman LS (1979) Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. Cell 16:191–200

Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65:4630–4636

Garcia-Martinez J, Acinas SG, Anton AI, Rodriguez-Valera F (1999) Use of the 16S–23S ribosomal genes spacer region in studies of prokaryotic diversity. J Microbiol Methods 36:55–64

Garcia-Martinez J, Bescos I, Rodriguez-Sala JJ, Rodriguez-Valera F (2001) RISSC: a novel database for ribosomal 16S–23S RNA genes spacer regions. Nucleic Acids Res 29:178–180

Ghozzi R, Morand P, Ferroni A, Beretti JL, Bingen E, Segonds C, Husson MO, Izard D, Berche P, Gaillard JL (1999) Capillary electrophoresis single-strand conformation polymorphism analysis for rapid identification of *Pseudomonas aeruginosa* and other gram-negative non-fermenting bacilli recovered from patients with cystic fibrosis. J Clin Microbiol 37:3374–3379

Gonzalez BE, Hulten KG, Kaplan SL, Mason EO Jr (2004) Clonality of *Streptococcus pneumoniae* serotype 1 isolates from pediatric patients in the United States. J Clin Microbiol 42:2810–2812

Graves LM, Swaminathan B (2001) PulseNet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. Int J Food Microbiol 65:55–62

Grimont F, Grimont PA (1986) Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. Ann Inst Pasteur Microbiol 137B:165–175

Groessner-Schreiber B, Hannig M, Duck A, Griepentrog M, Wenderoth DF (2004) Do different implant surfaces exposed in the oral cavity of humans show different biofilm compositions and activities? Eur J Oral Sci 112:516–522

Grundmann HJ, Towner KJ, Dijkshoorn L, Gerner-Smidt P, Maher M, Seifert H, Vaneechoutte M (1997) Multicenter study using standardized protocols and reagents for evaluation of reproducibility of PCR-based fingerprinting of *Acinetobacter* spp. J Clin Microbiol 35:3071–3077

Gürtler V (1993) Typing of *Clostridium difficile* strains by PCR-amplification of variable length 16S–23S rDNA spacer regions. J Gen Microbiol 139:3089–3097

Gürtler V, Barrie HD, Mayall BC (2001) Use of denaturing gradient gel electrophoresis to detect mutation in VS2 of the 16S–23S rDNA spacer amplified from *Staphylococcus aureus* isolates. Electrophoresis 22:1920–1924

Gürtler V, Barrie HD, Mayall BC (2002) Denaturing gradient gel electrophoretic multilocus sequence typing of *Staphylococcus aureus* isolates. Electrophoresis 23:3310–3320

Haga S, Hirano Y, Murayama O, Millar BC, Moore JE, Matsuda M. (2003) Structural analysis and genetic variation of the 16S–23S rDNA internal spacer region from *Micrococcus luteus* strains. Lett Appl Microbiol. 37:314–317

Hahm BK, Maldonado Y, Schreiber E, Bhunia AK, Nakatsu CH (2003) Subtyping of foodborne and environmental isolates of *Escherichia coli* by multiplex-PCR, rep-PCR, PFGE, ribotyping and AFLP. J Microbiol Methods 53:387–399

Hain T, Ward-Rainey N, Kroppenstedt RM, Stackebrandt E, Rainey FA (1997) Discrimination of *Streptomyces albidoflavus* strains based on the size and number of 16S–23S ribosomal DNA intergenic spacers. Int J Syst Bacteriol 47:202–206

Hakanen A, Jalava J, Kotilainen P, Jousimies-Somer H, Siitonen A, Huovinen P (2002) *gyrA* polymorphism in *Campylobacter jejuni*: detection of *gyrA* mutations in 162 *C. jejuni* isolates by single-strand conformation polymorphism and DNA sequencing. Antimicrob Agents Chemother 46:2644–2647

Hamid ME, Roth A, Landt O, Kroppenstedt RM, Goodfellow M, Mauch H (2002) Differentiation between *Mycobacterium farcinogenes* and *Mycobacterium senegalense* strains based on 16S–23S ribosomal DNA internal transcribed spacer sequences. J Clin Microbiol 40:707–711

Harmsen D, Dostal S, Roth A, Niemann S, Rothganger J, Sammeth M, Albert J, Frosch M, Richter E (2003) RIDOM: comprehensive and public sequence database for identification of *Mycobacterium* species. BMC Infect Dis 3:26

Harvey J, Norwood DE, Gilmour A (2004) Comparison of repetitive element sequence-based PCR with multi-locus enzyme electrophoresis and pulsed fied gel electrophoresis for typing of *Listeria monocytogenes* food isolates. Food Microbiol 21:305–312

Hassan AA, Khan IU, Abdulmawjood A, Lammler C (2003) Inter- and intraspecies variations of the 16S–23S rDNA intergenic spacer region of various streptococcal species. Syst Appl Microbiol 26:97–103

Hayashi K (1992) PCR-SSCP: a method for detection of mutations. Genet Anal Tech Appl 9:73–79

Healy M, Huong J, Bittner T, Lising M, Frye S, Raza S, Schrock R, Manry J, Renwick A, Nieto R, Woods C, Versalovic J, Lupski JR (2005) Microbial DNA typing by automated repetitive-sequence-based PCR. J Clin Microbiol 43:199–207

Hein I, Mach RL, Farnleitner AH, Wagner M (2003) Application of single-strand conformation polymorphism and denaturing gradient gel electrophoresis for *fla* sequence typing of *Campylobacter jejuni*. J Microbiol Methods 52:305–313

Heuer H, Smalla K (1997) Application of denaturing gradient gel electrophoresis and temperature gradient gel electrophoresis for studying soil microbial communities. In: van Elsas JD, Trevors JT, Wellington EMH (eds) Modern soil microbiology. Dekker, New York, pp 353–373

Heuer H, Krsek M, Baker P, Smalla K, Wellington EM (1997) Analysis of actinomycete communities by specific amplification of genes encoding 16S rRNA and gel-electrophoretic separation in denaturing gradients. Appl Environ Microbiol 63:3233–3241

Heuer H, Hartung K, Wieland G, Kramer I, Smalla K (1999) Polynucleotide probes that target a hypervariable region of 16S rRNA genes to identify bacterial isolates corresponding to bands of community fingerprints. Appl Environ Microbiol 65:1045–1049

Hill KE, Davies CE, Wilson MJ, Stephens P, Lewis MA, Hall V, Brazier J, Thomas DW (2002) Heterogeneity within the gram-positive anaerobic cocci demonstrated by analysis of 16S–23S intergenic ribosomal RNA polymorphisms. J Med Microbiol 51:949–957

Horz HP, Rotthauwe JH, Lukow T, Liesack W (2000) Identification of major subgroups of ammonia-oxidizing bacteria in environmental samples by T-RFLP analysis of *amoA* PCR products. J Microbiol Methods 39:197–204

Hulton CS, Higgins CF, Sharp PM (1991) ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol Microbiol 5:825–834

Huys G, Coopman R, Janssen P, Kersters K (1996) High-resolution genotypic analysis of the genus *Aeromonas* by AFLP fingerprinting. Int J Syst Bacteriol 46:572–580

Jackson CR, Roden EE, Churchill PF (1998) Changes in bacterial species composition in enrichment cultures with various dilutions of inoculum as monitored by denaturing gradient gel electrophoresis. Appl Environ Microbiol 64:5046–5048

Janssen P, Dijkshoorn L (1996) High resolution DNA fingerprinting of *Acinetobacter* outbreak strains. FEMS Microbiol Lett 142:191–194

Janssen P, Coopman R, Huys G, Swings J, Bleeker M, Vos P, Zabeau M, Kersters K (1996) Evaluation of the DNA fingerprinting method AFLP as an new tool in bacterial taxonomy. Microbiology 142:1881–1893

Jensen MA, Hubner RJ (1996) Use of homoduplex ribosomal DNA spacer amplification products and heteroduplex cross-hybridization products in the identification of *Salmonella* serovars. Appl Environ Microbiol 62:2741–2746

Jensen MA, Straus N (1993) Effect of PCR conditions on the formation of heteroduplex and single-stranded DNA products in the amplification of bacterial ribosomal DNA spacer regions. PCR Methods Appl 3:186–194

Jensen MA, Webster JA, Straus N (1993) Rapid identification of bacteria on the basis of polymerase chain reaction-amplified ribosomal DNA spacer polymorphisms. Appl Environ Microbiol 59:945–952

Jersek B, Gilot P, Gubina M, Klun N, Mehle J, Tcherneva E, Rijpens N, Herman L (1999) Typing of *Listeria monocytogenes* strains by repetitive element sequence-based PCR. J Clin Microbiol 37:103–109

Jonas D, Spitzmuller B, Weist K, Ruden H, Daschner FD (2003) Comparison of PCR-based methods for typing *Escherichia coli*. Clin Microbiol Infect 9:823–831

Jonas D, Spitzmuller B, Daschner FD, Verhoef J, Brisse S (2004) Discrimination of *Klebsiella pneumoniae* and *Klebsiella oxytoca* phylogenetic groups and other *Klebsiella* species by use of amplified fragment length polymorphism. Res Microbiol 155:17–23

Jones, CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, van de Wiel C, Bredemeijer G, Vosman B, Matthes M, Daly A, Brettschneider R, Bettini P, Buiatti M, Maestri E, Malceschii A, Marmirolli N, Aert R, Volkaert G, Rueda T, Linacero R, Vazques A, Karp A (1997). Reproducibility testing of RADP, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3:381–390

Kang HP, Dunne WM (2003) Stability of repetitive-sequence PCR patterns with respect to culture age and subculture frequency. J Clin Microbiol 41:2694–2696

Karakawa WW, Fournier JM, Vann WF, Arbeit R, Schneerson RS, Robbins JB (1985) Method for the serological typing of the capsular polysaccharides of *Staphylococcus aureus*. J Clin Microbiol 22:445–447

Kessler C (ed) (1992) Nonradioactive labelling and detection of biomolecules. (Springer laboratory series.) Springer, Berlin Heidelberg New York

Kim W, Hong YP, Yoo JH, Lee WB, Choi CS, Chung SI (2002) Genetic relationships of *Bacillus anthracis* and closely related species based on variable-number tandem repeat analysis and BOX-PCR genomic fingerprinting. FEMS Microbiol Lett 207:21–27

King S, McCord BR, Riefler RG (2005) Capillary electrophoresis single-strand conformation polymorphism analysis for monitoring soil bacteria. J Microbiol Methods 60:83–92

Kingsley MT, Straub TM, Call DR, Daly DS, Wunschel SC, Chandler DP (2002) Fingerprinting closely related xanthomonas pathovars with random nonamer oligonucleotide microarrays. Appl Environ Microbiol 68:6361–6370

Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001) rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Res 29:181–184

Kowalchuk GA, Stephen JR, De Boer W, Prosser JI, Embley TM, Woldendorp JW (1997) Analysis of ammonia-oxidizing bacteria of the beta subdivision of the class Proteobacteria in coastal sand dunes by denaturing gradient gel electrophoresis and sequencing of PCR-amplified 16S ribosomal DNA fragments. Appl Environ Microbiol 63:1489–1497

Lanoot B, Vancanneyt M, Dawyndt P, Cnockaert M, Zhang J, Huang Y, Liu Z, Swings J (2004) BOX-pCR fingerprinting as a powerful tool to reveal synonymous names in the genus *Streptomyces*. Emended descriptions are proposed for the species *Streptomyces cinereorectus*, *S. fradiae*, *S. tricolor*, *S. colombiensis*, *S. filamentosus*, *S. vinaceus* and *S. phaeopurpureus*. Syst Appl Microbiol 27:84–92

Leblond-Bourget N, Philippe H, Mangin I, Decaris B (1996) 16S rRNA and 16S to 23S internal transcribed spacer sequence analyses reveal inter- and intraspecific *Bifidobacterium* phylogeny. Int J Syst Bacteriol 46:102–111

Lee DH, Zo YG, Kim SJ (1996) Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. Appl Environ Microbiol 62:3112–3120

van Leeuwen W, Verbrugh H, van Leeuwen N, Heck M, van Belkum A (1999) Validation of binary typing for *Staphylococcus aureus* strains. J Clin Microbiol 37:664–674

Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol 63:4516–4522

Lüdemann H, Arth I, Liesack W (2000) Spatial changes in the bacterial community structure along a vertical oxygen gradient in flooded paddy soil cores. Appl Environ Microbiol 66:754–762

Lueders T, Friedrich M (2000) Archael population dynamics during sequential reduction process in rice fiefld soil. Appl Environ Microbiol 66:2732–2742

Lukow T, Dunfield, PF, Liesack W (2000) Use of T-RFLP technique to assess spatial and temporal changes in the bacterial community structure within agricultural soil planted with transgenic and non-transgenic potato plants. FEMS Microbiol Ecol 32:241–247

Lupski JR, Weinstock GM (1992) Short, interspersed repetitive DNA sequences in prokaryotic genomes. J Bacteriol 174:4525–4529

MacPherson JM, Eckstein PE, Scoles GJ, Gajadhar AA (1993) Variability of the random amplified polymorphic DNA assay among thermal cyclers, and effects of primer and DNA concentration. Mol Cell Probes 7: 293–299

Marsh TL, Saxman P, Cole J, Tiedje J (2000) Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. Appl Environ Microbiol 66:3616–3620

Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R (1992) A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res 20:3479–3483

Martinez-Murcia AJ, Acinas SG, Rodriguez-Valera F (1995) Evaluation of prokaryotic diversity by restrictase digestion of 16S rDNA directly amplified from hypersaline environments. FEMS Microbiol Ecol 17:247–256

Masco L, Huys G, Gevers D, Verbrugghen L, Swings J (2003) Identification of *Bifidobacterium* species using rep-PCR fingerprinting. Syst Appl Microbiol 26:557–563

Massol-Deya AA, Odelson DA, Hickey, RF, Tiedje JM (1995) Bacterial community fingerprinting of amplified 16S and 16–23S ribosomal DNA gene sequences and restriction endonuclease analysis (ARDRA). In: Akkermans ADL, van Elsas JD, de Bruijn FJ (eds) Molecular microbial ecology manual 3.3.2, Kluwer, Dordrecht, pp 1–8

McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, Hakenbeck R, Hryniewicz W, Lefevre JC, Tomasz A, Klugman KP (2001) Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. J Clin Microbiol 39:2565–2571

Meacham KJ, Zhang L, Foxman B, Bauer RJ, Marrs CF (2003) Evaluation of genotyping large numbers of *Escherichia coli* isolates by enterobacterial repetitive intergenic consensus-PCR. J Clin Microbiol 41:5224–5226

Melles DC, Gorkink RF, Boelens HA, Snijders SV, Peeters JK, Moorhouse MJ, van der Spek PJ, van Leeuwen WB, Simons G, Verbrugh HA, van Belkum A (2004) Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. J Clin Invest 114:1732–1740

Mendoza M, Meugnier H, Bes M, Etienne J, Freney J (1998) Identification of *Staphylococcus* species by 16S–23S rDNA intergenic spacer PCR analysis. Int J Syst Bacteriol 48:1049–1055

Menke M, Liesack W, Stackebrandt E (1991) Ribotyping of 16S and 23S rRNA and organization of *rrn* operons in members of the bacterial genera *Gemmata*, *Planctomyces*, *Thermotoga*, *Thermus* and *Verrucomicrobium*. Arch Microbiol 155:263–271

Milyutina IA, Bobrova VK, Matveeva EV, Schaad NW, Troitsky AV (2004) Intragenomic heterogeneity of the 16S rRNA–23S rRNA internal transcribed spacer among *Pseudomonas syringae* and *Pseudomonas fluorescens* strains. FEMS Microbiol Lett 239:17–23

Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Appl Environ Microbiol 65:3518–3525

Moore JE, Millar BC, Jiru X, McCappin J, Crowe M, Elborn JS (2001) Rapid characterization of the genomovars of the *Burkholderia cepacia* complex by PCR-single-stranded conformational polymorphism (PCR-SSCP) analysis. J Hosp Infect 48:129–134

Mora D, Ricci G, Guglielmetti S, Daffonchio D, Fortina MG (2003) 16S–23S rRNA intergenic spacer region sequence variation in *Streptococcus thermophilus* and related dairy streptococci and development of a multiplex ITS-SSCP analysis for their identification. Microbiology 149:807–813

Murchan S, Kaufmann ME, Deplano A, de Ryck R, Struelens M, Zinn CE, Fussing V, Salmenlinna S, Vuopio-Varkila J, El Solh N, Cuny C, Witte W, Tassios PT, Legakis N, van Leeuwen W, van Belkum A, Vindel A, Laconcha I, Garaizar J, Haeggman S, Olsson-Liljequist B, Ransjo U, Coombes G, Cookson B (2003) Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. J Clin Microbiol 41:1574–1585

Murray AE, Hollibaugh JT, Orrego C (1996) Phylogenetic compositions of bacterioplankton from two California estuaries compared by denaturing gradient gel electrophoresis of 16S rDNA fragments. Appl Environ Microbiol 62:2676–2680

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. Curr Opin Microbiol 2:317–322

Muyzer G, Smalla K (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek 73:127–141

Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol 59:695–700

Muyzer G, Teske A, Wirsen CO, Jannasch HW (1995) Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. Arch Microbiol 164:165–172

Nick G, Jusssila M, Hoste B, Maarit R, Kaijalainen S, De Lajudie P, Gillis M, de Bruijn FJ, Lindström K (1999) Rhizobia isolated from root nodules of tropical leguminous trees characterized using DNA–DNA dot-blot hybridization and rep-PCR genomic fingerprinting. System Appl Microbiol 22:287–299

Niemann S, Dammann-Kalinowski T, Nagel A, Puhler A, Selbitschka W (1999) Genetic basis of enterobacterial repetitive intergenic consensus (ERIC)-PCR fingerprint pattern in *Sinorhizobium meliloti* and identification of *S. meliloti* employing PCR primers derived from an ERIC-PCR fragment. Arch Microbiol 172:22–30

Nübel U, Engelen B, Felske A, Snaidr J, Wieshuber A, Amann RI, Ludwig W, Backhaus H (1996) Sequence heterogeneities of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J Bacteriol 178:5636–5643

Nübel U, Schmidt PM, Reiss E, Bier F, Beyer W, Naumann D (2004) Oligonucleotide microarray for identification of *Bacillus anthracis* based on intergenic transcribed spacers in ribosomal DNA. FEMS Microbiol Lett 240:215–223

O'Shea YA, Reen FJ, Quirke AM, Boyd EF (2004) Evolutionary genetic analysis of the emergence of epidemic *Vibrio cholerae* isolates on the basis of comparative nucleotide sequence analysis and multilocus virulence gene profiles. J Clin Microbiol 42:4657–4671

Olive DM, Bean P (1999) Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol 37:1661–1669

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86:2766–2770

Ott SJ, Musfeldt M, Timmis KN, Hampe J, Wenderoth DF, Schreiber S (2004) In vitro alterations of intestinal bacterial microbiota in fecal samples during storage. Diagn Microbiol Infect Dis 50:237–245

Overweg K, Hermans PW, Trzcinski K, Sluijter M, de Groot R, Hryniewicz W (1999) Multidrug-resistant *Streptococcus pneumoniae* in Poland: identification of emerging clones. J Clin Microbiol 37:1739–1745

Øvreas L, Forney L, Daae FL, Torsvik V (1997) Distribution of bacterioplankton in meromictic Lake Saelenvannet, as determined by denaturing gradient gel electrophoresis of PCR-amplified gene fragments coding for 16S rRNA. Appl Environ Microbiol 63:3367–3373

Pace NR, Stahl DA, Lane JL, Olsen GJ (1986) The analysis of natural microbial populations by ribosomal RNA sequences. Adv Microb Ecol 9:1–55

Pastrik K-H, Elphinstone JG, Pukall R (2002) Sequence analysis and detection of *Ralstonia solanacearum* by multiplex PCR amplification of 16S–23S ribosomal intergenic spacer region with internal positive control. Eur J Plant Pathol 108:831–842

Penner GA, Bush A, Wise R, Kim W, Domier L, Kasha K, Laroche A, Scoles G, Molnar SJ, Fedak G (1993) Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. PCR Methods Appl 2:341–345

Perez-Luz S, Fernandez J, Rodriguez-Valera F, Pascual L, Moreno C, Amo A, Apraiz D, Catalan V (2002) Sequence diversity of the internal transcribed spacer (ITS) region of the rRNA operons among different serogroups of *Legionella pneumophila* isolates. Syst Appl Microbiol 25:212–219

Pradella S, Hans A, Sproer C, Reichenbach H, Gerth K, Beyer S (2002) Characterisation, genome size and genetic manipulation of the myxobacterium *Sorangium cellulosum* So ce56. Arch Microbiol 178:484–492

Pradella S, Allgaier M, Hoch C, Pauker O, Stackebrandt E, Wagner-Dobler I (2004) Genome organization and localization of the *pufLM* genes of the photosynthesis reaction center in phylogenetically diverse marine alpha-Proteobacteria. Appl Environ Microbiol 70:3360–3369

Pukall R, Brambilla E, Stackebrandt E (1998) Automated fragment length analysis of fluorescently-labeled 16S rDNA after digestion with 4-base cutting restriction enzymes. J Microbiol Methods 32:55–63

Rademaker JLW, Louws FJ, de Bruijn FJ (1998) Characterization of the diversity of ecologically important microbes by rep-PCR genomic fingerprinting. In: Akkermans ADL, van Elsas JD, de Bruijn FJ (eds) Molecular microbial ecology manual. Kluwer, Dordrecht, pp 1–27

Rademaker JLW, Louws FS, Rossbach U, Vinuesa P, de Bruijn FJ (1999) Computer-assisted pattern analysis of molecular fingerprints and database construction. In: Akkermans ADL, van Elsas JD, de Bruijn FJ (eds) Molecular microbial ecology manual 7.1.3. Kluwer, Dordrecht, pp 1–33

Rademaker JLW, Hoste B, Louws FJ, Kersters K, Swings J, Vauterin L, Vauterin P, de Bruijn FJ (2000) Comparison of AFLP and rep-PCR genomic fingerprinting with DNA–DNA homology studies: *Xanthomonas* as a model system. Int J Syst Evol Microbiol 50:665–677

Reboli AC, Houston ED, Monteforte JS, Wood CA, Hamill RJ (1994) Discrimination of epidemic and sporadic isolates of *Acinetobacter baumannii* by repetitive element PCR-mediated DNA fingerprinting. J Clin Microbiol 32:2635–2640

Regnault B, Grimont F, Grimont PA (1997) Universal ribotyping method using a chemically labelled oligonucleotide probe mixture. Res Microbiol 148:649–659

Reysenbach AL, Giver LJ, Wickham GS, Pace NR (1992) Differential amplification of rRNA genes by polymerase chain reaction. Appl Environ Microbiol 58:3417–3418

Ritacco FV, Halti B, Janso FE, Greenstein M, Bernan VS (2003) Dereplication of Streptomyces soil isolates and detection of specific biosynthetic genes using an automated ribotyping instrument. J Ind Microbiol Biotechnol 30:472–479

Rölleke S, Muyzer G, Wawer C, Wanner G, Lubitz W (1996) Identification of bacteria in a biodegraded wall painting by denaturing gradient gel electrophoresis of PCR-amplified gene fragments coding for 16S rRNA. Appl Environ Microbiol 62:2059–2065

Rosenbaum V, Riesner D (1987) Temperature-gradient gel electrophoresis. Thermodynamic analysis of nucleic acids and proteins in purified form and in cellular extracts. Biophys Chem 26:235–46

Rumpf RW, Griffen AL, Leys EJ (2000) Phylogeny of *Porphyromonas gingivalis* by ribosomal intergenic spacer region analysis. J Clin Microbiol 38:1807–1810

Sakamoto M, Hayashi H, Benno Y (2003) Terminal restriction fragment length polymorphism analysis for human fecal microbiota and its application for analysis of complex bifidobacterial communities. Microbiol Immunol 47:133–142

Savelkoul PH, Aarts HJ, de Haas J, Dijkshoorn L, Duim B, Otsen M, Rademaker JL, Schouls L, Lenstra JA (1999) Amplified-fragment length polymorphism analysis: the state of an art. J Clin Microbiol 37:3083–3091

Schmalenberger A, Tebbe CC (2003) Bacterial diversity in maize rhizospheres: conclusions on the use of genetic profiles based on PCR-amplified partial small subunit rRNA genes in ecological studies. Mol Ecol 12:251–262

Schwartz DC, Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37:67–75

Schmalenberger A, Schwieger F, Tebbe CC (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl Environ Microbiol 67:3557–3563

Schwieger F, Tebbe CC (1998) A new approach to utilize PCR-single-strand-conformation polymorphism for 16S rRNA gene-based microbial community analysis. Appl Environ Microbiol 64:4870–4876

Sekiguchi H, Tomioka N, Nakahara T, Uchiyama H (2001) A single band does not always represent single bacterial species in denaturing gradient gel electrophoresis analysis. Biotechnol Lett 23:1205–1208

Sharples GJ, Lloyd RG (1990) A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. Nucleic Acids Res 18:6503–6508

Shaver YJ, Nagpal ML, Rudner R, Nakamura LK, Fox KF, Fox A (2002) Restriction fragment length polymorphism of rRNA operons for discrimination and intergenic spacer sequences for cataloging of *Bacillus subtilis* sub-groups. J Microbiol Methods 50:215–223

Snelling AM, Gerner-Smidt P, Hawkey PM, Heritage J, Parnell P, Porter C, Bodenham AR, Inglis T (1996) Validation of use of whole-cell repetitive extragenic palindromic sequence-based PCR (REP-PCR) for typing strains belonging to the *Acinetobacter calcoaceticus–Acinetobacter baumannii* complex and application of the method to the investigation of a hospital outbreak. J Clin Microbiol 34:1193–1202

Söller R, Hirsch P, Blohm D, Labrenz M (2000) Differentiation of newly described antarctic bacterial isolates related to *Roseobacter* species based on 16S–23S rDNA internal transcribed spacer sequences. Int J Syst Evol Microbiol 50:909–915

Song Y, Kato N, Liu C, Matsumiya Y, Kato H, Watanabe K (2000) Rapid identification of 11 human intestinal *Lactobacillus* species by multiplex PCR assays using group- and species-specific primers derived from the 16S–23S rRNA intergenic spacer region and its flanking 23S rRNA. FEMS Microbiol Lett 187:167–173

Spigaglia P, Mastrantonio P (2003) Evaluation of repetitive element sequence-based PCR as a molecular typing method for *Clostridium difficile*. J Clin Microbiol 41:2454–2457

Stackebrandt E, Brambilla E, Cousin S, Dirks W, Pukall R (2004). Culture-independent analysis of bacterial species from an anaerobic mat from Lake Fryxell, Antarctica: prokaryotic diversity revisited. Cell Mol Biol 50: 517–524

Stern MJ, Ames GF, Smith NH, Robinson EC, Higgins CF (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell 37:1015–1026

Szczuka E, Kaznowski A (2004) Typing of clinical and environmental *Aeromonas* sp. strains by random amplified polymorphic DNA PCR, repetitive extragenic palindromic PCR, and enterobacterial repetitive intergenic consensus sequence PCR. J Clin Microbiol 42:220–228

Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B (1995) Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239

Teske A, Sigalevich P, Cohen Y, Muyzer G (1996) Molecular identification of bacteria from a coculture by denaturing gradient gel electrophoresis of 16S ribosomal DNA fragments as a tool for isolation in pure cultures. Appl Environ Microbiol 62:4210–4215

del Tufo JP, Tingey SV (1994) RAPD assay. A novel technique for genetic diagnostics. Methods Mol Biol 28:237–241

Tyler KD, Wang G, Tyler SD, Johnson WM (1997) Factors affecting reliability and reproducibility of amplification-based DNA fingerprinting of representative bacterial pathogens. J Clin Microbiol 35:339–346

Tynkkynen S, Satokari R, Saarela M, Mattila-Sandholm T, Saxelin M (1999) Comparison of ribotyping, randomly amplified polymorphic DNA analysis, and pulsed-field gel electrophoresis in typing of *Lactobacillus rhamnosus* and *L. casei* strains. Appl Environ Microbiol 65:3908–3914

Vallaeys T, Topp E, Muyzer G, Macheret V, Laguerre G, Rigaud A, Soulas G (1997) Evaluationg of denaturing gradient gel electrophoresis in the detection of 16S rDNA sequence variation in rhizobia and methanotrophs. FEMS Microbiol Ecol 24:279–285

Van Reenen CA, Dicks LM (1996) Evaluation of numerical analysis of random amplified polymorphic DNA (RAPD)-PCR as a method to differentiate *Lactobacillus plantarum* and *Lactobacillus pentosus*. Curr Microbiol 32:183–187

Vaneechoutte M, Rossau R, de Vos P, Gillis M, Janssens D, Paepe N, De Rouck A, Fiers T, Claeys G, Kersters K (1992) Rapid identification of bacteria of the Comamonadaceae with amplified ribosomal DNA-restriction analysis (ARDRA). FEMS Microbiol Lett 72:227–233

Vaneechoutte M, De Beenhouwer H, Claeys G, Verschraegen G, De Rouck A, Paepe N, Elaichouni A, Portaels F (1993) Identification of *Mycobacterium* species by using amplified ribosomal DNA restriction analysis. J Clin Microbiol 31:2061–2065

Ventura M, Zink R (2002a) Rapid identification, differentiation, and proposed new taxonomic classification of *Bifidobacterium lactis*. Appl Environ Microbiol 68:6429–6434

Ventura M, Zink R (2002b) Specific identification and molecular typing analysis of *Lactobacillus johnsonii* by using PCR-based methods and pulsed-field gel electrophoresis. FEMS Microbiol Lett 217:141–154

Ventura M, Zink R (2003) Comparative sequence analysis of the *tuf* and *recA* genes and restriction fragment length polymorphism of the internal transcribed spacer region sequences supply additional tools for discriminating *Bifidobacterium lactis* from *Bifidobacterium animalis*. Appl Environ Microbiol 69:7517–7522

Ventura M, Meylan V, Zink R (2003) Identification and tracing of *Bifidobacterium* species by use of enterobacterial repetitive intergenic consensus sequences. Appl Environ Microbiol 69:4296–4301

Versalovic J, Koeuth T, Lupski JR (1991) Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Res 19:6823–6831

Versalovic J, Schneider M, de Burijn FJ, Lupski JR (1994) Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. Methods Mol Cell Biol 5:25–40

Wawer C, Jetten MS, Muyzer G (1997) Genetic diversity and expression of the [NiFe] hydrogenase large-subunit gene of *Desulfovibrio* spp in environmental samples. Appl Environ Microbiol 63:4360–4369

Webster CA, Towner KJ (2000) Use of RAPD-ALF analysis for investigating the frequency of bacterial cross-transmission in an adult intensive care unit. J Hosp Infect 44:254–260

Webster CA, Towner KJ, Humphreys H, Ehrenstein B, Hartung D, Grundmann H (1996) Comparison of rapid automated laser fluorescence analysis of DNA fingerprints with four other computer-assisted approaches for studying relationships between *Acinetobacter baumannii* isolates. J Med Microbiol 44:185–194

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Widjojoatmodjo MN, Fluit AC, Verhoef J (1994) Rapid identification of bacteria by PCR-single-strand conformation polymorphism. J Clin Microbiol 32:3002–3007

Wieser M, Busse HJ (2000) Rapid identification of *Staphylococcus epidermidis*. Int J Syst Evol Microbiol 50:1087–1093

Willems A, Munive A, de Lajudie P, Gillis M (2003) In most *Bradyrhizobium* groups sequence comparison of 16S–23S rDNA internal transcribed spacer regions corroborates DNA–DNA hybridizations. Syst Appl Microbiol 26:203–210

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Wintzingerode F von, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213–229

Witte W (2002) Molekulare Typisierung bakterieller Infektionserreger. Chemother J 3:95–101

Woese C (1987) Bacterial Evolution. Microbiol Rev 51:221–271

Xu D, Cote JC (2003) Phylogenetic relationships between *Bacillus* species and related genera inferred from comparison of 3′ end 16S rDNA and 5′ end 16S–23S ITS nucleotide sequences. Int J Syst Evol Microbiol 53:695–704

Yasuda M, Shiaris MP (2005) Differentiation of bacterial strains by thermal gradient gel electrophoresis using non-GC-clamped PCR primers for the 16S–23S rDNA intergenic spacer region. FEMS Microbiol Lett 243:235–242

Yeung PS, Kitts CL, Cano R, Tong PS, Sanders ME (2004) Application of genotypic and phenotypic analyses to commercial probiotic strain identity and relatedness. J Appl Microbiol 97:1095–1104

Yukphan P, Potacharoen W, Nakagawa Y, Tanticharoen M, Yamada Y (2004) Identification of strains assigned to the genus *Gluconobacter*, Asai 1935 based on the sequence and the restriction analyses of the 16S–23S rDNA internal transcribed spacer regions. J Gen Appl Microbiol 50:9–15

Zoetendal EG, Akkermans AD, De Vos WM (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl Environ Microbiol 64:3854–3859

4 Multiple Locus VNTR (Variable Number of Tandem Repeat) Analysis

Gilles Vergnaud, Christine Pourcel

4.1 Introduction

The present chapter will review the current state of the art in the field of bacterial strain typing through the use of tandem repeat polymorphism. We will first go through a brief overview of multiple locus VNTR analysis (MLVA) typing and then describe how to set-up or enrich a MLVA assay. We will also review representative examples of the currently proposed MLVA assays and discuss the methods used for MLVA data analysis. Finally, we will compare MLVA to other approaches, and discuss issues related to standardisation and possibilities offered by the internet in terms of shared databases for MLVA (MLVA web services).

4.2 MLVA Origins

The recognition of tandem repeats as often highly polymorphic loci is more than 20 years old. In the early 1980s, a number of laboratories trying to develop the first drafts of the human genetic map were characterising so-called restriction fragment length polymorphisms (RFLPs). Southern blots carrying DNA from large human families were systematically hybridised with DNA probes recognising a single locus in the human genome. RFLPs were bi-allelic and the maximum polymorphism information content (PIC) index (calculated as 1.0 minus the sum of the squares of allelic frequencies) was 0.5. One probe yielded an astonishing result, with multiple alleles and a PIC value well above 0.5. Detailed molecular analysis demonstrated that the observed polymorphism was the result of variations in the number of units in a tandem repeat. The first tandem repeats characterised were satellite DNAs. These tandem repeats cover megabases of DNA; and they

Gilles Vergnaud: Division of Analytical Microbiology, Centre d'Etudes du Bouchet, B.P. 3, 91710 Vert le Petit, France, E-mail: gilles.vergnaud@igmors.u-psud.fr

Christine Pourcel: GPMS laboratory, Institute of Genetics and Microbiology, University Paris XI, 91405 Orsay cedex, France

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

represent a sufficiently large portion of some eukaryote genomes to be able to produce a "satellite" band on caesium chloride density gradients, as soon as the repeat unit has a nucleotide composition slightly different from the genome average. For this historical reason, the small tandem repeats (in the kilobase range) analysed by Southern blotting were called minisatellites and, later, even smaller structures were called microsatellites. Tandem repeat structures cover a number of different situations in terms of origin, mode of evolution, mutation rate and function (when identified). When used for typing purposes, one key feature is the associated length polymorphism. Polymorphic tandem repeats are most often called VNTRs, which includes polymorphic mini- and microsatellites (for a review, see Vergnaud and Denoeud 2000). Towards the end of the 1980s, the advent of the PCR technology made possible the large-scale typing of the shorter tandem repeats to the extent that eventually the human genetic linkage map was essentially based upon microsatellite typing (Weissenbach et al. 1992). The second immediate application of highly polymorphic markers was individual identification; and tandem repeats polymorphism is still the basis of current forensic methods for DNA-based identification in humans. The assay is strictly speaking a multiple locus VNTR analysis, but the MLVA acronym was coined years later in the field of microbiological molecular epidemiology and forensics.

4.3 MLVA Set-up and Enrichment

Tandem repeats were also identified in prokaryotes during the 1980s and the polymorphism associated with a few specific genes, investigated for other reasons, was described. Multiple locus tandem repeats variability was shown to be promising for bacteria typing by Southern blotting and hybridisation with a GC-rich tandem repeat probe (Ross et al. 1992) or even an oligonucleotide probe (Marshall et al. 1996) as previously done in human genomics (Vergnaud 1989). It is the availability of large-scale sequence data which opened the way to PCR-based MLVA assays. The method was applied initially to *Haemophilus influenzae* (van Belkum et al. 1997) with an assay comprising five tetranucleotide microsatellites. However, all bacterial species are not equally amenable to MLVA typing and the first step in setting-up a MLVA assay is to evaluate the potential of MLVA typing for the species of interest.

4.3.1
Evaluation of the Potential Interest of MLVA for a Given Species

The identification of tandem repeats from sequence data is easily achieved owing to the availability of genome sequence data and software for sequence analysis (Benson 1999) and even unfinished, low-coverage genome sequence data can be used. Taking advantage of these resources, Le Flèche et al. (2001, 2002) and Denoeud and Vergnaud (2004) have developed and made available a tandem repeat database as part of a first "MLVA web service". The initial release in year 2001 contained 36 bacterial genomes, compared to the more than 200 genomes available in the latest update. In addition, the database also includes genome comparison results when two or more strains from the same (or sufficiently genetically close) species have been sequenced: the tandem repeats with a different size in the two strains are automatically identified (Denoeud and Vergnaud 2004). This greatly facilitates the identification of candidate polymorphic loci, as shown for instance by Ramisse et al. (2004). In order to avoid the duplication of work by independent groups and to limit the giving of different names to the same locus (as recalled by, for instance, Le Flèche et al. 2002), the database includes links to tandem repeats which have already been investigated and given names in the literature. These resources are accessible over the internet (http://minisatellites.u-psud.fr) and can also be set-up locally.

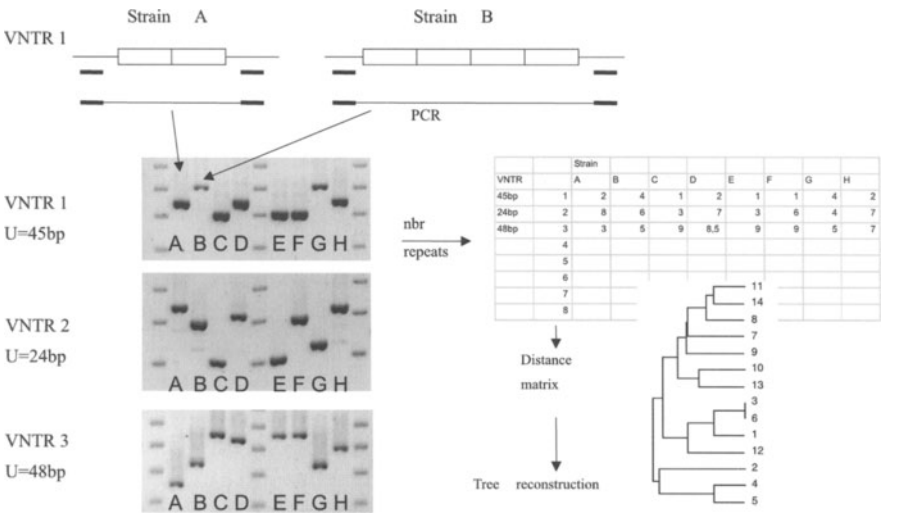


Fig. 4.1. Schematic representation of a MLVA scheme. Primers are chosen on both sides of VNTR loci and PCR products are electrophoresed (here on agarose gel) together with size markers. The amplicon size is converted into a repeat number. Multiple markers are analysed in the same way, a distance matrix is generated and a clustering analysis is produced

Candidate tandem repeats can then be tested on a few diverse strains. Less than ten strains will usually be sufficient and polymorphism can be easily evaluated on agarose gels, so that tens of loci can be quickly tested at low cost in a couple of weeks (Fig. 4.1).

4.3.2 MLVA Validation

After this quick screening has been achieved, it is necessary to precisely identify the need and to define and collect an appropriate reference strain collection. Ideally, the reference collection to be used should have been already characterised and typed using the currently recognised typing methods, so that MLVA can immediately be compared in terms of typeability, reproducibility, relevance and discriminatory efficiency. In particular, different distance coefficients and clustering methods can be evaluated and the dendrograms obtained can be compared with the known epidemiological relations between the strains. Often a few tens of relevant strains will be sufficient for this phase of setting-up an assay. Then the strength and validity of the assay increases as many more strains are genotyped and similarity coefficients and clustering methods are fully tested and validated. Once strains have been selected, the PCR-amplification of tandem repeat loci using primers flanking the array and the measuring of the PCR product length are relatively standard (summarised in Fig. 4.1). Any equipment able to measure a DNA fragment length with sufficient resolution depending on the repeat unit size can be used. Maximum resolution means that the method used must be able to confidently resolve PCR products differing by one repeat unit. Sophisticated equipments such as DNA sequencing machines are able to do this; and such machines may even be necessary for typing arrays with short or very short repeat units, or relatively long alleles. The majority of current needs can be satisfied by methods with a lower resolution, in particular agarose gels and ethidium bromide staining. The MLVA assay can then be run with very ordinary equipment and at very low cost in terms of consumables and equipment. A typical agarose gel MLVA typing set-up consists of a control strain and size marker, each loaded a number of times on each gel (usually 4–7 times, depending upon gel size) in order to be able to take into account and compensate for both intra- and inter-gel electrophoresis variations, as described for instance by Pourcel et al. (2004). In any case, and whatever method is used, the resulting data can be compared and merged only if the appropriate quality control procedures, common reference strains and identical allele assignment conventions are used. Figure 4.2 illustrates a typical MLVA set-up based on agarose gel electrophoresis.

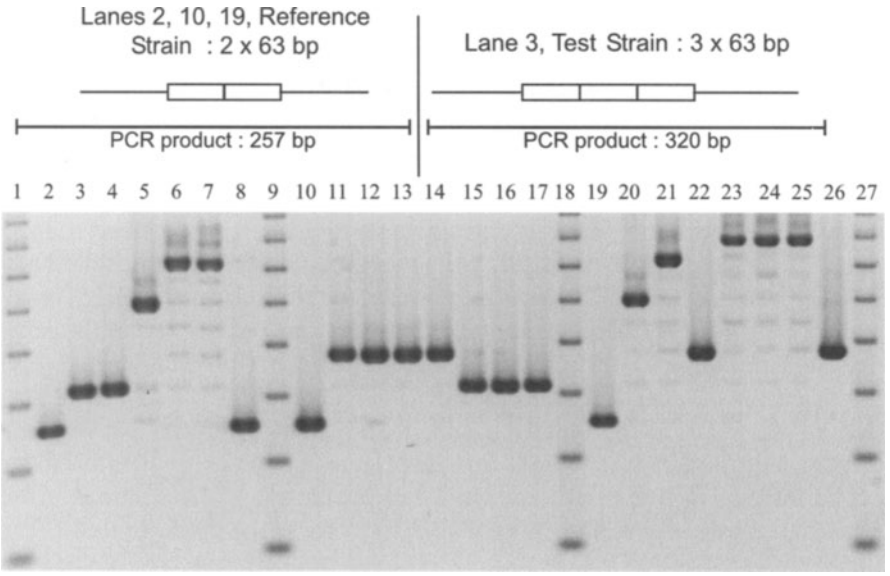


Fig. 4.2. MLVA typing on agarose gel. A tandem repeat from *Brucella* was amplified across 20 strains. Seven lanes are dedicated to controls, i. e. a size marker (four lanes: 1, 9, 18, 27) and a reference strain (three lanes: 2, 10, 19). The size marker used here is a 100-bp ladder (bands from 100 bp to 800 bp are shown). The repeat unit is 63 bp long. Such gels can be easily read manually: six different alleles are observed, comprising 2–8 repeat units

4.3.3 Data Management

The end-product of the assay is typing data, expressed in repeat copy number. These very simple files can be easily merged to produce integrated databases from different sources. When running small-scale projects, limited to a few tens of strains and/or when the biological characterisation of strains is limited to MLVA typing, there is little need for a real database management system. Careful double-checked manual reading and typing into a text file are appropriate (Fig. 4.2). However, when running larger projects and when different kinds of data must be stored and eventually merged for analysis, dedicated data management software is needed. The most widely used such software is Bionumerics (Applied-Maths) which acts as a warehouse for the storage of any biological data and also contains a collection of powerful tools for data analysis.

4.4 Existing First-generation MLVA Assays

MLVA is still quite new. So far, no official standard has been defined for any of the bacteria which will be presented below or are listed in Table 4.1. One reason for this is that MLVA assays are still in the development phase, in terms of the number of markers and strains tested, but it is very likely that, in the coming years, such standards will emerge, at least for the most actively investigated bacteria. Another reason is that the resolution of a MLVA assay can be increased by adding markers (Fig. 4.3), but requirements in terms of resolution depend upon the epidemiological question being asked. The investigation of local outbreaks for instance will benefit from the use of tandem repeats with a high mutation rate, in addition to a routine MLVA assay for strain typing. In other words, the single term MLVA assay will often cover probably two or three complementary panels of markers. In some cases, the use of a few markers will be quite sufficient to cover the need. Table 4.1 lists the bacteria for which MLVA assays have been published so far. In many instances, only one study has been reported, often including only a few markers and a limited number of strains. In other cases, much more work has already been done. Interestingly, a significant fraction of the more thorough investigations is related to pathogens which represent potential biological warfare agents. In this area of technological development, as in others before, it may be so that defence-driven projects related to microbial forensics will contribute to and speed up the development of epidemiological tools for many other pathogens which represent significant human health issues.

▶ **Fig. 4.3.** Comparison of the discrimination power of a MLVA analysis with 5, 13 or 19 VNTR. A collection of 50 strains from the *M. tuberculosis* complex (MTBC) were typed using 19 markers; and phylogenetic trees were produced using the data from either five markers (ETRs, *left panel*), 13 markers (5 ETRs + 8 MIRUs, *middle panel*) or 19 markers (5 ETRs + 8 MIRUs + 6 Mtubs, *right panel*). The strains were independently assigned to a MTBC group by classic biochemical assays and microdeletion typing (codes: *CAN* "*M. canettii*" strains, *EAI* East Africa/India, *BOV-AFRI M. bovis* and some *M. africanum* strains, *AFRI1* the rest of *M. africanum* type 1 strains, *BEIJ* Beijing strains, *MOD-CDC* the group of modern *M. tuberculosis* strains, including the reference CDC1551 strain, *MOD-H37* the group of modern *M. tuberculosis* strains, including the reference H37Rv strain).When all 19 markers are used in the analysis, 50 genotypes are identified (numbered from 1 to 50). The clustering fits with the independent classification. When 13 markers are used, the discrimination is slightly reduced (46 different genotypes identified). The clustering achieved is still reasonable, with a few inconsistencies: genotype 16 (EAI strain) is grouped with "*M. canettii*" strains and three genotypes from modern *M. tuberculosis* strains are incorrectly assigned (genotypes 44, 49, 40) to the Beijing group of strains. When only five markers are used (*left panel*), 36 different genotypes are identified, which is still relatively high, but the clustering achieved is of little value

We will discuss in more details the application of MLVA for epidemiology- or phylogeny-related investigations of five representative species: *Mycobacterium tuberculosis, Bacillus anthracis, Yersinia pestis, Brucella* sp, and *Legionella pneumophila*, for which enough data exist to assess the validity of the technique, or which illustrate specific points of interest.

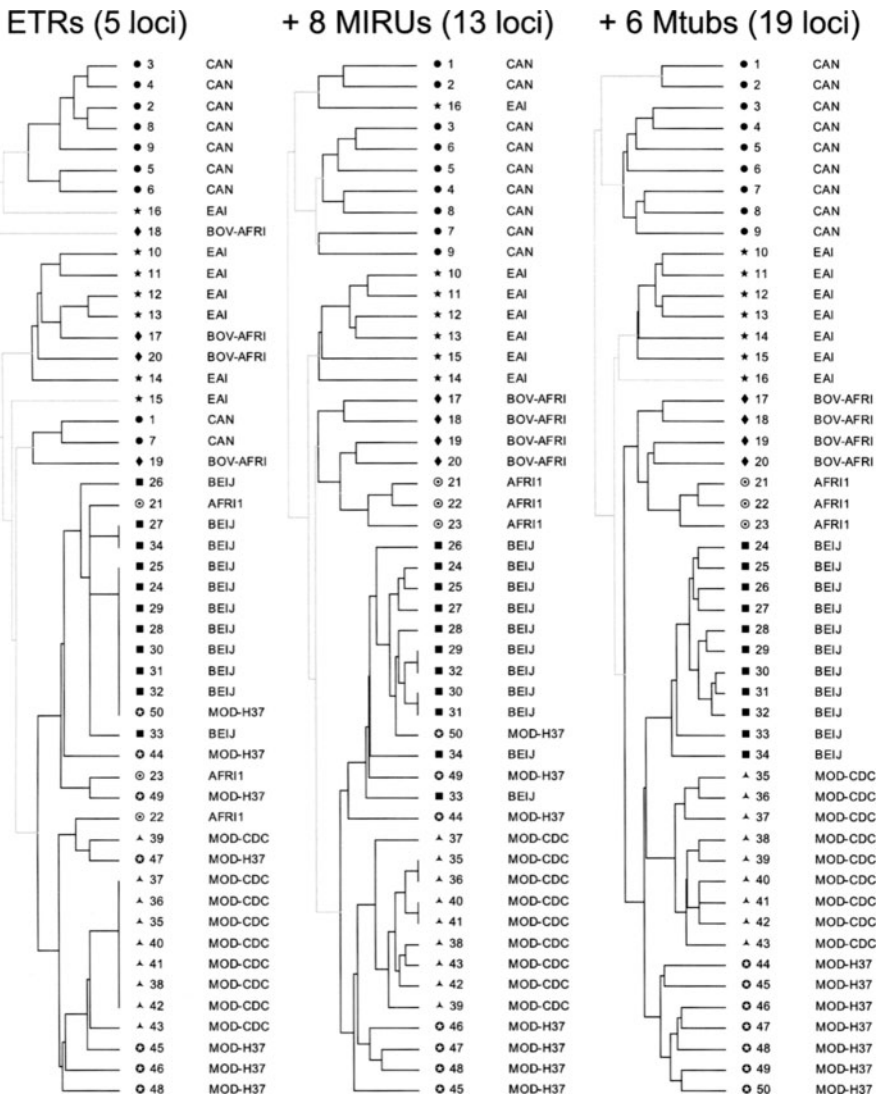


Table 4.1. List of MLVA development reports. Methods: *seq* gel-based sequencing machine, *capi* capillary electrophoresis, *aga* agarose gel

Bacteria[a]	VNTR loci[b]	Repeat (bp)[c]	Isolates	Method	Reference
Bacillus anthracis	8	2–36	426	seql	Keim et al. (2000)
	24 (18)	9–78	32	agarose	Le Flèche et al. (2001)
Bordetella pertussis	6	5–15	198	seq	Schouls et al. (2004)
Borrelia sp.	10	2–21	41	seq	Farlow et al. (2001)
	8	8	22	capi	Bricker et al. (2003)
Candida albicans[a]	3	4	100	seq	Botterel et al. (2001)
Enterococcus faecalis	7	141–393	83	aga	Titze-de-Almeida et al. (2004)
Enterococcus faecium	6	121–279	392	aga	Top et al. (2004)
Escherichia coli O157	7	6–18	81	sequencing	Noller et al. (2003)
	7	6–30	73	capi	Lindstedt et al. (2003)
Francisella tularensis	6	2–21	56	seq	Farlow et al. (2001)
	25	2–23	192	seq	Johansson et al. (2004)
Hemophilus influenzae	5	3–6	20	aga	van Belkum et al. (1997)
Legionella pneumophila	6	18–125	78	aga	Pourcel et al. (2003)
Leptospira interrogans	7	34–77	51	aga	Majed et al. (2005)
Mycobacterium avium	6	53	73	aga	Bull et al. (2003)
	5	20–70	50	aga	Overduin et al. (2004)
M. leprae	5	2–3	12	sequencing	Truman et al. (2004)
	9	1–27	4	seq	Groathouse et al. (2004)
M. tuberculosis	7	15–79	25	aga	Frothingham and Meeker-O'Connell (1998)
	12 (10)	53	31	aga	Supply et al. (2000)
	6	69	100	aga	Skuce et al. (2002)
	21 (8)	9–58	90	aga	Le Flèche et al. (2002)
Pseudomonas aeruginosa	7	6–115	89	aga	Onteniente et al. (2003)

Table 4.1. (continued)

Bacteria[a]	VNTR loci[b]	Repeat (bp)[c]	Isolates	Method	Reference
Salmonella typhimurium/ typhi	8	6–189	102	capi	Lindstedt et al. (2003)
	5	7–26	61	aga	Liu et al. (2003)
	10 (7)	3–20	99	aga	Ramisse et al. (2004)
Staphylococcus aureus	7	48–159	16	aga	Hardy et al. (2004)
Xylella fastidiosa	7	7–9	27	aga	Coletta-Filho et al. (2001)
Yersinia pestis	25	9–60	3+180	aga	Le Flèche et al. (2001), Pourcel et al. (2004)
	42 (35)	1–45	24+156	seq	Klevytska et al. (2001), Achtman et al. (2004)

[a] with the exception of *C. albicans*

[b] number of loci proposed for MLVA (number of new loci)

[c] repeat unit size range explored in the report

4.4.1 *Mycobacterium tuberculosis*

This is the bacterium for which MLVA has been the most extensively used to date and for which a large body of data is available. VNTR markers have been described by different teams and used alone or in combination. Particularly interesting markers were the exact tandem repeats (ETRs; Frothingham and Meeker-O'Connell 1998), multiple interspersed repetitive units (MIRUs; Supply et al. 2000), QUBs (for Queen's University of Belfast; Skuce et al. 2002) and Mtubs (Le Flèche et al. 2002). ETRs are 53–79 bp long and only ETRA is located within an ORF. The allelic profiles are reproducible and stable and VNTR typing was proposed to be useful for strain differentiation and evolutionary studies. MIRUs are tandem duplications of 53 bp except for MIRU04, a 77 bp repeat, which in fact corresponds to ETRD. MIRU31 corresponds to ETRE. Most are present in regions separating genes. In contrast, QUBs are mostly located inside genes. ETRA, QUb11a and QUB11b are present in the same protein, pUCB, a protein of the PPE family (O'Brien et al. 2000). They show a very high level of polymorphism. Additional informative markers were described by Le Flèche et al. (2002), in particular Mtub21 and Mtub39, both localised in intergenic regions. The size of the repeats in these different VNTRs is such that agarose gel-based MLVA can be performed. However, automated procedures are commonly used (Supply et al. 2001; Spurgiesz et al. 2003) and their utility

in clinical mycobacteriology analysis was recently demonstrated (Allix et al. 2004).

MLVA was compared to classic typing methods for *M. tuberculosis*: spoligotyping, which investigates the polymorphism of a single locus, the DR locus, and IS typing, usually performed by RFLP analysis (Sun et al. 2004). The most recent studies concluded that the resolution of MLVA compares favourably with the other techniques when a sufficient number of informative markers are used, i. e. more than the most frequently used set of 12 MIRUs. For instance, MLVA appears to be the best method to investigate the diversity inside the important "Beijing" family, a recently emerged group of strains. MLVA assay was also a key assay in describing the group of "*M. canettii*" as a single entity (Fabre et al. 2004). ETRA, a very informative marker for the complete *M.tuberculosis* complex, shows a single allele in "*M. canettii*," an allele which has been found only in two *M. tuberculosis* strains belonging to the more ancient family from East Africa/India (Pourcel, unpublished data). This family can be identified on the basis of a specific allele of MIRU24, an otherwise very poorly informative marker (Sun et al. 2004). However, although many reports suggest that MLVA may be the new gold standard technique for typing inside the *M. tuberculosis* complex, more needs to be done to define a common assay allowing comparison of data between laboratories. Some markers are commonly used (with sometimes different names), whereas others are only used by some laboratories. In addition, there is a wrong assumption that some markers, because they do not seem informative inside a subgroup, should not be used although they are clearly useful when a large population of strains is studied. In contrast, a marker such as QUB-11a, a highly polymorphic repetition, can be useful in epidemic situations because of rapid modifications but is probably not stable enough for phylogenetic studies. In recent reports, it was proposed that VNTR typing should be used in combination with IS6110RFLP, a rather cumbersome technique necessitating the preparation of high quality DNA. Instead, the addition of several VNTR markers, already described in the literature, bringing the total number to 19, should be sufficient for high resolution analysis. Figure 4.3 shows a clustering analysis performed on a collection of 50 strains of the *M. tuberculosis* complex using either five loci (the five ETRs), 13 loci (the five ETRs and eight MIRUs) or 19 loci (the previous markers plus six Mtubs). The strains were selected from our collection to contain representatives of the major *M. tuberculosis* families ("Modern", "Beijing", ancient East Africa/India), plus some *M. bovis*, *M. africanum*, and "*M. canettii*"strains; and a similar pattern was observed even when more strains were used. Interestingly, these major groups are well defined by biochemical assays or by the independent tools (micro-deletion typing) described by Marmiesse et al. (2004). Typing with five markers is clearly not robust, even if the discriminatory power

is already very good (36 genotypes resolved). With 13 markers (the ten most relevant MIRUs, the ETRs), a much nicer clustering is achieved with still some inconsistencies: one East Africa/India strain is grouped with the *M. canettii* group and the Beijing and Modern clusters are poorly defined. Forty-six genotypes are resolved. The panel of 19 markers proposed by Fabre et al. (2004) correctly clusters the strains and 50 genotypes are resolved. If necessary, the typing assay could be extended to 25 easily typable VNTRs, by using some QUB markers and additional markers uncovered by sequence comparison between the *M. bovis* genome and *M. tuberculosis* (Le Flèche, unpublished data).

4.4.2 *Bacillus anthracis*

B. anthracis is a highly monomorphic species, recently emerged from the *B. cereus/thuringiensis* group through the acquisition of two virulence plasmids. The main reason for the development of a MLVA assay in this dangerous pathogen is microbial forensics. This bacterium is no longer a significant health problem but is a potential bioterrorist agent, as illustrated by the 2001 events. MLVA-based genotype databases have been the key tools to identify the precise strain which was used in the bioterrorist event. The first MLVA assay was built upon a number of contributions. An extensive search for DNA polymorphisms eventually led to the finding that tandem repeats were a major source of polymorphism in this organism. The assay comprised eight markers, two of which were located on the virulence plasmids (Keim et al. 2000); and 426 isolates were typed. A number of these isolates were collected during a single outbreak, or corresponded to reference strains conserved for a number of years in different laboratories. With very few exceptions, the genotypes were indeed identical, which demonstrated that most tandem repeats are sufficiently stable to define strains. Eighty-nine genotypes were resolved. Whereas some genotypes were restricted to geographic regions, others were found to be widely distributed.

Taking advantage of the availability of large-scale sequence data, the MLVA assay was later expanded by adding 18 new markers (Le Flèche et al. 2001; http://bacterial-genotyping.igmors.u-psud.fr/). All of these markers are located on the chromosome within ORFs. Some of the encoded proteins are components of the outer layers of the spore. Bams13, for instance, is a 9-bp repeat located within the *BclA* gene which shows a 500-bp size difference between the largest and the smallest alleles (Le Flèche et al. 2001). The collagen-like BclA protein is the main component of the *B. anthracis* exosporium; and the Bams13 length polymorphism is directly related to the

exosporium size (Sylvestre et al. 2003). The typing of 32 isolates confirmed the existence of the two main clusters, A and B, identified by (Keim et al. 2000) and showed the existence of additional clearly distinct branches, represented by isolates from West Africa. Much work still needs to be done, using MLVA in combination with other DNA analysis methods, on both *B. anthracis* and *B. cereus* to, for instance, identify the geographic origin of *B. anthracis*. The currently proposed assay comprises 24 loci which can be typed by agarose gels or by capillary electrophoresis, has a much higher resolution than the earlier 8-markers assay and represents a good first-level typing assay for phylogenetic investigations. To investigate local outbreaks, microsatellites (i.e. tandem repeats with the shortest repeat units) might constitute a second-level MLVA set. Eventually, it can only be hoped that all new isolates identified in the world will be genotyped and the genotypes submitted to common databases, as illustrated by the prototype (http://bacterial-genotyping.u-psud.fr).

4.4.3 *Yersinia pestis*

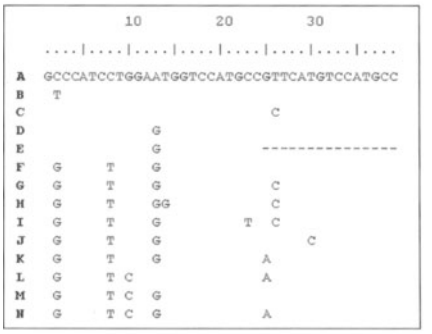
The first report of the analysis of VNTR polymorphism at one locus in *Y. pestis* by Adair et al. (2000) suggested that these sequences could be a useful source of polymorphism in this very monomorphic species. Indeed, the works of Le Flèche et al. (2001) and Klevytska et al. (2001) confirmed that a MLVA scheme could be used to efficiently genotype *Y. pestis* strains. The choice of markers by the two teams was very complementary, in part due to the different electrophoresis techniques used and only seven markers were common to the two sets (Pourcel et al. 2004). Markers with small repetitions, of the microsatellite class, were favoured by the Keim laboratory, whereas in our laboratory we chose markers with repetitions larger than 12 bp, to allow for agarose gel separation of alleles. The selection of markers on the basis of the repetition size can have consequences on their discriminatory efficiency, as the mechanisms of variability of microsatellite and minisatellite (more than 9 bp long) can be different. More recently, significantly larger and more diverse collection of strains have been analysed using the two sets of markers (Achtman et al. 2004; Pourcel et al. 2004). Previous classifications were essentially based upon biochemical assays, which define the three classically recognised biovars: Antiqua, Medievalis and Orientalis. The MLVA clustering is usually in agreement with this rough classification but provides a much higher discrimination. The results obtained clearly distinguish between Antiqua strains from Asia and Antiqua strains from Africa. Interestingly, a few abnormalities were uncovered, with some Medievalis strains clustering among Antiquas. Fur-

ther investigations demonstrated that the Medievalis phenotype resulted in these strains from different mutation events inactivating the *napA* gene. Although other very high resolution typing methods had been used for some years before, including IS typing by Southern blotting, these methods were unable to detect these inconsistencies, which illustrates the power of MLVA typing.

4.4.4 *Brucella sp.*

The single study published so far on MLVA typing in the *Brucella* genus illustrates some aspects of MLVA set-up and marker selection. Bricker et al. (2003) investigated the polymorphism associated with a family of octameric tandem repeats located within a mobile element present in multiple copies in the *Brucella* genome. The mobile element is small enough to use at least one primer located outside of it, so that each locus can be amplified and analysed independently from the others. The *Brucella* genus is separated in a number of species, not for genetic reasons (the genus is very highly homogeneous) but because of some features of the associated disease, a strong host specificity within mammals and varying virulence in human (Moreno et al. 2002). Each species can be further separated in a few biovars by biotyping, which is a combination of biochemical assays, phage typing, serotyping and growth in the presence of specific dyes. Biotyping data necessitates the manipulation of live bacteria, has a low resolution and is sometimes ambiguous, so that alternative, DNA-based assays would clearly be of interest. The 8-loci MLVA assay proposed by Bricker et al. (2003) is very discriminatory, highly reproducible and all strains investigated can be fully typed for the eight loci. However, the clustering of strains deduced from the MLVA typing data does not fit with the biotype or even with the species assignment. The best explanation for this behaviour and inconsistency is that the tandem repeat loci used have such a high mutation rate, within a limited allele size range, that many alleles have an identical size in spite of a different evolutionary origin (homoplasy). Because the repeat array is perfect, such alleles will be strictly identical and cannot be distinguished even by sequencing, as can be done when internal heterogeneity exists (see paragraph below; Fig. 4.4). Such an assay cannot replace the existing tools and is limited to the investigation of local outbreaks.

Hopefully, the existence of many additional tandem repeat sequences in the *Brucella* genome indicates that a MLVA assay will eventually be developed for this species (Le Flèche et al., in preparation).



B

Ms77 allele size	coding
05C020_520bp_6U	AFFFGM
22D032_481bp_5U	ALKKL
12A241_481bp_5U	AFGMN
05A398_481bp_5U	ALGMN
05A593_481bp_5U	BLFGN
04C010_442bp_4U	AFGM
PAO1_77_442bp_4U	**AFGJ**
19A211_403bp_3U	ALF
10A353_403bp_3U	ALM
22E042_403bp_3U	ALM
11C010_403bp_3U	ALM
08D038_403bp_3U	AHM
09A086_403bp_3U	AHM
04A163_388bp_2.5U	EIM
05A400_388bp_2.5U	EFL
03C030_388bp_2.5U	EFL
01A129_364bp_2U	AF
22A019_364bp_2U	AF
18A403_364bp_2U	CM
07C038_364bp_2U	CM
20C099_364bp_2U	DL
05A345_349bp_1.5U	EK

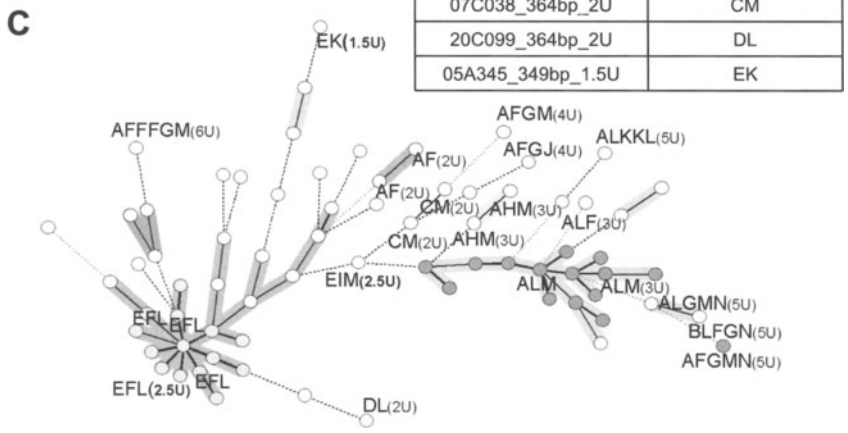


Fig. 4.4. Analysis of the internal variation inside the *P. aeruginosa* ms77 marker. **a** Sequence of the different motifs and letter code. **b** Coding of the ms77 allele in 22 different strains, showing the internal variability. **c** Phylogenetic relationship between strains using the minimum spanning tree representation and corresponding ms77 alleles (Onteniente et al. 2003; Onteniente 2004)

4.4.5 *Legionella pneumophila*

VNTR analysis in *L. pneumophila* led to the description of three informative markers that could be used for strain comparison in epidemic situations and the description of some additional markers that were amplified only in a subset of strains (Pourcel et al. 2003). Indeed, comparison of the genome sequence of the strains Philadelphia, Paris and Lens revealed that 13% of the DNA was strain-specific (Cazalet et al. 2004; Chien et al. 2004). By comparison of repeated sequences in the three sequenced genomes, new VNTRs were selected and primers chosen to match all the strains (http://minisatellites.u-psud.fr/comparison/; Pourcel, in preparation). Thus for species such as *L. pneumophila* which show a very high intraspecies variability, the availability of several sequenced genomes is necessary to set a MLVA assay. However, the number of available markers for a *L. pneumophila* MLVA assay remains limited. Sequencing of a large collection of alleles for three markers shows an important internal variability, resulting in homoplasy (Pourcel et al. 2003; Pourcel, unpublished data). Sequence data add to the resolution of the assay and open the way to an analysis of the mechanism of evolution of repeated sequences.

4.4.6 Other Bacteria

For a number of bacteria such as *Mycobacterium avium* subsp. *paratuberculosis*, *Pseudomonas aeruginosa*, *Salmonella enterica* subsp. *enterica* (including *typhi* and *typhimurium*), *Staphylococcus aureus*, etc (see Table 4.1) VNTR markers were identified and tested on sometimes relatively small collections of strains; and much still needs to be done before MLVA can become a standard procedure. However, in many instances and although only a limited number of markers were used, the authors believe that the resolution of the assay was comparable to that of other more complex assays. The major problem is the frequent use of microsatellites (2-bp to 8-bp repeat units) which tend to be unstable, as reported in several studies, with especially high homoplasy levels. In addition, they necessitate the use of sequencing gels or methods with equivalent resolution.

4.5 Validating and Analysing MLVA Data

A number of aspects specific for tandem repeat analysis must be kept in mind. Firstly, tandem repeat loci can be very variable in terms of muta-

tion rates, some loci having an extreme mutation rate while others are monomorphic. At present, this behaviour cannot be predicted from the sequence itself and will have to be experimentally measured by eventually typing hundreds of strains, as was done previously for human forensics-related projects. Regarding human forensics and paternity analyses for instance, hundreds of individuals of different ethnic origins have been genotyped in order to estimate both reliable allele frequencies and mutation rates for each locus employed in an assay. A number of different processes have been shown to drive mutation events in tandem repeats, including replication slippage and double-strand break repair (Debrauwère et al. 1999; Vergnaud and Denoeud 2000). Highly polymorphic markers which often result from a higher rate of mutation events will usually have a high homoplasy level. Such markers are sometimes called "highly informative", which is not necessarily correct. On the contrary, a MLVA assay based solely on such markers would probably be unable to cluster strains according to their true historical proximity, as illustrated previously with *Brucella*. Diversity indexes such as Simpson's index, promoted by Hunter and Gaston (1988), although very useful for comparing the discriminatory power of assays, do not measure the relevance of the discrimination which is achieved by a given marker, or combination of markers. Eventually, it will probably make sense to consider that two strains which differ at one highly variable marker are more similar than two strains differing at a moderately variable marker. Such more sophisticated distance coefficients cannot be developed until many strains have been typed, so that it should not come as a surprise to get a feeling that MLVA typing is at least in some instances not yet mature. MLVA will clearly take a major place among the epidemiological tools available to type a number of major bacterial pathogens, but this field of investigation is at present a very quickly evolving and competitive area of research and development.

Because many strains from different countries will have to be typed, it is essential that the MLVA data be carefully validated with appropriate controls, so that data sets from different laboratories can be merged. The main reason for this is that PCR products containing tandem repeats may occasionally show an electrophoretic behaviour which can give incorrect size measurements. Abnormalities are not random, but are locus-dependant, i. e. only a few loci may contain a repeat unit sequence bias or length which may modify the migration behaviour. The effect will be different if PCR product sizes are run as double-strand DNA (as for instance on an agarose gel) or single-strand DNA (as usually done on sequencing machines). The effect may also be proportional to the number of repeat units. For instance, marker Bams01 from *Bacillus anthracis* (Le Flèche et al. 2001) comprises a 21-bp very highly purine-rich repeat unit and runs significantly more slowly than expected from sequencing data. This suggests that the repeat

unit is slightly kinked and that this is amplified for larger alleles because a size of 21 bp represents exactly two DNA helix turns.

This difficulty is easily circumvented, by using a few reference DNAs with well characterised repeat copy numbers. When such sets are not easily sharable, or have not been defined, they can be organised locally once and for all by sequencing alleles from a few representative strains.

MLVA data can then be held and exchanged in simple text files. Each strain is described by a succession of values corresponding either to numbers of repeat units or to allele sizes expressed in base pairs. Although the former format is usually preferred, the latter can sometimes not be avoided for some rare tandem repeats in which combinations of repeat unit lengths coexist. This is observed for instance in *Legionella pneumophila* ms4 (Pourcel, unpublished data). The allele size will then depend on the set of primers used, so that the data should be carefully corrected if different primer sets are employed by different research groups.

Expressing the data in terms of repeat copy numbers is for this reason usually more appropriate. However, even in this case, rules have to be defined, because tandem repeat arrays often do not contain a perfect copy number. The true copy number can be for instance 2.5, 3.5, 4.5, etc., which for simplicity will be coded as 2, 3, 4, or alternatively 3, 4, 5. This is illustrated by ETRD (alias MIRU04), for instance, from *Mycobacterium tuberculosis*. Since different conventions were used initially, published data must be converted before data from different groups can be merged. In publications, the conventions used need to be clearly described. It is convenient to refer to a sequenced genome, especially when the corresponding strain is widely available, which is usually the case. Then the comprehensive description of a tandem repeat can be summarised for instance by Bams30_9bp_727bp_57U in which Bams30 is the locus name (Le Flèche et al. 2001), 9 bp is the repeat unit length, 727 bp is the PCR product size expected in the reference strain using the primers referred to in the given report and 57U is the corresponding repeat unit number.

The resulting data matrix can be imported into data-mining tools or into more conventional biology-oriented clustering methods. The currently preferred method to measure similarities between two strains is the simple counting of the number of markers at which the two strains differ (divided by the total number of markers and expressed as a percentage). This is a very crude similarity measure which gives the same weight to all markers. It also considers that alleles which differ by one repeat unit are not evolutionarily closer than alleles which differ by many repeat units. The two assumptions are often wrong but, in spite of this, the resulting clustering analyses make sense (Fig. 4.3, right). This is because the use of multiple markers compensates for variable homoplasy levels at individual markers.

Figure 4.4 shows the sequence variability of *Pseudomonas aeruginosa* marker Ms77 (Fig. 4.4a) and the encoding of the different alleles (Fig. 4.4b). The clustering analysis shown in Fig. 4.4c uses the minimum spanning tree method and clearly demonstrates that strains possessing alleles with the same number of repeats but with different codes are correctly clustered when several VNTRs are typed (Onteniente 2004). This also shows that additional information can be obtained by sequencing alleles in species with important variability. As larger MLVA data sets will be available, containing hundreds of genotypes, it is likely that different similarity measures will be developed to take more precisely into account, first, the evolutionary rate and homoplasy level (which can be indirectly deduced in part from the HGDI values) and, second, the mode of evolution of each individual marker.

Once MLVA data has been produced and collected, it is easy to set-up shared internet resources, for instance MLVA web services, as exemplified at http://bacterial-genotyping.igmors.u-psud.fr/ (Le Flèche et al. 2002) and http://www.mlva.umcutrecht.nl (Top et al. 2004).

4.6 MLVA Compared to Other Methods

MLVA does not provide a molecular clock. It is clear at least in some instances that the mutation of tandem repeats directly influences the phenotype of the corresponding strains, so that these mutations are not neutral and probably contribute to the adaptation of the species to its environment, in a reversible way. When MLVA is to be used for evolutionary studies, other sequence-based methods with a lower resolution will usually be employed in combination, as illustrated by Achtman et al. (2004) and Fabre et al. (2004). MLVA applies to sub-species typing. In many pathogens of interest, tandem repeat polymorphism analysis, including MLVA, will complement the existing tools. In some instances, it will even become the gold standard, which does not mean that it will replace existing methods, each one often providing a different and complementary point of view. The *M. tuberculosis*, *B. anthracis*, *Y. pestis* studies and a few others are clearly among these. One key feature of MLVA typing is that its low cost opens the possibility of an almost systematic typing, not limited to the few hundred of strains (at best) included in research projects. In addition to the importance of this aspect for clinical epidemiology, the possibility to quickly check the identity of a strain is also very important for the maintenance of strain collections, in particular when dangerous pathogens or precious strains are involved.

Acknowledgements.

Work on the typing of bacterial pathogens in our laboratory is supported by University Paris XI and by the French Délégation Générale pour l'Armement.

References

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci USA 101:17837–17842

Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P (2000) Diversity in a variable-number tandem repeat from *Yersinia pestis*. J Clin Microbiol 38:1516–1519

Allix C, Supply P, Fauville-Dufaux M (2004) Utility of fast mycobacterial interspersed repetitive unit-variable number tandem repeat genotyping in clinical mycobacteriological analysis. Clin Infect Dis 39:783–789

van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H (1997) Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. Infect Immun 65: 5017–5027

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Botterel F, Desterke C, Costa C, Bretagne S (2001) Analysis of microsatellite markers of *Candida albicans* used for rapid typing. J Clin Microbiol 39:4076–4081

Bricker BJ, Ewalt DR, Halling SM (2003) *Brucella* 'hoof-prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs). BMC Microbiol 3:15

Bull TJ, Sidi-Boumedine K, McMinn EJ, Stevenson K, Pickup R, Hermon-Taylor J (2003) Mycobacterial interspersed repetitive units (MIRU) differentiate *Mycobacterium avium* subspecies *paratuberculosis* from other species of the *Mycobacterium avium* complex. Mol Cell Probes 17:157–164

Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C (2004) Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet 36:1165–1173

Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SM, Asamani G, Hill K, Nuara J, Feder M, Rineer J, Greenberg JJ, Steshenko V, Park SH, Zhao B, Teplitskaya E, Edwards JR, Pampou S, Georghiou A, Chou IC, Iannuccilli W, Ulz ME, Kim DH, Geringer-Sameth A, Goldsberry C, Morozov P, Fischer SG, Segal G, Qu X, Rzhetsky A, Zhang P, Cayanis E, De Jong PJ, Ju J, Kalachikov S, Shuman HA, Russo JJ (2004) The genomic sequence of the accidental pathogen *Legionella pneumophila*. Science 305:1966–1968

Coletta-Filho HD, Takita MA, de Souza AA, Aguilar-Vildoso CI, Machado MA (2001) Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis. Appl Environ Microbiol 67:4091–4095

Debrauwère H, Buard J, Tessier J, Aubert D, Vergnaud G, Nicolas A (1999) Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. Nat Genet 23:367–371

Denoeud F, Vergnaud G (2004) Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. BMC Bioinform 5:4

Fabre M, Koeck JL, Le Fleche P, Simon F, Herve V, Vergnaud G, Pourcel C (2004) High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". J Clin Microbiol 42:3248–3255

Farlow J, Smith KL, Wong J, Abrams M, Lytle M, Keim P (2001) *Francisella tularensis* strain typing using multiple-locus, variable-number tandem repeat analysis. J Clin Microbiol 39:3186–3192

Frothingham R, Meeker-O'Connell WA (1998) Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 144:1189–1196

Groathouse NA, Rivoire B, Kim H, Lee H, Cho SN, Brennan PJ, Vissa VD (2004) Multiple polymorphic loci for molecular typing of strains of *Mycobacterium leprae*. J Clin Microbiol 42:1666–1672

Hardy KJ, Ussery DW, Oppenheim BA, Hawkey PM (2004) Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. Microbiology 150:4045–4052

Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26:2465–2466

Johansson A, Farlow J, Larsson P, Dukerich M, Chambers E, Bystrom M, Fox J, Chu M, Forsman M, Sjostedt A, Keim P (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J Bacteriol 186:5808–5818

Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J Bacteriol 182:2928–2936

Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, Keim P (2001) Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J Clin Microbiol 39:3179–3185

Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. BMC Microbiol 1:2

Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G (2002) High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC Microbiol 2:37

Lindstedt BA, Heir E, Gjernes E, Kapperud G (2003) DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar *typhimurium* with emphasis on phage type DT104 based on variable number of tandem repeat loci. J Clin Microbiol 41:1469–1479

Liu Y, Lee MA, Ooi EE, Mavis Y, Tan AL, Quek HH (2003) Molecular typing of *Salmonella enterica* serovar *typhi* isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats. J Clin Microbiol 41:4388–4394

Majed Z, Bellenger E, Postic D, Pourcel C, Baranton G, Picardeau M (2005) Characterization of *Leptospira interrogans* sensu stricto serovars by VNTR polymorphism analysis. J Clin Microbiol 43:539–545

Marmiesse M, Brodin P, Buchrieser C, Gutierrez C, Simoes N, Vincent V, Glaser P, Cole ST, Brosch R (2004) Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. Microbiology 150:483–496

Marshall DG, Coleman DC, Sullivan DJ, Xia H, O'Morain CA, Smyth CJ (1996) Genomic DNA fingerprinting of clinical isolates of *Helicobacter pylori* using short oligonucleotide probes containing repetitive sequences. J Appl Bacteriol 81:509–517

Moreno E, Cloeckaert A, Moriyon I (2002) *Brucella* evolution and taxonomy. Vet Microbiol 90:209–227

Noller AC, McEllistrem MC, Pacheco AG, Boxrud DJ, Harrison LH (2003) Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. J Clin Microbiol 41:5389–5397

O'Brien R, Danilowicz BS, Bailey L, Flynn O, Costello E, O'Grady D, Rogers M (2000) Characterization of the *Mycobacterium bovis* restriction fragment length polymorphism DNA probe pUCD and performance comparison with standard methods. J Clin Microbiol 38:3362–3369

Onteniente L (2004) Etude du polymorphisme associé aux répétitions en tandem pour le typage de bactéries pathogènes: *Pseudomonas aeruginosa* et *Staphylococcus aureus*. PhD thesis, University of Evry, Val d'Essonne

Onteniente L, Brisse S, Tassios PT, Vergnaud G (2003) Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. J Clin Microbiol 41:4991–4997

Overduin P, Schouls L, Roholl P, van den Zanden A, Mahmmod N, Herrewegh A, van Soolingen D (2004) Use of multilocus variable-number tandem-repeat analysis for typing *Mycobacterium avium* subsp. *paratuberculosis*. J Clin Microbiol 42:5022–5028

Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C (2003) Characterization of a tandem repeat polymorphism in *Legionella pneumophila* and its use for genotyping. J Clin Microbiol 41:1819–1826

Pourcel C, Andre-Mazeaud F, Neubauer H, Ramisse F, Vergnaud G (2004) Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol 4:22

Ramisse V, Houssu P, Hernandez E, Denoeud F, Hilaire V, Lisanti O, Ramisse F, Cavallo JD, Vergnaud G (2004) Variable number of tandem repeats in *Salmonella enterica* subsp. *enterica* for typing purposes. J Clin Microbiol 42:5722–5730

Ross BC, Raios K, Jackson K, Dwyer B (1992) Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J Clin Microbiol 30:942–946

Schouls LM, van der Heide HG, Vauterin L, Vauterin P, Mooi FR (2004) Multiple-locus variable-number tandem repeat analysis of Dutch *Bordetella pertussis* strains reveals rapid genetic changes with clonal expansion during the late 1990s. J Bacteriol 186:5496–5505

Skuce RA, McCorry TP, McCarroll JF, Roring SM, Scott AN, Brittain D, Hughes SL, Hewinson RG, Neill SD (2002) Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets. Microbiology 148:519–528

Spurgiesz RS, Quitugua TN, Smith KL, Schupp J, Palmer EG, Cox RA, Keim P (2003) Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS6110 isolates. J Clin Microbiol 41:4224–4230

Sun YJ, Lee AS, Ng ST, Ravindran S, Kremer K, Bellamy R, Wong SY, van Soolingen D, Supply P, Paton NI (2004) Characterization of ancestral *Mycobacterium tuberculosis* by multiple genetic markers and proposal of genotyping strategy. J Clin Microbiol 42:5058–5064

Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C (2000) Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol Microbiol 36:762–771

Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C (2001) Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol 39:3563–3571

Sylvestre P, Couture-Tosi E, Mock M (2003) Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. J Bacteriol 185:1555–1563

Titze-de-Almeida R, Willems RJ, Top J, Rodrigues IP, Ferreira RF II, Boelens H, Brandileone MC, Zanella RC, Felipe MS, van Belkum A (2004) Multilocus variable-number tandem-repeat polymorphism among Brazilian *Enterococcus faecalis* strains. J Clin Microbiol 42:4879–4881

Top J, Schouls LM, Bonten MJ, Willems RJ (2004) Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of *Enterococcus faecium* isolates. J Clin Microbiol 42:4503–4511

Truman R, Fontes AB, De Miranda AB, Suffys P, Gillis T (2004) Genotypic variation and stability of four variable-number tandem repeats and their suitability for discriminating strains of *Mycobacterium leprae*. J Clin Microbiol 42:2558–2565

Vergnaud G (1989) Polymers of random short oligonucleotides detect polymorphic loci in the human genome. Nucleic Acids Res 17:7623–7630

Vergnaud G, Denoeud F (2000) Minisatellites: mutability and genome architecture. Genome Res 10:899–907

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second-generation linkage map of the human genome. Nature 359:794–801

5 Bacterial Phylogeny Reconstruction from Molecular Sequences

Shigeaki Harayama, Hiroaki Kasai

5.1 Introduction

Systematics is a hierarchical system of nomenclature of living organisms linked to evolutionary theory and modern systematics aims at a classification based on phylogenetic relationships. In the 1960s, the application of protein sequence data to systematics became prevalent and scientists recognized that protein amino acid sequences contain useful information regarding phylogeny. However molecular systematics only recently became popular after the development of rapid DNA sequencing methods and the advent of DNA amplification using polymerase chain reaction (PCR). While small subunit ribosomal RNAs (SSU rRNAs) became the molecules of choice for molecular systematics studies, nucleotide sequences of protein-encoding genes and amino acid sequences deduced from the nucleotide sequences also proved to be valuable in phylogenetic research.

Molecular sequences have also been used in the exploration of the divergent evolution of early life; however, as has been discussed by Kurland et al. (2003), elucidating the evolutionary relationships between major groups of prokaryotes at or above the phylum level (i. e. establishing the branching order of deep branches in the phylogenetic tree of prokaryotes) is difficult. In our opinion, no solid method yet exists to reveal it. The main purpose of this chapter, therefore, is to describe DNA and protein sequence methodologies used to analyze the diversity within major taxonomic groups of prokaryotes at ranks lower than phylum or kingdom, but not to address them as they are used to clarify early events in their evolution.

In this review, methodologies for analyzing the diversity of major taxonomic groups within the domain Bacteria are described, although the same methodologies would also be applicable for analyzing archaeal strains. This review is not intended to make a complete inventory of studies on molec-

Shigeaki Harayama: Department of Biotechnology, National Institute of Technology and Evaluation, 2-5-8 Kazusa-Kamatari, Kisarazu-shi,Chiba 292–0818, Japan, E-mail: harayama-shigeaki@nite.go.jp

Hiroaki Kasai: Marine Biotechnology Institute, 3-75-1 Heita, Kamaishi, Iwate 026–0001, Japan, E-mail: hiroaki.kasai@mbio.jp

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

ular systematics, but rather to focus on approaches and on the power and limitations that these approaches have.

5.2 Species Definition

A principal aim of systematics is to detect and classify diverse living organisms. Traditionally, the species is the fundamental unit of diversity. In highly sexual organisms, frequent genetic exchange hinders genetic divergence among its members; and thus, a species can be defined as a reproductive community whose members have the potential to interbreed and produce fertile offspring. Although genetic exchange systems exist in bacteria and are involved in the rapid propagation of adaptive alleles such as drug resistance genes, the bacterial world is generally considered to be asexual. For this reason, systematists have not yet reached a consensus for the defininition of a bacterial species (Cohan 2002). For some, the whole concept of bacterial species as natural distinct entities is becoming questionable.

Because of this unsettled state, a variety of definitions of bacterial species have been proposed. Wayne et al. (1987) defined a species as an entity that includes strains sharing approximately 70% or greater DNA–DNA relatedness and with a difference of less than 5 °C in the DNA melting temperature between homologous and heterologous DNA hybrids. This definition for bacterial species is generally accepted among taxonomists and DNA hybridization is acknowledged as the reference method for establishing relationships within and between species. Nevertheless, the DNA–DNA hybridization method did not achieve widespread adherents, probably because of theoretical as well as practical problems. First, there is no firm theoretical basis for setting the value of 70% relatedness as a boundary for species designation, although Johnson (1973) found that strains from the "same species" – as classified by phenotypic traits – nearly always shared 70% or more genomic DNA homology, while strains from "different" species nearly always shared less than 70% homology. Second, the DNA–DNA hybridization value is not invariable, as different methods provide different values (Springer and Krajewski 1989). Third, to carry out reciprocal DNA–DNA hybridization, it is necessary to collect all relevant strains and isolate their DNA (Vauterin et al. 1995). Fourth, DNA hybridization experiments are cumbersome because many physicochemical parameters must be carefully controlled (Grimont et al. 1980). Finally, the determination of the degree of DNA hybridization does not provide any information concerning the phylogenetic relationships. The value of 70% is not a well defined standard, but merely an indicative value. For ex-

ample, DNA–DNA relatedness could be expressed in three categories: high DNA relatedness (indicating the same species), low but significant DNA relatedness (indicating the same genus), and non-significant DNA relatedness (indicating different genera; Vandamme et al. 1996). The relationship of the extent of DNA–DNA hybridization to the 16S rRNA homology was examined by Stackebrandt and Goebel (1994). They found that strains exhibiting less than 97% homology in 16S rRNA gene sequences were nearly always members of different species as determined by DNA–DNA hybridization. Currently, the combination of the 16S rRNA gene analysis and DNA–DNA hybridization analysis is most frequently used to discern closely related strains. For descriptions of new species and genera, an integration of phylogenetic relationships with phenotypic marker analysis, which is referred to as polyphasic taxonomy, is highly recommended (Vandamme et al. 1996).

In comparison, Rossello-Mora and Amann (2001) defined a species to be a "monophyletic and genomically coherent cluster of individual organisms showing a high degree of overall similarity in many characteristics." In practice, many bacteriologists recognize that most newly isolated bacteria are classified into discrete phenotypic and genetic clusters, which are separated by large phenotypic and genetic gaps. Therefore, this definition is natural and agreeable for them; however, the existence of bacterial clusters clearly separated by neighboring clusters has not yet been rigorously proven; and several clusters will fuse into single cluster as data for more strains become available.

The concept of "periodic selection" of "ecotypes" is central to a third definition proposed by Cohan (2004). An ecotype is a subgroup of a genomically coherent bacterial group that differs genetically from other subgroups by adaptation to local ecological conditions. In asexual organisms, a derivative of an ecotype which has acquired favorable mutations may out-compete original members of the same ecotype because they occupy the same ecological niche and compete for limited resources. The successful mutant will be brought to fixation and purge other members from the niche; and thus, the genetic diversity within the population of each ecotype is periodically reset to zero. Based on this conclusion, it has been proposed that bacterial species can be defined as equivalent to the ecotype (Cohan 2004).

As shown below, the genomic definition of species by Wayne et al. (1987) (approximately 70% similarity by DNA–DNA hybridization) is the most frequently applied in the literature.

5.3 Bacterial Diversity

Bacterial diversity revealed by molecular phylogenetic analyses may be a reflection of bacterial metabolic diversity. Bacterial growth and survival are primarily determined by the availability of inorganic and organic compounds which are used as sources for energy and the ability to adapt to physicochemical conditions such as temperature, pH or pressure. Bacteria can exploit nearly every redox-coupled reaction and fill all available metabolic niches. Certainly, contemporary bacteria have adapted to nutritional and physical requirements by evolving required functions. This adaptation will mainly occur by: (1) mutations of existing genes (changes in regulatory circuits, substrate specificities of regulatory proteins or enzymes, stability or turnover rates of enzymes, etc.), often associated with the duplication of relevant genes, and (2) gene recruitment by horizontal gene transfer (HGT).

The enormous diversity in terms of morphology, physiology, and genome sequences in bacteria leads to a fundamental question: What is the extent of bacterial diversity? This question is not only of scientific interest, but is also relevant to the industrial application of biological resources. Difficulty in answering this question arises for two reasons. First is the problem of defining what the diversity of bacteria is. In this chapter, the term "diversity" is used to indicate species richness, or the number of operational taxonomic units (OTUs). Second, a standardized methodology for measuring bacterial diversity is still not established. Below we discuss some recent studies that exemplify existing approaches to diversity estimates.

The first reliable estimate of the diversity of bacteria in soil was published by Torsvik et al. (1990a, 1990b). In their studies, DNA isolated from the "bacterial fraction" of soil was heated for separation into single-stranded DNA, and the reassociation kinetics of the DNA were used to estimate the diversity of the DNA molecules. The results indicated that at least 4,000 bacterial genomes were found in DNA isolated from 30 g of soil. From the reassociation kinetics, it was also suggested that more than 99% of hybridized DNA molecules obtained in this experiment were heteroduplexes consisting of two DNA strands from two different species; The melting temperature (T_m) of the hybridized DNA molecules was lower than that of homoduplexed DNA by 5 °C or more. The diversity described above, therefore, may be underestimated by 100-fold, and the real number could be approximately 4×10^5 genomes (Dykhuizen 1998). The reassociation kinetics approach also revealed that the bacterial diversity in soils and sediments was much higher than that in water columns (Torsvik et al. 2002).

One of the observations of ecological science is that the abundance of species in animal, plant, and insect communities can be described using

a lognormal distribution. Assuming that bacterial species abundance also follows a lognormal distribution, Curtis et al. (2002) calculated prokaryotic diversity in different environments. The total number of species was estimated from two parameters, N_{max}, which is the number of individuals in the most abundant species, and N_T, which is the total number of individuals in the community. N_T can be estimated as the total microscopic count, while N_{max} can be determined, for example, by quantitative fluorescent *in situ* hybridization. Bacterial diversities were thus estimated to be 160 species/ml of seawater, 6,400 – 38,000 species/ g of soil, and 70 species/ml of sewage.

Hagström et al. (2002) observed that the rate of discovery of new 16S rRNA sequences of marine plankton is dropping in the public databases, suggesting that the inventory appeared to be nearly complete at about 1,117 unique ribotypes (species), which were grouped by using the cutoff at 97% identity. More recently, Schloss and Handelsman (2004) analyzed 16S rRNA gene sequences deposited in databases. In this analysis, too, an operational taxonomic unit (species) was defined as a group of sequences that are more than 97% identical to each other. A rarefaction curve, which plots the total number of samples versus the total number of species (Gotelli and Colwell 2001), was made for each bacterial phylum or for all bacteria to assess the current state of sampling. The result was a curve that increased steeply at first, then gradually leveled off. This method estimated the species richness to be in the order of 10^6, which is much smaller than another estimate of 10^9 (see below).

A recent shotgun survey of environmental DNA sampled from the Sargasso Sea (Venter et al. 2004) found, among 1,045 Gbp of nonredundant sequences, about 1,400 16S rRNA sequences in which 148 sequences were judged to be derived from new species (defined by the 97% cutoff). The rate of discovery of new species at 0.1 (10%) is the range expected from the rarefaction curve made by Schloss and Handelsman (2004; see also Chap. 9, "Metagenome Analyses").

Because 16S rRNA gene sequences from dominant populations are overrepresented in the public databases, the rarefaction curve represents only the rate of acquisition of new sequences from abundant species. It is likely that the current sampling strategies do not allow detection of minor populations of communities, which require intensive sequencing of many clones (Curtis and Sloan 2004).

Dunbar et al. (2002) constructed 200-member 16S rRNA gene clone libraries of four bacterial soil communities from two locations in Arizona, in which the 16 rRNA sequences were classified into nearly 500 species groups. Assuming the lognormal distribution of species abundance, they calculated that between 4,000 and 8,000 species inhabited the four Arizona samples. Under the assumption of 4,000 species as community members, they calculated that the isolation and sequencing of 25,000 independent 16S

rRNA gene clones is required to detect half of the members (2,000 different species).

Lunn et al. (2004) developed a nonparametric method to estimate bacterial biodiversity from clone libraries without making any assumption concerning species distribution (such as lognormal distribution). They used a data set of 100 unique clones from a sample of Amazonian soil and determined that the species richness in the soil sample was probably higher than 10^5.

Recently, Acinas et al. (2004a) analyzed microbial diversity in seawater, using PCR to clone and sequence 16S rRNA genes with high coverage. In their experiments, they took care to reduce PCR artifacts (nucleotide misincorporation errors or formation of chimeras and heteroduplex molecules; see below), and sequenced 1,000 rRNA genes from a single community. When clusters sharing at least 99% sequence identity were defined as OTUs, the estimated diversity was 520 OTUs; and, by reducing the clustering threshold to 97% identity, diversity was reduced to 450 OTUs. This number was higher (but only 3-fold) than that estimated by Curtis et al. (2002): 160 species/ml of seawater. More examples of bacterial diversity estimates in different environments are given by Hughes et al. (2002a, b).

Thus, revelations of the abundance of species in terms of DNA homology within limited numbers of ecosystems has commenced. It is likely that different species inhabit different ecosystems, while similar species are recovered from similar environments. Therefore, for an estimation of the full complement of bacterial biodiversity, it is necessary to estimate the diversity of natural bacterial habitats. Unfortunately, very little is known about spatial and temporal variability of bacterial community structures (number and taxonomic positions of species and their population sizes; Kirk et al. 2004), but 2,000 different bacterial communities with a species richness of 4×10^5 (Torsvik's result reestimated by Dykhuizen 1998) may be a modest estimate; and even with this conservative estimate, the number of bacterial species on the earth is estimated to be approximately 10^9.

5.4 Phylogenetic Analysis Based on 16S rDNA Sequences

In the past decade, spectacular developments in taxonomy were accomplished mainly by the introduction of new techniques, including nucleotide and protein sequencing. These techniques revolutionized insights into phylogeny by reducing confusion and increasing taxonomic precision. rRNA sequence data especially have proved to be useful in establishing the division of all living organisms into three primary domains, the Archaea, the Bacteria, and the Eucarya (Woese et al. 1990). Nowadays, the taxo-

nomic classification of living organisms, in particular bacteria and archaea, has mainly been achieved by sequence comparisons among rRNA gene sequences. This tendency has arisen through intensive investigation of rRNA molecules during the past three decades. The sequencing of 5S rRNA molecules gradually resulted in an accumulation of data for numerous bacteria; and the comparison of the 5S rRNA sequences has been used to establish bacterial lineages (Hori and Osawa 1986; Specht et al. 1997). Also, a limited number of SSU rRNA gene sequences became available by sequencing after cloning. The use of reverse transcriptase with universal primers has allowed a rapid increase in identified SSU rRNA sequences (Lane et al. 1985). More recently, the advent of PCR technology has allowed the direct sequencing of genes for SSU rRNA without cloning (Edwards et al. 1989; Medlin et al. 1988). Because these sequences provided a phylogenetic framework for bacterial molecular taxonomy, 16S rRNA (bacterial SSU rRNA) sequences became the favored method of bacterial classification for many scientists.

There are many reasons why rRNA molecules have been selected as standard molecules for molecular taxonomy. They are constituents of all organisms. They exist in abundance and therefore can readily be isolated and sequenced by reverse transcriptase. For sequence comparison, many conserved regions of rRNA molecules allow alignment between distantly related organisms, while variable regions are useful for the distinction of closely related organisms (Gutell et al. 1994; Van de Peer et al. 1996). Furthermore, there is little evidence for horizontal transfer of rRNA gene (Kurland et al. 2003; Sneath 1993; van Berkum et al. 2003), although many other genes are expected to have been transferred from one species to other distantly related species. At present, rRNA sequences are accumulating rapidly (> 105,000 in February 2005) and they are accessible via an international database (ribosomal database project II; RDP-II, http://rdp.cme.msu.edu/index.jsp; Cole et al. 2003). Public databases even contain 16S rRNA sequences of uncultured bacteria (Amann et al. 1995).

Figure 5.1 shows the steps required for determining the phylogenetic position of a bacterium in the 16S rRNA tree using the neighbor-joining method. Listed below are some useful hints for conducting a proper 16S rRNA-based phylogenetic analysis, followed by a discussion of some of the problems with such analyses.

PCR amplification. Based on conserved sequences in 16S rRNA, a set of "universal" primers was designed and used to PCR-amplify the rRNA genes *in vitro*. The direct sequencing of the amplified DNA could provide almost complete rRNA gene sequences, although many designed primers were not complementary to the conserved regions of all published sequences. Sets of primers containing deoxyinosine residues were thus designed for the

Work sequence for the 16S-rRNA-based phylogenetic analysis using the neighbor-joining method

1. DNA isolation from a bacterium of interest
2. PCR amplification of a partial rRNA gene sequence
3. Alignment of the obtained rRNA sequence with other rRNA sequences in databases
4. Estimation of distances between each pair of sequences using one of the evolution models
5. Reconstruction of a phylogenetic tree from these distances following a particular algorithm (neighbor-joining)
6. Bootstrap analysis

Fig. 5.1. Work sequence for the 16S rRNA-based phylogenetic analysis using the neighbor-joining method (Saitou and Nei 1987)

amplification of a broader selection of 16S r vanRNA genes (Watanabe et al. 2001a). Even with these primers, however, amplification of all bacterial 16S rRNA genes is not guaranteed (Baker et al. 2003).

One problem with the amplification process is that Taq DNA polymerase lacks exonuclease-dependent proofreading activity and therefore the error rate in DNA replication is relatively high (10^{-5} per basepair per extension). If 1,500-bp rRNA gene fragments are amplified by 30 cycles of PCR, the probability of errors in the PCR products is significant. Using enzymes with proofreading activities and a smaller number of amplification cycles may reduce PCR artifacts, but PCR product yield may also be reduced. One way to check PCR-provoked sequencing errors is to examine the secondary structure conservation (Field et al. 1997).

16S rRNA gene sequence analysis is also a powerful tool for assessing genetic diversity in environmental samples. PCR amplification followed by cloning of 16S rRNA genes from environmental DNA has detected new lineages of uncultured microorganisms (DeLong and Pace 2001). PCR amplification of 16S rRNA genes from mixed DNA samples, however, may form chimeric structures at an appreciable frequency. A chimera is a sequence composed of two or more distinct parental sequences and seems to be formed by copying different parental sequences during template switches. Chimeras thus are composed of two or more phylogenetically distinct parent sequences and falsely mirror phylogenetic novelty. A large number of chimeric 16S rDNA sequences are found in the public databases (Hugenholtz and Huber 2003); and therefore, care should be taken to discard chimeric sequences from phylogenetic analyses.

In mixed-template PCR, heteroduplex formation is another problem. In the annealing step of PCR, annealing of two heterologous single-stranded DNAs may occur (heteroduplex formation). When the heteroduplex molecules are cloned in *Escherichia coli*, mismatch repair systems of the host can convert a heteroduplex into a single non-natural hybrid sequence.

A method to avoid the cloning of heteroduplex molecules has been proposed (Thompson et al. 2002).

Alignment. The alignment of rRNA gene sequences is very important for inferring phylogenetic relationships correctly. The presence of insertions and deletions (indel sequences) may make the alignment less accurate, especially when the homology is low. The use of the secondary structure information thus becomes essential to localize the indel sequences. The 16S and 23S rRNA secondary structure models were constructed by searching coordinated base substitutions (covariation) among a set of aligned sequences. When covariation is found, the covariable pair is considered to interact by forming a helix structure. The current 16S and 23S rRNA secondary structure models are in agreement with recently determined high-resolution crystal structures of the 30S and 50S ribosomal subunits (Ban et al. 2000; Schluenzen et al. 2000; Wimberly et al. 2000): nearly all of the predicted helices were present in the crystal structures (Gutell et al. 2002). Several software packages have been developed to optimize the alignments, taking into account both primary and secondary structures (Notredame et al. 1997).

Not all aligned positions of rRNA sequences are equally informative for phylogenetic inference, as the rates of substitution differ at individual positions (Wuyts et al. 2001). Invariant and conserved residues are useful for the accurate alignment of rRNA sequences, moderately variable residues are used to establish phylogenetic relationships of distantly related bacteria, and more variable regions are valuable for the discrimination of closely related strains. The inclusion of residues of hypervariable regions for phylogeny construction is not recommended – especially in analyses of distantly related strains – because it increases noise for the following two reasons: (1) the alignment of residues in hypervariable regions is often difficult because of a very low degree of sequence homology and (2) hypervariable regions of 16S rRNA are mainly located in helices and therefore contain multiple base changes, including compensatory mutations to keep the helical structure. Because the probability of such compensatory mutations may be very high under strong selection pressure, these mutations should not be considered equivalent to mutations in other regions.

In fact, when Yamamoto and Harayama (1998) conducted a phylogenetic analysis of 20 *Pseudomonas* strains using the nucleotide sequences of the genes for 16S RNA, the DNA gyrase B subunit (*gyrB*), and RNA polymerase σ70 factor (*rpoD*), the phylogenetic tree reconstructed from the 16S rRNA sequences, excluding sequences in variable regions, was congruent with the *gyrB*- and *rpoD*-based trees. However, in the 16S rRNA-based tree, including sequences in variable regions, *P. putida* biovar A and B strains were not separated into two independent clusters.

The root of a phylogenetic tree is usually determined by using an outgroup. The outgroup should be similar to – but also less related to – any other sequences. For the reconstruction of a phylogenetic tree, different outgroup sequences should be tested to avoid false results. The addition of new related sequences often changes tree topology. Addition of new data generally improves the tree structure, but the addition of incomplete or incorrect sequences adversely affects phylogenetic reconstruction. Branches represented by a single sequence can often be incorrectly positioned in phylogenetic trees (Ludwig and Schleifer 1994).

Reconstruction of a phylogenetic tree. Probabilistic methods of phylogenetic analysis, such as maximum likelihood (Felsenstein 1981) and neighbor-joining (Saitou and Nei 1987), are based on an evolutionary model that defines probabilities for the transition from one base (or amino acid, in the case of protein sequences) to another. Traditionally, the Jukes–Cantor model (Jukes and Cantor 1969) or Kimura's 2-parameter model (Kimura 1980) has been used for nucleotide substitutions. Recently, substitution rates in rRNA have been estimated by counting the relative substitution probabilities in rRNA databases (Smith et al. 2003); and a substitution matrix for rRNA was constructed and incorporated into the Phylip software package (http://evolution.genetics.washington.edu/phylip.html).

As has been discussed so far, 16S rRNA analyses are generally believed to be the best way to obtain significant information on the taxonomic position of bacteria, especially for new or atypical isolates. However, the resolution of 16S rRNA sequence analyses seem too low to distinguish closely related bacteria. Comparative analysis of DNA–DNA similarities and 16S rRNA gene sequence homology indicates that organisms sharing more than 97% 16S rRNA identity may belong to different species, even at a level of 99.5% identity (Stackebrandt and Goebel 1994). Certainly, because of an inherently slow speed of divergent evolution of 16S rRNA, the resolution of 16S rRNA sequence analysis between closely related organisms is generally lower than that of the DNA hybridization analysis.

For example, the 16S rRNA sequences of members of genus *Aeromonas* were very similar to each other, with a range of identity from 98% to 100%. From these sequences, diagnostic signature sequences were discerned that could differentiate most *Aeromonas* species. However, the phylogenetic interrelationships deduced from the 16S rRNA sequences were markedly different from the results of chromosomal DNA–DNA hybridization (Martinez-Murcia et al. 1992).

In contrast, the variation of 16S rRNA sequences in different strains within the same species can be unexpectedly high (Clayton et al. 1995). Sources of variation may be either sequencing errors or strain misidentification. But intraspecies variation in so-called "hypervariable regions"

of 16S rRNA may also be high; and even in a single strain, multiple 16S rRNA genes may have sequences that are not identical (Acinas et al. 2004b). Comparisons of paralogous 16S rRNA sequences of related strains may give an overestimation of intraspecies variation of 16S rRNA (Cilia et al. 1996). Thus, although 16S rRNA sequences are highly useful for taxonomy, low sequence variability in 16S rRNA genes may limit their usefulness in distinguishing related strains, while high sequence variability in their hypervariable regions may limit their use in grouping related strains.

The rate of nucleotide substitution in 16S rRNA sequences is estimated to be approximately 0.05 per site per 250 million years (Myr; Ochman et al. 1999), and thus we can roughly estimate divergence time from 16S rRNA sequence divergence. For example, the distance value of 0.03 that is thought to differentiate at the species level corresponds to a divergence time of 150 Myr.

5.5 Phylogenetic Analysis Based on Protein Sequences

5.5.1 Selection of Target Proteins

Because the paucity of the divergence of 16S rRNA sequences between two closely related bacteria obstructs the reconstruction of their phylogenetic trees, phylogenetic analyses using protein-encoding gene sequences have recently been performed by many research groups. The use of protein-encoding genes has two main advantages over the use of rRNA genes. (1) Protein-encoding genes are known to evolve much faster than rRNA genes, especially at the third positions of codons, where nucleotide substitutions result in mostly silent (synonymous) mutations; and therefore, these genes seem to be more appropriate for phylogenetic analysis of closely related bacteria. (2) The alignment of protein-encoding genes can be done using translated sequences comprising 20 amino acid species; and therefore, the alignment of protein sequences is easier and more accurate than that of rRNA genes.

Potentially, many protein-encoding genes can be used for phylogenetic analysis if they fulfill the following conditions: (1) they are not subject to HGT, (2) they are present in all bacteria, (3) preferentially there is a single copy on each genome, and (4) at least two regions are highly conserved to allow the design of appropriate PCR primers (Yamamoto and Harayama 1996).

Jain et al. (1999) reported that extensive HGT has occurred between bacteria, especially in genes for metabolic functions (such as biosynthesis

of phospholipids, etc.), but rarely in genes participating in transcription and translation (informational genes). They proposed that a major factor limiting HGT in informational genes is that their products are members of complex systems interacting with other proteins and therefore are difficult to integrate into new hosts (the complexity hypothesis). If this is the case, protein-encoding genes to be used as phylogenetic markers should be selected with caution. Accordingly, Brown et al. (2001) examined 23 genes from 45 species and concluded that only 14 genes have very unlikely undergone HGT. The tree reconstructed from the combined sequences of these 14 proteins was highly congruent with the 16S rRNA tree.

More recent studies, however, indicated that selecting orthologues with certain characteristics may be a key, and that HGTs are rare among single-copy orthologous genes (in other words, HGT occurs in genes other than single-copy orthologous genes). This conclusion was drawn from the observations that the topologies of phylogenetic trees constructed by different orthlogous genes were congruent with each other (Daubin et al. 2002; Lerat et al. 2003). It seems to be important to include only orthologous genes having a single significant match per genome for the analysis, rather than using circular or reciprocal "best BLAST hit" relationships (Altschul et al. 1997) for the selection of orthologues. The latter procedure may include hidden paralogues instead of real orthologues (Daubin et al. 2003). In fact, when orthologous sets of bacterial genes (COGs) consisting of orthologous and possibly paralogous proteins were used to construct phylogenetic trees, 30% of them showed substantial anomalies in tree topology. However, in all the trees from the 108 COGs that were single-copy orthologues, such anomalies were not observed. Interestingly, genes for certain ribosomal proteins and tRNA synthetases are not appropriate for use in phylogenetic analyses (Novichkov et al. 2004). Note, however, that Zhaxybayeva et al. (2004) recently argued that phylogentic reconstruction is not influenced by different orthologue selection procedures, but that the selection of genomes may influence the results because the extent of mosaicism may differ among genomes.

In summary, bacterial genome projects have provided abundant information concerning genetic diversity in bacteria. Comparative genomics uncovered many genome variations in closely related bacteria and brought into question the validity of the tree-like history of the whole genome by demonstrating the high frequency of HGTs. Nevertheless, it is still likely that conserved core genes not involved in HGT are common to all bacterial genomes and that these core genes can be used as convenient phylogenetic markers individually or as concatenated sequences.

Thus, carefully selected protein-encoding genes can be used for the classification of bacteria at the species level as an alternative approach to DNA–DNA hybridization. Recently, an ad hoc committee for the reevalu-

ation of the species definition in bacteria proposed that a small set (e.g. five) of protein-encoding genes can be used for quantitative evaluation of taxonomic relatedness, and issued a call for the identification of such genes (Stackebrandt et al. 2002).

Several research groups have already used multiple molecular markers to delineate bacterial taxonomic relationships. Maiden et al. (1998), using a strategy called "multilocus sequence typing", demonstrated that a small set of protein-encoding genes could reliably establish phylogenetic relationships of bacterial species. Primer sets for the amplification of 11 housekeeping genes from *Neisseria meningitidis* were designed and amplified genes were sequenced and analyzed. The dendrograms constructed from the pairwise differences in multilocus allelic profiles were consistent with clonal groupings previously determined by multilocus enzyme electrophoresis. A subset of six genes was sufficient to retain the resolution achieved using all 11 loci.

Zeigler (2003) examined the nucleotide sequences of 32 proteins in 44 strains belonging to 16 different genera and compared to whole-genome sequence identities of these strains. He demonstrated that whole genome sequence identity correlated well with genome similarity measurements obtained by DNA–DNA hybridization. He also showed that even single genes could predict overall phylogenetic relatedness with high precision, although the use of multiple genes for analysis did increase the resolution.

Below, we describe several protein-encoding genes that may be useful for phylogenetic analyses of bacteria. Obviously, this is not a complete list, but a list of our personal preferences.

RecA. RecA is a multifunctional protein involved in homologous recombination, DNA repair, and the SOS response. It binds single-stranded DNA and unwinds duplex DNA. Moreover, RecA bound to single-stranded DNA acts as an allosteric effector that induces the proteolytic (self-cleavage) activities of the LexA and UmuD proteins of *E. coli* and the *c*I protein of lambda phage. It is ubiquitous in bacteria (Dew-Jager et al. 1995). Lloyd and Sharp (1993) compared 25 bacterial RecAs and concluded that the topology of the RecA tree is very similar to that of the 16S rRNA tree. More recently, 62 bacterial RecA protein sequences were compared by determining pairwise similarity scores (Karlin et al. 1995); and although the RecA tree was not constructed, the grouping of the RecA sequences generally agreed with the pattern obtained from the 16S rRNA tree. Phylogenetic analyses of *Vibrio* strains using *recA* gene sequences was also recently conducted by Thompson et al. (2004). The RecA protein family website is found at http://www.tigr.org/~jeisen/RecA/RecA.html.

Chaperonins. Chaperonins are a class of proteins that assist protein folding *in vivo*. One class of chaperonins is composed of two subunits of 10 kDa

and 60 kDa called either CPN10 and CPN60, HSP10 and HSP60, or GroES and GroEL. The homologues of CPN10 and CPN60 are found in almost all bacteria and some archaea, and in eukaryotic cell organelles such as mitochondria and chloroplasts (Hill et al. 2004). The genes for CPN60s are useful for phylogenetic studies as well as for specific detection and identification of particular organisms (Jian et al. 2001; Kwok and Chow 2003; Mikkonen et al. 2004; Viale et al. 1994). Universal degenerate PCR primers for the amplification of approximately 550-bp CPN60 genes have been developed and can be used for diverse bacterial strains (Goh et al. 1996). A curated database of the gene sequences of CNP60 genes is available at http://cpndb.cbr.nrc.ca.

CPN70 (HSP70, DnaK) is another chaperonin with a total mass of 70 kDa. It is ubiquitous among bacteria and may be the most conserved bacterial protein. Both amino acid and nucleotide sequences of CPN70s have been widely used in phylogenetic studies (Stepkowski et al. 2003). However some CNP70 genes have undergone HGT; for example, the archaeal homologue of CNP70 may have derived from bacterial donors. HGT between two bacteria is also suggested from analysis of CNP70 sequences (Gribaldo et al. 1999).

RNA polymerase subunits. DNA-directed RNA polymerase catalyzes the synthesis of RNA by copying a DNA template. The core enzyme of RNA polymerase consists of four different subunits, alpha (α), beta (β), beta′ (β'), and omega (ω) in the configuration $\alpha 2\beta\beta'\omega$. The structural gene for the β-subunit, *rpoB*, has been successfully used for phylogenetic analyses of several bacteria. In common with other protein-encoding genes, *rpoB* differentiates between closely related strains better than 16S rDNA sequences (Mollet et al. 1997).

For the taxonomic classification of rapidly growing mycobacteria (RGM), the complete sequences of the 16S rRNA gene (gene sizes were between 1,483 bp and 1,489 bp), *rpoB* (3,486–3,495 bp), *recA* (1,041–1,056 bp), partial sequences of the *hsp65* (HSP60 analogue in mycobacteria, amplified length was 420 bp) and *sodA* (the structural gene for superoxide dismutase, 441 bp) were determined in 19 species of RGM. Phylogenetic trees based on each gene sequence and those based on combined datasets were constructed and compared. Bootstrap values were highest at the nodes in the *rpoB*-based tree followed by those in the *recA*- and 16S rRNA gene-based trees, while some nodes in the *hsp65*- and *sodA*-based trees were poorly supported by the bootstrap sampling. Because of this difference, the authors suggested a superiority of *rpoB* and *recA* over *hsp65* and *sodA* in phylogenetic analysis (Adekambi and Drancourt 2004). A higher statistical significance observed with the *rpoB* sequence, however, may merely be the result of a larger sample size (longer sequence length).

The β′-subunit of RNA polymerase is encoded by *rpoC*. The gene has not frequently been used to characterize the taxonomic classification of bacteria. It has been suggested, based on an rRNA-based study, that *Oenococcus oeni* is fast-evolving, i.e. the length of the branch of this strain in the rRNA-based tree was longer than other branches. The long branch of *O. oeni* found in the rRNA-based tree, however, was not reproduced in the *RpoC*-based tree, although the branching order of the *RpoC*-based tree was similar to that of 16S rRNA-based tree. Thus, the hypothesis that *O. oeni* is a fast-evolving bacterium was not supported by the analysis using *rpoC* (Morse et al. 1996).

Elongation factor G. Elongation factor G (EF-G) catalyzes the translocation of tRNAs and mRNA on the ribosome. The sequence of *fus*, the structural gene for EF-G, has not been used frequently in phylogenetic analysis. The taxonomic positions of *Aquifex pyrophilus* and *Thermotoga maritima*, both hyperthermophilic bacteria, were investigated using the sequences of *fus*, *rpoB*, *rpoC*, etc. The tree showed that *A. otriphilus* and *T. maritima* are on the two deepest branches of the bacterial tree (Bocchetta et al. 2000). Recently, we showed that Fus is a useful phylogenetic marker for the genus *Enterococcus* (Sato and Harayama, manuscript in preparation).

GyrB. DNA gyrase is a type II topoisomerase and composed of two subunits which are encoded by *gyrA* and *gyrB* (see the next section for more detailed information concerning gyrase). Using *gyrB* sequences, the phylogenetic relationships of 46 *Acinetobacter* strains, which have previously been classified into 18 genomic species by DNA–DNA hybridization studies were investigated. The phylogenetic grouping of *Acinetobacter* strains based on *gyrB* genes was almost congruent with that based on DNA–DNA hybridization studies, indicating that *gyrB* sequence comparison can be used to resolve the taxonomic positions of bacterial strains at the level of genomic species (Yamamoto et al. 1999).

The *Bacillus cereus* group is a clade including *B. anthracis* (the causative agent of anthrax), *B. cereus* (a food-borne pathogen), and *B. thuringiensis* (the producer of BT toxin). Laboratory and environmental strains in this clade were differentiated better by *gyrB* than by 16S rRNA genes. The classification of these strains by DNA–DNA hybridization resulted in a grouping which was almost identical to that obtained using *gyrB* (La Duc et al. 2004). The authors concluded that *gyrB*-based phylogenetic analysis is as powerful as DNA–DNA hybridization.

The phylogenetic relationships of all known species of the genus *Aeromonas* were investigated using *gyrB* sequences. The *gyrB*-based grouping was consistent with an established taxonomic classification of all *Aeromonas* species, mainly determined by DNA–DNA hybridization (Yanez et al. 2003).

In addition, slowly growing *Mycobacterium* species can be discriminated from each other by using *gyrB* sequences (Kasai et al. 2000). Based on these results, a microarray system based on *gyrB* sequences was developed for rapid identification of *Mycobacterium* species (Fukushima et al. 2003).

gyrB is also useful as a probe to monitor environmental microorganisms. In activated sludge fed with phenol, the formation of flocs (bacterial aggregates) is important for its settleability. When nonflocculating bacteria outgrow the sludge, the activated sludge flows out, and the process breaks down. One of the major populations in activated sludge is *Aquaspirillum*. The *Aquaspirillum* population has been found both in stable (flocculating) as well as unstable (nonflocculating) activated sludges. The *gyrB* analysis of the *Aquaspirillum* population – but not the 16S rRNA analysis – separated the population into two subpopulations. One subpopulation could form flocs while the other could not. A competitive PCR analysis in which specific *gyrB* sequences were used as the primers was able to monitor a population shift from flocculating *Aquaspirillum* to nonflocculating *Aquaspirillum* during the shift of activated sludge from settleable to nonsettleable (Watanabe et al. 1999).

The *gyrB* sequences thus far characterized are stored in a database called the Identification and classification of bacteria (ICB) database (Kasai et al. 1998; Watanabe et al. 2001b; http://www.mbio.jp/icb/).

Other proteins. In this section, we have mainly discussed "multitalented proteins" that have versatile utilities in bacterial taxonomy and its application to biotechnology. The proteins described above may be useful for many purposes: classification, phylogenetic analysis, taxonomic identification, and the diagnostic detection of bacterial strains. For the selection of such proteins, Yamamoto and Harayama (1996) proposed four criteria which are described at the beginning of this section; and Santos and Ochman (2004) applied an almost identical selection method: they selected genes (1) whose orthologues are present in a single copy in nearly all completely sequenced bacterial genomes, and (2) whose sequences contain at least two highly conserved regions separated by at least 100 amino acids. It should be noted that all the primers described to be universal failed to amplify target gene sequences of some test organisms selected from the six phyla.

Clearly, some other proteins which are not selected by these criteria may also be useful for the identification of bacteria and/or their functions. The elongation factor EF-Tu, which loads the amino acyl tRNA molecule onto the ribosome during translation, is an essential bacterial protein. It belongs to the Ras protein family and exhibits GTPase activity. The structural gene for EF-Tu, *tuf*, is duplicated in many strains of bacteria, but the duplication is not universal. This observation was interpreted to suggest that the duplication was an early event in the evolution of bacteria and

that the ancient duplication has been differentially lost and maintained in different lineages of bacteria (Lathe and Bork 2001). If this interpretation is correct, *tuf* is not a convenient marker for establishing a universal bacterial tree. However, this gene was useful for specific detection/identification of several bacterial strains (Ludwig et al. 1993; Picard et al. 2004).

Nonubiquitous genes are of use to detect specific strains and their functions. For example, the detection of Shiga-like toxin (SLT) gene from an isolate indicates that the source of the isolate is contaminated by Shiga-like-toxin-producing *E. coli* (Begum et al. 1993). Similarly, the detection of ketosynthase genes in actinomycetes provides information regarding antibiotic production capabilities, but not taxonomic information (Metsa-Ketela et al. 2002). FliC (a major component of bacterial flagellar filament) and OspC (an outer membrane protein) are not ubiquitous in all bacteria, but are useful for identification/detection of some pathogens. The evolution rates of these proteins are rapid as a result of acquiring adaptive mutations to avoid host defense mechanisms (Amhaz et al. 2004; Bellingham et al. 2001; Lin et al. 2002; Wang et al. 2003). Catabolic genes may also not be appropriate for phylogenetic inference because many of them are transferred from one host to the other by HGT (Jain et al. 1999). However, these genes are generally more specific for strain identification/detection.

5.5.2 Design of PCR Primers for the Amplification of Protein-encoding Genes: A Case Study with *gyrB*

The design of primers for the amplification of a specific gene from many different species is not straightforward because the individual gene sequences can be highly divergent. In this section, the design of PCR primers to amplify *gyrB*, the structural gene for the DNA gyrase B protein, is described. Similar approaches can be used to design primers for other protein-encoding genes.

DNA topoisomerases are enzymes essential for DNA replication, transcription, recombination, and repair. They control the level of supercoiling by cleaving and resealing the phosphodiester backbone of DNA. The topoisomerases are classified into type I (EC 5.99.1.2) and type II (EC 5.99.1.3), according to their enzymatic properties. The bacterial DNA gyrase is a type II topoisomerase that can introduce negative supercoils into a relaxed, closed, circular DNA molecule. This reaction is coupled to ATP hydrolysis, but DNA gyrase can also relax supercoiled DNA without ATP hydrolysis. DNA gyrase comprises two proteins in the quaternary structure of A2B; the A protein (GyrA) is approximately 100 kDa, and the B protein (GyrB) is either 90 kDa or 70 kDa. Comparison of the structures of the 90 kDa and

70 kDa classes of GyrBs revealed that the 90 kDa type has an insertion of about 170 amino acids commencing from residue 560 in the 70-kDa-type sequence. The N-terminal portion of GyrB is thought to catalyze the ATP-dependent supercoiling of DNA, while the C-terminal portion is thought to support complex formation with the A protein and ATP-independent relaxation.

Topoisomerase IV is a bacterial enzyme that appears to be closely related to DNA gyrase and required for partitioning of the bacterial chromosome (Kato et al. 1990). The role of this enzyme may be to unlink the catenated daughter chromosomes prior to partition. Topoisomerase IV cannot catalyze DNA supercoiling; and it catalyzes supercoil relaxation by a mechanism that requires ATP hydrolysis (Roca 1995, 2004; Wigley 1995). The B protein of topoisomerase IV, a paralogue of GyrB, is called ParE. The crystal structures of the N-terminal 43-kDa domains of GyrB and ParE have been determined (Bellon et al. 2004; Lamour et al. 2002; Wigley et al. 1991).

For two reasons, the design of primers for the PCR amplification of *gyrB* was difficult. First, in GyrB sequences, there are few highly conserved regions (seven amino acids for typical primer length) convenient for primer design. Primers are designed with all of the possible combinations of codons corresponding to the amino acid sequences of the conserved regions. This approach often results in an unusually high number of degenerate primers. Second, primers designed for the amplification of *gyrB* also amplify *parE* because these two sequences are very similar to each other in highly conserved regions.

The first set of universal primers, UP1/UP2r, was designed from two conserved regions of the amino acid sequences of GyrBs from *E. coli*, *P. putida*, and *B. subtilis* (Yamamoto and Harayama 1995). The two conserved amino acid sequences were reverse-translated, and a 41-nt N-terminal PCR primer (UP1; 5′-GAA GTC ATC ATG ACC GTT CTG CAY GSN GGN GGN AAR TTY GA-3′) and a 44-nt C-terminal PCR primer (UP2r; 5′-AGC AGG GTA CGG ATG TGC GAG CCR TCN ACR TCN GCR TCN GTC AT-3′) designed. The nucleotide sequences of the first 23 residues at the 5′ ends of both primers are not degenerate and, therefore, may not necessarily be complementary to the target *gyrB* sequences. These 23 residues, however, can be used as the hybridization sites of the sequencing primers, UP1s (5′-GAA GTC ATC ATG ACC GTT CTG CA-3′) and UP2rs (5′-AGC AGG GTA CGG ATG TGC GAG CC-3′). The remaining 18 and 21 nucleotides, respectively, of the UP1 and UP2r primers are degenerate and each of them makes 512 variations. PCR amplification of *gyrB* was carried out using DNA from bacteria of different taxonomic groups and PCR products with a size predicted from the known *gyrB* sequences (1.2 kb) were amplified from the various strains. Thus, by using a set of primers as presented above, it was possible to amplify the *gyrB* genes from a broad range of bacteria.

In many cases, however, *gyrB* amplification was more difficult than 16S rRNA amplification, resulting in low yields of specific amplification products, probably because of competitive inhibition as a result of high primer degeneracy. A universal base that can substitute for any of the four natural bases in DNA would be of great utility in PCR because using it could significantly reduce the complexity of degenerate oligonucleotide mixtures (Loakes 2001). We compared the efficiency of PCR between the *gyrB* primers containing degenerate nucleotides (UP1E, UP2r) and primers containing deoxyinosine (UP1Ei, UP2ri). The sequence of degenerate primer UP1E (which has broader specificity than UP1) was 5′-GAA GTC ATC ATG ACC GTT CTG CAY GSN GGN GGN AAR TTY RA-3′, while those of deoxyinosine primers UP1Ei and UP2ri were 5′-GAA GTC ATC ATG ACC GTT CTG CAY GSI GGI GGI AAR TTY RA-3′ and 5′-AGC AGG GTA CGG ATG TGC GAG CCR TCI ACR TCI GCR TCI GTC AT-3′, respectively. It was shown that yields of *gyrB* fragments increased by using the deoxyinosine primers, as described by Rossolini et al. (1994).

For further evaluation of *gyrB* primers, we retrieved *gyrB*-related sequences from the available whole-genome sequences. To collect the *gyrB* sequences, a BLAST search (http://www.ncbi.nih.gov/BLAST/Genome/EnvirSamplesBlast.html; McGinnis and Madden 2004) was performed against 203 microbial genome sequences available on the NCBI website (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi/) using *gyrB* of *E. coli* and *B. subtilis* as queries. Interestingly, a bacterial *gyrB*-related sequence was found in some genomes of Euryarchaeota, but not in other archaeal taxa. The *gyrB* sequences thus obtained were aligned and analyzed. The amino acid sequences in the UP2r site are highly conserved, while those in the UP1 site are less conserved (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi/). Thus, *gyrB* in ten of 203 bacterial genomes – most of which are "uncommon" bacteria – may not be amplified by using UP1Ei, and accordingly, new primers were designed for amplification of *gyrB* from these genomes. Otherwise, current universal primers, i. e. UP1Ei or UP1Gi (see below) and UP2ri, can be used for most other bacterial strains.

Recently, metagenome sequences of microbial communities in the Sargasso Sea (Venter et al. 2004) and biofilm in an extremely acidic mine drainage (Tyson et al. 2004) were released, thereby allowing the retrieval of *gyrB* sequences from uncultured microorganisms. From the metagenome sequences, we first collected complete or nearly complete gene sequences for *gyrB* or its *parE* paralogue (> 1,350 bp). Fifty-three *gyrB*/*parE* sequences were identified and their translated sequences used to construct a phylogenetic tree (Fig. 5.2). Thirty-nine sequences were classified as GyrB, while the remaining 14 sequences were designated as ParE. As shown in Fig. 5.2, 25 GyrB sequences were affiliated with proteobacteria, while some clusters were constituted uniquely by metagenomic GyrBs. The UP1 and

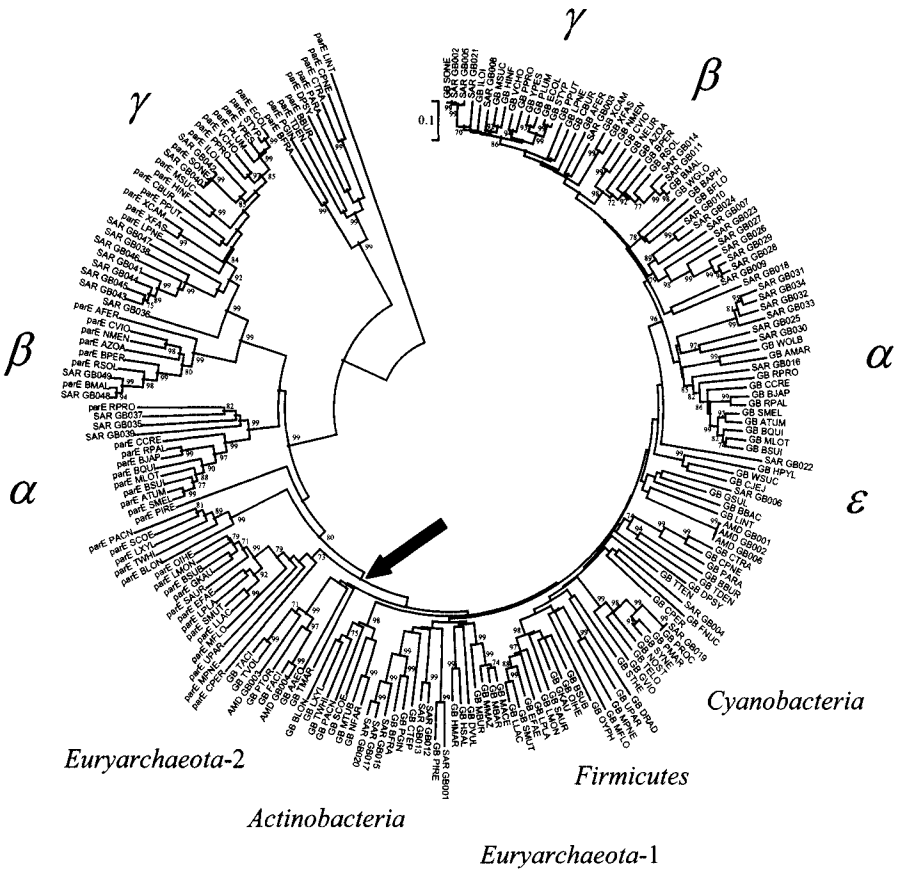


Fig. 5.2. Unrooted tree based on the amino acid sequences of GyrB and toposiomerase IV subunit B. Multiple alignment of the amino acid sequences was created using ClustalX (Thompson et al. 1997). The BLOSUM matrix was used for weight matrix parameters. The gap-open penalty was set to 20 and the gap-extension penalty to 0.1 for multiple alignment. The neighbor-joining tree was constructed based on the Poisson correction distance model by using MEGA ver. 2.1 (Kumar et al. 2001). The tree was constructed using the neighbor-joining method. Bootstrap values calculated from 1,000 trees are represented as percentages and given at each branch-point. Only values greater than 70 are shown. GyrB sequences are indicated by *GB*, while ParE sequences are indicated by *parE*. Sequences derived from metagenomes are indicated with *SAR* for the Saragasso Sea metagenome (Venter et al. 2004) and *AMD* for the acid mine drainage metagenome (Tyson et al. 2004). The *arrow* indicates the branching point between GyrB and ParE. Details of each sequence are given in the ICB database (http://www.mbio.jp/icb/; Kasai et al. 1998; Watanabe et al. 2001b)

UP2r regions in these 25 proteobacterial GyrBs matched the consensus sequences.

About 700 short *gyrB* sequences (< 1,350 bp) were additionally collected from the metagenome sequences. In these sequences, a new substitution

was found in the UP1 and the UP2r regions, respectively. However, most *gyrB* retrieved from the metagenome sequences encoded proteins whose sequences matched the consensus sequences of GyrB. The results of this survey, including the alignment of GyrB and ParE, are available at the ICB database website (http://www.mbio.jp/icb/).

From these analyses, we concluded that the current universal primers are useful for the majority of bacteria, and if PCR fails using these primer sets, the design and use of other primer sets should be considered, using the NCBI website (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi/) as a reference and guide.

The CODEHOP designer (http://blocks.fhcrc.org/codehop.html) was recently developed to design "consensus-degenerate hybrid oligonucleotide" primers (Rose et al. 2003). Each primer designed by their strategy consists of a short 3′ degenerate core region and a longer 5′ consensus clamp region. Only three to four highly conserved amino acid residues are necessary to design the core, whose annealing to template molecules is stabilized by the clamp sequences. During later rounds of amplification, the nondegenerate clamp permits stable annealing to product molecules (Rose et al. 1998). This method may be worth trying: Santos and Ochman (2004) successfully developed and used primer sets for the amplification of ten different proteins which are conserved in most bacterial genomes.

For the reconstruction of an accurate phylogenetic tree based on GyrB, it is essential to make the distinction between the two paralogues, GyrB and ParE, or their genes, *gyrB* and *parE*. In almost all other microbial genomes, genes for both *gyrB* and *parE* exist. However, *parE* is missing from the genomes of Corynebacterineae, *Clostridia*, Mollicutes, Rickettsiales, δ- and ε-proteobacteria, some insect symbionts of γ-proteobacteria (e. g. *Buchnera aphidicola*; Shigenobu et al. 2000), *Wigglesworthia glossinidia* (Akman et al. 2002), and *Blochmannia floridanus* (Gil et al. 2003). In general, orthologous and paralogous genes can be distinguished by creating a phylogenetic tree that includes both genes. As shown in Fig. 5.2, GyrB and ParE can be differentiated phylogenetically.

In Bacillales, the degenerate *gyrB* primers described above cannot be used for specific amplification of *gyrB* because these primers also anneal *parE*, whose translated sequences are identical to those of GyrB at the UP1 (HAGGKFG in the majority of *Bacillus* strains) and UP2r (MTDADVD) sites. For specific amplification of *gyrB* in Bacillales, another conserved sequence of GyrB, PGKLADC (from position 408 to position 414 of *B. subtilis* GyrB), which differs from the corresponding ParE sequence, was used to design a new degenerate primer (5′-CAR TCI GCI ARY TTI CCI GG-3′). This primer (5′-GAA GTC ATC ATG ACC GTT CTG CAY GSI GGI GGI AAR TTY RG-3′; specific to *gyrB* in the majority of *Bacillus* strains) in combination with UP1Gi or UP1Ei was successfully used for the specific

amplification of 900-bp *gyrB* fragments. These primer pairs can also be used for the specific amplification of *gyrB* in Lactobacillales and Mollicutes.

In α-proteobacteria, the amino acid sequences at UP1, UP2r, and PGK-LAD (used for a Bacillales primer) are completely identical between GyrB and ParE, and specific amplification of *gyrB* is difficult. GyrB is larger than ParE, however, in all Proteobacteria because of an insertion at a specific region of GyrB; and thus, the PCR product of *gyrB* is longer than that of *parE* and the difference of about 500 bp in length is large enough to separate the *gyrB* fragment from the *parE* fragment by agarose gel electrophoresis. The *gyrB* fragment can subsequently be isolated from the gel and sequenced.

In Actinobacteria, UP1E/UP1G and UP2r are available to amplify *gyrB*. In some novobiocin-resistant strains, however, additional *gyrB* which shows resistance to novobiocin is found in the novobiocin biosynthetic gene cluster (Steffensky et al. 2000; Thiara and Cundliffe 1988). Similarly, additional *gyrB* which is resistant to coumermycin A1 has been identified near the biosynthetic gene cluster of coumermycin A1 (Schmutz et al. 2003; Wang et al. 2000). In these cases, PCR amplifies two types of *gyrB* and therefore cloning of the amplified *gyrB* fragments followed by sequencing of several clones is required. Instead of universal primers, primer sets applicable to limited lineages of bacteria can also be designed (Hatano et al. 2003; Richert et al. 2005).

Currently, the ICB database (http://www.mbio.jp/icb/) stores more than 1,000 sequences of *gyrB* and several sets of universal primers.

5.6 Limitations in Reconstructing Phylogenetic Trees

No "right" method for estimating a phylogenetic tree exists, because all methods rely on a number of assumptions and approximations (Brocchieri 2001). In addition to the limitations imposed by analytical methods, "unusual" patterns of evolution of protein-encoding genes and/or a limited number of informative sites on the marker molecules are problems associated with phylogenetic tree reconstruction.

Gene duplication and gene transfer. Phylogenetic analyses based on different nucleotide or protein sequences often lead to contradictory results, and several hypotheses involving HGT or unrecognized gene duplications have been proposed to explain these discrepancies (Brocchieri 2001). Unidentified HGT may hamper the reconstruction of phylogenetic trees; indeed, an average 6% of genes in bacterial genomes are estimated to have been acquired by HGT (Ochman et al. 2000). Although recent lateral transfer of DNA would be recognized by a biased codon usage (Harayama 1994), the detection of ancient gene transfer events may be extremely difficult.

Gene duplications are probably widespread and many paralogous gene families may exist. If one duplicated family became extinct in one lineage and if the distinction between alternative families is difficult to discern from protein sequences, the reconstruction of gene phylogenies may produce contradictory results.

However, ancient gene duplications allow tracing back to the common ancestor of all organisms, identifying the root of the tree of life. When gene duplication has occurred in an ancestor, relatedness between two paralogous genes in the same descendents becomes lower than that between two orthologous genes in different descendents. This concept has been applied to identify parts of the tree of life where no suitable outgroup organisms exist, and has clarified the relationships among the three major lineages: Bacteria, Archaea, and Eukarya (Iwabe et al. 1989).

Number of informative sites. Equivalent gene sequences of two distantly related organisms may contain many sites where base substitutions have occurred. The phylogenetic information from these sites, however, is lost if these sites have suffered multiple mutations. Because mutation rates are much higher in synonymous (amino acid nonsubstituting) sites than in nonsynonymous (amino acid substituting) sites, synonymous sites are the first to be saturated with mutations. Because the inclusion of mutation-saturated sites in an analysis does not enhance resolution but increases noise, only the first and second (but not the third) positions of codons are often used for phylogenetic analyses of genes from distantly related organisms.

For more distantly related genes, the bias of G+C content influences nucleotide substitution rates. Furthermore, biases in dinucleotide and tetranucleotide frequencies (Karlin et al. 1997) can also provide constraints to free substitution of nucleotides. In such a case, it may be better to analyze the amino acid sequences of their products rather than their nucleotide sequences (however, note that protein sequences are secondarily affected by nucleotide compositional bias; see Foster and Hickey 1999).

Another advantage to using amino acid sequences instead of nucleotide sequences in phylogenetic analyses of distantly related organisms is the lower substitution rate of amino acid sequences compared to nucleotide sequences. If necessary, variable positions where higher amino acid changes are observed could be discarded from the analysis to reduce the noise. Such manipulation also reduces the number of useful sequences, however, and increases statistical errors. Thus, one should not believe that protein sequence analysis always provides useful phylogenetic information: proteins with low sequence conservation do not.

Most methods for inferring phylogenetic relationships only regard nucleotide and amino acid substitutions. As indel sequences are difficult to

align accurately, these are generally neglected; and gaps in alignments are either removed from the analysis, or arbitrarily treated. But indel sequences share a subset of homologous proteins that are very useful phylogenetic markers, because all strains possessing the markers can be considered as descendents from a common ancestor. A stretch of amino acid sequence that is strictly conserved in a subset of proteins may also provide valuable information about their phylogenetic relationships (Rivera and Lake 1992). These specific changes observed in the primary structures of proteins in one or more taxa but not in other taxa are called "signature sequences" and used to delineate many taxa (Gupta 1998). The statistical significance of such signature sequences should be tested, however, by a computational method (e. g. Karlin and Altschul 1990) before any conclusion can be drawn.

Limitations of analytical tools. Although several algorithms for the alignment of multiple sequences have been developed, the results of alignment are often not satisfactory in the eyes of experts and the aligned sequences are then corrected manually before construction of the phylogenetic tree. For example, the very popular Clustal W does not guarantee finding the best alignment. Any bias in the alignment can modify the topology of phylogenetic tree.

Many algorithms have also been developed for the reconstruction of phylogenies of life. However, certain algorithms are based on simplified assumptions. For example, there are the assumptions that genetic divergence occurs by accumulation of single nucleotide substitutions, that the rates of base changes are constant throughout gene sequences, or that evolution rates in different organisms are equal. Yet base substitutions may not be provoked solely by single mutation mechanisms, but by multiple mutation mechanisms involving duplication, deletion, transposition, or gene conversion that may also play important roles in divergent evolution (Averof et al. 2000; Harayama and Rekik 1993). The rates and patterns of base substitutions are not uniform even in a single gene (Vawter and Brown 1993). They are also influenced by G+C content and by the degree of gene expression (Rocha and Danchin 2004).

The evolution rate is the power of the mutation rate and the frequency of the fixation of mutation. The mutation rate depends on DNA replication fidelity, the efficacy of repair systems, and the degree of exposure to mutagens, while the chance of fixation of any mutation may depend on survival constraints on the mutations. During adaptation to a new environment, organisms may require the development of new sets of enzymes. In such adaptive processes, constraints for some mutations may be relaxed, while those for other mutations may be imposed, thus changing the probabilities of fixation of specific mutations (Moran 1996). It is likely that the assumption of an equal evolution rate may not apply to many living organisms. It is

known that the topology of phylogenetic trees is influenced when evolution rates of involved organisms are not equal (Felsenstein 1978).

Accordingly, results of analyses using any algorithms should be cautiously interpreted with an awareness of all possible assumptions of specific evolution models. One should not accept automatically the concept that the calculated evolutionary distance is a molecular clock – a measure of the time elapsed after the separation of two organisms. Readers are directed to an excellent review on this subject by Brocchieri (2001).

Compositional bias in nucleotide sequences. Most amino acids are determined by multiple codons; and degeneracy permits synonymous substitutions, which do not change the encoded amino acid. Because synonymous mutations are largely free from natural selection – in contrast to nonsynonymous mutations which are under selective pressure – the rate of fixation of synonymous substitutions is much higher than that of nonsynonymous substitutions. Synonymous substitutions in many organisms are, however, not random and codon usage in these organisms is biased. Although we do not yet fully understand molecular mechanisms leading to biased codon usage, it is influenced by genomic G+C content. In GC-rich organisms, the third codon position is rich in GC, and GCs are found primarily in this position. It is also known that preferred codons generally correspond to the most abundant tRNA species for each amino acid. The degree of codon bias is related to growth rate, gene expression, and relative tRNA abundance and seems to be important for efficient and accurate translation (Ikemura 1981; Rocha 2004; Rocha and Danchin 2004).

If variations in codon bias exist across the tree, incorrect phylogenetic trees will be constructed using any of the commonly used phylogenetic analysis methods (Chang and Campbell 2000). Accordingly, third positions are problematic in many data sets because base compositional bias is generally concentrated in these positions. Under such circumstances, use of the first and second codon positions is recommended (Jin and Nei 1990a,b).

5.7 Conclusion and Future Perspective

The bacterial diversity under a variety of environmental conditions estimated by various methods yielded values ranging from <100 to 400,000 "species". Of the 1,000 clones containing the 16S rRNA genes isolated from a 2.2-l seawater sample, approximately 60% were unique in their sequence (ribotype) and the number of ribotypes in the entire population was estimated to be 1,633. When clones harboring homology higher than

99% were grouped, the number was reduced to 520. In other words, two-thirds [(1,633 – 519)/1,633] or more of the ribotypes found were variants that could be grouped together within the tight "99% identity clusters." If clones sharing homology higher than 97% were grouped together, the number marginally decreased to 450, which indicated that most of the ribotypes or "operational taxonomic units" consist of individuals with high homology (> 99% identity) that can clearly be distinguished from others (Acinas et al. 2004a). In this sense, therefore, the definition of species by Rossello-Mora and Amann (2001), i.e. monophyletic and genomically coherent cluster, seems to have a valid basis. The existence of these taxonomically coherent clusters appears to support the ecotype concept (Cohan 2002, 2004) that predicts that the existence of highly homologous operational taxonomic units resulted from periodic selection. However, since approximately 50 Myr are calculated to be required to acquire 1% divergence in the 16S rRNA gene sequence, the existence of "microdiverse" clusters perhaps suggests two possibilities: ineffectiveness in the periodic selection (e.g., by weak intra-specific competition resulted from rapid environmental fluctuations), or rapid migration and mixing of microdiverse populations adapted to different microdiverse niches.

Despite these controversies, which should be addressed further, the picture of the bacterial world has become much clearer than it was 10 years ago. It is possible that the 99% identity clusters observed by Acinas et al. (2004a) correspond to fundamental entities of taxonomic groups of bacteria, or species. Furthermore, several lines of evidence support the idea that bacterial diversity is enormous (typically estimated to comprise 10^9 species) and hence their classification may become more and more complicated and impractical as the number of described species increases. This is good and bad news for taxonomists: good news because we have an almost inexhaustible supply of new bacteria and it will take another several hundred years to describe all the bacterial species. However, the value of describing new species undoubtedly diminishes as the number of described species increases and the contribution by "contemporary" taxonomists to the development of new scientific concepts will become less important in the future. This is bad news. It would become a big problem if a revolutionary technique for the isolation of novel bacteria becomes available, since the number of bacterial species would then reach a level that exceeds our ability to name them. Under such circumstances, how could microbial taxonomists cope with the increasing numbers of species?

Taxonomists have been seeking the way to delineate the history of life that includes several overlapping components: clustering of organisms based on variations among them (classification), deduction of causes and consequences of the variation, establishment of systems to organize clustered organisms into hierarchical categories (phylogenetic analysis), and provi-

sion of methods to assign organisms into specific clusters (identification, detection). This field is important because it provides a standard for the classification of organisms to solve many practical problems, as exemplified by the detection of pathogens and the implementation of nature preservation plans. The significance of taxonomy will not diminish but rather increase as the number of described species grows, if the taxonomy can co-evolve with other taxonomy-related scientific disciplines, including biosecurity and biotechnology.

Nonetheless, taxonomists should, in our opinion, sooner or later stop their routine work of describing new species with new names. Rather, it may be time to consider and implement a new way of cataloging bacterial species by adopting a new nomenclature rule: for example, assigning systematic numbers to each genera and species, while maintaining the conventional rules for the nomenclature of all taxa above family level. Many readers may think that this proposal is too radical. However, take astronomy as an example: recent cataloguing of stars will be generated by computer in combination with high-resolution telescopes, allowing the description of more than 10^9 distinct objects, a number that corresponds to the estimated diversity of bacteria.

At present, the lack of taxonomists is an urgent problem to solve and the situation in the future may become worse because of a disturbing decline in the number of students and young researchers in this discipline. We hope that this chapter may help readers to discover, or to know better, the wonderful aspects of bacterial taxonomy.

Acknowledgements. This work is supported by the New Energy and Industrial Technology Development Organization (NEDO). We would like to thank Drs. Katsumi Isono and Paul Kretchmer for their careful reading of the manuscript.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004a) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551–554

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004b) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. J Bacteriol 186:2629–2635

Adekambi T, Drancourt M (2004) Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, *hsp65*, *sodA*, *recA* and *rpoB* gene sequencing. Int J Syst Evol Microbiol 54:2095–2105

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet 32:402–407

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Amhaz JM, Andrade A, Bando SY, Tanaka TL, Moreira-Filho CA, Martinez MB (2004) Molecular typing and phylogenetic analysis of enteroinvasive *Escherichia coli* using the *fliC* gene sequence. FEMS Microbiol Lett 235:259–264

Averof M, Rokas A, Wolfe KH, Sharp PM (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science 287:1283–1286

Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905–920

Begum D, Strockbine NA, Sowers EG, Jackson MP (1993) Evaluation of a technique for identification of Shiga-like toxin-producing *Escherichia coli* by using polymerase chain reaction and digoxigenin-labeled probes. J Clin Microbiol 31:3153–3156

Bellingham NF, Morgan JA, Saunders JR, Winstanley C (2001) Flagellin gene sequence variation in the genus *Pseudomonas*. Syst Appl Microbiol 24:157–165

Bellon S, Parsons JD, Wei Y, Hayakawa K, Swenson LL, Charifson PS, Lippke JA, Aldape R, Gross CH (2004) Crystal structures of *Escherichia coli* topoisomerase IV ParE subunit (24 and 43 kilodaltons): a single residue dictates differences in novobiocin potency against topoisomerase IV and DNA gyrase. Antimicrob Agents Chemother 48:1856–1864

van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD (2003) Discordant phylogenies within the *rrn* loci of Rhizobia. J Bacteriol 185:2988–2998

Bocchetta M, Gribaldo S, Sanangelantoni A, Cammarano P (2000) Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J Mol Evol 50:366–380

Brocchieri L (2001) Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol 59:27–40

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285

Chang BS, Campbell DL (2000) Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. Mol Biol Evol 17:1220–1231

Cilia V, Lafay B, Christen R (1996) Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. Mol Biol Evol 13:451–461

Clayton RA, Sutton G, Hinkle PS Jr, Bult C, Fields C (1995) Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int J Syst Bacteriol 45:595–599

Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56:457–487

Cohan FM (2004) Concepts of bacterial biodiversity for the age of genomics. In: Fraser CM, Read T, Nelson KE (ed) Microbial genomes. Humana, Totowa, pp 175–194

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31:442–443

Curtis TP, Sloan WT (2004) Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. Curr Opin Microbiol 7:221–226

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99:10494–10499

Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12:1080–1090

Daubin V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. Science 301:829–832

DeLong EF, Pace NR (2001) Environmental diversity of bacteria and archaea. Syst Biol 50:470–478

Dew-Jager K, Yu WQ, Huang WM (1995) The *recA* gene of *Borrelia burgdorferi*. Gene 167:137–140

Dunbar J, Barns SM, Ticknor LO, Kuske CR (2002) Empirical and theoretical bacterial diversity in four Arizona soils. Appl Environ Microbiol 68:3035–3045

Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek 73:25–33

Edwards U, Rogall T, Blocker H, Emde M, Bottger EC (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res 17:7843–7853

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Field KG, Gordon D, Wright T, Rappe M, Urback E, Vergin K, Giovannoni SJ (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. Appl Environ Microbiol 63:63–70

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284–290

Fukushima M, Kakinuma K, Hayashi H, Nagai H, Ito K, Kawaguchi R (2003) Detection and identification of *Mycobacterium* species isolates by DNA microarray. J Clin Microbiol 41:2605–2615

Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Holldobler B, van Ham RC, Gross R, Moya A (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci USA 100:9388–9393

Goh SH, Potter S, Wood JO, Hemmingsen SM, Reynolds RP, Chow AW (1996) HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. J Clin Microbiol 34:818–823

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 4:379–391

Gribaldo S, Lumia V, Creti R, de Macario EC, Sanangelantoni A, Cammarano P (1999) Discontinuous occurrence of the hsp70 (*dnaK*) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181:434–443

Grimont PAD, Popoff MY, Frimond F, Coynault C, Lemelin M. (1980) Reproducivility and correlation study of three deoxiribonucleic acid hybridization procedures. Curr Microbiol 4:325–330

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435-1491

Gutell RR, Larsen N, Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol Rev 58:10–26

Gutell RR, Lee JC, Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol 12:301–310

Hagström Å, Pommier T, Rohwer F, Simu K, Stolte W, Svensson D, Zweifel UL (2002) Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. Appl Environ Microbiol 68:3628–3633

Harayama S (1994) Codon usage patterns suggest independent evolution of two catabolic operons on toluene-degradative plasmid TOL pWW0 of *Pseudomonas putida*. J Mol Evol 38:328–335

Harayama S, Rekik M (1993) Comparison of the nucleotide sequences of the meta-cleavage pathway genes of TOL plasmid pWW0 from *Pseudomonas putida* with other meta-cleavage genes suggests that both single and multiple nucleotide substitutions contribute to enzyme evolution. Mol Gen Genet 239:81–89

Hatano K, Nishii T, Kasai H (2003) Taxonomic re-evaluation of whorl-forming *Streptomyces* (formerly *Streptoverticillium*) species by using phenotypes, DNA–DNA hybridization and sequences of *gyrB*, and proposal of *Streptomyces luteireticuli* (ex Katoh and Arai 1957) corrig., sp. nov., nom. rev. Int J Syst Evol Microbiol 53:1519–1529

Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM (2004) cpnDB: a chaperonin sequence database. Genome Res 14:1669–1675

Hori H, Osawa S (1986) Evolutionary change in 5S rRNA secondary structure and a phylogenic tree of 352 5S rRNA species. Biosystems 19:163–172

Hugenholtz P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int J Syst Evol Microbiol 53:289–293

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2002a) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 67:4399–4406

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2002b) Erratum. Appl Environ Microbiol 68:448

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Jian W, Zhu L, Dong X (2001) New approach to phylogenetic analysis of the genus *Bifidobacterium* based on partial HSP60 gene sequences. Int J Syst Evol Microbiol 51:1633–1638

Jin L, Nei M (1990a) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Jin L, Nei M (1990b) Erratum. Mol Biol Evol 7:201

Johnson JL (1973) Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. Int J Syst Bacteriol 23:308–315

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 21–132

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Karlin S, Weinstock GM, Brendel V (1995) Bacterial classifications derived from *recA* protein sequence comparisons. J Bacteriol 177:6881–6893

Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913

Kasai H, Watanabe K, Gasteiger E, Bairoch A, Isono K, Yamamoto S, Harayama S (1998) Construction of the *gyrB* database for the identification and classification of bacteria. Genome Inform Ser Workshop Genome Inform 9:13–21

Kasai H, Ezaki T, Harayama S (2000) Differentiation of phylogenetically related slowly growing mycobacteria by their *gyrB* sequences. J Clin Microbiol 38:301–308

Kato J, Nishimura Y, Imamura R, Niki H, Hiraga S, Suzuki H (1990) New topoisomerase essential for chromosome segregation in *E. coli*. Cell 63:393–404

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H, Trevors JT (2004) Methods of studying soil microbial diversity. J Microbiol Methods 58:169–188

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA 100:9658–9662

Kwok AY, Chow AW (2003) Phylogenetic study of *Staphylococcus* and *Macrococcus* species based on partial *hsp60* gene sequences. Int J Syst Evol Microbiol 53:87–92

La Duc MT, Satomi M, Agata N, Venkateswaran K (2004) *gyrB* as a phylogenetic discriminator for members of the *Bacillus anthracis-cereus-thuringiensis* group. J Microbiol Methods 56:383–394

Lamour V, Hoermann L, Jeltsch JM, Oudet P, Moras D (2002) An open conformation of the *Thermus thermophilus* gyrase B ATP-binding domain. J Biol Chem 277:18947–18953

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82:6955–6959

Lathe WC 3rd, Bork P (2001) Evolution of *tuf* genes: ancient duplication, differential loss and gene conversion. FEBS Lett 502:113–116

Lerat E, Daubin V, Moran NA. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. PLoS Biol 1:E19

Lin T, Oliver JH Jr, Gao L (2002) Genetic diversity of the outer surface protein C gene of southern *Borrelia* isolates and its possible epidemiological, clinical, and pathogenetic implications. J Clin Microbiol 40:2572–2583

Lloyd AT, Sharp PM (1993) Evolution of the *recA* gene and the molecular phylogeny of bacteria. J Mol Evol 37:399–407

Loakes D (2001) Survey and summary: the applications of universal DNA base analogues. Nucleic Acids Res 29:2437–2447

Ludwig W, Schleifer KH (1994) Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev 15:155–173

Ludwig W, Neumaier J, Klugbauer N, Brockmann E, Roller C, Jilg S, Reetz K, Schachtner I, Ludvigsen A, Bachleitner M, et al (1993) Phylogenetic relationships of Bacteria based on comparative sequence analysis of elongation factor Tu and ATP-synthase beta-subunit genes. Antonie Van Leeuwenhoek 64:285–305

Lunn M, Sloan WT, Curtis TP (2004) Estimating bacterial diversity from clone libraries with flat rank abundance distributions. Environ Microbiol 6:1081–1085

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145

Martinez-Murcia AJ, Benlloch S, Collins MD (1992) Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. Int J Syst Bacteriol 42:412–421

McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32:W20–W25

Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71:491–499

Metsa-Ketela M, Halo L, Munukka E, Hakala J, Mantsala P, Ylihonko K (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various *Streptomyces* species. Appl Environ Microbiol 68:4472–4479

Mikkonen TP, Karenlampi RI, Hanninen ML (2004) Phylogenetic analysis of gastric and enterohepatic *Helicobacter* species based on partial HSP60 gene sequences. Int J Syst Evol Microbiol 54:753–758

Mollet C, Drancourt M, Raoult D (1997) *rpoB* sequence analysis as a novel basis for bacterial identification. Mol Microbiol 26:1005–1011

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci USA 93:2873–2378

Morse R, Collins MD, O'Hanlon K, Wallbanks S, Richardson PT (1996) Analysis of the beta′ subunit of DNA-dependent RNA polymerase does not support the hypothesis inferred from 16S rRNA analysis that *Oenococcus oeni* (formerly *Leuconostoc oenos*) is a tachytelic (fast-evolving) bacterium. Int J Syst Bacteriol 46:1004–1009

Notredame C, O'Brien EA, Higgins DG (1997) RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res 25:4570–4580

Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, Koonin EV (2004) Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 186:6575–6585

Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci USA 96:12638–12643

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Picard FJ, Ke D, Boudreau DK, Boissinot M, Huletsky A, Richard D, Ouellette M, Roy PH, Bergeron MG (2004) Use of *tuf* sequences for genus-specific PCR detection and phylogenetic analysis of 28 *streptococcal* species. J Clin Microbiol 42:3686–3695

Richert K, Brambilla E, Stackebrandt E (2005) Development of PCR primers specific for the amplification and direct sequencing of *gyrB* genes from microbacteria, order Actinomycetales. J Microbiol Methods 60:115–123

Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76

Roca, J (1995) The mechanisms of DNA topoisomerases. Trends Biochem Sci 20:156–160

Roca J (2004) The path of the DNA along the dimer interface of topoisomerase II. J Biol Chem 279:25783–25788

Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14:2279–2286

Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108–116

Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res 26:1628–1635

Rose TM, Henikoff JG, Henikoff S (2003) CODEHOP (consensus-degenerate hybrid oligonucleotide primer) PCR primer design. Nucleic Acids Res 31:3763–3766

Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67

Rossolini GM, Cresti S, Ingianni A, Cattani P, Riccio ML, Satta G (1994) Use of deoyinosine-containing primers vs degenerate primers for polymerase chain reaction based on ambiguous sequence information. Mol Cell Probes 8:91–98

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Santos SR, Ochman H (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environ Microbiol 6:754–759

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691

Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A (2000) Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. Cell 102:615–623

Schmutz E, Muhlenweg A, Li SM, Heide L (2003) Resistance genes of aminocoumarin producers: two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. Antimicrob Agents Chemother 47:869–877

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407:81–86

Smith AD, Lui TW, Tillier ER (2003) Empirical models for substitution in ribosomal RNA. Mol Biol Evol 21:419–427

Sneath PH (1993) Evidence from *Aeromonas* for genetic crossing-over in ribosomal sequences. Int J Syst Bacteriol 43:626–629

Specht T, Szymanski M, Barciszewska MZ, Barciszewski J, Erdmann VA (1997) Compilation of 5S rRNA and 5S rRNA gene sequences. Nucleic Acids Res 25:96–97

Springer M, Krajewski C (1989) DNA hybridization in animal taxonomy: a critique from first principles. Q Rev Biol 64:291–318

Stackebrandt E and Gobel BM (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44:846–849

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampfer P, Maiden MC, Nesme X, Rossello-Mora R, Swings J, Truper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047

Steffensky M, Muhlenweg A, Wang ZX, Li SM, Heide L (2000) Identification of the novobiocin biosynthetic gene cluster of *Streptomyces spheroides* NCIB 11891. Antimicrob Agents Chemother 44:1214–1222

Stepkowski T, Czaplinska M, Miedzinska K, Moulin L (2003) The variable part of the *dnaK* gene as an alternative marker for phylogenetic studies of rhizobia and related alpha Proteobacteria. Syst Appl Microbiol 26:483–494

Thiara AS, Cundliffe E (1988) Cloning and characterization of a DNA gyrase B gene from *Streptomyces sphaeroides* that confers resistance to novobiocin. EMBO J 7:2255–2259

Thompson CC, Thompson FL, Vandemeulebroecke K, Hoste B, Dawyndt P, Swings J (2004) Use of *recA* as an alternative phylogenetic marker in the family Vibrionaceae. Int J Syst Evol Microbiol 54:919–924

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res 30:2083–2088

Torsvik V, Goksoyr J, Daae FL (1990a) High diversity in DNA of soil bacteria. Appl Environ Microbiol 56:782–787

Torsvik V, Salte K, Sorheim R, Goksoyr J (1990b) Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. Appl Environ Microbiol 56:776–781

Torsvik V, Ovreas L, Thingstad TF (2002) Prokaryotic diversity – magnitude, dynamics, and controlling factors. Science 296:1064–1066

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Van de Peer Y, Chapelle S, De Wachter R (1996) A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res 24:3381–3391

Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Vauterin L, Host B, Kersters K, Swings J (1995) Reclassification of *Xanthomonas*. Int J Syst Bacteriol 45:472–489

Vawter L, Brown WM (1993) Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics 134:597–608

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Viale AM, Arakaki AK, Soncini FC, Ferreyra RG (1994) Evolutionary relationships among bacterial groups as inferred from GroEL (Chaperonin) sequence comparisons. Int J Syst Bacteriol 44:527–533

Wang L, Rothemund D, Curd H, Reeves PR (2003) Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. J Bacteriol 185:2936–2943

Wang ZX, Li SM, Heide L (2000) Identification of the coumermycin A(1) biosynthetic gene cluster of *Streptomyces rishiriensis* DSM 40489. Antimicrob Agents Chemother 44:3040–3048

Watanabe K, Teramoto M, Harayama S (1999) An outbreak of nonflocculating catabolic populations caused the breakdown of a phenol-digesting activated-sludge process. Appl Environ Microbiol 65:2813–2819

Watanabe K, Kodama Y, Harayama S (2001a) Design and evaluation of PCR primers to amplify 16S ribosomal DNA fragments used for community fingerprinting. J Microbiol Methods 44:253–262

Watanabe K, Nelson J, Harayama S, Kasai H (2001b) ICB database: the *gyrB* database for identification and classification of bacteria. Nucleic Acids Res 29:344–345

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stachebrandt E, Starr MP, Truper HG (1987) Report of the ad hoc committee on reconcilation of approaches to bacterial systematics. Int J Syst Bacteriol 37:463–464

Wigley DB (1995) Structure and mechanism of DNA topoisomerases. Annu Rev Biophys Biomol Struct 24:185–208

Wigley DB, Davies GJ, Dodson EJ, Maxwell A, Dodson G (1991) Crystal structure of an N-terminal fragment of the DNA gyrase B protein. Nature 351:624–629

Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonhein C, Hartsch T, Ramakrishnan V (2000) Structure of the 30 S ribosomal subunit. Nature 407:327–339

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Wuyts J, Van de Peer Y, De Wachter R (2001) Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. Nucleic Acids Res 29:5017–5028

Yamamoto S, Harayama S (1995) PCR amplification and direct sequencing of *gyrB* genes with universal primers and their application to the detection and taxonomic analysis of *Pseudomonas putida* strains. Appl Environ Microbiol 61:1104–1109

Yamamoto S, Harayama S (1996) Phylogenetic analysis of *Acinetobacter* strains based on the nucleotide sequences of *gyrB* genes and on the amino acid sequences of their products. Int J Syst Bacteriol 46:506–511

Yamamoto S, Harayama S (1998) Phylogenetic relationships of *Pseudomonas putida* strains deduced from the nucleotide sequences of *gyrB*, *rpoD* and 16S rRNA genes. Int J Syst Bacteriol 48:813–819

Yamamoto S, Bouvet PJ, Harayama S (1999) Phylogenetic structures of the genus *Acinetobacter* based on *gyrB* sequences: comparison with the grouping by DNA–DNA hybridization. Int J Syst Bacteriol 49:87–95

Yanez MA, Catalan V, Apraiz D, Figueras MJ, Martinez-Murcia AJ (2003) Phylogenetic analysis of members of the genus *Aeromonas* based on *gyrB* gene sequences. Int J Syst Evol Microbiol 53:875–883

Zeigler DR (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol 53:1893–1900

Zhaxybayeva O, Lapierre P, Gogarten JP (2004) Genome mosaicism and organismal lineages. Trends Genet 20:254–260

6 Integrated Databasing and Analysis

Luc Vauterin, Paul Vauterin

6.1 Introduction

Unlike higher organisms such as animals and plants, microorganisms cannot adequately be described in terms of morphological features. Microbiologists have been forced to use alternative features to characterize and describe the organisms they study. This has led to the exploration of a variety of phenotypic and genotypic treats, including biochemical and physiological properties, chemotaxonomical markers, protein patterns, DNA restriction and amplification fragment patterns, and DNA sequence analysis. Remarkably, although bacterial taxonomy is a young discipline compared to the taxonomies of higher organisms, it has quickly grown to be one of the most progressive in terms of the application and exploration of advanced molecular techniques.

A modern classification technique typically yields large amounts of data. For example, a microplate system containing carbon sources such as the Biolog system (Biolog, Hayward, Calif., USA) produces 96 optical density (OD) values per reading, and by means of HPLC, more than 30 quantifiable fatty acids can be detected in a single bacterial strain (MIDI, Newark, Del., USA). More extremely, a single densitometric scanning record of an electrophoresis profile can be composed of several thousands of densitometric values and a sequenced gene such as the 16S ribosomal RNA gene typically is up to 1,500 nucleotide bases long. It is obvious that, even for the comparison between two bacterial strains, such amounts of data cannot be interpreted objectively without the aid of computers and software. This leads us to a whole universe of techniques and computer algorithms for resemblance estimation, dimension reduction, and clustering, commonly referred to as *numerical analysis*, or in the context of classification, *numerical taxonomy*. One of the most valuable reference works on numerical taxonomy is written by Sneath and Sokal (1972). In spite of the fact that a large number of new data mining, clustering and identification techniques have been introduced in taxonomy since this early work, there is

Luc Vauterin, Paul Vauterin: Applied Maths BVBA, Keistraat 120, 9830 Sint-Martens-Latem, Belgium, E-mail: Luc_Vauterin@applied-maths.com

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

no recent book that covers the many aspects of numerical taxonomy in a comprehensive way.

While most other chapters in this book deal with various genomic characterization and typing techniques, the goal of this chapter is to go more deeply into databasing and analysis of different typing data. It is not within the scope of the chapter to provide a review of clustering and identification techniques, which would require a whole book rather than a chapter. Instead, we will focus on some basic concepts on typing data and issues like reproducibility, portability, normalization, and analysis tools for different data types. Finally, we will examine the problem of consensus clustering and highlight some existing solutions.

Unless explicitly referenced, all of the terminology, methods, algorithms, and examples provided in this chapter are based upon the BioNumerics software package (Applied Maths, Sint-Martens-Latem, Belgium). Throughout the chapter, the term *entry* will be used for the data obtained from one single organism studied. We prefer not to use the terms organism, strain or isolate, as there is no symmetry between the two: an organism, strain or isolate can be stored as several entries, for example if repeated experiments are conducted. We will also use the term *association coefficient* to denote both similarity and distance coefficients: it is relatively easy to convert a distance into a similarity and vice versa. Likewise, we will use the term *resemblance matrix* to denote both similarity and distance matrices.

6.2 Classes of Data

Just like we can infer classifications of microorganisms, we can classify the data types that are used to study these organisms. As is the case with classifications in general, the result depends both on the criteria used and on the intended purpose of the classification. One classification that is often used by taxonomists is according to the source of the data: genotypic and phenotypic data (Vandamme et al. 1996). Phenotypic data can be further subdivided into morphological, biochemical, and chemotaxonomic data; and genotypic can be subdivided into fragment analysis data and sequencing data. From our perspective, a more useful classification is obtained from the type of data: data of the same type can be analyzed using the same approaches and algorithms, and can be relatively easy assembled into *composite* data sets (see further). As such, six major classes of data can be distinguished, which we will call *data types*: the character type, the fingerprint type, the sequence type, the matrix type, the trend curve type, and the 2D gel type. These data types will be defined and described in the following paragraphs.

6.3 Character Type Data

6.3.1 Definition

Any experimental measurement that results in an array of named characters for each organism studied, can be classified as a character type. Character data can easily be applied for numerical analysis in a sense that the data requires little or no manipulation and can be presented as an $m \times n$ data matrix. The early reports on numerical classification all used character type data as input.

Within the character type data, we can make two further subdivisions that have implications on the way the data is analyzed: *open* versus *closed* character data sets, and *binary*, *numerical* or *categorical* data sets.

Open and Closed Data Sets

A closed data set is generated when the investigator analyzes a set of features that is well defined before the study is initiated. Regardless of the number of organisms studied, the character set remains the same. Closed data sets are typically generated when biochemical, physiological and morphological features are recorded. Other examples of closed character sets are antibiotic resistance spectra, phage typing profiles, and of course, the commercial phenotypic test panels such as BioMérieux (Marcy l'Etoile, France) and Biolog (Hayward, Calif., USA).

A typical example of a data type where a closed data set is not suitable is the analysis of cellular fatty acids. Whereas hundreds of different fatty acid compounds have been discovered in bacteria (Sasser 1990), it is unlikely to identify more than a few dozens in a particular genus. Considering the fatty acids as a closed data set would mean that the investigator is forced to create an initial data set containing more than 100 detectable fatty acid species and then, as the study proceeds, filling the values for the entries added to the data set. The result would be a data set for which most characters are zero for all entries which is, in many aspects, not desirable.

An open data set allows the investigator to "discover" the features during the study. The study starts with a character set of zero characters, and as the study proceeds by adding new entries, new characters that are identified are added to the data set. In the case of fatty acid profiling, the final data set only contains those fatty acids that are present in at least one of the entries studied.

Binary, Numerical, and Categorical Data

Binary character data are the simplest form of information, where a character is interpreted as either positive or negative, and usually recorded as 1 or 0. This type of encoding is suitable for features that are either present or absent, for example formation of spores, presence of a plasmid, motility, etc. Many features, however, cannot easily be interpreted as just present or absent, because "absence" and "presence" are often observed on a continuous scale from undetectable to very strong. For example, an enzymatic activity can be undetectable in one bacterial strain, in which case it can be recorded as absent. In a second strain, there can be a weak activity for the same enzyme, whereas in a third strain, the activity measured can be very strong. By recording the two activities with a simple "1", useful information is actually lost. Worse, a very weak activity might be interpreted as positive in one observation and negative in another. It is therefore more interesting as well as more reliable to record such features as numerical values, ranging, for example, from zero to 100%. It is obvious that objective numerical recordings can only be obtained by using a measuring instrument, which can be a scanner, a charge-coupled device (CCD) camera, a microplate reader, etc. Note that numerical data can be converted into binary data "on the fly", anytime during the operations. Several parameters can be applied for such conversions, such as a specific threshold as a percentage of the maximum value, or a percentage of the average or the median value.

In addition to binary and numerical encoding, there is a third class of character encoding, which we call *categorical.* As the name suggests, categorical encoding is used for characters that exhibit several categories. Therefore, the name *multi-state* characters is also used (Sneath and Sokal 1973). A simple example is color: if you are to describe the color of a bacterial colony, you can define several categories, such as yellow, white, cream, grey, pink, etc. In a categorical encoding, each color will usually be assigned an integer number, for example yellow = 1, white = 2, etc. In contrast to numerical values, the numbers have no rank order associated: in the above example, white cannot be considered "bigger" than yellow. Therefore, association coefficients used for numerical data cannot be used for categorical data. A categorical association coefficient considers two values only as a match if they are the same, otherwise they are considered as a mismatch. Note that binary data can be considered as a special case of categorical data, where each character contains only two categories. Categorical data can always be converted into binary data, by representing each category of a character as a new, binary character. In the above example, each color would represent a binary character. A disadvantage of this approach is that the characters are not independent from each other: if a bacterial colony is white, it cannot be yellow at the same time and vice versa. A number of statistical methods requires the characters to be independent.

Typical examples of categorical character data techniques used in bacterial molecular typing, population genetics, and taxonomy are multi-locus sequence typing (MLST) and variable number tandem repeats (VNTR). These techniques are discussed further in Sects. 6.5.6 and 6.4.4, respectively.

6.3.2 Data Transformation

Data transformation is usually the first manipulation that is performed on the data after it is logged into a computer. It includes all of the manipulations performed on the numerical input values before they are subjected to numerical analysis. This can include averaging from repeated experiments, log transformation, standardization, regression analysis, imputing missing values, etc. For most character type data, the need for transformation, if any, is usually limited. Microarray data form a special case, where data transformation is an essential part of the analysis process.

A number of clustering algorithms and statistical tools require a *data matrix* as input. A data matrix can be assembled from any set of common characters for a number of entries. The character values for each entry together form a data array or *data vector*; and the set of all data vectors together form the data matrix (see Fig. 6.1).

Standardization

Standardization of the data vectors is the most elementary step in the transformation of a data matrix. It can consist of two functions: *centering* and *scaling*.

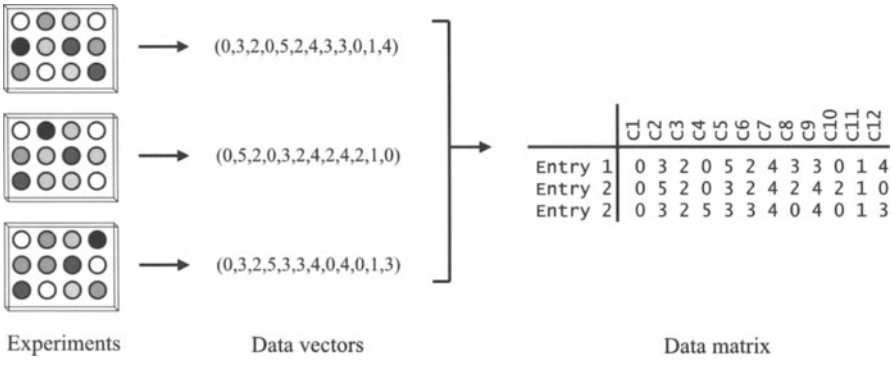


Fig. 6.1. Creation of a data matrix from digitized character experiments

By centering an array of values, a constant value, C, is subtracted from every value of that array:

$$x_i' = x_i - C \tag{6.1}$$

Usually, the mean value of the array, $\bar{x}$ is taken as C:

$$x_i' = x_i - \bar{x} \tag{6.2}$$

After centering according to (6.2), the average (or the sum) of all values of the array (the *offset*) is zero. Therefore, this standardization is called *centering around zero*:

$$\frac{\sum (x_i - \bar{x})}{n} = 0 \tag{6.3}$$

Figure 6.2 shows the effect of centering around zero of a data vector.

By *scaling* is meant that the values of an array are divided by a constant value:

$$x_i'' = \frac{x_i}{S} \tag{6.4}$$

Usually, the *root mean square* (RMS) value of the array is taken as S. The RMS is defined as follows:

$$RMS(x) = \sqrt{\frac{\sum x_i^2}{n}} \tag{6.5}$$

When the array is centered around zero, i. e. the mean value is subtracted from the array in (6.2), the RMS value in (6.5) becomes the *standard*

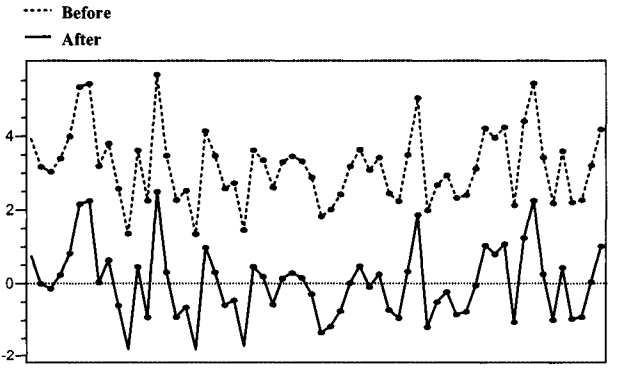


Fig. 6.2. Character array displayed as a curve, without centering (*dotted line*), and centered around zero (*solid line*)

deviation (SD):

$$SD(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \tag{6.6}$$

Since a data matrix consist of rows and columns, it can also be standardized along the columns (Fig. 6.3).

In general, one can assign the following interpretation to the use of the different standardization methods for columns and rows:

A. Rows (Entries)

1. Centering around zero: differences in background are neutralized. This standardization is recommended when background levels are known or expected to be variable between entries (e.g. more or less purified strains, DNA, enzymes, ...).
2. Scaling to RMS: compensates for differences in overall intensity between the arrays of different entries. Use this standardization when you expect different entries to yield stronger or weaker overall reaction.

In some special cases, the standardization should be switched off (fully or in part) to reveal certain features. For example, some entries may display lower overall metabolic activity and gene expression activity due to experimental circumstances. It is obvious that this feature can no longer be discovered after division by the RMS value.

B. Columns (Characters)

1. Centering around zero: the background is removed per character, considered over all the entries included. This standardization is useful to

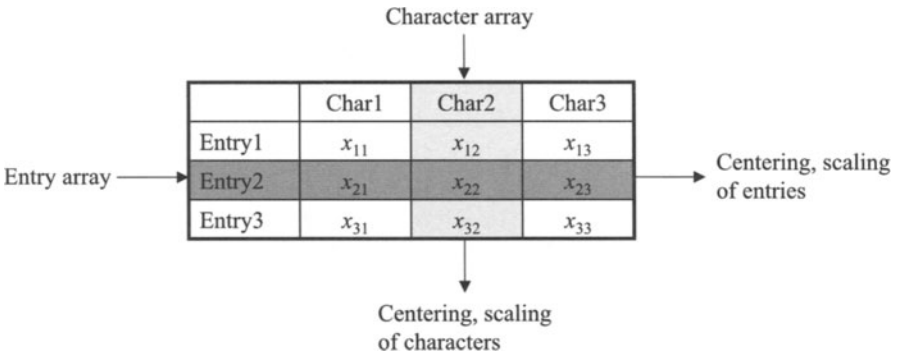


Fig. 6.3. Data matrix showing the meaning of average and RMS correction at the rows and columns level

compensate for different background levels inherent to different characters included. This type of standardization can seriously distort the result of entry clustering.

2. Scaling to RMS: overall intensity differences are compensated for between characters from a collection of entries. This standardization can be useful to neutralize differences in intensity caused by more efficient reaction of some characters as compared to others (e. g. in case one probe hybridizes better than another). Also, this scaling is useful to weigh the characters equally. Taking again the example of fatty acids, it is known that some fatty acids are present in very abundant amounts, whereas others are found only in very small percentages. Most association coefficients will weigh differences in the abundant fatty acids much heavier than those found in minor fatty acids. As the latter can be equally relevant for classification and identification, scaling of the characters can be appropriate.

While column standardization can be useful as explained above, it can seriously alter the result of a clustering of entries (rows) and may produce unexpected results. In addition it is worth noting that, while entry standardization is a process that acts on individual entries, character standardization is dependent on the collection of entries studied. Hence, the result of a character standardization will be different when entries are added or deleted from the study.

If both entries and characters are standardized, the sequence of standardization is also important. The normal procedure is to standardize the entries first and the characters next.

Dealing with Missing Values

In the course of a study, it may happen that a number of character observations are ambiguous or unreliable, so that the investigator prefers not to include them in the data matrix. The result is an incomplete data matrix, which is an unsuitable data source for most coefficients and algorithms. A simple way to solve this problem is by generating a new data matrix from the subset of characters for which every entry has a valid record. This method, however, can allow a lot of useful information to be lost, by removing characters for which just one or a few values are missing. The problem is particularly accute for microarrays, where extremely large numbers of genes are usually studied. There are two possibilities to bypass the problem of incomplete data matrices without losing useful information.

1. Most association coefficients act on pairs of character arrays and do not need a complete data matrix as input (see further). For such coefficients,

the subset of common characters for each pair of entries can be used as input.

2. An incomplete data matrix can be converted into a complete data matrix by "predicting" missing values on the basis of other values from the data matrix, a process called *imputing*. It can be based on simple calculations, such as imputing the average or median of the row or column array, or both. A more sophisticated method is to calculate the average of the k nearest neighbors, k being a variable that depends on the number of entries.

6.3.3 Cluster Analysis of Character Type Data

A data matrix of binary or numerical characters is the most basic type of input for comparison and cluster analysis. Therefore, the analysis of character data will be discussed in Sect. 6.10 (Hierarchical cluster analysis).

6.4 Fingerprint Type Data

6.4.1 Definition

Any array of intensity values recorded as a one-dimensional profile of peaks or bands can be considered as a fingerprint type. Examples are electrophoresis patterns which can be produced on slab gels, capillary electrophoresis systems, or sequencers. Also gas chromatographic or HPLC profiles, spectrophotometric curves, MALDI and SELDI profiles, etc. are one-dimensional peak profiles which belong to the fingerprint type class. Fingerprint types can also be derived from image files (such as TIFF, JPEG). In that case, one of the first steps in the image preprocessing is to convert the two-dimensional bitmap image into a set of one-dimensional densitometric arrays.

Note that in specific cases character type data (Sect. 6.3) also can be derived from densitometric curves. In case of fatty acid analysis for example, the fatty acid content of a bacterial strain is derived from a HPLC peak profile. In general, if the peaks of a densitometric profile can be identified as named characters (for example fatty acids), they can be considered as character type data. Fingerprint type data typically are densitometric profiles for which the peaks have not been identified as named characters.

Note that molecular weights, isoelectric points, or numbers of bases are estimations of physical properties but not names. There is always some error associated with such estimations, and therefore, comparisons between fingerprint patterns have to be based upon certain assumptions (see below,Sect. 6.4.3).

6.4.2 Preprocessing of Fingerprint Data

Fingerprint type data can be obtained in two different formats:

1. Two-dimensional image files (TIFF, GIF, JPEG, etc.) of gels, which are produced by, e. g. a flatbed scanner or CCD camera.
2. Densitometric curves, which can be derived from automated sequencers, laser densitometers, gas chromatographs or HPLC instruments, spectrophotometers, etc.

Starting from scanned two-dimensional gel images, the entire preprocessing scheme involves four steps (Fig. 6.4): (1) import and preprocessing of the gel image, (2) extraction of densitometric curves, (3) normalization, and (4) band detection. When densitometric curves are imported, step 1 is skipped. The preprocessing then starts at step 2 or step 3, depending on whether background subtraction and filtering of densitometric curves is required or not.

Step 1. Import and Preprocessing of the Gel Image

Bitmap images can be stored in different formats. In a *true RGB* TIFF file, each color component, red, green, and blue, of a pixel on the bitmap is represented by one byte (8 bits), i. e. 256 levels of color intensity. The format is therefore also called a *24-bit* TIFF file. For densitometric analysis

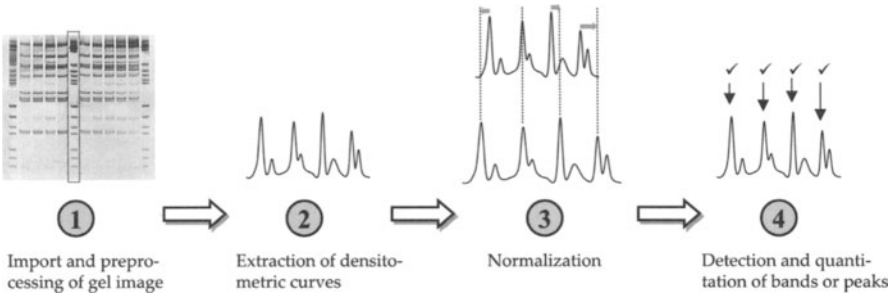


Fig. 6.4. Main steps in the preprocessing of electrophoresis patterns

of banding patterns, however, only one channel can be used and true 24-bit TIFF images must therefore be converted to *8-bit* or *grayscale* TIFF images. The average of the three color channels is usually taken, but it is also possible to create a grayscale TIFF file from just one channel, for example red only. Many scanners, cameras, and densitometers produce TIFF files of a higher OD range, which can for example be 10-bit (1,024 gray levels), 12-bit (4,096 gray levels) or 16-bit (65,536 gray levels). In all these cases, the images are stored as 16-bit monochrome TIFF files. In the GIF format, which is a compressed 8-bit image format, the three color channels are mapped onto one palette of 256 colors. For color images, this format usually causes some loss of quality. This is, however, not a problem with grayscale gel images, and since the compression used in the GIF format does not result in loss of pixel quality, this format is very well suited for storing gel images. The JPEG format uses 24-color depth, but compresses the image in a way that its pixel definition is slightly affected. The loss of quality depends on the strength of the compression applied, but is usually

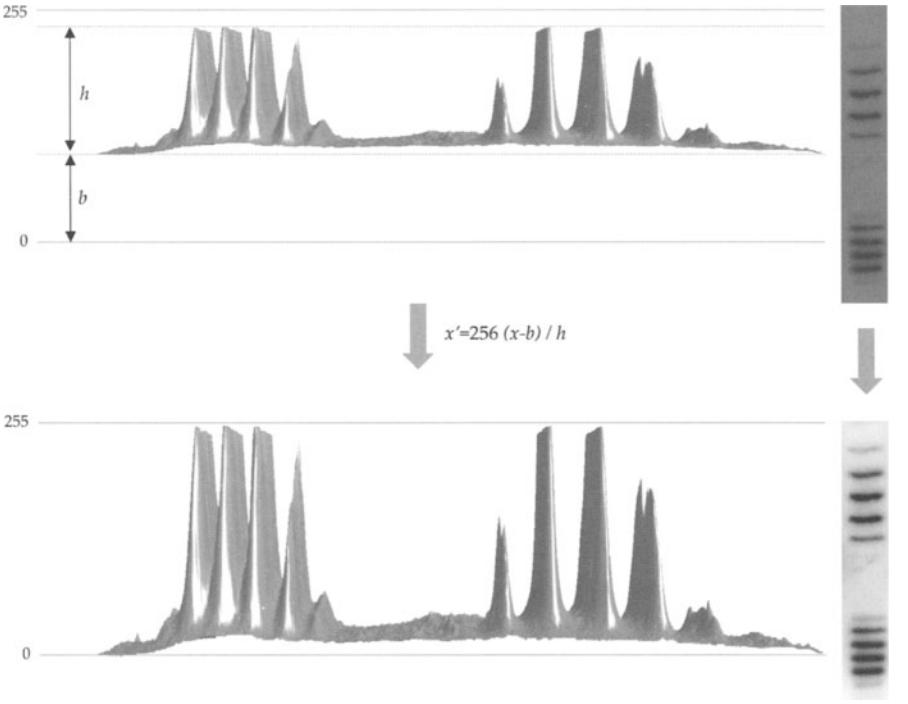


Fig. 6.5. The effect of brightness and contrast adjustment: brightness is increased so that the lowest background is white; and the contrast is then increased until the image covers the full 256 grayscale gamma

not problematic for gel images and photographs, which are by nature not characterized by sharp contours. It is, however, not recommended to edit and save JPEG images multiple times.

It is recommended to optimize the brightness and contrast of the image prior to further editing. The lowest pixel value b and the highest pixel value $b + h$ of the bitmap are identified and a linear transformation $x' = 256(x-b)/h$ is performed on every pixel (Fig. 6.5). As a result, the image has the lowest background and maximal contrast within the OD range used.

This background and scaling adjustment is a simple linear, and in principle reversible, transformation which can be applied to any bitmap image. However, it does not correct for local background differences. To remove

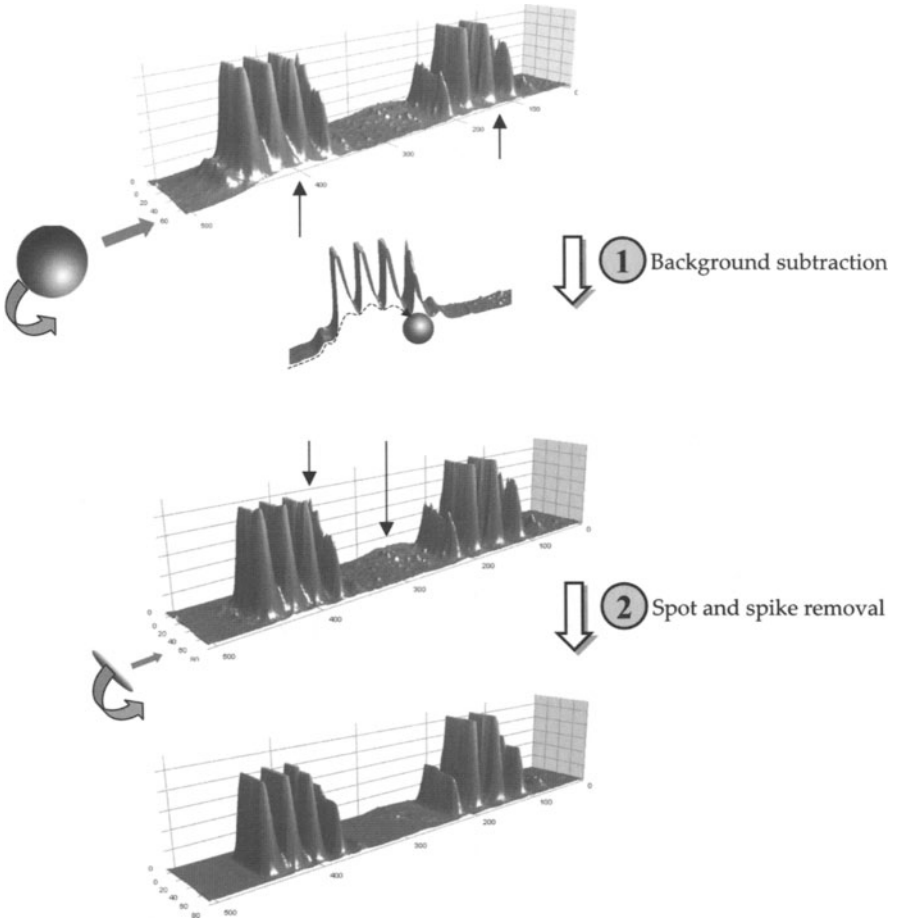


Fig. 6.6. Preprocessing of 2D gel images: 1 non-linear background subtraction, 2 spot and spike removal

local differences in background, the *rolling ball* mechanism can be used. A sphere is pushed against the underside of the image surface (Fig. 6.6, part 1) and rolled so that the entire surface is traversed. The surface formed by the highest points reached anywhere by the sphere is subtracted from the image's surface. The diameter of the sphere is to be chosen large enough so that it cannot roll into the cavities formed by the bands.

A more or less opposite mechanism can be used for removing spots and spikes from the image. To that end, a small ellipsoid is pushed against the underside of the image surface (Fig. 6.6, part 2) and rolled in the direction of the pattern. The ellipsoid will fit perfectly into the elongate cavities formed by the bands, but will not fit in small round cavities formed by spikes and spots. The surface formed by the highest points reached anywhere by the ellipsoid is used in place of the original image. The size of the ellipsoid is critical: if it is too large, the resulting image will be heavily distorted.

One other important action in step 1 is to define a bounding box around the relevant part of the image (Fig. 6.7) and to delineate the contours of the patterns on the gel image, so that these can be stored as individual image strips (*gelstrips*), normalized (see step 3), and shown independently in combination with dendrograms and other comparisons. Optionally, the contours of the bounding box can be used to apply a correction for distortion and smiling effects on the gel (Fig. 6.7). An advantage of this manip-

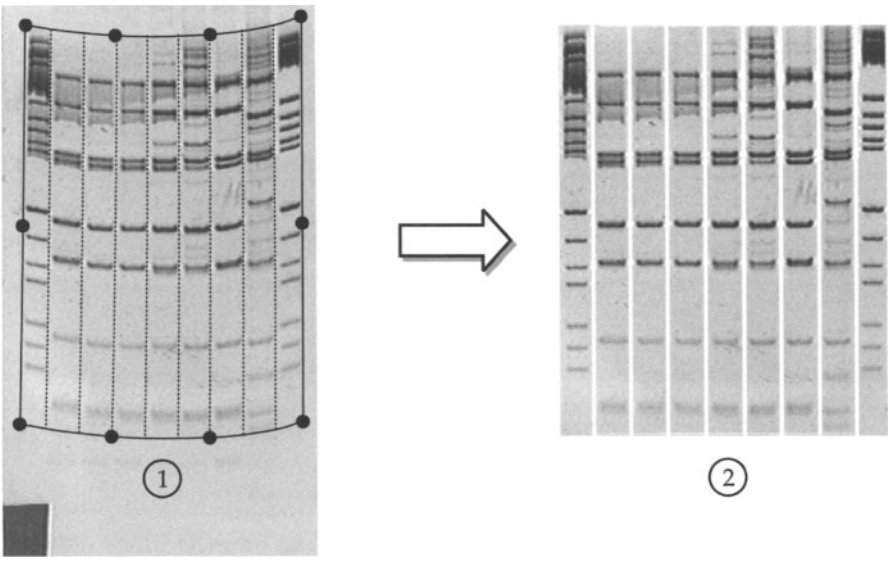


Fig. 6.7. Using a bounding box with distortion nodes to border the relevant part of the gel and to correct gel distortion and "smiling" effects (shown in 1). Patterns are delineated on the gel (1) and converted into separate gelstrips (2)

ulation is that smiling and distortion on bands are removed, which results in sharper peaks on densitometric curves (see step 2). A disadvantage, however, is that it involves a deformation of the original image, which one might want to avoid, honoring the principle that one should try to stay as close to the original data as possible.

Step 2. Extracting Densitometric Curves from the Gel Image Lanes

As shown in Fig. 6.8, a densitometric curve is calculated from a gel lane by averaging the pixel values on the same horizontal line within an *averaging window*. The size of the averaging window depends on several factors; obviously the resolution of the image, but also the shape of the bands. If "smiling" or halter-shaped bands occur, it is probably more reliable to include only the center of the bands, so as to avoid the distortion at the edges to be reflected in the curves.

In Fig. 6.8, the densitometric curve has been calculated using the *arithmetic average* (sum of the pixel values divided by the number) and the *median average*. In median averaging, the values are ranked according to height. If the number of values is odd, the value in the center is used; if the number is even, the average of the two center values is used. The interesting characteristic of the median is that the average is not distorted by small numbers of really excessive values (*outliers*). In the case of banding pat-

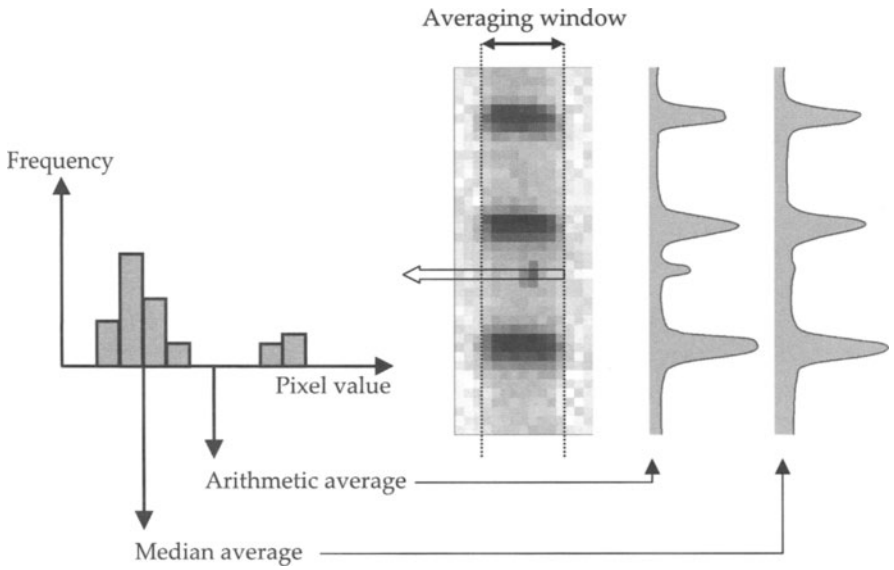


Fig. 6.8. Calculating densitometric curves from a scanned gel image using arithmetic averaging and median averaging

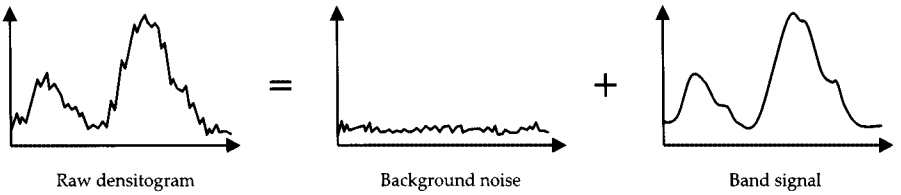


Fig. 6.9. Decomposition of raw densitometric curve into signal and noise

terns, a small spot does not result in a peak on the derived densitometric curve (see Fig. 6.8).

Some additional enhancement techniques that can be applied to densitometric curves are described below.

Noise filtering. Images obtained using CCD technology (CCD cameras, flatbed scanners) can contain quite some random background scatter or "noise". Typically, a densitometric curve is composed of two types of signal: the broad signal from the bands and the small, random scatter from the background noise (Fig. 6.9).

Although averaging of the densitometric profiles as described above can reduce the noise to a large extent (especially arithmetic averaging), it may be necessary to apply a noise filter to smoothen the curves. A very efficient filter is the so-called *least-square filter*. This filter consists of two parameters: a *cut-off value* specifying a bandwidth below which the signal is filtered out as noise, and a *power*, determining the strength of the filter, i. e. the sharpness of the transition between signal and noise. The cut-off value can be specified as a percentage of the length of the curve.

Deconvolution. This is a method to deblur (sharpen) one- and two-dimensional arrays. The function sharpens and enhances the contrast of peaks in the densitometric curves. While the peaks become sharper, noise also increases. Deconvolution actually does the opposite of least-square filtering.

Background subtraction. Two-dimensional background subtraction as explained in step 1 is a computing-intensive operation. If the purpose is to obtain densitometric curves without background, it is more time-efficient and sensible to perform a one-dimensional background subtraction on the densitometric curves. The principle is the same as the rolling ball method, but a disk is used instead of a ball (Fig. 6.10). The size of the disk is inversely proportional to the amount of background subtracted.

Step 3. Normalization

In a fingerprint type experiment, the position of bands on a gel or peaks on a densitometric profile is the only information available for comparing

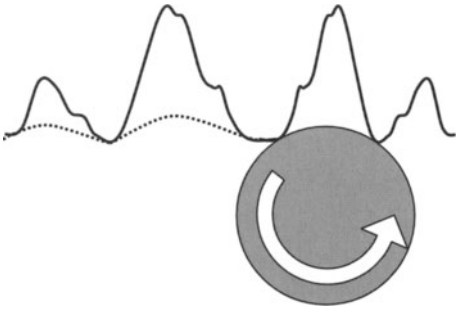


Fig. 6.10. Rolling disk background subtraction applied to densitometric curves

different patterns. These positions are influenced by all kinds of differences and fluctuations in experimental conditions, such as voltage, temperature, concentration of buffers, quality of reagents, running time, etc. It is therefore usually not possible to compare banding patterns from different gels without a prior normalization step. Even within the same gel, shifts may occur between bands on different lanes.

Normalization is usually achieved by running dedicated reference patterns between the data patterns. On a gel, this is done by loading the same reference sample at given intervals (Fig. 6.11). In gas chromatography, HPLC, or spectrophotometric analysis, the same result can be obtained by running a reference sample each time after a fixed number of data samples. On automated sequencers, however, it is possible to run a reference sample inside each lane, using a different color dye.

The principle of normalization is to define and save a set of *reference positions*. These positions correspond to the bands on the reference pattern. The reference positions can be derived from a physical reference pattern on a good-looking gel, but can also be defined by the user, for example by taking the average positions calculated from many reference patterns. A reference position is defined by two characteristics: its running distance on the normalized gel (e.g. as a percent distance from the top of the gel) and its *metric*, which is the physical property of the band, e.g. the molecular weight, length in base pairs, or isoelectric point. The set of reference positions with their characteristics together form the basis for normalization, which we call the *reference system*.

Once a reference system is defined for an electrophoresis system, all bands on the reference patterns of that gel are associated to the corresponding reference positions (Fig. 6.11, upper part). The association of reference bands can happen automatically, using a pattern recognition algorithm. After visual inspection and manual correction (if necessary) of the associations, the gel is aligned to the reference positions (Fig. 6.11,

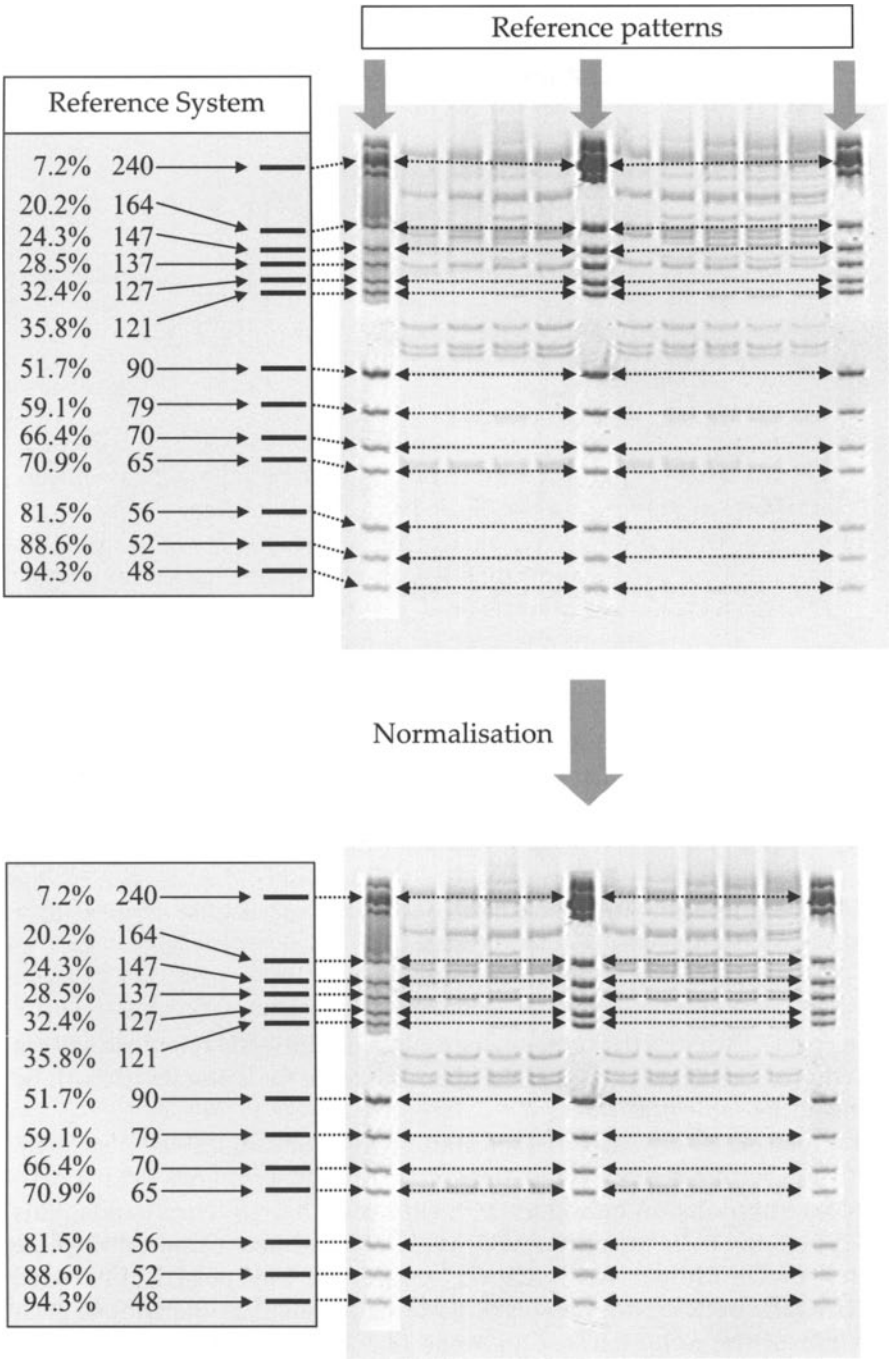


Fig. 6.11. The use of reference patterns and a globally defined reference system to normalize a gel. The reference system consists of reference positions, each characterized by a percentage run length (*left value*) and, optionally, a metric (*right value*). In the example, the metric is defined by the length of the fragments in number of bases

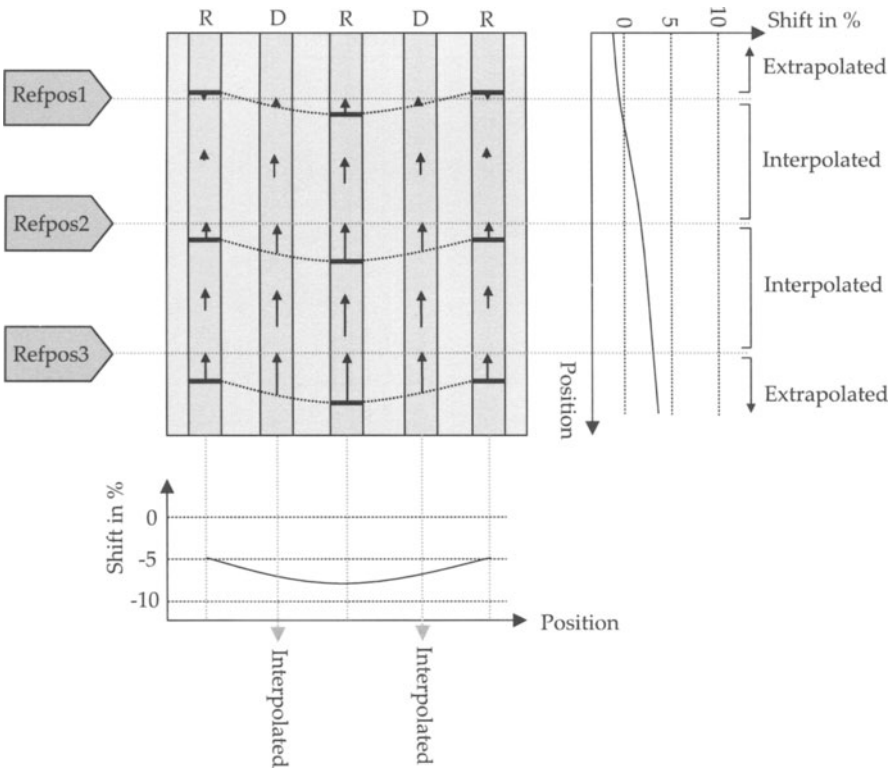


Fig. 6.12. Scheme for normalization of a gel with three reference patterns (*R*) and two non-reference (data) patterns (*D*). Horizontally (*between lanes*) and vertically (*within each lane*), a cubic spline interpolation is calculated through the reference points; and the shift in each position of a data lane is calculated from the combination of both regressions. Shifts beyond the outermost reference points are extrapolated by linearly extending the slope defined by the last two reference points

lower part). Two gels that are normalized using the same reference system are compatible with each other, i. e. bands having the same metric will be found at the same position.

Normalization should be looked at in a two-dimensional way (Fig. 6.12). Vertically, within a reference pattern, the shift for each position is calculated by interpolation between the positions of the reference bands. This interpolation can be linear (Pot et al. 1989) or non-linear using cubic spline regression (BioNumerics, see Fig. 6.12). A second interpolation is needed horizontally between the reference patterns to calculate the shift in each position of the non-reference patterns that fall between the references. Likewise, this interpolation can be done linearly (Vauterin and Vauterin 1992) or by cubic-spline regression (BioNumerics; shown in Fig. 6.12).

After the two-way interpolation, each array of densitometric values has a corresponding array of transposition vectors. Based upon the transposition vectors, a new, corrected densitometric curve is calculated by interpolation between the original densitometric values. The same transposition vectors are used to normalize the two-dimensional TIFF images of the lanes.

Step 4. Detection of Bands/Peaks

Strictly seen, this step is not mandatory, as one can analyze electrophoresis patterns by means of a correlation coefficient that compares the densitometric curves rather than the band positions. In a number of genotyping methods, however, the comparison of band positions will lead to more accurate and meaningful results than the comparison of densitometric curves (see further).

In the case of a slab gel, the bands can be detected directly on the gel image. However, as calculations on two-dimensional images are often very involving, the bands are usually detected by searching for peaks on the corresponding densitometric curves. This actually causes no loss of information and might even be more accurate if the curves are defined using the right parameters and filter settings (see step 3). Detection of peaks on densitometric curves can happen in an automatic way, using a peak detection algorithm. As many small peaks are usually artifacts caused by incomplete digestion, false amplification, unspecific hybridization, etc., there should be a filter that sets a threshold below which peaks are not considered as valid signals. This is probably the most critical and subjective step in the preprocesing of electrophoresis fingerprints. An automated peak search action on a gel requires careful inspection by the user. Within a gel, the user will be able to edit the automated peak assignments in a fairly consistent way. Over different gels however, the user will often change his/her own intuitive threshold according to the gel observed, which inevitably leads to systematic differences in band assignment behavior.

Several mechanisms exist to detect peaks and shoulders in densitometric curves. We will describe a combination of tools that have proven to be satisfactory for most types of electrophoretic genotyping methods.

Peak searching and filtering. A peak search algorithm first locates every peak on the densitometric curve. A Gaussian fit is then done through each peak to determine the height and the area (Fig. 6.13). Optionally a deconvolution can be applied prior to the peak search (e. g. see *Deconvolution* in Step 2 in Sect. 6.4.2) to deconvolute shoulders and peak doublets into separate peaks (see Fig. 6.13). Note that the peak height derived from its Gaussian fit does not necessary correspond to the height of the peak on the densitometric

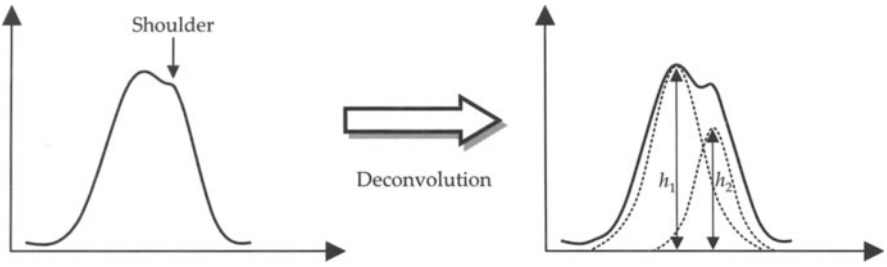


Fig. 6.13. Deconvolution and decomposition of peaks and shoulders into Gaussian curves

curve. The real height of two adjacent peaks may be less than the observed height, due to partial overlap (see h_2 in Fig. 6.13).

From the Gaussian peaks, the height of a peak and its surface can easily be calculated; and one of these peak parameters, or both, can be used as a threshold for peak detection. In practice, the peak height as the only threshold seems to work best for most systems. As the overall intensity of the profile can be quite different in different lanes, it is useful to relate the height or the area to a percentage of the maximum height or overall area of the profile.

Peak intensity regression. A problem often encountered in the electrophoresis of DNA fragments is that the intensity of bands is a function of their size. Commonly used staining compounds, such as ethidium bromide and Acridine orange, for example, the molecules of which intercalate into the DNA strands, are more effective at staining large molecules than small molecules. The result is that there is a gradual decrease of peak height from the top of the patterns (high molecular weight) to the end of the patterns (low molecular weight). Both area- and height-based peak detection filters, which assume a constant peak intensity over the patterns, will provide unsatisfactory results, either detecting noise as peaks in the high molecular weight area, or skipping relevant peaks in the low molecular weight area. This intensity issue can be compensated for by performing a regression on a (large) number of peaks from previously processed patterns (Fig. 6.14). From the regression curve, a position-dependent correction factor can be inferred, which is applied to the peak intensities before passing through the intensity filter. A condition for this approach is that a sufficient number of patterns has been processed with careful manual editing prior to establishing the peak intensity regression.

Uncertain bands or peaks. A way to deal with the uncertainty whether peaks around the threshold are relevant or not is to flag such peaks as "uncertain". Rather than setting a single threshold (e. g. minimum height), below which peaks are ignored and above which peaks are used, two threshold values

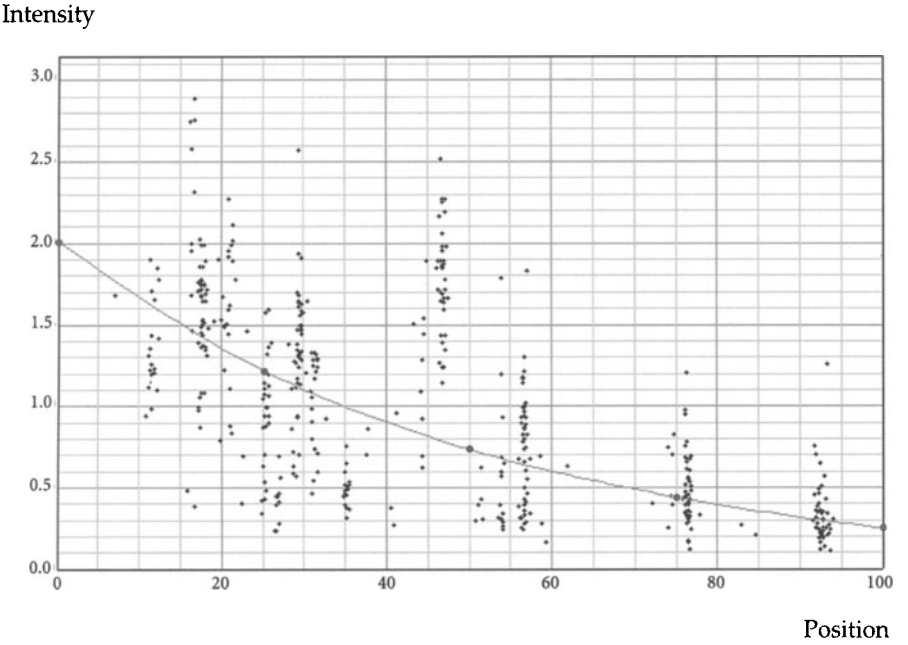


Fig. 6.14. Plot of peak intensities based on a large number of ethidium bromide-stained electrophoresis patterns

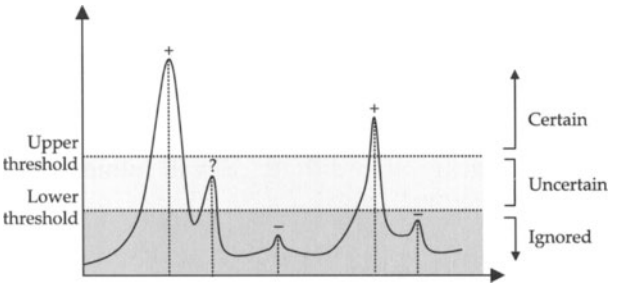


Fig. 6.15. Defining "certain" and "uncertain" bands using upper and lower thresholds

are set (Fig. 6.15), between which peaks are marked as uncertain. The advantages of working with uncertain bands for band scoring analysis is explained further in Sect. 6.4.3.

6.4.3 Comparison of Fingerprint Data

Usually, a comparison of fingerprint data relies on a similarity or distance value between pairs of fingerprints. Since fingerprint data can be treated

as densitometric curves on the one hand, or as sets of peak positions with metric information on the other hand, there are two quite different approaches for comparative analysis of the data. In the first approach, the global similarity between the densitometric curves is determined using a correlation coefficient. In the second approach, the similarity or distance is calculated between the banding patterns based upon the number of matching and non-matching bands. The strong and weak points of both approaches are discussed below, as well as some guidelines as to which approach to follow for specific techniques and purposes.

Comparison of Densitometric Curves

The densitometric curves of fingerprints are arrays of n values, between which a correlation value can be calculated. Usually, the Pearson (1926) product-moment correlation r is calculated:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}} \tag{6.7}$$

This coefficient essentially measures the *goodness of fit* between two arrays of values based upon a linear regression (see Fig. 6.16).

Interestingly, the correlation coefficient is insensitive to differences in intensity and differences in background. If the linear regression is given by the formula $y = ax + b$, the intensity differences in the two profiles will influence the factor a, whereas background differences will influence the

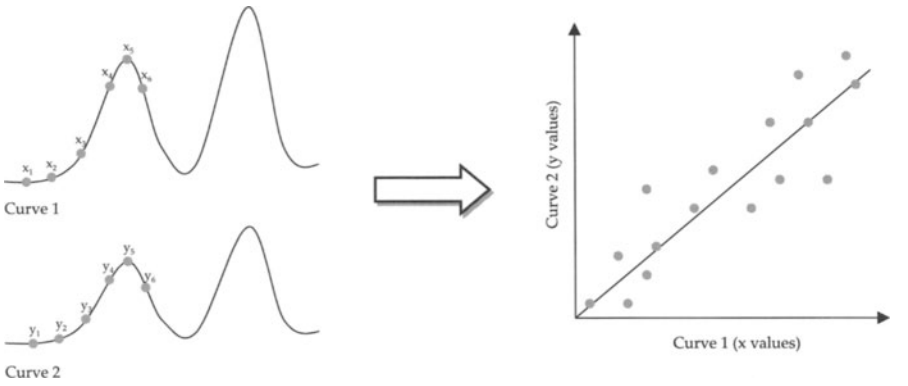


Fig. 6.16. Graphical representation of the mechanism of the Pearson product-moment correlation: the fit of all the dots to a linear regression in a x, y scatterplot determines the correlation value

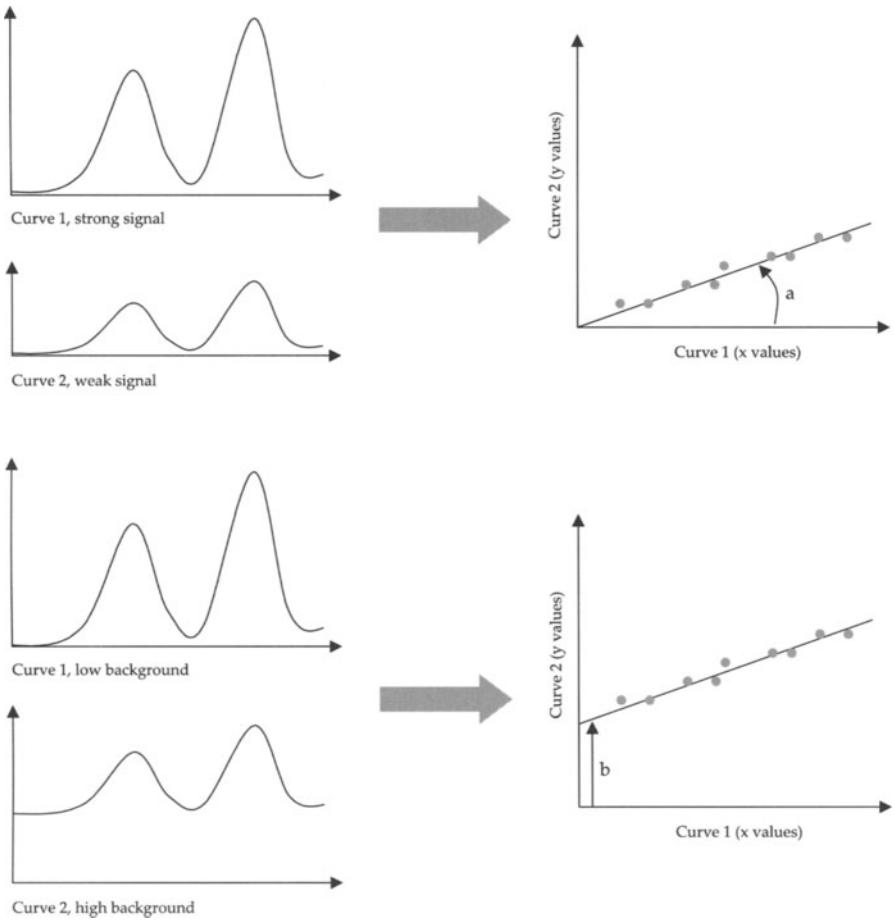


Fig. 6.17. Insensitivity of the correlation coefficient to differences in intensity and background

offset b. Figure 6.17 shows two examples comparing profiles having different intensities and backgrounds, respectively. The resulting regression clearly illustrates that the correlation is not influenced by these differences. The above features only count for linear differences between profiles, i. e. if one profile is a function of the other, in the form $y_i = ax_i + b$. It should be mentioned that the correlation coefficient is strongly sensitive to local differences in background and local differences in intensity. The latter feature is an important characteristic of the correlation coefficient: it penalizes differences in intensities of individual bands. Therefore, the Pearson product-moment correlation is a suitable similarity measure if differences in band intensities are a relevant datum in the investigation.

The Pearson correlation ranges between +1 and −1. A correlation value of zero indicates that there is not the least correlation between the compared patterns. Correlation values below zero indicate that there is an "anti-match" between the patterns, i. e. peaks on one pattern correspond to dips on the other. Since this has no biological meaning, correlation values are in practice often cut-off at zero.

Another correlation coefficient, which is very related to Pearson product-moment correlation, is the *cosine* coefficient:

$$r_c = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{6.8}$$

This coefficient also measures the fit of an (x, y) scatterplot to a linear regression, which passes, unlike the Pearson correlation, through the origin of the plot. This implies that the cosine correlation is influenced by overall differences in the offset (background) between the curves. If a background subtraction is performed on the curves, the cosine correlation coefficient is an interesting alternative to the Pearson correlation.

Comparison of Band/Peak Positions

This approach involves more steps than the comparison of densitometric curves. A first requirement is that all bands (peaks) are defined on the densitometric curves. As we will discuss below, this is often the most tedious and subjective step in the comparison of banding patterns. The comparison between a pair of patterns is then a two-step mechanism:

1. Matching is performed between the bands of the two profiles.
2. The similarity or distance is calculated between the profiles, based upon the number of matching and/or non-matching bands.

The band matching step, shown in Fig. 6.18, relies on an important parameter, i. e. the maximum distance (d_{max}) allowed between two bands in order to be considered matching. We will call this parameter the *position tolerance*. Only if two bands with positions P_A and P_B, respectively, are within a distance that is equal to or less than d_{max}, i. e. $P_A - P_B \leq d_{max}$, can they be matched.

Note that, under this criterion, two bands on one pattern can be eligible for matching with the same band on the other pattern (Fig. 6.19). The solution that comes out depends on the algorithm used. The *closest band matching* algorithm will always match the two bands that have the shortest distance, whereas the *first band matching* algorithm will always match

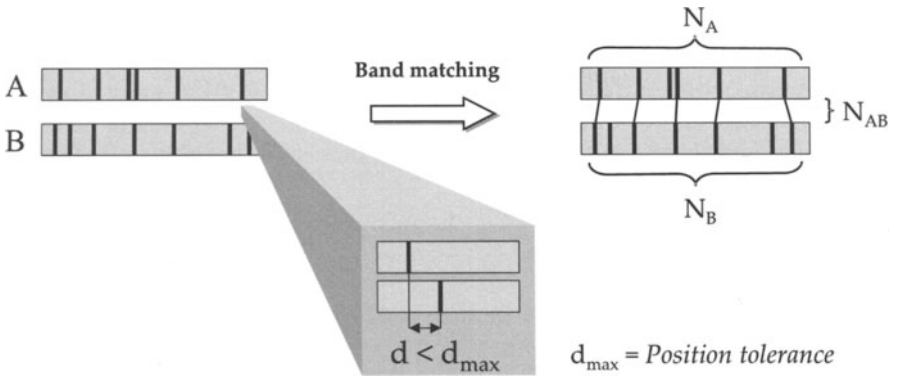


Fig. 6.18. Pairwise band matching in the comparison of banding patterns

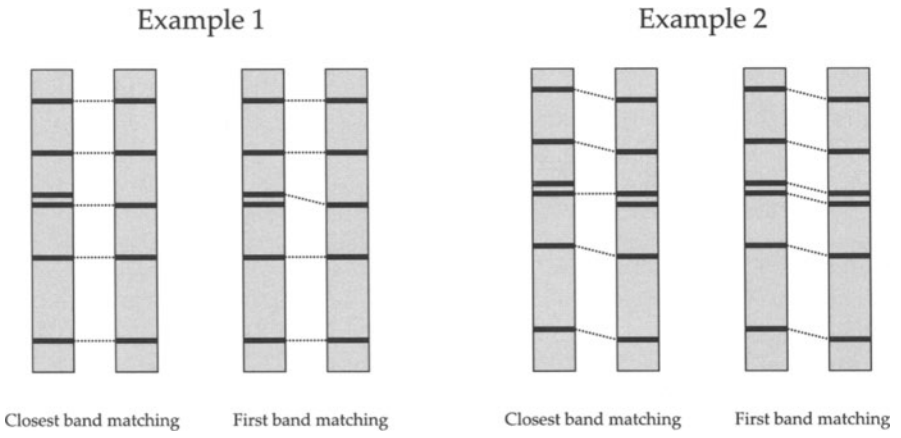


Fig. 6.19. Example of the result of two pairwise band matching algorithms: the closest band matching and the first band matching. Differences can be observed if band doublets occur

the first candidates encountered during a progressive scan from band 1 to bands N_A and N_B (Dawyndt 2004). As illustrated in Fig. 6.19, the closest band matching method may give a more correct representation of a match in the case of a band doublet (example 1). The similarity however, is not influenced. In contrast, the first band matching algorithm may provide a more correct similarity in the case of band doublets that are not perfectly aligned (example 2). The first band matching algorithm is used in the BioNumerics software. However, the software performs a correction for instances where the first band matched is not the closest (Fig. 6.19, example 1). In the BioNumerics software, a so-called *fuzzy logic* variant of the position tolerance can also be chosen. Under this option, the program lets the scoring value of two bands gradually decrease with the distance between the bands.

Association coefficients. Based upon pairwise band matching, the resemblance can be calculated using an appropriate coefficient. The most commonly used band matching association coefficients are the Jaccard coefficient (Jaccard 1908) and the Dice coefficient (Dice 1945). The Jaccard coefficient divides the number of characters present in both samples by the total number of characters:

$$s_J = \frac{N_{\text{Common}}}{N_{\text{Total}}} \tag{6.9}$$

In the case of banding patterns, the total number of characters should be interpreted as the total number of different bands, in other words two matching bands are considered as the same band. Therefore, using the notation in Fig. 6.18, the Jaccard coefficient can also be written as:

$$s_J = \frac{N_{AB}}{N_A + N_B - N_{AB}} \tag{6.10}$$

The coefficient of Dice is very similar to the Jaccard coefficient, putting more weight on common bands:

$$s_D = \frac{2N_{\text{Common}}}{N_{\text{Total}} + N_{\text{Common}}} \tag{6.11}$$

Using the notation in Fig. 6.18, the coefficient can also be written as:

$$s_D = \frac{2N_{AB}}{N_A + N_B} \tag{6.12}$$

In this formulation, the coefficient is the same as the estimator that Nei and Li (1979) proposed for measuring the genetic distance between restriction endonuclease patterns. Note that the Dice coefficient can be rewritten as a simple function of the Jaccard coefficient:

$$S_D = \frac{2S_J}{S_J - 1} \tag{6.13}$$

While the branch lengths of dendrograms obtained using both coefficients are different, the topologies of the trees are always the same.

Two other coefficients that are sometimes used for measuring the similarity between banding patterns are the Jeffreys X coefficient (Jeffreys and Pena 1993) and the coefficient of Ochiai (1957):

$$\text{Jeffreys X} \quad S_X = \frac{1}{2}\left(\frac{N_{AB}}{N_A} + \frac{N_{AB}}{N_B}\right) \tag{6.14}$$

$$\text{Ochiai} \quad S_O = \frac{N_{AB}}{\sqrt{N_A N_B}} \tag{6.15}$$

As opposed to the Jaccard and Dice coefficients, these two coefficients have the interesting feature that they are sensitive to the proportion of different bands in both patterns: the similarity is higher when the non-matching bands occur on one pattern than when they are equally spread over both patterns.

Distance coefficients. There is only one distance measure which makes sense and is commonly used in the frame of pairwise band matching: the number of different bands. This can simply be the total number of non-matched bands in both patterns ($N_{\text{Unmatched}}$), or a distance scaled between zero and +1, by dividing $N_{\text{Unmatched}}$ by the total number of band instances:

$$D = \frac{N_{\text{Unmatched}}}{N_{\text{Total}}} \tag{6.16}$$

Using the notation in Fig. 6.18, the same coefficient can be written as:

$$D = \frac{N_A + N_B - 2N_{AB}}{N_A + N_B - N_{AB}} \tag{6.17}$$

Dealing with Uncertain Bands

As explained in Sect. 6.4.2 (Step 4), flagging bands as uncertain can make the difficult step of marking bands a little less subjective and critical. Following the reasoning that the presence of an uncertain band is left undetermined, uncertain bands are never included in the similarity calculation, whether the other pattern contains a matching band or not. As illustrated in Fig. 6.20, two possible cases exist:

1. A certain band on a pattern matches with an uncertain band on another. This band is left out from the comparison so that it does not influence the final similarity.
2. An uncertain band on one pattern has no corresponding band on the other pattern. The band is equally left out from the comparison, so that there is no mismatch.

Optimization of Pattern Alignment

Although the position tolerance can solve most of the problems associated with non-perfect matching between patterns or individual bands, there are instances where an additional optimization step may be useful. To that end, one pattern is shifted pixel by pixel in both directions with respect to the other. For each single pixel shift, the matching is calculated between the two patterns; and the highest similarity value thus obtained is used. The use of an optimization window offers the advantage that the position tolerance

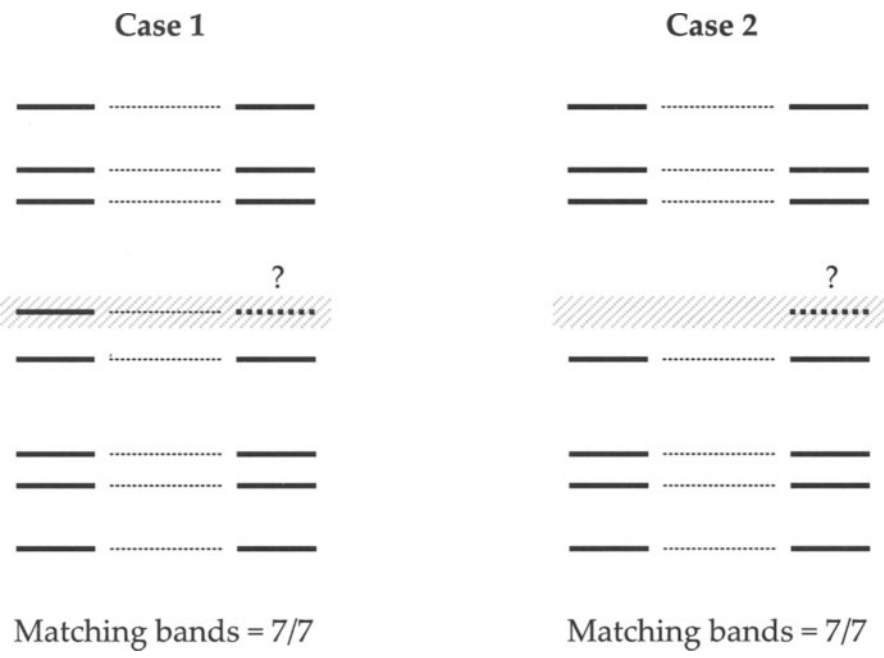


Fig. 6.20. Comparison of patterns with uncertain bands (flagged with *question marks*)

window can be kept smaller, yet obtaining optimal alignments between patterns. Needless to mention that the smaller the position tolerance window can be set, the more false matchings can be avoided.

Choosing the Most Appropriate Coefficient

The most critical decision is whether to use a curve-based or a band-based association coefficient. Both approaches have their advantages and disadvantages, so that it will often depend on the specific needs and priorities of the researcher which approach is the most suitable. The Pearson correlation calculated on densitometric curves is often suitable as a first quick analysis tool to explore new data sets. It is a very robust coefficient, offering a number of advantages over band-based coefficients:

1. Directly applicable to densitometric curves, avoiding a tedious and often subjective, error-prone band detection step
2. Largely insensitive to differences in pattern intensity and background
3. no need to identify position tolerance, again avoiding a source of subjectivity and error

Another characteristic of the Pearson coefficient, is its sensitivity to the intensity differences of individual bands. This may be an advantage or a disadvantage, depending on the type of analysis. One important consequence is that the Pearson correlation never shows perfect matches, even between visually identical patterns. For example, if macro-restriction fragments separated by pulsed-field electrophoresis are compared, one knows that in theory each fragment should be present in the same molarity. In many applications, for example epidemiological typing, one does not want to see such artifacts reflected in the similarity. In contrast, two patterns that are visually the same will score 100% using a band-based association coefficient if all bands are properly defined and if the position tolerance is well chosen. Therefore, a pairwise binary band matching is much preferred for epidemiological research (Tenover et al. 1995).

Figure 6.21 shows a comparison between the Pearson correlation and the Dice similarity obtained from two patterns that contain the same sample. Visually, the patterns look identical, and after careful band assignment and using proper tolerance settings, the Dice coefficient is 1.00 which indicates that the patterns are identical. However, the densitometric curves are not perfectly identical due to accumulated error during the different steps of the experiment. Although a correlation of 94% is very high, it cannot be easily interpreted as an indication of identity.

The most critical step in band-based comparison of patterns is the assignment of bands. Therefore, one should let the choice of a band-based approach depend on the feasibility to define bands consistently over different gels and with reasonable user input. Some researchers prefer to use band-based comparison because it provides a better control on the results obtained. Indeed, if two patterns should cluster together, it is possible to add and/or delete bands until they match 100%. While such manipulations can be correct in cases of obvious misassignment of bands, we should make the user aware of the fuzzy limits between what one actually sees and what one likes to see. Therefore, as a general guideline, we can state that only electrophoresis types that provide sharply defined and well separated peaks of equal intensity are suitable for analysis by pairwise band matching. Techniques that yield many overlapping peaks and/or peaks of different intensities are more suitable for the analysis of densitometric curves.

Translating this to DNA fragment pattern analysis, one can safely state that techniques in which restriction endonuclease cleavage is the last step before electrophoretic separation are suitable for analysis using band matching. This guarantees that fragments are present in equimolar amounts (supposing that cleavage is complete), which is an essential requirement for consistent band assignment. Examples of such techniques that are frequently applied for molecular subtyping are restriction fragment-length polymorphism (RFLP; van Embden et al. 1993), pulsed-field gel elec-

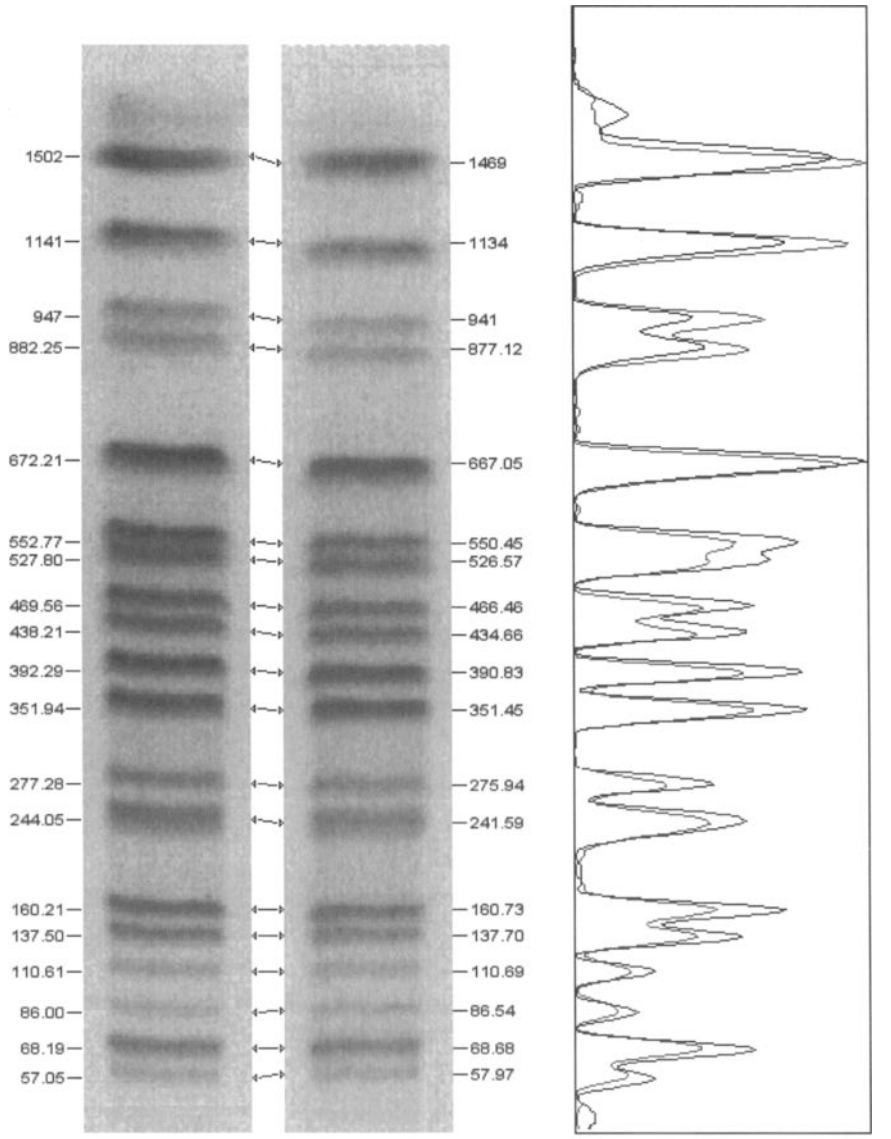


Fig. 6.21. Comparison between Dice similarity on band matching data and Pearson correlation on densitometric curves for PFGE macro-restriction fragments

trophoresis (PFGE) of macro-restriction fragments (Tenover et al. 1995), amplified rDNA restriction analysis (ARDRA; Vaneechoutte and Heyndrickx 2001), and ribotyping (Grimont and Grimont 1986).

A variety of other techniques rely on PCR amplification as a final step before electrophoretic separation. These techniques are less suitable for band matching analysis, since concurrent PCR amplification of a large number of fragments usually yields bands of different intensity. However, the uniformity of the amplification relies on the stringency of the PCR conditions, and therefore, some techniques are better suited than others. In decreasing order of PCR stringency (and hence suitability for band matching), we can mention AFLP (Janssen et al. 1996), Rep-PCR (de Bruijn 1992), and RAPD (Williams et al. 1990).

While in AFLP the PCR amplification happens at high annealing temperatures, irregular amplification of fragments is still observed and can hamper the ability to reliably assign bands over multiple gels. Moreover, AFLP often yields large numbers of fragments per DNA sample, which makes visual inspection of assigned bands very laborious, and leads to many overlapping peaks in the profile. As the number of fragments can be fine-tuned by careful choice of restriction enzymes and adaptor extension bases, AFLP protocols exist that produce less complex and more reproducible patterns (Fry et al. 2002) which can be successfully analyzed by pairwise band matching.

In Rep-PCR, the annealing temperature is lower (between 40 °C and 53 °C, depending on the primer; Rademaker and de Bruijn 1997), which reduces the reproducibility as compared to AFLP. Rep-PCR patterns are therefore usually compared by calculating correlation between the densitometric curves (Rademaker and de Bruijn 1997).

With annealing temperatures as low as 36 °C, RAPD fingerprinting is the least reproducible of the aforementioned fingerprinting techniques. Since fragments are amplified with all ranges of efficiency, the resulting patterns are very complex to interpret and automatic or manual band assignment is impossible. RAPD patterns can therefore only be compared by robust curve matching using the Pearson correlation (Grundmann et al. 1997).

Some other electrophoresis techniques, such as DGGE (Muyzer et al. 1993), TGGE (Muyzer and Smalla 1998), and TRFLP (Moeseneder et al. 1999), are used for assessing the microbial diversity in complex populations and ecosystems, based upon 16S rDNA. As these techniques also produce bands of different intensity, they will probably provide more satisfactory results when analyzed using densitometric curves.

6.4.4
Fingerprint Techniques That Require Special Analysis Methods

Variable Number Tandem Repeats

The availability of whole-genome sequences has made it possible to find a large number of regions on the genome that are useful as markers for typing purposes. A number of typing techniques are based on the occurrence of VNTRs, i. e. short tandem repeats exhibiting variation in length among individuals or strains. VNTRs have been widely observed in prokaryotic and eukaryotic genomes (van Belkum et al. 1998). Variations in the number of repeats result in different alleles for a given locus. The target loci are amplified using specific upstream and downstream PCR primers; and the resulting PCR products are analyzed by electrophoresis. The number of repeats in a VNTR locus can be deduced from the size of the PCR product (in base pairs). When multiple loci are analyzed, the allelic patterns can be discriminatory at the clonal or individual level. The technique known as *microsatellite analysis* is commonly used for mapping, linkage analysis, and to trace inheritance patterns, for example in forensic identification of humans (for a review, see Goldstein and Schlötterer 1999).

More recently, VNTR-based methods have been exploited for molecular subtyping of epidemic bacteria (van Belkum et al. 1997). The combined analysis of multiple VNTR loci is sometimes referred to as multi-locus VNTR analysis (MLVA; Keim et al. 2000). A schematic overview of the method of multi-locus VNTR analysis of bacterial strains is given in Fig. 6.22. Each VNTR locus is amplified using a specific PCR primer set, of which the forward and reverse primer are upstream and downstream, respectively, of the target VNTR locus. The amplification products are sep-

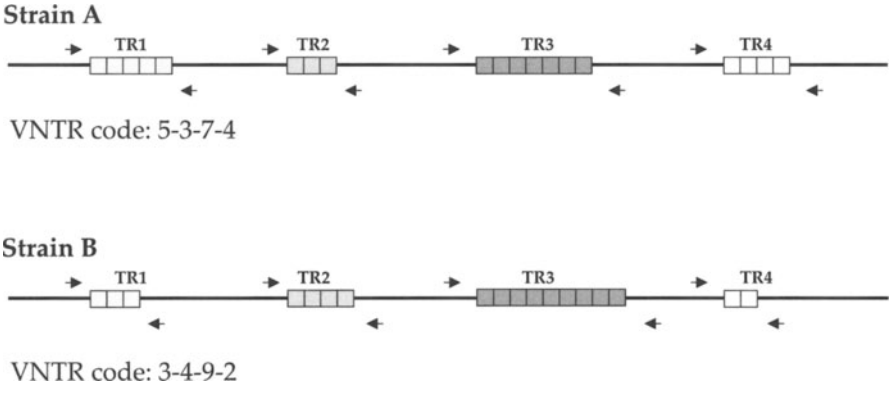


Fig. 6.22. MLVA analysis of four VNTR targets using specific outbound PCR primers

arated electrophoretically, usually on an automated sequencer that offers the possibility to load four or more patterns in the same lane using different color dyes. The length of the fragments can be calculated (in base pairs) from the electrophoretic mobilities. Of course, due to the fact that the primers were chosen upstream and downstream from the VNTR locus, the two offsets should be subtracted to obtain the length of the VNTR allele. By dividing this length by the repeat unit length, the number of repeat units in the VNTR allele is obtained. In the case when four VNTR loci are analyzed (Fig. 6.22), a VNTR profile of four numbers is generated for each strain analyzed, which can be considered as a character set of four characters.

Due to the simplicity of the data, cluster analysis of VNTR data usually provides many equivalent solutions for the same problem, i. e. one data set can be clustered into many trees with different topologies but equally valid according to the criterion used (the *degeneracy* of a tree, see further). It is therefore necessary to reduce the number of possible trees to those that have the most probable evolutionary interpretation. The priority rules applied for MLST analysis (see Sect. 6.5.6) can also be used for VNTR analysis. However, the origin of VNTR alleles is more complex than can be exlained with the recombination model used for MLST. Variations in the number of repeat units per locus may result from polymerase inadequacy such as slipped-strand mispairing as well as from recombination processes (van Belkum et al. 1998). The relative contribution of each ot these processes is difficult to estimate and may depend on the organisms studied. It is, however, important to understand the way VNTR types originate, in order to use the data correctly for population genetics. When slipped-strand mispairing accounts for most of the allelic variation, one can assume that the distance between strains increases with the difference in the number of repeat units in a given locus. In other words, two strains respectively having 17 and 20 repeat units in a locus have further evolved from each other than two strains respectively having 17 and 18 repeat units. The data are not categorical, so a distance or association coefficient should be applied. Conversely, if recombination accounts for most of the allelic variation, it is probably more correct to treat the data as categorical, the same way MLST data are treated (see Sect. 6.5.6). For the analysis of VNTR data, see also Sect. 6.10.3.

6.5 Sequence Type Data

6.5.1 Definition

Sequence type data is the easiest data type to circumscribe. It includes DNA (RNA) sequences and protein sequences. Most of the complexity of sequence analysis lies in the alignment: from the construction of consensus sequences from sequencer trace files to clustering, phylogeny, and fast database screening, alignment is the key to succesful sequence analysis.

6.5.2 Assembling Sequencer Trace Files into Consensus Sequences

Automated sequencers typically generate readings of 400 – 800 bases in a single trace. In many cases, the target sequences under study are longer, so that two or more overlapping regions need to be sequenced. Moreover, to obtain a higher certainty at the consensus level, short sequences are usually sequenced on the two complementary strands.

The following steps are usually involved in assembling sequences:

1. Read four-channel chromatogram files from automated sequencer.
2. Perform base-calling (usually performed by the sequencer software).
3. Assign a quality score to each base, based upon information derived from the chromatograms (Ewing and Green 1998; Ewing et al. 1998).
4. Trim-off bad ends of the sequence, using the base quality scores and the percentage of unresolved positions on the sequence traces.
5. Mark internal regions of insufficient resolution as inactive, i. e. shown in the alignment but not contributing to the consensus.
6. Optionally, remove vector sequence from sequence traces.
7. Perform multiple alignment on the trace sequences to obtain consensus sequence.
8. Display problem positions on consensus and allow for automated or manual problem correction.

Multiple alignment is discussed further in this section. For the calculation of consensus sequences, however, special parameter settings are required, as one can assume that mismatches as well as gaps are rare, actually only caused by sequencing errors.

6.5.3 Alignment of Sequences

If sequences from a specific target gene are compared, even when obtained using conserved primers and from closely related organisms, they are likely be out of frame. The mutational event that is responsible is called a deletion or insertion. Deletions/insertions can range from single bases to large segments. It is therefore necessary, before sequences can be compared, to align them to each other. Alignment is very important in all aspects in sequence analysis: cluster analysis, phylogeny, functional analysis, motif search, gene identification, and database screening. Since sequence databases are often very large, a lot of research has been done on this topic, which has resulted in a number of widely used, fast alignment algorithms.

The goal of an alignment algorithm is to reconstruct the deletions and insertions that have happened on the sequences compared. In practice, this is obtained by searching for stretches of high homology on both sequences and creating gaps in either sequence so that all the homologous stretches match each other (Fig. 6.23). The way alignment algorithms work is to optimize a score function by introducing gaps, whereby each matching position on the two sequences is assigned one score unit, which may depend on the type of match. However, if an algorithm is allowed to introduce gaps without any restriction, the score based upon matching residues will be maximal, but the resulting aligned sequences will be fragmented in an unrealistic way (see Fig. 6.23, alignment 1). Therefore, a penalty is usually assigned to each gap created (the *open gap penalty*). Alignment 2 in Fig. 6.23 shows a score optimization with a gap penalty which is equal to minus the match score (common setting). The resulting alignment looks more realistic, as the two homologous stretches have been aligned by introducing only two gaps. In practical computer implementations, single residues should not be allowed for nucleotide sequence alignments, even with a zero open gap penalty setting.

In addition to assigning a penalty for introducing a gap, it can sometimes be interesting to assign a penalty to each position by which a gap is increased (the *unit gap penalty*). In targets where one expects that insertions and deletions are merely single residue events (e. g. 16S rRNA genes), it might be useful to apply a significant unit gap penalty. In targets where recombination is frequently encountered (e. g. housekeeping genes of epidemic bacteria), however, a single insertion or deletion will usually include a large number of residues, so that setting a unit gap penalty could result in unsatisfactory alignments.

In the case of nucleic acid sequences, the calculation of matching and non-matching bases is usually simple: the same base is matching, a mutation is non-matching. A difference is sometimes made between *transitions*

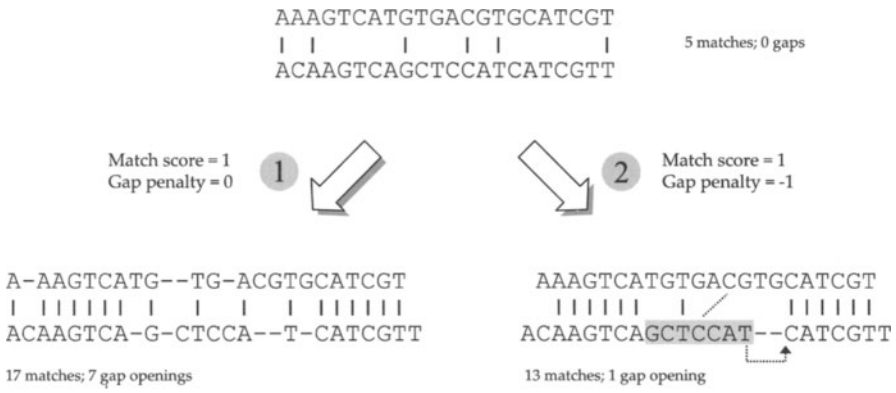


Fig. 6.23. Alignment of sequences: optimizing a score function

[purine to purine (A,G) or pyrimidine to pyrimidine (C,T)] and *transversions* (purine to pyrimidine or vice versa). Transitions are found to occur more frequently than transversions (Wakeley 1996). A similar complication occurs when ambiguous positions are found in a sequence (i. e. where the consensus is not univocal). The IUPAC code for nucleic acid allows such residues to be denoted using a special symbol (Table 6.1).

For example, if a residue on one sequence is "A" and the corresponding residue on the other sequence is "R", there is 50% chance that the position is a match and 50% that there is a transition. A score for this alignment could be calculated as the average of a match and a transition. It might thus be appropriate to use a global scoring table specifying the score factor for each possible mutation event. In the case of amino acids, the need for a scoring matrix is even more obvious, as amino acids can be classified in groups of similar properties (Table 6.2). Within a group, amino acids are easily mutated, and hence should have a much higher score than between groups. Substitution matrices with scores for all possible exchanges between amino

Table 6.1. IUPAC notation for nucleotides

A: adenine	R: A or G (purine)	B: C or G or T, not A
C: cytosine	Y: C or T (pyrimidine)	D: A or G or T, not C
G: guanine	M: A or C (amino)	H: A or C or T, not G
T: thymine	K: G or T (keto)	V: A or C or G, not T
	S: C or G, strong (3 H bonds)	
	W: A or T, weak (2 H bonds)	
N: A or C or G or T (i. e. any nucleotide)		

Table 6.2. Amino acids, with abbreviated names and classification

Group	Name	Short name	Symbol
Hydrophobic	Alanine	ala	A
	Glycine	gly	G
	Isoleucine	ile	I
	Leucine	leu	L
	Valine	val	V
Hydrophilic	Asparagine	asn	N
	Glutamine	gln	Q
	Serine	ser	S
	Threonine	thr	T
Neutral	Cysteine	cys	C
	Methionine	met	M
	Proline	pro	P
Aromatic	Phenylalanine	phe	F
	Tryptophan	trp	W
	Tyrosine	tyr	Y
Acidic	Aspartic acid	asp	D
	Glutamic acid	glu	E
Basic	Histidine	his	H
	Lysine	lys	K
	Arginine	arg	R

acids have been published for different purposes, e.g. probability accepted mutation (PAM) matrices (Dayhoff et al. 1978) and BLOcks SUbstitution Matrices (BLOSUM; Henikoff and Henikoff 1992).

One of the most widely applied sequence alignment algorithms is the method of Needleman and Wunsch (1970). The method is applicable for nucleic acid and protein sequences, as it makes use of a scoring table for each substitution. If two sequences A and B of respectively length m and n are aligned, a matrix of size $m \times n$ is constructed. The matrix is filled with the scores for each position of sequence A with each position of sequence B (Fig. 6.24, step 1). The whole score matrix is then iterated from element (1,1) until element (m, n) as follows: for each element (i, j) of the matrix, the value is incremented with the highest score found in elements $(i-1, 1)$ to $(i-1, j-1)$ and $(1, j-1)$ to $(i-2, j-1)$ (Fig. 6.24, step 2). In the incremental score matrix thus obtained, the path is calculated that has the highest total score, as shown in Fig. 6.24, step 3. Optionally, open gap penalties and unit gap penalties can be subtracted wherever the path skips to another diagonal. The alignment can be derived from the path followed.

A variant of this method, the Smith–Waterman algorithm (Smith and Waterman 1981), does not calculate a global alignment of the two full

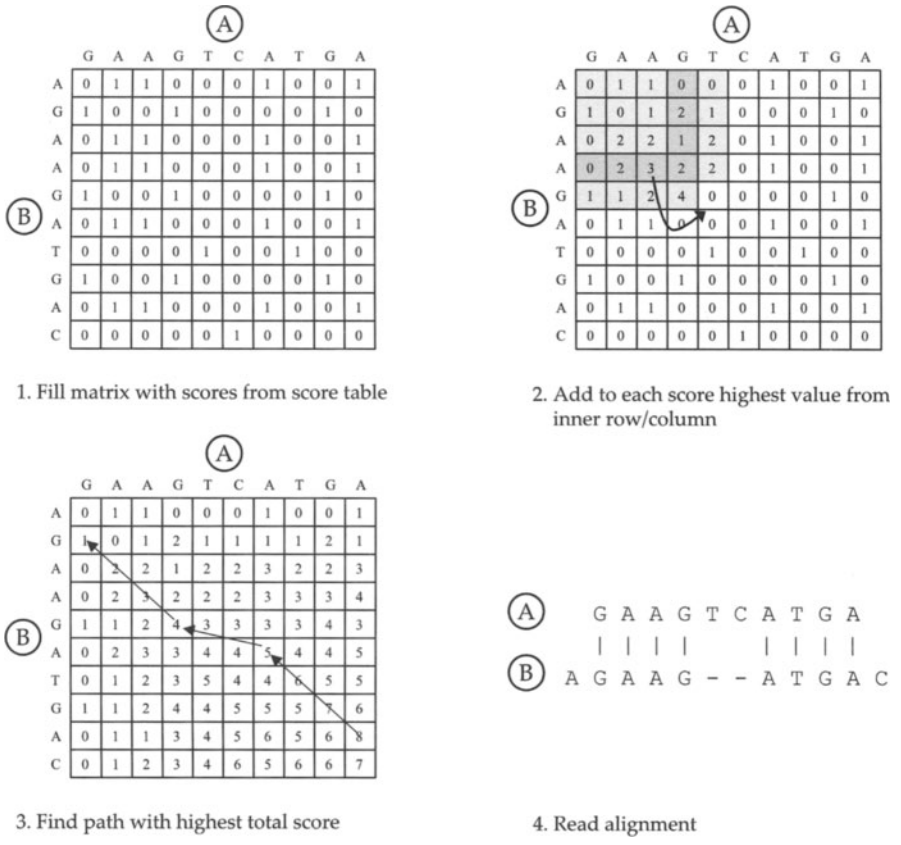


Fig. 6.24. Needleman and Wunsch alignment of sequences

sequences, but shorter, localized paths on the score matrix, corresponding to regions of high homology.

Although the Needleman and Wunsch algorithm is simple and universally applicable, its major drawbacks are the calculation time and the memory needed. To align two sequences of 10,000 bp, for example, a matrix of a 100×10^6 elements is constructed, which can easily lead to computer memory overflow. FASTA (Lipman and Pearson 1985) and BLAST (Altschul et al. 1990) are two important shortcut mechanisms that improve both speed and memory management, and for which numerous variants exist. One of the optimizations used is to create a lookup table of words of residues. For example, if four nucleotides are taken together, 4^4 or 256 combinations are possible, which can still be represented in one byte. Such a lookup table is used as a basis for fast hit-searching (BLAST) and searching for regions of high homology (FASTA). Another commonly applied optimization is to

reduce the number of diagonals of the score matrix. First, the scores of the two sequences over all frame shifts are calculated, without introducing gaps. After sorting the scores, the t frame shifts with the highest scores are withheld. One can thus avoid the construction of a full score matrix and work with t diagonals contributing to the highest score. The choice of the parameter t depends on the purpose of the alignment: for fast database screening, t can be set to 1, which means that only one stretch (with highest homology) is considered (FASTA). For alignment purposes, the choice of t is determined on the basis of the length and the diversity of the sequences to be aligned.

6.5.4 Multiple Alignment

If two sequences are aligned, it is possible to calculate a similarity value from the relative number of matching residues using a score table, and optionally, by penalizing the gaps that were introduced by the alignment. In a study comprising n sequences, for example 16S rRNA gene sequences obtained from n bacterial strains, it is thus possible to construct an $n \times n$ resemblance matrix which can be used as the input for cluster analysis. This workflow is simple and straightforward but has limitations for phylogenetic purposes. A first limitation is that the investigator has no control on the outcome of the $[n \times (n-1)]/2$ alignments. Second, most phylogenetic methods do not rely on a resemblance matrix as input. One exception, though, is the neighbor joining algorithm (Saitou and Nei 1987). This method is discussed further in Sect. 6.10.1.

For phylogenetic study, a *multiple alignment* is therefore usually constructed, in which more than two sequences are aligned to each other so that a table is generated which is comparable to a character data matrix (see Sect. 6.3.2) having the sequences as rows and the base or amino acid positions as columns. A multiple alignment has the advantage that one can derive a lot of information by simple visual inspection of the alignment table: conserved and variable regions, motifs and function prediction of proteins, target positions for primers or probes, etc. It also allows the investigator to inspect the alignments obtained and correct them using evidence from structural characteristics, e.g. three-dimensional conformation of proteins or secondary structure of rRNA molecules. In addition, a multiple alignment is the input for widely used phylogenetic clustering methods such as *maximum parsimony* and *maximum likelihood*. The fact that a multiple alignment is a character table also allows for cluster significance statistics such as the *bootstrap* method. Phylogenetic clustering methods are discussed elsewhere in this book and will not be treated here.

In theory, multiple alignment can be achieved using an algorithm similar to that of Needleman and Wunsch (1970). Rather than starting from a two-dimensional score matrix, one can use an *n*-dimensional matrix for aligning *n* sequences. It is obvious that this approach is extremely inefficient in terms of time–memory space. There exist optimized variants of this method, e. g. the Carillo–Lipman method (Carillo and Lipman 1988), but even these are only applicable to few and short sequences.

The alternative is a heuristic progressive alignment based on a dendrogram of pairwise alignments. This approach is used by the CLUSTAL program (Higgins and Sharp 1988) and most current multiple alignment programs. First, all possible pairs of sequences are aligned in a pairwise manner according to Needleman and Wunsch (1970) or an optimized method as described in Sect. 6.5.3. The pairwise resemblance matrix is then used to construct a tree using either the unweighted pair group method using arithmetic averages (UPGMA) or the neighbor joining method. This tree serves as a guide for a progressive multiple alignment, starting from the tips of the branches. Once two sequences have been aligned, their relative alignment is no longer changed, although gaps may be introduced on both sequences to have them match with other sequences on the tree. The method is illustrated schematically in Fig. 6.25.

6.5.5 Phylogenetic Clustering

The most widely used phylogenetic clustering methods are *maximum parsimony* (Fitch 1971) and *maximum likelihood* (Felsenstein 1981). They are discussed in detail in Chap. 5 of this book and are also mentioned in Sect. 6.10.2.

A general workflow for constructing a phylogenetic tree from sequences is shown in Fig. 6.26.

6.5.6 Multi-locus Sequence Typing

Multi-locus sequence typing, usually denoted as MLST, is a technique whereby a number of well chosen housekeeping genes (*loci*) are sequenced, usually in part (Maiden et al. 1998). The technique relies on the proven concepts of multi-locus enzyme electrophoresis (MLEE; Selander et al. 1990), but alleles are defined directly on the nucleotide sequences rather than deriving them from the electrophoretic mobility of the enzymes. In a typical MLST approach, one does not look at the total sequence similarity between bacterial strains. Instead, each sequence for a given locus is screened for

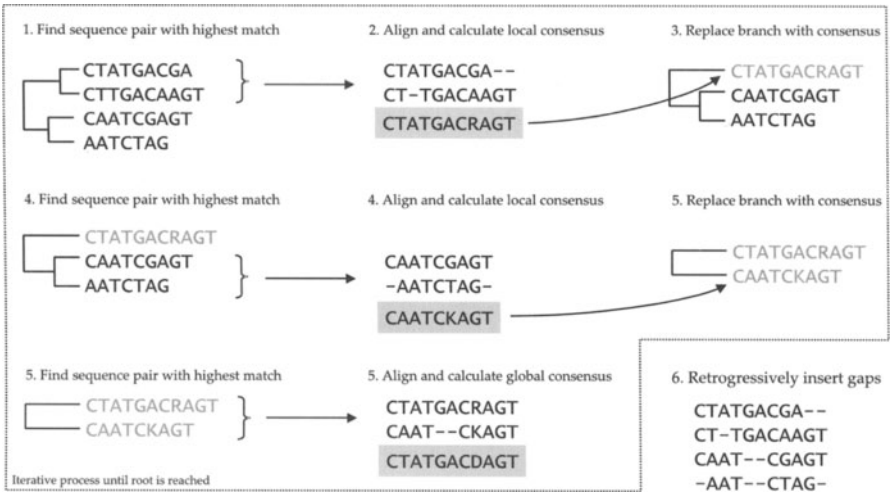


Fig. 6.25. Tree-based progressive multiple sequence alignment

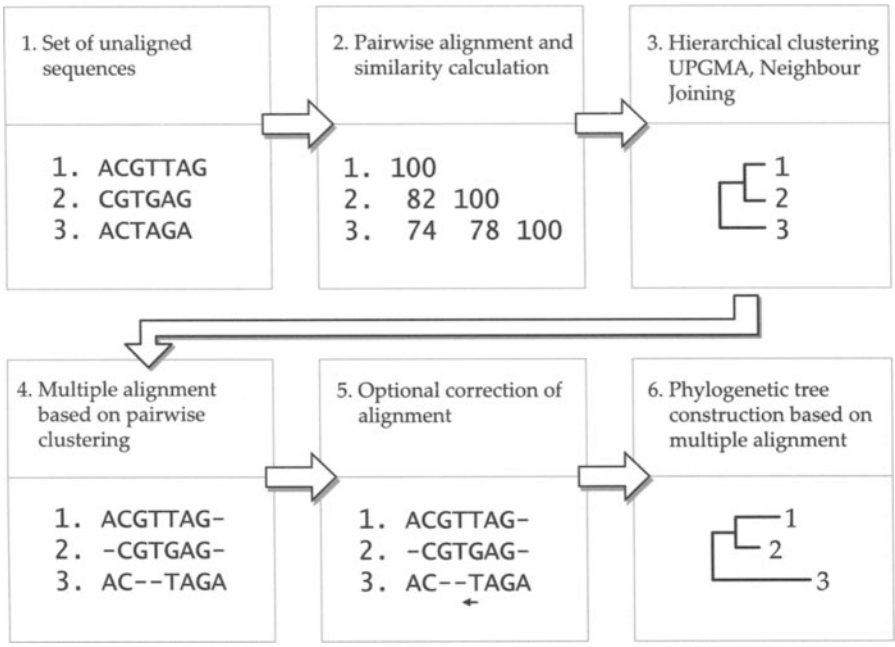


Fig. 6.26. Workflow of phylogenetic clustering using maximum parsimony or maximum likelihood

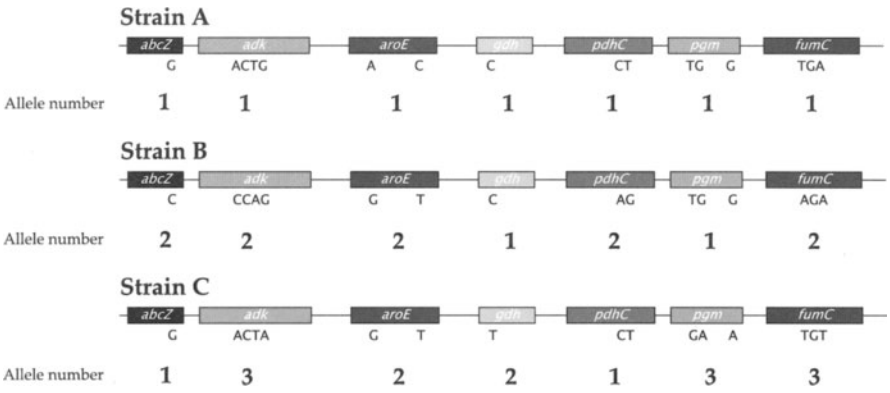


Fig. 6.27. Deriving allelic profiles from partial sequences of housekeeping genes in the MLST approach. As an example, the allelic profile of Strain C is 1–3–2–2–1–3–3

identity with already known sequences for that locus. If the sequence is different, it is considered to be a new allele and is assigned a unique (arbitrary) allele number (see Fig. 6.27). In a case where seven housekeeping genes are studied, each strain is thus characterized by a profile of seven allele numbers. The allelic profiles can be considered as a character set of seven *categorical* characters (see Sect. 6.3.1).

Interestingly, whether an allele of a given locus differs from another allele in 20 bases or just one base is not taken into account: an allele can only be the same or different. The rationale for this approach is that the origin of different alleles is primarily based upon recombination rather than on point mutations (Maiden et al. 1998). In a recombination model, one single gene transfer event can lead to an allele with one or many base differences with the same likeness. Prior to applying this technique, it is therefore necessary to determine the degree of horizontal gene transfer in the bacteria studied; and this should be a multiple of the degree of point mutations.

Analysis of MLST Data

The term MLST is often used for the sequencing of multiple housekeeping genes in general, whereby the analysis is not necessarily based on allele numbering but on the calculation of total sequence similarity. To avoid the confusion between sequence analysis of multiple loci in general and MLST *sensu strictu* as described by Maiden et al. (1998), we suggest the term multilocus sequence analysis (MLSA) for the first activity. MLSA then includes the approach where strains are clustered based upon total sequence identity between all the loci investigated, whereas MLST is reserved for the approach where allelic types are derived.

MLST has been used successfully to study population genetics and reconstruct the micro-evolution of epidemic bacteria, based upon MLST data (for a review, see Feil and Spratt 2001). Since a MLST data matrix is extremely simple (typically seven categorical characters are generated per entry), a clustering algorithm usually provides many equivalent solutions for the same problem, i.e. one data set can be clustered into many trees with different topologies but equally valid according to the criterion used (the *degeneracy* of a tree, see Sect. 6.12). Therefore, a number of *priority rules*, with respect to the linkage of types in a tree, have been proposed (Feil et al. 2003) to reduce the number of possible trees to those that have the most probable evolutionary interpretation. These rules assign priority, in decreasing order, to types that have: (1) the highest number of single locus variants (SLVs) associated, (2) the highest number of double locus variants (DLVs) associated (in the case of equivalent solutions), and (3) the highest number of samples belonging to the type. These priority rules have been implemented in the BURST program available on the MLST website (http://www.mlst.net). In the BioNumerics software, the *most frequent alleles* can also be used as a priority rule. BioNumerics provides a *mimimum spanning tree* implementation to reconstruct the evolution of populations from MLST data (see Sect. 6.10.3).

6.6 Matrix Type Data

Some experiments do not provide a set of characters or a fingerprint per organism or sample studied, but provide the result of a comparison between two organisms or samples studied. This result can be a similarity or a distance value. A typical example is *DNA hybridization* or reassociation. When DNA is heated to denaturation temperatures to form single strands and then cooled, double helices will re-form (renaturation) at regions of sequence complementarity. This technique is widely used for determining the sequence similarity between the DNA genomes of two different organisms, in which case the two DNA samples are mixed and the amount of hybridization after renaturation is measured. It has the advantage over most other DNA genomic techniques that it measures the global degree of homology between the entire genomes. In bacterial taxonomy, it is therefore still regarded as a gold standard for the delineation of species (Stackebrandt et al. 2002).

As stated earlier, techniques like DNA hybridization do not provide a character set or fingerprint for an organism, but a measure of similarity. Consequently, when more than two organisms are studied, the data can only be stored in a resemblance matrix. We therefore introduced the *matrix*

type data, capable of storing similarity values obtained between pairs of samples or organisms. Since the obtained resemblance matrix is similar to resemblance matrices obtained from other data types, the similarity-based cluster analysis techniques described in Sect. 6.10.1 can be applied. In the case of larger DNA hybridization studies, however, incomplete resemblance matrices are usually generated, for the simple reason that conducting $[n \times (n-1)]/2$ hybridization experiments can be a tremendous amount of work when the number of organisms studied (n) is large. Therefore, a modified clustering algorithm should be used which is able to cluster incomplete resemblance matrices.

6.7 Trend Type Data

A single measurement at one point in time is not always sufficient to describe the behavior of an organism. More particularly, reactions to certain substrates or conditions are sometimes recorded in multiple readings as a function of time, as *kinetic* readings. Examples are the kinetic analysis of metabolic and enzymatic activity (e.g. Bochner et al. 2001), real-time PCR (Livak 1995), or time-course experiments using microarrays. Although multiple readings per experiment are mostly done as a function of time, they can also depend on another factor. An example where readings are done as a function of different concentrations is the BioPlex 2200 system (Bio-Rad, Hercules, Calif., USA).

These different data types have in common that they measure a trend of one parameter as a function of another. We therefore call them *trend type* data. Analysis is usually done by fitting a model curve through the measurement points and comparing the characteristics of the curves rather than the original measurement points. Bacterial growth or activity is usually analyzed using a *logistic growth* fit or *Verhulst equation* (after the inventor Pierre-François Verhulst; see Quetelet 1866). A number of parameters can be calculated from the curve fit (Fig. 6.28), including the time to 5% growth increase (T_{05}), 50% growth increase (T_{50}), and 95% growth increase (T_{95}), the maximum slope (S_{max}), the time at maximum slope (TS_{max}), the initial value (MIN), the final value (MAX), the initial exponential growth rate (r), and the initial doubling time (T_{doubl}).

Depending on the data type, other fit models may be used, such as linear, logarithmic, exponential, hyperbolic, Gaussian, Gompertz, power function, etc., each resulting in specific parameters that describe the fit.

Some commercial phenotypic test panel systems allow the kinetic reading of a large number of reactions, e.g. enzymatic or metabolic activities (e.g. OmniLog ID from Biolog, Hayward, Calif., USA, or PhenePlate

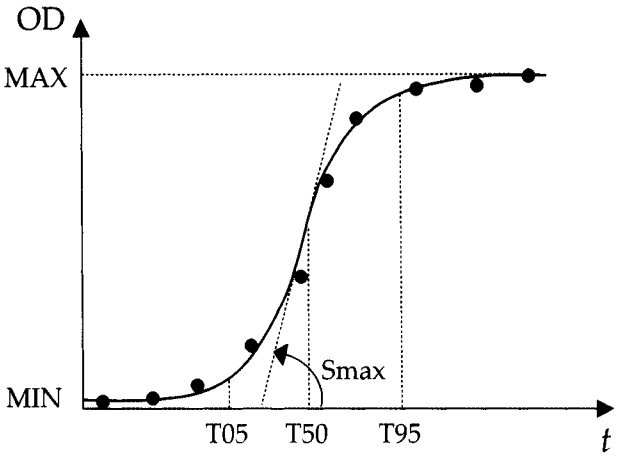


Fig. 6.28. Trend curve that follows the logistic growth model and some derived parameters. *T*05, *T*50, and *T*95 are the times at 5%, 50%, and 95% growth increase, respectively. *Smax* is the maximum slope, *MIN* is the initial value, and *MAX* is the final value

from PhPlate, Stockholm, Sweden). The kinetic reading of enzymatic or metabolic activity is thought to be both more informative and more reliable than measuring the degree of activity at one point in time. The analysis and comparison of curve type data can be done on one or more parameters derived from the curve fit. For example, if one uses S_{max} and MAX, each curve is translated into two character values. Figure 6.29 illustrates in a schematic way how a hypothetical test panel (in the example, containing six tests) is processed into a data matrix. Each test results in five readings (1), through which a curve is fit, using an appropriate model (2). The *logistic growth* model is used in the example. For a given model, one or more characteristic parameters can be derived from the curves. In the example, the *maximum slope* S_{max} and the *final value* MAX are calculated (3). This leads to two data matrices, each containing one value per test and per organism or sample (4).

For taxonomy or typing purposes, one might be interested in combining the data from multiple parameters into one clustering or identification. In the BioNumerics software, it is possible to specify a comparison coefficient for each used parameter separately. The software then averages the respective similarity values into one similarity value per pair of entries compared. An important issue is that the parameters used can have different ranges, as is the case in the example in Fig. 6.29. If a coefficient is chosen that has no inherent *scaling*, e. g. Euclidian distance, an appropriate range should be specified for each parameter, so that the weights of the different parameters are standardized when they are combined by averaging (see Sect. 6.3.2, Standardization).

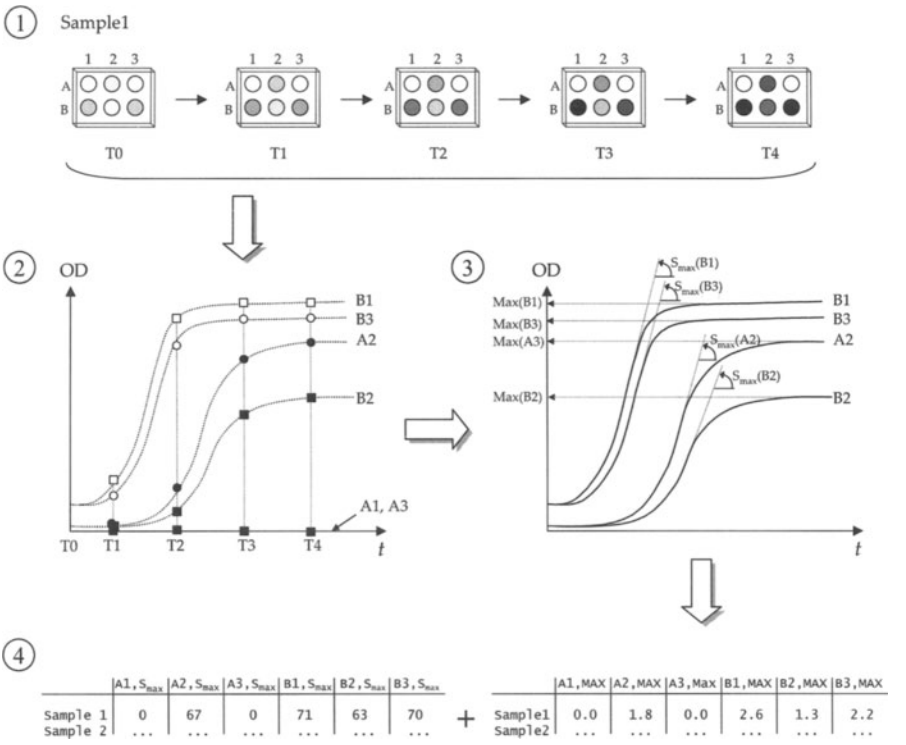


Fig. 6.29. Example of the processing of kinetic readings of a phenotypic test panel. *Step 1* Readings are done at different times *T0–T4*. *Step 2* A curve model is fit through the values obtained for each well in the test panel (in the example, *logistic growth*). *Step 3* One or more specific parameters are derived from the curves [in the example, the final value (*Max*) and the maximum slope (S_{max})] *Step 4* A data matrix is constructed from a curve parameter obtained for each well, including all the samples analyzed. In the example, two data matrices are generated because two parameters were chosen

6.8 Two-dimensional Gel Type Data

Two-dimensional (2D) gel electrophoresis includes all gel electrophoresis techniques in which macromolecules are separated in two dimensions and according to different physico-chemical properties.

The most obvious 2D gel application is the 2D protein gel electrophoresis. This technique separates proteins based on their iso-electric points (pI values) in a so-called first dimension performed in a carrier that contains a pH gradient created using ampholytes, followed by a second dimension in a carrier that separates on molecular weight in a traditional electrophoresis process. Two-dimensional separation, detection, quantifica-

tion, and comparison of proteins is a core technique in modern proteomics research.

Although 2D gel electrophoresis is usually associated with protein separation, it should be mentioned that a few reports exist on the 2D separation of DNA molecules as well. In one dimension, the DNA is separated according to size and conformational differences, which are due to mispairing, insertions or deletions, hairpins, methylations, etc. In the other dimension, the DNA is separated purely on the basis of size (Gunnarsson et al. 2004). This technique can be used for heteroduplex analysis in disease diagnostics.

2D gel electrophoresis shares a number of gel processing algorithms with 1D electrophoresis (Sect. 6.4.2). However, spot detection and normalization are both two-dimensional, which make the algorithms much more complex and slow. Step 1 outlined in Sect. 6.4.2 for fingerprint type images, i. e. image smoothing, background subtraction, and noise filtering, is very similar for 2D gel images. The further main processing steps for 2D gels are illustrated in Fig. 6.30. These include: Step 2 – automatic spot detection, where spots are defined by means of their contours, in order to accurately quantify them, Step 3 – normalization, where all detected spots on data gels are aligned to corresponding spots on a reference gel, Step 4 – calculation of metrics in both directions, i. e. pI in one direction and molecular weight in the other, and Step 5 – querying and comparison between 2D gels in order to screen for proteins that are significantly overexpressed or underexpressed in one gel compared to others.

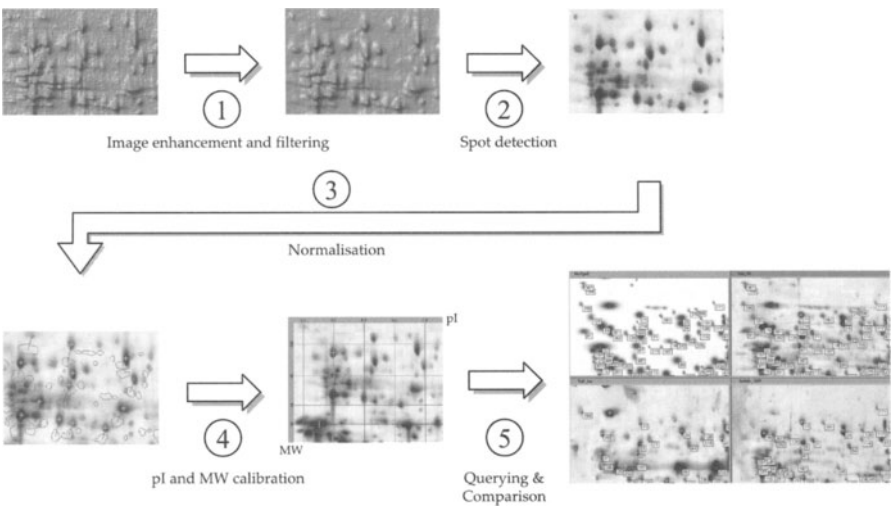


Fig. 6.30. Main steps in the processing of 2D protein gels

6.8.1 Analyzing 2D Gels

One of the main purposes of analyzing 2D gels is to detect proteins that are invariantly expressed or differentially expressed in different circumstances. Another application could be to compare patterns of protein expression between different organisms, in the same circumstances. All these applications require that spots representing the same protein are linked to each other. This is done by normalizing different gels to a common reference system and linking spots of the gel to the homologous reference spots (see above).

During the normalization procedure, two spots from different gels may be linked to the same reference spot (Fig. 6.31). For each protein spot on each gel, a unique identifier is stored. The spots on the reference system also have an identifier. When a spot is linked to a reference spot, it gets the same identifier as that reference spot, so that it is recognized as the same protein. When a spot on another gel is linked to the same reference spot, it also gets the same identifier, so that the spots on both gels are recognized as the same using a simple transitivity rule.

A comparison between a number of 2D gels can be transformed into a character matrix. All known spots are presented as characters of which quantified amounts are filled in for the gels. Thus obtained protein expression matrices are very similar to gene expression matrices obtained from microarray data. Consequently, a number of data mining and exploration tools that have been used for the analysis of microarray data (Amaratunga and Cabrera 2004) are also applicable to 2D protein gel data.

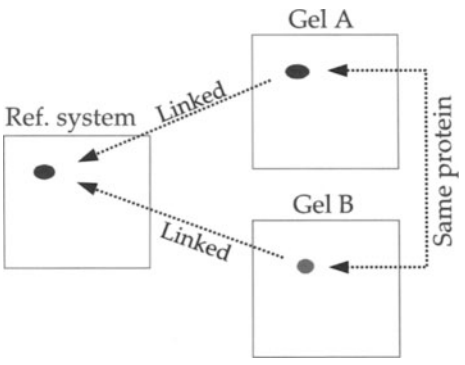


Fig. 6.31. Indirect linking of spots via reference spots: two spots linked to the same reference spot are recognized as the same protein

6.9 The Integrated Database

An aspect of growing importance in typing and taxonomy is data management and databasing. It is obvious that, with growing amounts of data, databasing and data management are becoming crucial issues. Today, the generation of large integrated databases including thousands of strains and data from many different typing techniques is common practice. To successfully store and manage different kinds of data in an integrated database, a carefully designed, expandable database structure is indispensible. Figure 6.32 gives an overview of the database design in the BioNumerics software. Parts of the database relating to the different classes of data are indicated.

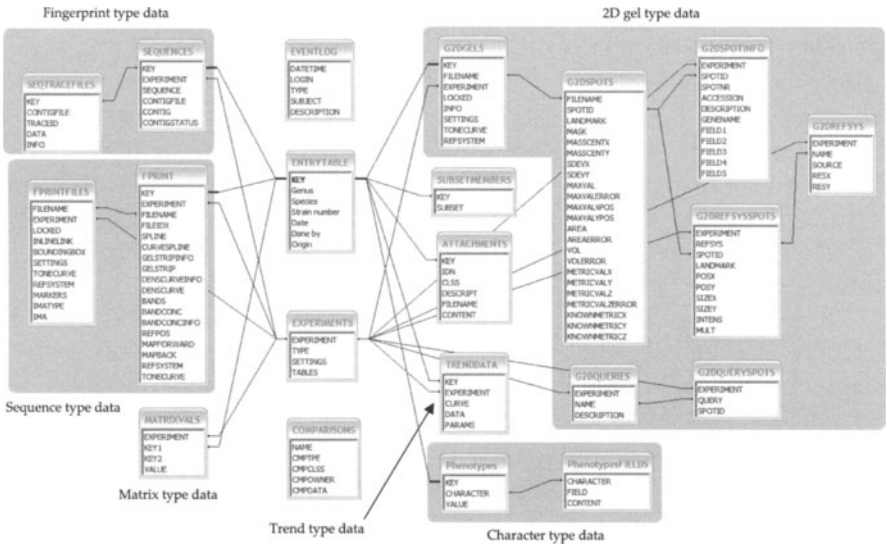


Fig. 6.32. Overview of database design in the BioNumerics software

6.9.1 Distributed Databases and Portability of Data

There are several factors that contribute to the interest of distributed databasing and data exchange. The increased speed and accessibility of the Internet, for example, has made it possible for laboratories to exchange data and to set up server databases containing typing and taxonomic data from many individual sites. The earliest examples of server databases of biological data are public sequence databases such as EMBL and GenBank. The ribosomal database project (RDP; http://rdp.cme.msu.edu) is

a taxonomy-oriented initiative that provides aligned 16S bacterial ribosomal RNA sequences. An exciting recent initiative is the global biodiversity information facility project (GBIF; www.gbif.org). As the name suggests, the mission of this ambitious project is to provide a global platform for information on biodiversity, including all living organisms.

Another factor that has led to an accute need for networked databases of microbial typing data is the mondialization of epidemics. The modern traveling behavior of man has taken down the natural geographical barriers that kept infectious diseases and epidemiological outbreaks local. The problem is particularly threatening with multidrug-resistant strains of disease agents, such as *Staphylococcus aureus* (Lowry 1998), *Mycobacterium tuberculosis* (Cohn et al. 1997), or *Escherichia coli* (Levy et al. 1988), to name just a few. Effective control of such disease agents can only be realized when national and international surveillance networks are available. Examples of some major existing networks are: PulseNet (www.cdc.gov/pulsenet), a United States initiative to perform surveillance on foodborne and enteric pathogenic bacteria, CAonTB (https://hypocrates.rivm.nl/bnwww/index.html), an international concerted action on tuberculosis typing and epidemiology, and MedVetNet (www.medvetnet.org), a European network on the prevention and control of zoonoses, including foodborne diseases. These networks have associated databases containing molecular typing data and epidemiological information on many thousands of pathogenic bacterial strains.

The need for exchanging data between laboratories and establishing international databases automatically confronts us with the critical issue of *comparability* (or compatibility) of the data. Comparability of data is determined at two levels: the level of *reproducibility* of the techniques, and the level of *transformation* of the data. The degree of reproducibility is inherent to a specific technique: although the use of standardized protocols, highly purified chemicals etc. can improve the reproducibility of a given technique, it is clear that some typing techniques are by nature more reproducible than others. Sequencing, for example, is highly reproducible: when different laboratories are to sequence the same 16S rRNA gene independently, exactly the same sequence will normally be returned, regardless of the instruments and protocols used by the respective laboratories. At the other extreme, for example, there is the quantification of chemotaxonomic markers such as cellular fatty acids. The fatty acid profile of a bacterium is strongly dependent on growth conditions, such as temperature, medium, the history, and the age of the colonies; and furthermore, the quantification of the profile is dependent on the extraction procedure followed and the HPLC instrument used. Clearly, reproducibility is a major concern in fatty acid profiling, and consequently, this technique is not very attractive for setting up microbial typing databases on an interlaboratory basis.

Database networking involves three major aspects:

1. Data acquisition and preprocessing. This part usually happens locally, i. e. on a client computer. Critical issues with respect to the data are the degree of *standardization*, the *reproducibility* inherent to a technique, the *portability* of the data, and the connectivity (e. g. transfer speed).
2. Storage and distributed databasing. This part happens centrally, i. e. on a server computer containing the central database. Critical issues here are the database *organization* and the data *structure* (cf. the example of a BioNumerics database in Fig. 6.32), the *size* of the data, and of course, the *security* of the database. The latter issue is particularly important in case of sensitive data, e. g. in clinical environments.
3. Data access, querying, and analysis. This part may happen both locally and centrally. Important factors here are the data *accessibility*, depending again on database structure and organization, but also on connectivity, *querying* possibilities, and *remote* or *local analysis* tools.

The suitability of a technique or data type for distributed databasing can be circumscribed as its *portability*. Based on all these aspects, an evaluation of the portability can be made for the different data types that exist. The portability for the main data types described earlier in this chapter is summarized in Table 6.3. It should be emphasized that there is no relation between portability and resolving power. A technique can be highly portable but offer only a poor taxonomic or epidemiology resolving power, or vice versa.

Note in this respect that fingerprint type data is quoted with a low portability, whereas fingerprint type techniques such as PFGE are used in the majority of epidemiological surveillance networks. One of the main motives to use PFGE for epidemiological typing is that this technique is universally applicable and has a high resolving power in virtually all epidemiologi-

Table 6.3. Comparison of portability between different data types

Data type	Reproducibility	Size of data	Overall portability
Fingerprint	Low, high degree of standardization required	Considerable (gel images, densitograms, peak data)	Low
2D gel	Very low	Tremendous	Very poor
Character	Moderate; higher in automated systems	Compact to considerable (microarrays)	Moderate to high, depending on technique
Sequence	Very high	Efficient	High
MLST	Very high	Extremely compact	Very high
VNTR	High	Extremely compact	High

cally important bacteria. Epidemiological surveillance networks such as PulseNet have achieved such a high degree of standardization that the patterns have become reproducible enough to set up international exchange networks. Moreover, Table 6.3 does not include the cost of the techniques, which is low for most fingerprint type techniques, including PFGE.

6.10 Hierarchical Cluster Analysis

Hierarchical cluster analysis is one of the most popular ways of revealing and visualizing hierarchical structure in complex data sets. The term is a collective noun for a variety of methods that have the common feature of visualizing the hierarchical relatedness between samples by grouping them in a *dendrogram* or *tree*. The tree usually allows the samples to be classified in groups based upon the clusters produced by the method. Apart from this common goal, the approaches and algorithms used, as well as the purposes, are very different. Cluster analysis *sensu latu* has therefore been subdivided into three categories in this chapter:

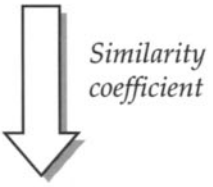
1. Similarity-based hierarchical cluster analysis is carried out on a matrix of similarities between samples. The algorithm calculates bifurcating dendrograms to cluster the samples.
2. Phylogenetic clustering methods are methods which attempt to create trees that optimize a specific phylogenetic criterion. With the exception of neighbor joining, these methods start from the data set directly rather than from a resemblance matrix.
3. Minimum spanning trees are trees calculated from a distance matrix. They possess the property of having a summed branch length that is as small as possible.

6.10.1 Similarity- or Distance-based Clustering Techniques

The most universally applied clustering methods are pairwise clustering algorithms that use a distance or resemblance matrix as input (see Fig. 6.33). The unweighted pair group method using arithmetic averages (UPGMA), complete linkage (furthest neighbor), single linkage (nearest neighbor), Ward's method, and neighbor joining are examples of such methods. Of these, UPGMA is by far the most popular clustering technique, due to its simplicity, intuitivity, and universal applicability. The advantage of

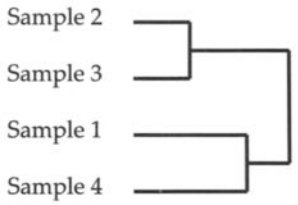
	Char 1	Char 2	Char 3	Char 4
Sample 1	x_{11}	x_{12}	x_{13}	x_{14}
Sample 2	x_{21}	x_{22}	x_{22}	x_{22}
Sample 3	x_{31}	x_{32}	x_{33}	x_{34}
Sample 4	x_{41}	x_{42}	x_{43}	x_{44}

Data matrix



Sample 1	100			
Sample 2	d_{12}	100		
Sample 3	d_{13}	d_{23}	100	
Sample 4	d_{14}	d_{24}	d_{34}	100

Similarity/distance matrix



Dendrogram

Fig. 6.33. Steps in similarity-based cluster analysis

similarity-based clustering methods is that they can be applied to any type of data, as long as there exists a suitable association or distance coefficient that can generate a resemblance matrix from the data. As such, similarity-based clustering can be applied to incomplete data sets or data that are not presented in the form of a data matrix (e.g. electrophoresis band sizes, see Sect. 6.4.3; or matrix type data, see Sect. 6.6).

Within the similarity-based clustering methods, a subdivision should be made on the basis of the algorithm used, i.e. the pairwise clustering methods on the one hand and the neighbor joining method on the other hand.

UPGMA and Related Clustering Algorithms

These methods start from a resemblance matrix of size n (n being the number of samples) and n clusters, each sample being one cluster. The algorithm is a repetitive process of merging clusters and thus reducing the resemblance matrix, until the matrix consists of one single cell, corresponding to the root node of the dendrogram. The workflow of the algorithms is illustrated in Fig. 6.34.

The different methods (UPGMA, complete linkage, single linkage, Ward) differ in the way the similarity is updated after merging two clusters, i.e. how the similarity is calculated between the newly joined cluster and the other existing clusters. In UPGMA, the arithmetic average is calculated from all the individual similarities between the samples of the new cluster on the one hand and the existing cluster on the other hand. In the single linkage variant, the highest similarity value is used; and the method is therefore sometimes referred to as *nearest neighbor* clustering. In complete linkage, the lowest similarity value is used (see Fig. 6.35), which has led to the synonym *furthest neighbor* clustering.

The method of Ward (1963) has a somewhat more complex statistical interpretation. Unlike UPGMA and single/complete linkage which use a criterion of maximal similarity for joining clusters, the Ward method uses a criterion of minimum *incremental sum of squares* (ISQ). For each cluster A with n_A entries in a dendrogram, one can calculate an average array of size m (with j ranging from 1 to m):

$$\bar{x}_{A,j} = \frac{1}{n_A} \sum_{i=1}^{n_A} x_{A,i,j} \tag{6.18}$$

The "*within sum of squares*" value for that cluster (WSQ) is then defined as:

$$\mathrm{WSQ}_A = \sum_{j=1}^{m} \sum_{i=1}^{n_A} \left(x_{A,i,j} - \bar{x}_{A,j}\right)^2 \tag{6.19}$$

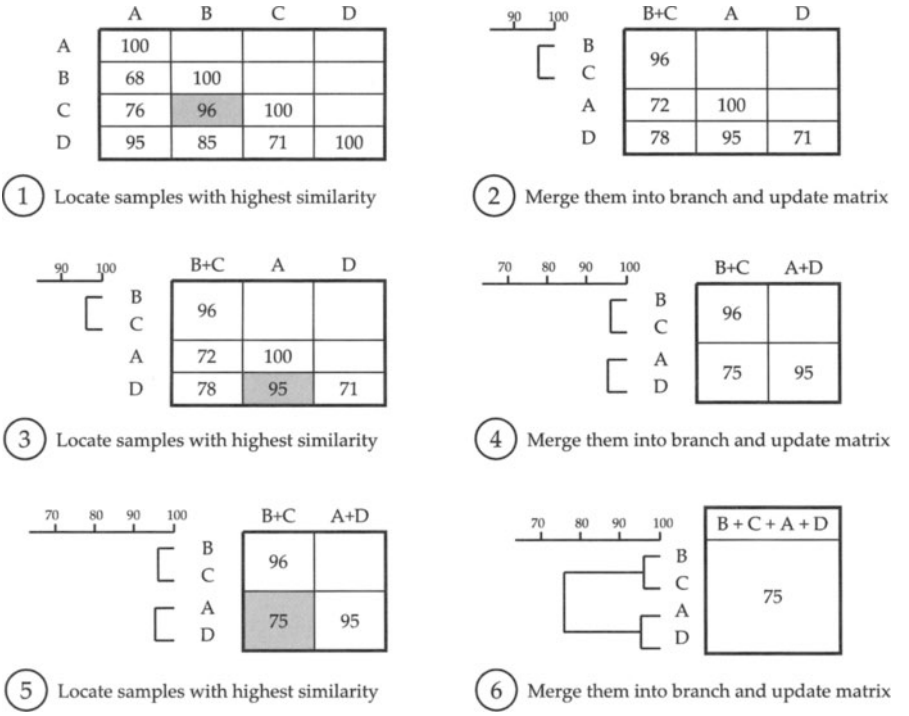


Fig. 6.34. Repetitive process of matrix reduction in similarity-based pairwise clustering algorithms

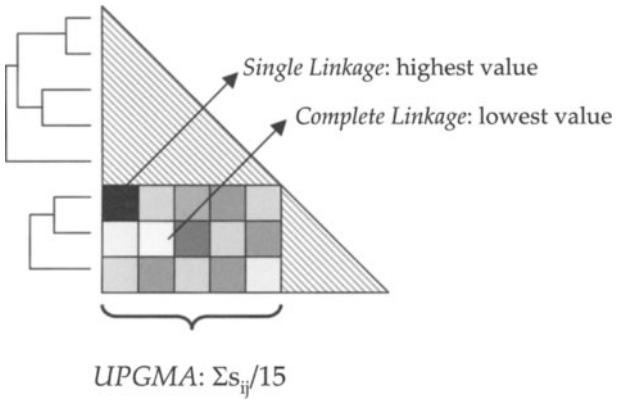


Fig. 6.35. Calculation of similarity between partial clusters in UPGMA, complete linkage, and single linkage. Similarities are represented as *shaded* blocks

The WSQ value is a measure of the variation of the arrays within a cluster. If WSQ_A is the variation of the arrays in cluster A, and WQS_B is the variation of the arrays in cluster B, then Ward's method is to merge clusters where the ISQ is minimal:

$$ISQ_{A,B} = WSQ_{A,B} - WSQ_A - WSQ_B \tag{6.20}$$

It can be shown that when the resemblance matrix is obtained using a correlation coefficient, a matrix of ISQ values can easily be derived from the correlation matrix. Consequently, although Ward's clustering method can be applied to any resemblance matrix, the elegant statistical interpretation is only valid if applied on a correlation matrix obtained from numerical arrays.

Neighbor Joining Technique

The method of neighbor joining (Saitou and Nei 1987) is a phylogenetic clustering method which, unlike parsimony and maximum likelihood, relies not on the data set but on a distance matrix. One of the characteristics of a tree, rooted or unrooted, is that there is exactly one path between any two of its entries (see Fig. 6.36).

The neighbor joining algorithm generates an unrooted tree for which the distance $D_{T,i,j}$ from any entry i to any other entry j approximates as closely as possible the distance $D_{M,i,j}$ between these entries given by the distance matrix, i. e. $\sum_{i,j} \left(D_{T,i,j} - D_{M,i,j}\right)^2$ should be as small as possible.

Compared to a rooted tree with aligned branch tips, such as produced by UPGMA, an unrooted tree has an extra degree of freedom in terms of branch lengths and hence can more faithfully approximate the distance matrix. The most difficult question with the principle of neighbor joining, however, is finding the tree that has the right topology. The method

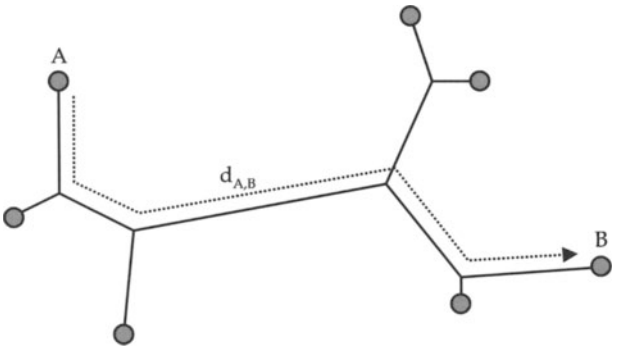


Fig. 6.36. Unrooted tree with the distance between two entries *A* and *B* indicated

uses the tree that has the shortest total branch length. This principle of *minimum evolution trees*, i. e. explaining the total evolution in as few mutational events as possible, is found in other clustering approaches used for phylogeny as well, for example maximum parsimony (Sect. 6.5.5) and minimum spanning trees (Sect. 6.10.3). In theory, one could let an algorithm explore all possible tree topologies for a given set of entries and calculate the total branch length for each case. In practice, however, the number of possible trees grows tremendously as a function of the number of entries, making this approach only possible for very small sets. Therefore a shortcut is used in the neighbor joining algorithm, finding a sub-optimal solution. In practice, the neighbor joining method appears to be a reliable algorithm for phylogenetic clustering that produces trees very similar to parsimony and maximum likelihood, in spite of the very different input.

An unrooted tree as output from the neighbor joining method or other phylogenetic clustering methods, such as parsimony and maximum likelihood, is often difficult to interpret. Therefore, a very distant entry is often added to the set and clustered along with the other entries. The root is then selected from the branch connecting the outgroup with the rest of the tree (see Fig. 6.37). This makes it possible to present the tree in one of the more conventional rooted forms, which are easier to interpret (Fig. 6.37).

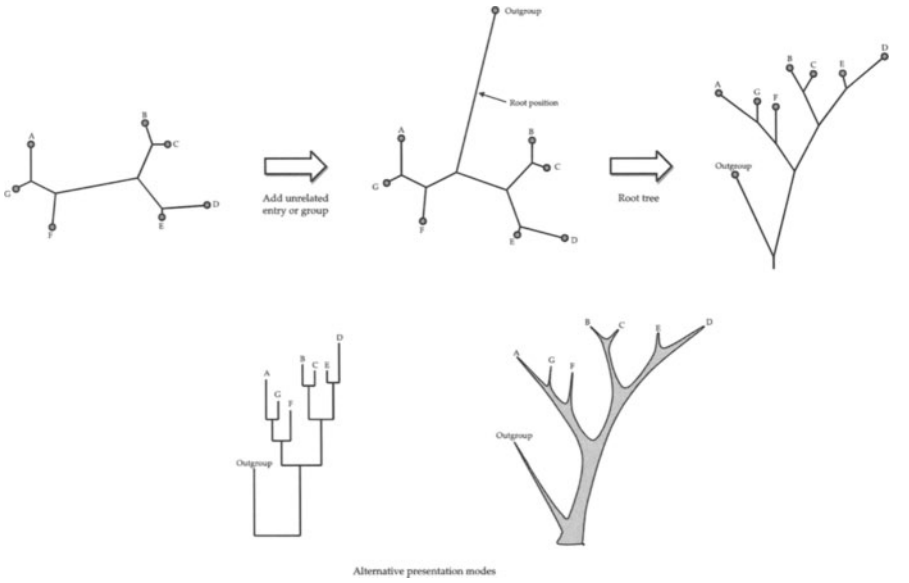


Fig. 6.37. Unrooted tree as produced by a phylogenetic clustering algorithm. An "outgroup" is added to root the tree, resulting in different presentation modes as shown

6.10.2 Phylogenetic Clustering Methods

The most widely used phylogenetic clustering methods are *maximum parsimony* (Fitch 1971) and *maximum likelihood* (Felsenstein 1981). They are also discussed in Chap. 5 of this book. In both methods, an evolution is reconstructed based upon sequence data by optimizing a certain criterion. Parsimony tries to find a tree that explains the sequence diversity with a minimum number of total mutations needed (i. e. the most *parsimonious* tree). The branch lengths of the tree reflect the number of mutations along the branches. The maximum likelihood method is based on a probabilistic model for base substitution. A tree is searched for that has the highest *likelihood*, i. e. the probability that the given sequences are the result of an evolution along that tree, following the assumed probabilistic model. The branch lengths of the tree correspond to evolutionary time. Both methods, but maximum likelihood in particular, have the disadvantage that they are extremely slow. Several heuristic methods exist that allow a sub-optimal tree to be found. These methods are typically applied for the clustering of DNA and protein sequences, but they can in principle be extended to binary or categorical data sets in general. Parsimony is sometimes applied to binary band matching tables of DNA restriction fragment patterns.

Note that the neighbor joining method (Sect. 6.10.1) can also be classified under phylogenetic clustering methods. We have classified it elsewhere because of its property of using a distance matrix as input.

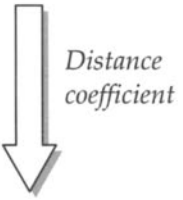
6.10.3 Minimum Spanning Trees

Minimum spanning trees (MSTs) have long been known in the context of mathematical topology. When a set of distances is given between n entries, a minimum spanning tree is a tree that connects all entries in such a way that the summed distance of all branches of the tree is the shortest possible (Fig. 6.38).

In a biological context, the MST principle and the maximum parsimony principle (Sect. 6.10.2) share the idea that evolution should be explained with as few events as possible. There are, however, major differences between parsimony and MST. The parsimony method allows the introduction of hypothetical samples, i. e. samples that are not part of the data set. Such hypothetical samples are created to construct the internal branches of the tree, whereas the real samples from the data set occupy the branch tips. The phylogenetic interpretation of the internal branches is that they are supposed to be common ancestors of current entries, which no longer

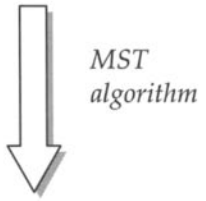
	Char 1	Char 2	Char 3	Char 4
Sample 1	x_{11}	x_{12}	x_{13}	x_{14}
Sample 2	x_{21}	x_{22}	x_{22}	x_{22}
Sample 3	x_{31}	x_{32}	x_{33}	x_{34}
Sample 4	x_{41}	x_{42}	x_{43}	x_{44}

Data matrix



Sample 1	100			
Sample 2	d_{12}	100		
Sample 3	d_{13}	d_{23}	100	
Sample 4	d_{14}	d_{24}	d_{34}	100

Distance matrix



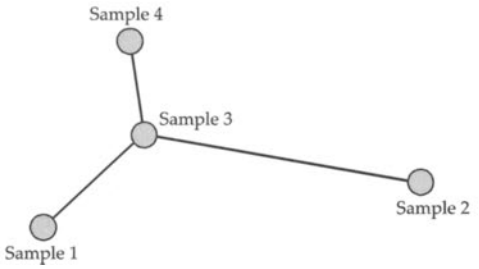


Fig. 6.38. Principle of minimum spanning trees: given a set of entries for which the distances have been calculated, the entries are connected so that the total branch length is as short as possible

exist but are likely to have existed in the past, under the criterion of parsimony.

The MST principle, in contrast, requires that all samples are present in the data set to construct the tree. Internal branches are also based upon existing samples. This means that, when a MST is calculated for evolutionary studies, there are two important conditions that have to be met: (1) the study must focus on a very short time-frame, assuming that all forms or states are still present, and (2) the sampled data set must be complete enough to enable the method to construct a valid tree, i. e. representing the full biodiversity of forms or states as closely as possible. Through these restricting conditions, the method of MST is only applicable for specific purposes, of which population modeling (micro-evolution) is a good example.

The trees resulting from parsimony on the one hand and MST on the other also have a topological difference. The parsimony method assumes that two (related) samples are evolved from one common ancestor through one or more mutations at either side. This normally results in a bifurcating (dichotomic) tree: the ancestor is at the connecting node, and the samples at the tip. A MST chooses the sample with the highest number of related samples as the root node and derives the other samples from this node. This may result in trees with star-like branches (see example in Fig. 6.38) and allows for a correct classification of population systems that have a strong mutational or recombinational rate, where a large number of single locus variants (SLVs) may evolve from one common type (Maynard Smith et al. 1993).

MSTs can only be calculated from a true distance matrix. A criterion for a true distance matrix is that, given three samples A, B, and C, the distance from A to C should never be longer than the summed distance from A to B and B to C. This restriction implies that MSTs are not compatible with all data types. For example, a distance matrix based upon pairwise compared fingerprint type patterns does not fulfill this criterion and hence cannot be used for MST analysis. In contrast, a distance matrix based upon a data matrix (in the case of fingerprint type data, a global band matching table), can be used. In theory, every distance coefficient applied on a data matrix produces a distance matrix suitable for analysis with the MST method. The most typical applications for use with MSTs, however, are categorical multi-locus sequence typing (MLST) data used in population genetics and epidemiological studies (see Sect. 6.5.6).

MSTs and Population Genetics

An implementation of MST for population genetics is found in the BioNumerics software. The MST method usually provides many equivalent solutions for the same problem, i. e. one data set can be clustered in to many

MSTs with different topologies but with the same total distance. Therefore, a number of priority rules, with respect to the linkage of types in a tree, have been adopted from the BURST program (see the MLST website at http://www.mlst.net; Feil et al. 2003) to reduce the number of possible trees to those that have the most probable evolutionary interpretation. These rules assign linkage priority, in decreasing order, to: (1) types that have the highest number of *single locus variants* (SLVs) associated, (2) the highest number of *double locus variants* (DLVs) associated (in case of equivalent solutions), and (3) the highest number of samples belonging to the type. These rules can easily be explained as a function of what happens during the evolution of clonal bacterial populations (see Fig. 6.39). As the parent population grows, a number of SLVs will gradually be formed. The growing SLV populations in turn will produce SLVs which are DLVs to the parent type. As such, a population that has a large number of SLVs is likely to be a parent type. In addition, populations that have a large number of DLVs as well as SLVs are indicative of being an old parent type. Following the same reasoning, it is clear that a population with a high number of entries is a parent type too. Intuitively, this criterion could even be used as a first priority rule. However, the number of entries is subject to sampling bias, which may lead to false assignments of parent types. Especially with hospital-acquired bacteria, this criterion needs to be considered with care, as antibiotic-resistant mutants may be acquired much more frequently than sensitive strains.

In BioNumerics, the *most frequent alleles* can also be used as a priority rule. Thereto, a frequency table is generated for each allele, based upon the number of allelic types where it is found. The product of all allele frequencies is then calculated for each allelic type (Fig. 6.40). Priority is assigned to the type having the highest product of allele frequencies. The biological background for this priority rule is that types having a high overall frequency of alleles are more likely to be ancestor types than types having a low overall frequency of alleles.

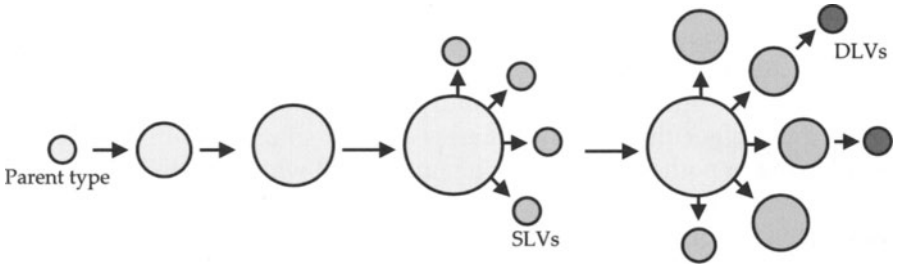


Fig. 6.39. Illustration of how the evolution of clonal bacterial populations is observed in MLST. See text for explanation

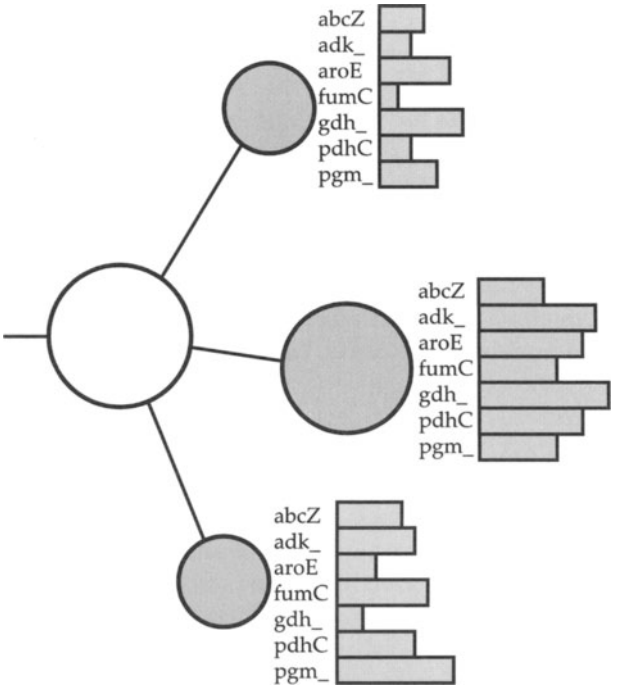


Fig. 6.40. Creation of a frequency table for each allelic type. Priority can be assigned to types having the highest product of frequencies

As discussed in the introduction, a pure minimum spanning tree assumes that all types needed to construct a correct tree are present in the sampled data. Conversely, algorithms like maximum parsimony will introduce hypothetical nodes for every internal branch, while the samples from the data set define the branch tips.

The major problem with the minimum spanning tree algorithm in this light is that it requires a very complete data set to obtain a probably correct tree topology. In reality, a number of existing types may not have been included in the sampled data set. If such missing samples represent central nodes in the "true" MST, their absence may cause the resulting tree to look very different, with a much larger total span.

The MST algorithm in BioNumerics offers a solution to this problem, by allowing hypothetical types to be introduced wherever they can cause the total span of the tree to decrease significantly. In the context of MLST, these usually correspond to missing types for which a number of SLVs are present in the data set. From an evolutionary point of view, it is very likely that such types indeed exist, explaining the existence of SLVs. Simulations with randomly sampled data sets from the large MLST database of

Neisseria ghonorrhae (available at http://www.mlst.net) have indicated that 92% of the introduced types correspond to real types in the entire database (unpublished data).

6.11 Consensus Grouping and Classification

Along with the exploration of novel genomic and phenotypic characterization techniques for typing and taxonomy, there has been a growing awareness among microbiologists that a single technique is usually not a reliable basis for building hierarchical trees and classification systems. Whereas in molecular typing and population gentics, a single technique often provides enough discrimination for the purposes of the investigation, taxonomic study usually cannot rely on a single phenotypic or genomic marker. The term polyphasic taxonomy (Colwell 1970) has been used to denote the evaluation of a variety of phenotypic and genotypic characterization methods to obtain more stable classifications (Vandamme et al. 1996). The combination of information from different characterization techniques, however, complicates the interpretation of the results. In a "monophasic" approach, i. e. where one technique is used to classify organisms, numerical analysis of the data set results in one single dendrogram or spatial distribution of the organisms studied. While the correct interpretation of a single dendrogram is not always easy, the comparison of multiple groupings is even much more complicated and confusing, especially if the suggested classifications are discrepant. Conversely, while polyphasic analysis is generally advanced as the most appropriate approach in classification and identification (Vandamme et al. 1996; Stackebrandt et al. 2002), it is an extremely pragmatic approach, leaving much room for personal interpretation and often burdening the investigator with uncertainty. The subjectivity of the approach, however, could be reduced by applying numerical and statistical algorithms that combine the information provided by different characterization techniques into consensus clusterings. So far, little work has been published in this challenging new area of numerical taxonomy. The spectacular advances in terms of automation, speed and accuracy of DNA sequencing, DNA fragment analysis, microarray analysis, and phenotypic and chemotaxonomic fingerprinting systems will force microbial taxonomists to explore novel algorithms that produce consensus classifications or "super-trees" from a variety of information sources.

When it comes to analyzing combined information from different experimental sources, one should make a distinction between homologous and non-homologous data. Homologous data always belong to the same data type as defined in Sect. 6.2 and further. Usually, homologous data are ob-

tained using variants of the same technique. For example, a PFGE pattern can be obtained from *E. coli* strains using *Xba*I as a fragment-generating restriction enzyme. Another PFGE pattern can be obtained from the same strains using *Avr*II. Essentially, the same technique is used, but with different restriction enzymes as variants. Two different fingerprints are generated per strain, which together are likely to provide more information than one single fingerprint. A similar example is MLSA of different house-keeping genes. Thus obtained homologous data can easily be concatenated to form one bigger data set. The main purpose of concatenating homologous experiments is to refine the taxonomic resolution and obtain more reliable groupings of the organisms studied. This approach is commonly used in epidemiology, population genetics, and phylogeny.

In polyphasic taxonomy, however, data from non-homologous techniques is co-evaluated as well, in order to obtain a global picture of complex taxonomic relationships. Data are sometimes combined from techniques that resolve at different taxonomic levels, e.g. DNA hybridization data and 16S rRNA gene sequences. Data can also be combined that reflect different facets of the organisms studied, e.g. genotypic and phenotypic data. It is clear that such non-homologous data sets cannot be concatenated as can be done with homologous data.

Combining different data sets into one dendrogram can happen at three levels (Fig. 6.41):

1. The two data sets are concatenated into one combined data set, and all further analysis steps are performed on the combined data set.

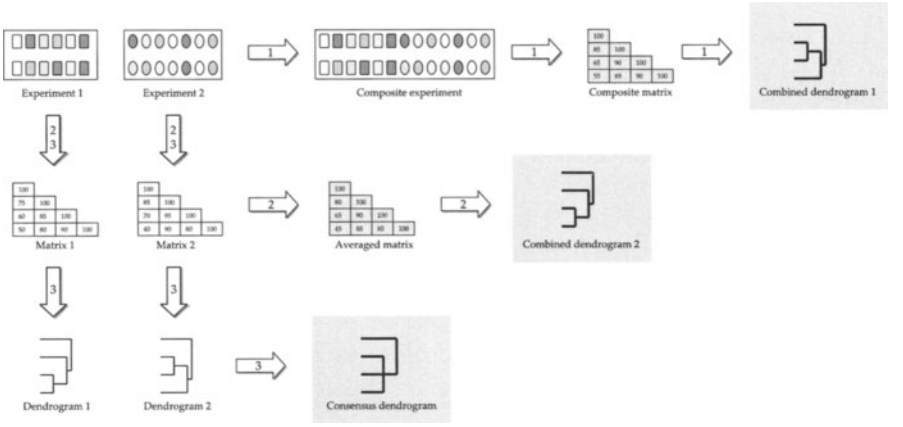


Fig. 6.41. Scheme showing the three main approaches to obtain a combined cluster analysis from different data sets: *approach 1* by concatenating the data sets, *approach 2* by averaging the resemblance matrices, and *approach 3* by merging dendrograms into a consensus representation

2. Resemblance matrices are obtained for each data set individually, which are combined into an averaged matrix, from which a dendrogram is calculated.
3. Dendrograms are calculated for each data set individually. A consensus dendrogram is then calculated from the individual dendrograms.

6.11.1 Concatenation of Data Sets

Each of the aforementioned approaches has its advantages and drawbacks. For suitable data, concatenating data sets is the most intuitive and objective approach. In addition, it allows a number of statistical techniques to be applied to the concatenated data set, for example PCA, bootstrap analysis, MANOVA, etc. However, it has a limited applicability, since concatenation can only be applied to data sets that are homologous, e.g. phenotypic character sets or sequences. In addition, the data sets should have the same range, or at least be scaled to the same range (Sect. 6.3.2). In the BioNumerics software, homologous data sets can be merged into so-called *composite data sets*, for which the investigator has the option to calculate a resemblance matrix using the same coefficients as can be applied to the individual data sets.

6.11.2 Averaging Resemblance Matrices

Clustering an averaged resemblance matrix has the advantage of being universally applicable, as it works for all data sets for which a resemblance matrix can be generated. A problem with this method is that, during the averaging, an important assumption is made with respect to the data sets and their characters. Using unweighted averaging, the implicit assumption is that each *data set* as a whole has an equal importance. If one data set consists of ten characters and the other consists of 100 characters, this assumption is likely to be incorrect. The BioNumerics software provides an option to weight the matrices in proportion to the number of characters contained in the respective data sets. If this option is enabled, the assumption is that each *character* has an equal importance. This is probably the most objective approach when homologous data sets are combined. However, a problem arises when non-homologous data are combined. Suppose that a MLST data set of seven characters is combined with a fatty acid data set containing 35 fatty acid characters in total. Using the option of weighting the matrices in proportion to the number of characters, the fatty

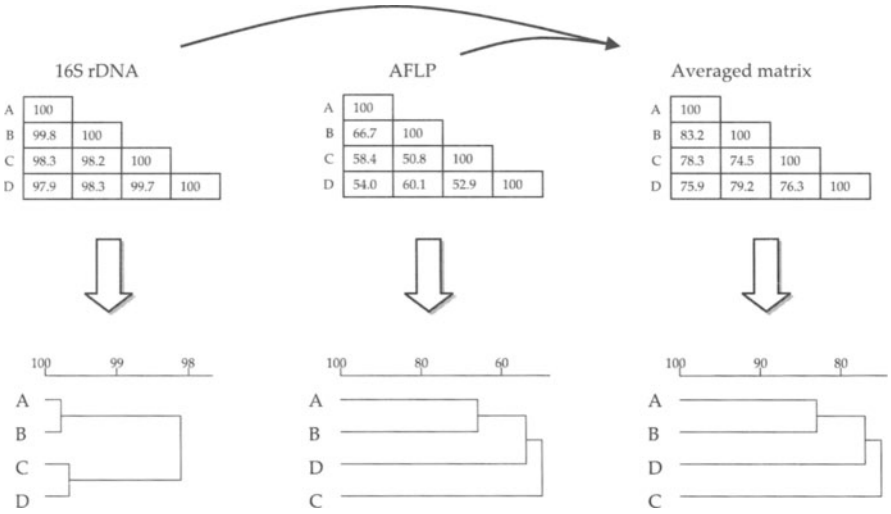


Fig. 6.42. The inadequacy of averaging resemblance matrices from techniques that have a different taxonomic range. The data that exhibit the highest diversity account for most of the topology on the combined dendrogram

acid data will have a 5-fold higher impact on the final dendrogram than the MLST data. It is obvious that weighted averaging does not solve the problems related to combining non-homologous data.

Another problem with averaging resemblance matrices is related to the taxonomic range of the techniques that are being combined. Usually the data set that exhibits the highest diversity will contribute most to the topology of the tree drawn from an averaged resemblance matrix. As the example in Fig. 6.42 illustrates, one technique that reveals deeper taxonomic relationships could exhibit small but very significant differences between entries (e.g. a 16S rRNA gene sequence comparison), whereas another technique that discriminates at a more clonal level (e.g. AFLP) might result in low similarities between all of the same entries. The latter can happen if the distance between the entries studied is beyond the reliable range of the technique. In a tree calculated from an averaged resemblance matrix, the smaller and significant differences from one technique will be completely masked by the bigger, but rather insignificant differences resulting from the other technique.

Harmonization of Distance Matrices

To address the problem of combining matrices resulting from techniques having different taxonomic ranges, the BioNumerics software uses a model-based approach to map distance matrices onto a hypothetical, uniform-

distance scale. It is assumed that all observed distance matrices, obtained by different characterization techniques, are functions of a common, hypothetical uniform-distance matrix. Let us assume that the observed distance for technique k between entries i and j is written as $d_{k,ij}$, where k ranges between 1 and the number of techniques, n. We further assume that the uniform biological distance matrix for the same set of entries is given by D_{ij}, and that, for each experiment k, a monotonically increasing function, f_k, can be applied to those values to give the observed distances:

$$d_{k,ij} \cong f_k\left(D_{ij}\right), \quad \forall i,j,k \tag{6.21}$$

Naturally, this model gives rise to a least-squares formalism:

$$\sum_{k=1}^{n} \sum_{i,j} \left[d_{k,ij} - f_k\left(D_{ij}\right)\right]^2 \tag{6.22}$$

where the summation over i and j runs over all distance values in the matrix. This expression should be minimized by choosing appropriate values D_{ij} and functions f_k. Obviously, we need a further parameterization for the functions f_k in order to be able to solve this problem. This can be achieved by assuming an appropriate model, such as a polynomial of degree p. Solutions can be obtained by applying an iterative, non-linear fit algorithm such as Levenberg–Marquardt.

It is also important to note that (6.22) does not need to be a unique solution: any remapping of the distances D_{ij} by a monotonically increasing function will yield a new solution of equal quality, provided that the functions f_k are adjusted accordingly. This simply reflects the facts that the real

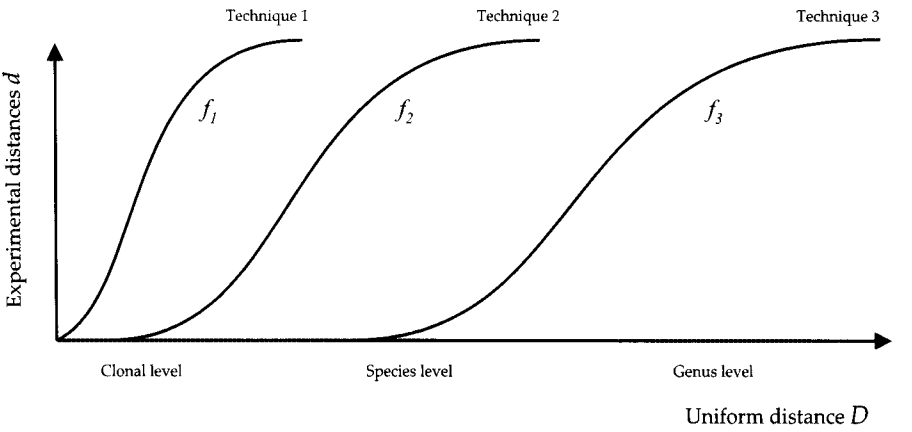


Fig. 6.43. Mapping distances from different techniques onto a theoretical uniform distance using a parameterized model, f_k

values of the uniform distances D_{ij} cannot be known; and the only information that can be learned from the observations is the relative ordering of the distances. Figure 6.43 shows a possible situation in a schematic way. Three characterization techniques are involved, each with its own dynamic range of distances. The obtained distances D span a dynamic range that consists of a union of the three techniques.

6.11.3 Consensus Trees

We will define a consensus tree as a tree that has a common topology between two or more individual trees that exhibit discrepancies. This corresponds to solution 3 in Fig. 6.41. An example of a consensus tree is given in Fig. 6.44. A more truthful representation of the relationships suggested by solution 1 and solution 2 in this example can only be obtained by respecting the indeterminacy resulting from the different branches. Using the conventional pairwise linkage dendrogram representation, this cannot be achieved; and therefore a dendrogram representation should be used that allows more than two entries or branches to be linked together. The resulting tree can be called a consensus tree because it allows all entries that are part of a discrepancy to be linked at one similarity level in a single consensus branch (Fig. 6.44).

Obviously, apart from differences in branching order, different trees may also exhibit differences in similarities at branching levels. The latter problem can easily be solved by averaging the similarities of the corresponding branches in an unweighted or weighted manner in the consensus tree.

6.12 Error on Dendrograms

In the analysis steps outlined in Fig. 6.33, one should consider the matrix of pairwise similarities (or distances) as the complete comparative information between all the entries analyzed. Obviously, for larger numbers of entries, interpreting a resemblance matrix becomes hardly simpler than looking at the original data. This is why a resemblance matrix is not usually calculated as a final result, but as an intermediate step for grouping algorithms such as cluster analysis. Both the power and the weakness of a dendrogram lie in its ability to present an easily interpreted, well structured, hierarchical grouping of the entries. Indeed, simplification means loss of information; and there is no way to present the data in a simple and easily interpretable way while holding all the information. As a consequence, every dendrogram resulting from a non-artificial data set will

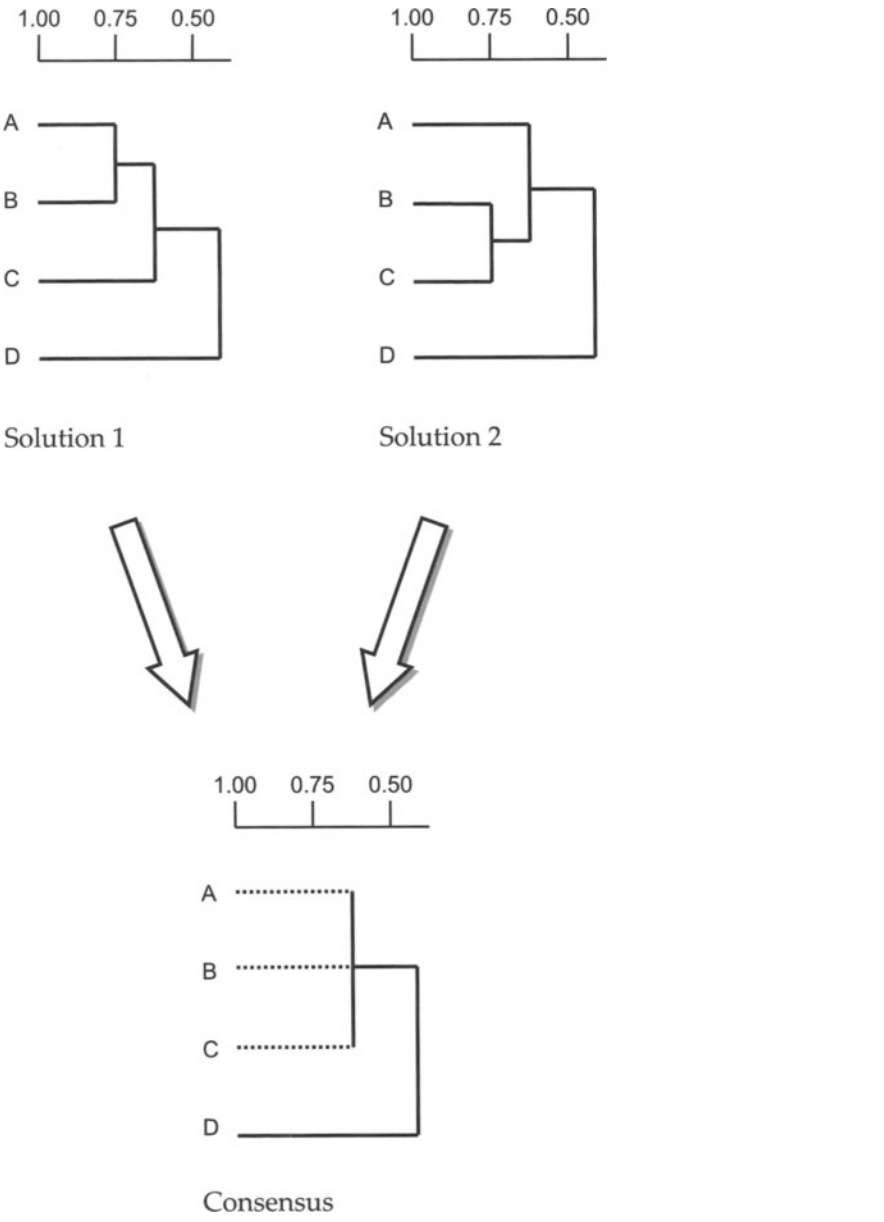


Fig. 6.44. Displaying different trees as a consensus tree

contain errors, the amount of error being proportional to the complexity of the resemblance matrix. A second source of error results from the fact that hierarchical clustering always imposes hierarchical structure, even if the data do not support it. The fact that even a randomly filled resemblance matrix results in a dendrogram with branches is a clear example of the danger that hierarchical clustering holds. Various statistical methods allow the error associated with dendrogram branches or their uncertainty in function of the resemblance matrix to be estimated, e.g. the standard deviation values and the cophenetic correlation. Other methods, such as bootstrap, allow the probability of dendrogram branches to be indicated as a result of the underlying data set (Felsenstein 1985).

6.12.1 Degeneracy of Dendrograms

Another problem with pairwise hierarchical clustering methods such as UPGMA is the degeneracy of the solution. Whereas UPGMA results in just one tree, in many cases there exist a number of equally good alternative solutions. Such degeneracies are very likely to occur in cases where the resemblance matrix contains multiple identical values. In practice, binary and categorical data sets treated as absent/present states result in the frequent occurrence of identical similarity values, whereas quantitative measurements registered as decimal numbers rarely yield identical similarity values. A number of commonly used molecular techniques are very sensitive to the problem of degeneracies, e.g. pairwise binary scoring of banding patterns, MLST and VNTR data, and binary character data.

To understand how the occurrence of identical similarity values can result in multiple possible trees, we consider the example of three banding patterns (Fig. 6.45). As can be seen from this simple example, s[A,B] and s[B,C] are both 0.75, whereas s[A,C] is 0.50. The way UPGMA constructs a dendrogram is by first searching for the highest similarity value in the matrix and then linking the two samples from which it results (see Sect. 6.10.1). In the present example, [A,B] and [B,C] are equivalent solutions, so that two partial dendrograms can be constructed: one with [A,B] linked at 75% (solution 1) and the other with [B,C] linked at 75% (solution 2). In the next step of UPGMA, the remaining sample is linked at the average of its similarity with the samples already grouped. In solution 1, this leads to C being linked at 62.5% to [A,B], whereas in solution 2, A is linked at 62.5% to [B,C]. Both dendrograms suggest a quite different hierarchical relatedness but actually neither of them truly reflects the relationships suggested by the data set and the resemblance matrix.

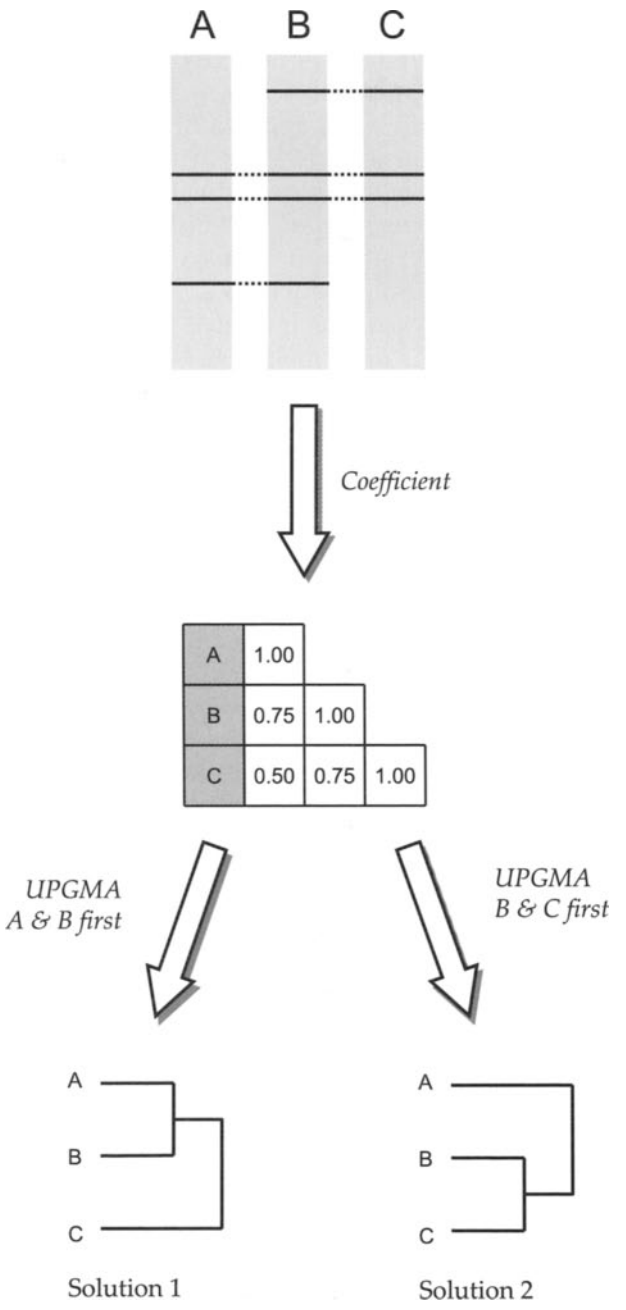


Fig. 6.45. A scenario of three banding patterns resulting in two possible UPGMA solutions

Another inconsistency in pairwise clustering results from the inability to deal with infringments upon the transitivity rule of identity. When sample A is identical to sample B, and sample B is identical to sample C, the transitivity rule predicts that A will also be identical to C. Infringments upon this rule are particularly found in the pairwise comparison of banding patterns, where the identity of bands is judged based upon their distance, using a position tolerance value that specifies a maximum distance between bands to be considered identical. The example in Fig. 6.46 illustrates the result of a UPGMA clustering of three banding patterns for which one band is slightly shifted. With a position tolerance as indicated on the figure, the pairs of patterns [A,B] and [A,C] will have a 100% score, whereas [A,C] will have only 75% similarity, as the distance between their lower bands is greater than the position tolerance specified. As explained above, the UPGMA algorithm has two choices to perform the first linkage; and the results are displayed as solution 1 and solution 2. Neither of the two dendrograms reflects the discrepancy indicated by the similarity values, but instead, each dendrogram falsely suggests a hierarchical structure that is not supported by the data.

6.12.2 Dealing with Dendrogram Degeneracies

A dendrogram that contains degeneracies cannot be presented correctly and may be misleading to the observer. Some computer implementations allow the degeneracies to be indicated on the dendrogram branches. A truthful representation of the examples given in Figs. 6.45, 6.46 can only be obtained by respecting the indeterminacy resulting from the identical similarity values. The BioNumerics software provides a solution similar to the consensus trees discussed in Sect. 6.11.3. All entries (or branches) involved in a degeneracy are linked at one common branch that respects the different possible solutions (see also Fig. 6.44).

An interesting way to deal with dendrogram degeneracies is to apply a *secondary criterion*. The *primary criterion* is the dendrogram-constructing algorithm, e. g. UPGMA, single linkage, complete linkage. The secondary criterion will be applied if two equivalent solutions emerge while iterating a resemblance matrix into the dendrogram. This is a principle similar to the secondary criteria used to reduce the degeneracy in minimum spanning trees calculated from MLST data (see Sect. 6.10.3). Possible secondary criteria to reduce degeneracies in similarity-based pairwise clustering algorithms that are implemented in BioNumerics are:

1. Highest overall similarity: the two clusters will be joined that result in the cluster with the highest overall similarity with all other members of the comparison.

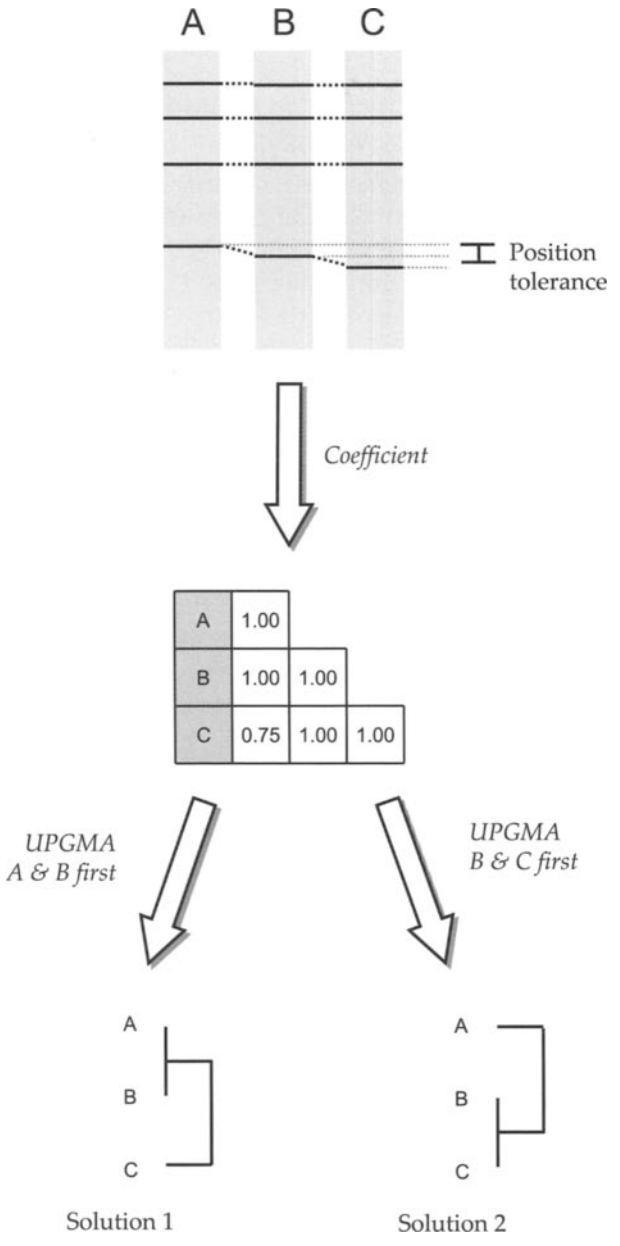


Fig. 6.46. Infringement upon the transitivity rule for sample identity and resulting dendrogram solutions

2. Largest number of entries: the two clusters will be joined that result in the cluster with the largest number of entries.
3. Most homogeneous cluster: the two clusters will be joined that result in a cluster that has the highest internal homogeneity.

Note that criteria 1 and 3 are complementary to each other, as criterion 1 will only consider the external similarity values of the resulting clusters, whereas criterion 3 will only consider their internal similarity values.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Amaratunga D, Cabrera J (2004) Exploration and analysis of DNA microarray and protein array data. Wiley-Interscience, Hoboken

van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh HA (1997) Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. Infect Immun 65:5017–5027

van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 62:275–293

Bochner BR, Gadzinsky P, Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Res 11:1246–1255

Bruijn FJ de (1992) Use of repetitive (repetitive extragenic palindromic and enterobacterial repetitive intergenic consensus) sequences and the polymerase chain reaction to fingerprint the genomes of *Rhizobium meliloti* isolates and other soil bacteria. Appl Environ Microbiol 58:2180–2187

Carrillo H, Lipman D (1988) The multiple sequence alignment problem in biology. SIAM J Appl Math 48:1073–1082

Cohn D, Bustreo F, Raviglione M (1997) Drug-resistant tuberculosis: review of the worldwide situation and the WHO/IUATLD global surveillance project. Clin Infect Dis 24[Suppl 1]:S121–S130

Colwell RR (1970) Polyphasic taxonomy of the genus *Vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. J Bacteriol 104:410–433

Dawyndt P (2004) Knowledge accumulation of microbial data aiming at a dynamic taxonomic framework. PhD thesis, University of Ghent, Ghent

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, suppl 3. National Biomedical Research Foundation, Washington, D.C., pp 345–352

Dice LR (1945) Measures of the amount of ecological association between species. J Ecol 26:297–302

van Embden JDA, Cave MD, Crawford JT, et al (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 31:406–409

Ewing B, Hillier L, Wendl MCM, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 55561–590

Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al (2003) How clonal is *Staphylococcus aureus*? J Bacteriol 185:3307–3316

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:666–670

Fitch WM (1971) Towards defining the course of evolution: minimum change for a specified tree topology. Syst Zool 20:406–416

Fry NK, Bangsborg JM, Bergmans A, Bernander S, Etienne J, et al (2002) Designation of European working group on *Legionella* infections amplified fragment length polymorphism types of *Legionella pneumophila* serogroup 1 and results of intercentre proficiency testing using a standard protocol. Eur J Clin Microbiol Infect Dis 21:722–728

Goldstein DB, Schlötterer C (1999) Microsatellites: evolution and applications. Oxford University, Oxford

Grimont F, Grimont PAD (1986) Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. Ann Inst Pasteur Microbiol 137B :165–175

Grundmann HJ, Towner KJ, Dijkshoorn L, Gerner-Smidt P, et al (1997) Multicenter study using standardized protocols and reagents for evaluation of reproducibility of PCR-based fingerprinting of *Acinetobacter* spp. J Clin Microbiol 35:3071–3307

Gunnarsson GH, Thormar HG, Gudmundsson B, Akesson L, Jonsson JJ (2004) Two-dimensional conformation-dependent electrophoresis (2D-CDE) to separate DNA fragments containing unmatched bulge from complex DNA samples. Nucleic Acids Res 32:23

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment in a microcomputer. Gene 73:237–244

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44 :223–270

Janssen P, Coopman R, Huys G, Swings J, Bleeker M, Vos P, Zabeau M, Kersters K (1996) Evaluation of the DNA fingerprinting method AFLP as an new tool in bacterial taxonomy. Microbiology 142:1881–1893

Jeffreys AJ, Pena SDJ (1993) Brief introduction to human DNA fingerprinting In: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys AJ (eds) DNA fingerprinting: the state of the science. Birkhauser, Basel, pp 1–19

Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J Bacteriol 182:2928–2936

Levy SB, Marshall B, Schluederberg S, Rowse D, Davies J (1988) High frequency of antimicrobial resistance in human fecal flora. Antimicrob Agents Chemother 32:1801–1806

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. PCR Methods Appl 4:357–362

Lowry FD (1998) *Staphylococcus aureus* infections. N Engl J Med 339:520–532

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388

Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Appl Environ Microbiol 65:3518–3525

Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol 59:695–700

Muyzer G, Smalla K (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek 73:127–141

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Ochiai A (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions (in Japanese, with English summary). Bull Jpn Soc Sci Fish 22:526–530

Pearson K (1926) On the coefficient of racial likeness. Biometrika 18:105–117

Pot B, Gillis M, Van de Velde A, Bekaert F, Kersters K, De Ley J (1989) Intra- and intergeneric relationships of the genus *Oceanospirillum*. Int J Syst Bacteriol 39:23–24

Quetelet A (1866) "Pierre-François Verhulst." In: Thiry H (ed) Sciences mathematiques et physiques chez les Belges au commencement du XIX siècle. Van Buggenhoudt, Bruxelles, pp 165–183

Rademaker JLW, Bruijn FJ de (1997) Characterization and classification of microbes by rep-PCR genomic fingerprinting and computer-assisted pattern analysis. In: Caetano-Anollés G, Gresshoff PM (eds) DNA markers: protocols, applications, and overviews. Wiley, New York, pp 151–171

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sasser M (1990) Identification of bacteria by gas chromatography of cellular fatty acids. (MIDI technical note 101.) Microbial ID, Newark

Selander RK, Beltran P, Smith NH, Helmuth R, Rubin FA, Kopecko DJ, Ferris K, Tall BD, Cravioto A, Musser JM (1990) Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. Infect Immun 58:2262–2275

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló R, et al (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047

Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B (1995) Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239

Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438

Vaneechoutte M, Heyndrickx M (2001) Application and analysis of ARDRA patterns in bacterial identification, taxonomy and phylogeny. In: Dijkshoorn D, Towner T, Struelens S (eds) New approaches for analysis of microbial typing data. Elsevier, New York, pp 211–247

Vauterin L, Vauterin P (1992) Computer-aided objective comparison of electrophoresis patterns for grouping and identification of microorganisms. Eur Microbiol 2:37–41

Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol Evol 11:158–163

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

7 Assessment of Microbial Phylogenetic Diversity Based on Environmental Nucleic Acids

Josh D. Neufeld, William W. Mohn

7.1 Introduction

"It is a Golden Age for the discovery of new organisms and for achieving a better understanding of the global ecosystem, which is, after all, based upon the microbial world" (Hugenholtz and Pace 1996). These sentiments reflect renewed confidence in microbial ecology, as the last decades of the twentieth century witnessed the circumvention of culture-based approaches by the advent of molecular methodology. Molecular methods have enabled novel insight into microbial community composition and expanded the range of ecological questions that may now be addressed. Despite tremendous advances, a major consideration remains the enormous and largely unexplored diversity of most microbial communities on Earth.

Prokaryotic diversity remains poorly characterized, yet microbial diversity and its controlling factors are major concerns of scientific and practical importance for society. A clear understanding of diversity is critical for understanding the relationship between community composition and function. There is an inevitable overlap between approaches intended to explore and characterize microbial diversity per se and approaches intended to monitor populations and determine their functional importance. From a practical perspective, understanding microbial diversity is critical in order to evaluate the impact of introduced species on pre-existing communities, the survival of pathogens in the environment, the impact of human activities, such as forestry, agriculture, and aquaculture, as well as the impact of climate change. Since many environmental problems and processes are linked to microbial communities, improved knowledge of

Josh D. Neufeld: Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL, UK, E-mail: jneufeld@warwick.ac.uk

William W. Mohn: Department of Microbiology and Immunology, University of British Columbia, 300–6174 University Boulevard, Vancouver, British Columbia, V6T 1Z3, Canada, E-mail: wmohn@interchange.ubc.ca

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

microbial diversity greatly benefits areas of health, pollutant biodegradation, wastewater treatment, ecosystem management, and biotechnology. The ability of biotechnology to efficiently exploit microbial catalysts and products is hinged on empirical knowledge of prokaryotic diversity in the environment. Finally, the existence of vast microbial diversity in its own right poses important research questions. Is this diversity functionally important? Or, is there great redundancy which is a superfluous consequence of microbial evolution? Answers to these questions must temper our knowledge of the practical importance of microbial communities.

Sampling the broad scope of microbial diversity requires the application and continued development of molecular methods suitable for rapid and efficient characterization. Here we discuss the impetus for molecular methodology, explore the ecological questions being addressed, and summarize techniques currently used for rapidly assessing microbial phylogenetic diversity in the environment. The wealth of molecular methods can generate confusion regarding the suitability and limitations of specific methods for particular applications. By presenting some of the background to molecular methodology in the context of microbial ecology, we hope to provide a helpful discussion in which the value and utility of individual methods, and combinations of methods, may be framed.

7.2 Microbial Phylogenetics and the 16S rRNA Gene

In the late 1800s, Robert Koch grew bacteria on solid culture medium for the first time. Since then, the physiology of a select few culturable bacteria (such as *Escherichia coli* and *Bacillus subtilis*) has become well understood, but it has gradually become apparent that microorganisms grown on defined media are not representative of the most abundant members of natural microbial populations. In fact, by comparing the number of stained viable cells observed microscopically to the number of colonies formed on plates, between 0.001% and 15% of the microorganisms in a given environment are estimated to be culturable using standard techniques (Amann et al. 1995). Culture-based studies are important for understanding the physiology and function of microorganisms (Palleroni 1997), but these approaches alone remain insufficient for monitoring the abundance and diversity of organisms within the environment and for describing their evolutionary interrelationships. Generations of microbiologists have been aware of the dichotomy between observations of the microbial world through microscopes and plate counts; and this problem was referred to as the "great plate count anomaly" (Staley and Konopka 1985). Objective evaluations of community composition, diversity, and dynamics were so elusive that

Rosswall and Kvillner (1978) described it thus: "All these factors, together with taxonomic difficulties, make a conventional description of the number and species composition of microorganisms from natural environments difficult, if not impossible".

The severe bias associated with cultivation approaches was partly circumvented as molecular sequence information for measuring evolutionary relationships began in the 1950s and was established by Zuckerkandl and Pauling (1965). They viewed microorganisms as "infomostats", analogous to chemostats or thermostats, since information in the form of macromolecules (DNA, RNA, proteins) are stored in each cell and passed along to subsequent generations. They reasoned that, by virtue of these molecules, a phylogenetic classification of microbial life would be possible. Incentive derived from necessity led microbiologists to identify the macromolecules most appropriate for taxonomic classification. Comparing ribosomal sequences or ribosomal gene sequences to one another provided a logical and rational manner by which broad-scale prokaryotic diversity could be organized into categories of similarity. The small subunit ribosomal RNA genes were ideal because of their universal distribution, structural conservation, the presence of conserved and variable regions, and resistance to lateral gene transfer (Olsen et al. 1986). By the late 1970s, microbiologists began using ribosomal RNA genes for measuring the phylogenetic relationships between microorganisms (Woese and Fox 1977). Bacterial taxonomy had become "a field fresh with the excitement of the experimental harvest" (Fox et al. 1980). Using such an approach, Carl Woese (1990) showed that life could be classified into three broad domains: *Bacteria*, *Eucarya*, and *Archaea*. Within the bacterial domain, Woese identified 12 divisions (Woese 1987; Woese et al. 1985), represented entirely by cultured isolates. Biotechnological innovations (particularly 16S rRNA gene cloning, sequencing, PCR) helped overcome major methodological hurdles (Lane et al. 1985; Saiki et al. 1985) and facilitated the collection of ribosomal sequences from cultured isolates and also from the environment (Giovannoni et al. 1990). Hugenholtz and coworkers (1998) credited a culture-independent approach with tripling the number of recognized domains to 36, of which 13 were represented only by environmental sequences. By 2003, the number of recognized divisions had jumped to 53 (Rappé and Giovannoni 2003), of which 26 divisions had no cultured representatives. While successful cultivation of organisms from diverse phylogenetic groups has resulted from recent advances in culturing techniques (Connon and Giovannoni 2002; Janssen et al. 2002; Joseph et al. 2003), the number of recognized bacterial divisions is also likely to grow as additional 16S rRNA gene sequences are collected at ever increasing rates (Rappé and Giovannoni 2003).

7.3
16S rRNA and the Environment

Ribosomal sequence collection from the environment began with hot springs in Yellowstone National Park and involved a collection of 5S rRNA gene sequences (Stahl et al. 1985). However, the description and comparison of ribosomal sequences shifted from 5S rRNA to 16S rRNA, since the increased length of the latter gene provided superior phylogenetic resolution. Environmental sequences have been deposited in databases, such as GenBank (Benson et al. 2000). The ribosomal database project (RDP-II; Cole et al. 2003) frequently retrieves and aligns ribosomal genes. The extremely rapid rate of 16S rRNA gene discovery from the environment is reflected in the progressive number of aligned and annotated sequences stored in the RDP-II. The total number of 16S rRNA gene sequences was 50,055 in September 2002; and it has more than doubled to 124,165, as of February 2005. A recent extrapolation from RDP-II alignments indicated that, depending on the criterion selected for identifying unique phylotypes, conservative estimates of total global diversity are in the range of 10,000 to 325,000 unique taxonomic units (Schloss and Handelsman 2004). Despite these tremendous advances, knowledge of bacterial diversity is obviously still under construction and will involve much more exploration and discovery.

The 16S rRNA gene has provided a helpful framework for describing novel microbial diversity. As mentioned, 16S rRNA is present in all organisms and contains regions of high sequence conservation interspersed with nine highly variable regions (Gutell et al. 1994). Conserved ribosomal domains provide "universal" regions (Giovannoni et al. 1988; Olsen et al. 1986; Zheng et al. 1996) suitable for probing, PCR priming sites, and for guiding sequence alignments. Bacterial species so far examined contain between one and 15 ribosomal RNA operons per genome (Schmidt 1997) and the number of operons is positively correlated with the ability to quickly respond to changing environmental conditions (Klappenbach et al. 2000). Unfortunately, the taxonomic specificity of 16S rRNA is not completely certain. One of the most important criteria for delineating a bacterial species is the similarity of its genome to the genomes of other organisms (Stackebrandt et al. 2002). In general, a DNA–DNA similarity of 70% is considered sufficient for delineating species, approximately coinciding with the threshold for phenotypic uniqueness. Organisms with $> 70\%$ genome similarity typically also have 16S rRNA gene similarities of $> 97\%$ (Stackebrandt and Goebel 1994); and this ribosomal similarity threshold is considered an additional criterion for species identity. However, a perfect correlation does not exist between DNA similarity and 16S rRNA gene similarity (Rossello-Mora and Amann 2001). The literature

contains examples of distinct species with highly similar or identical 16S rRNA genes (Fox et al. 1992; Jaspers and Overmann 2004; Martinez-Murcia et al. 1992) and strains of the same species containing highly divergent 16S rRNA genes (Nübel et al. 1996). Although a recent survey of 16S rRNA genes in 55 sequenced genomes demonstrated a maximum of 98.74% sequence heterogeneity between operons within the same organism (Coeyne and Vandamme 2003), there are occurrences of individual organisms with multiple highly divergent ribosomal operons (Wang et al. 1997; Yap et al. 1999).

The presence of divergent 16S rRNA gene sequences within the same organism suggests that conserved housekeeping genes may not be completely free from the impact of horizontal gene transfer (HGT) on prokaryotic evolution (Doolittle 1999). HGT is thought to have contributed to approximately 18% of the *Escherichia coli* genome over the past 100×10^6 years (Lawrence and Ochman 1998), which demonstrates an important role for HGT in shaping bacterial genomes. However, Ragan (2001) argued that 16S rRNA genes may be relatively "immune" to HGT, since the translational components of cells were established early in evolutionary history (Graham et al. 2000; Woese 2000), ribosomal RNA molecules are dependent on interactions with many other components of the ribosome (Jain et al. 1999), and phylogenetic trees constructed from microbial genomes generally agree with 16S rRNA gene trees (Fitz-Gibbon and House 1999; Snel et al. 1999; Tekaia et al. 1999). Although ribosomal genes are not perfect delineators of microbial species' identity and 16S rRNA gene phylogenies may poorly account for HGT effects, 16S rRNA gene sequence collections are extremely practical for phylogenetic discovery, surveys of environmental samples, and for initial classification of cultured isolates. Until genome sequencing becomes greatly simplified and more affordable (Shendure et al. 2004), for characterization of isolates and even for whole community analysis, the analysis of 16S rRNA sequences remains the most functional and practical basis for rapid phylogenetic assessment of microbial communities.

Other genes are also useful as phylogenetic markers for describing the diversity of microbial communities. Alternate phylogenetic markers share characteristics typical of ribosomal genes, such as high structure and sequence conservation. Such genes include those coding for DNA recombination systems, RNA polymerase subunits, elongation factors, sigma factors, heat-shock proteins, ATPases, and DNA gyrases. Analogous to studies in which 16S rRNA genes are PCR-amplified for fingerprinting or for cloning and sequencing, these alternative phylogenetic markers can be collected from the environment for assessments of community composition and diversity. Examples of these approaches include a DNA gyrase subunit (*gyrB*) collected for characterizing activated sludge (Watanabe et al. 1998), a chap-

eronin gene (*cpn60*) collected from pig feces (Hill et al. 2002), and an RNA polymerase subunit (*rpoB*) used to characterize environmental isolates (Dahllöf et al. 2000) and soil microbial communities (Peixoto et al. 2002). A recent genetic survey of the Sargasso Sea suggested that improved quantitative indications of phylogenetic group abundances may be obtained with multiple phylogenetic markers (Venter et al. 2004). This was attributed to the variable frequency of ribosomal operons in different species. However, a current drawback to the analysis of alternative genetic markers is their relatively poor database representation. For example, approximately 2,000 sequences were recently reported as stored in the chaperonin gene database (Hill et al. 2004), and this is in stark contrast to the ca. 125,000 ribosomal RNA sequences currently available in the RDP-II. We performed a search of GenBank for other frequently used phylogenetic markers and discovered similarly low coverage (as of February 2005): *rpoB* (2,670 sequences), *gyrB* (3,065 sequences), *recA* (2,424 sequences) and *HSP70* (570 sequences). The use of these markers for providing assessment of microbial diversity is expected to increase as database representation improves, particularly for complementing assessments generated by ribosomal markers. The methods discussed below focus primarily on 16S rRNA genes, but most molecular methods can theoretically be extended to the analysis of alternative phylogenetic markers, particularly those amenable to PCR by virtue of "universal" conserved sequence sites, sufficiently conserved among broad groups of microorganisms for primer binding.

7.4 Molecular Methodology in Microbial Ecology

Microbial communities are responsible for transformations of myriad compounds, such as plant and animal components, atmospheric gases, and anthropogenic pollutants. Microbial ecologists seek to understand the relationship between community composition and environmental factors. The questions ultimately being addressed include: What factors influence microbial diversity and community abundance? How do community changes affect changes in the activity or functional potential? How does community diversity affect resilience and responses to environmental changes? How do human activities impact microbial communities in the environment? And, how lasting are these impacts? Phylogenetic diversity is a reservoir of functions which affect the response of a community to changing conditions.

Complete descriptions of microbial community composition include phylogenetic and functional components. Together, both phylogenetic diversity (species richness and evenness) and functional roles reflect micro-

bial community structure. Understanding community structure as it relates to environmental factors is especially challenging for microbiologists, due to the extreme diversity of most microbial communities. As a result, focusing on a specific component of community structure (either functional or phylogenetic diversity) tends to offer the most tractable and realistic goals for individual studies in microbial ecology. Nonetheless, since microbial diversity is linked to physiological diversity, understanding the magnitude of overall genetic diversity in particular environments is an important prerequisite for predicting associated biochemical potential.

Much research has been dedicated to the estimation of microbial abundance and diversity. Some suggest that there may be on the order of $4-6 \times 10^{30}$ prokaryotes on earth, with most of these organisms living in the open ocean, soils, and in the earth's subsurface (Whitman et al. 1998). Whitman and coworkers also suggested that the total carbon contained within microbial cells might be almost equivalent to the carbon stored within all other living organisms. The abundance of prokaryotes on earth is extremely high; and the taxonomic diversity of these prokaryotes may be correspondingly high. While the entire species diversity of the ocean might be less that 2×10^6, which is only half that expected in a ton of soil (Curtis et al. 2002), estimates of the Earth's total prokaryotic diversity range as high as 10^{12} microbial species (Dykhuizen 1998).

The enormous magnitude of microbial diversity has posed a challenge even for modern microbial ecologists equipped with molecular methodology. Molecular methods generally provide phylogenetic assessment for only the most abundant community members. While complete descriptions and comparisons of diversity have been reported for simple community assemblages inhabiting hot spring microbial mats (Blank et al. 2002; Skirnisdottir et al. 2000; Ward et al. 1998), microbial diversity has never been completely described in any other natural environment on Earth (Curtis and Sloan 2004). High microbial complexity has limited the testing of ecological principles governing community structure in the environment. One ecologist framed the lack of knowledge regarding microbial diversity by commenting: "...I do not discuss the diversity of microorganisms at all. From the point of view of diversity, they are probably the most poorly known of taxa. Perhaps their diversity shows many patterns, but I am unaware of them" (Rosenzweig 1995). While Borneman and coworkers (1996) stated that: "An enormous amount of effort is being made worldwide by microbial ecologists to identify microorganisms in environmental samples", the abundance of studies did "not seem proportional to our understanding of the significance of biodiversity for ecological processes in the microbial world..." (Morris et al. 2002). Many basic questions regarding the factors affecting microbial community composition and diversity within the environment remain unanswered.

Limited sampling of diverse communities is not a problem unique to microbial ecology. Macro-ecologists have also struggled to sufficiently sample diverse communities and have developed mathematical models to extrapolate total diversity from incomplete surveys (Colwell and Coddington 1994). The application of similar statistical approaches to the description and comparison of datasets generated by molecular methods has furthered ecological hypothesis testing of microbial communities. Microbial ecologists began comparing the diversity and composition of 16S rRNA gene clone libraries from multiple samples using statistical methods such as rarefaction curves (Dunbar et al. 1999), taxonomic richness estimates (Kroes et al. 1999; Nübel et al. 1999b), and general diversity indices (McCaig et al. 1999; Nübel et al. 1999a). However, the appropriateness of these statistics was unknown since they were developed for the analysis of macro-diversity. Beginning in 2001, Hughes and coworkers (Bohannan and Hughes 2003; Hughes and Bohannan 2004; Hughes et al. 2001) investigated the justification for applying ecological diversity measures to 16S rRNA gene sequence libraries. The novel application of nonparametric diversity estimators such as Chao1 (Chao 1984) and ACE (Chao and Lee 1992) to clone library data demonstrated that even though environmental clone libraries are dominated by rare sequences (Kemp and Aller 2004), the distribution of phylotype frequencies can still provide enough information for estimating total bacterial diversity and comparing estimates from multiple samples. Since this initial demonstration, novel applications of ecological statistics toward the analysis of 16S rRNA gene clone libraries has become a popular approach for comparing the diversity (Curtis et al. 2002; Hill et al. 2003; Martin 2002; Nee 2003) and composition (Schloss et al. 2004; Singleton et al. 2001) of these libraries, even for circumstances when all sequenced clones in a library are unique (Lunn et al. 2004), such as those derived from Amazon soils (Borneman and Triplett 1997). Further, a new computer program called DOTUR (distance-based operational taxonomic unit and richness) is expected to facilitate diversity studies involving 16S rRNA gene clone libraries, since it simplifies an analysis that previously required multiple time-consuming and problematic steps (Schloss and Handelsman 2005).

Statistical approaches have also been adopted to evaluate diversity in studies that generate DNA "fingerprints" from individual communities (e. g., denaturing gradient gel electrophoresis). While the chief application of fingerprints is to rapidly compare the similarity of communities, fingerprints are also commonly used to evaluate diversity. By considering the number of bands in each fingerprint profile and measuring the relative intensity of each band, community diversity may be estimated for each sample and compared to the diversity of other patterns. A common approach is to calculate a Shannon diversity index (which incorporates diversity and evenness of bands) or evenness index for analyzing patterns from environ-

mental samples (Fromin et al. 2002). Statistical approaches have also been adopted for sample comparisons based on overall fingerprint similarity, quantified in a dendrogram (Kropf et al. 2004). Comparison of multiple fingerprints is done either by matching bands across multiple fingerprints (e.g. Griffiths et al. 2003) or by comparing overall fingerprint intensity profile correlations (e.g. Leckie et al. 2004). Fromin and coworkers (2002) recently summarized some of the statistical tools available for comparing the similarity of numerous fingerprints. Clustering and ordination methods can include environmental parameters to help evaluate the impact of different factors on community composition (Besemer et al. 2005).

Together with statistical approaches such as those described above, assessment of phylogenetic diversity using molecular methods has helped bridge the practical gap between hypothesis testing in microbial ecology and macro-ecology. Previously used ecological approaches focusing on macro-organism diversity as it relates to factors such as productivity (Chase and Leibold 2002), functional diversity (Tilman et al. 1997), stability (McCann 2000), and stress response (Hughes and Stachowicz 2004; Mulder et al. 2001) are becoming possible for microbial ecologists to employ, with the benefit of molecular tools (Horner-Devine et al. 2004a). Brendan Bohannan's group at Stanford has used 16S rRNA gene clone libraries to determine relationships between the magnitude of bacterial diversity and primary productivity that reflect relationships common for macro-communities and that differ depending on the taxonomic group being examined (Horner-Devine et al. 2003). Furthermore, contrary to some literature suggesting that the biosphere is composed of relatively few microbial species with cosmopolitan distributions (Finlay 2002), Bohannan's group demonstrated measurable local turnover (beta diversity) in community species composition, which was clearly correlated with environmental factors (Horner-Devine et al. 2004b). Similar community turnover was discovered for fungal diversity, although a correlation was not observed with measured environmental factors per se, but rather with geographical distance (Green et al. 2004). Other studies used separation of 16S rRNA gene fragments in a gel matrix to generate unique fingerprints, gauging the impact of pollutants on microbial community structure (stability) and recovery (resilience) in soil environments (Girvan et al. 2005; Griffiths et al. 2004). These are simply a few examples of a growing ability, due to molecular methodologies, to address major ecological questions in microbial ecology, well beyond a simple description of diversity.

Since microbial diversity is often extremely high, established and new methods that rapidly assess phylogenetic diversity will play critical roles in future studies. While measuring, describing, and comparing microbial diversity, the depth with which communities are profiled will help determine the strength of the analysis. Furthermore, the ability to efficiently

process multiple samples (sample throughput) strengthens the power of an analysis by permitting replication and comparison of multiple treatments or environments. Lack of replication is a serious limitation of many studies of microbial ecology. In the following sections we discuss state-of-the-art molecular methods that provide phylogenetic assessment of microbial communities. We focus specifically on those methods that enable rapid retrieval of sequence information or provide high sample throughput. Since the high conservation of 16S rRNA genes does not reveal unique ecological adaptations of surveyed phylotypes, rapid assessment of phylogenetic diversity describes only 'phylotypes' or operational taxonomic units (OTUs) from a community. OTUs in this context are clusters of similar 16S rRNA gene sequences or distinct bands in a community fingerprint profile. The diversity of OTUs measured from an environment is expected to reflect the species diversity of microbial communities being studied. However, protein-coding genes may provide a more ecologically relevant focus for some studies (Dahllöf et al. 2000; Neufeld et al. 2001; Palys et al. 1997) in which the interest may be on closely related organisms with functionally distinct roles. Some of the methods summarized below are also adaptable to the analysis of functional genes for an alternative survey of ecologically distinct populations, in which gauging community function is key. Unique theoretical considerations may apply, since many approaches rely on conserved regions for universal primers, which often do not exist in protein-encoding genes (i. e. they are less conserved). Also, while many proteins do have conserved and homologous domains, as for prosthetic group binding, these may be more conserved in related proteins of different function than in proteins of the same function.

7.5 General Considerations of Bias

Bias is associated with the characterization of OTU diversity from microbial communities, using molecular methodology. While bias specific to individual methods is discussed in subsequent sections, some general biases are discussed here. Bias is introduced at the level of sample storage (Rochelle et al. 1994), DNA extraction (von Wintzingerode et al. 1997), 16S rRNA gene copy number (Farrelly et al. 1995), PCR (Becker et al. 2000; Polz and Cavanaugh 1998; Qiu et al. 2001; Reysenbach et al. 1992; Schmalenberger et al. 2001), and cloning (Rainey et al. 1994). PCR provides a powerful tool which has revolutionized microbial ecology but has been associated with various artifacts. PCR generates chimeric sequences (Kopczynski et al. 1994) which are created by increasing cycles of amplification (Qiu et al. 2001; Wang and Wang 1996, 1997) and are accumulating in public databases (Hugenholtz

and Huber 2003). PCR also generates heteroduplexes (Thompson et al. 2002) formed by the annealing of 16S rRNA gene amplicons from different organisms that lead to additional bands in fingerprint analysis (Ward et al. 1998).

Database sequence deposition may also inadvertently contribute to bias on the level of PCR primer design. The primers used for "universal" or specific amplification of particular phylogenetic groups are based almost entirely on cultured isolates, which are certainly not a reliable representation of microbial diversity. Analyses of environmental DNA libraries, which do not require prior PCR amplification, have provided some initial indication that 16S rRNA gene primer design may need revisiting (Vergin et al. 1998). Further contributing to this problem, relaxed annealing temperatures may allow PCR primers to anneal where they do not perfectly match, causing infidelity in the primer region of the resulting amplicons. Consequently, primer regions should not be included in sequence submissions to databases such as GenBank. However, this is not a required or routine standard for submissions and universal primer design and reevaluation will be unnecessarily biased by primer submissions. This problem should be addressed using careful database curation, as was recently done for chimeric sequences (Hugenholtz and Huber 2003).

Microbial ecologists are becoming aware that spatial scale is a factor that affects the characterization of microbial communities. For many years, sampling from environments such as activated sludge, soil, and sediment proceeded by measuring and reporting diversity without attention to the potential influence of sample size. However, heterogeneous distributions at small spatial scales may bias the comparison of diversity data. Soil is a particularly clear example of heterogeneous distributions of microorganisms. The soil matrix consists of physical and biological gradients, forming distinct microenvironments on small spatial scales (Grundmann 2004). Higher diversity and distinct phylogenetic groups may be associated with small soil-size fractions and fungal grazing may be responsible for reducing the bacterial diversity associated with larger soil particles (Sessitsch et al. 2001). Collecting an adequate number of microenvironments for reliable representation requires different sample sizes depending on the environment and target microorganisms being studied. For example, larger soil samples may be required for representative profiles of fungal diversity (> 1 g), while small samples (e.g. 0.125 g) may adequately represent the bacterial communities associated with larger sample sizes (Ranjard et al. 2003). In well mixed aeration tanks of activated sludge treatment systems, a relatively small aliquot may adequately reflect the OTU composition of the entire basin (Smith et al. 2003). As a result, care must be associated with the choice of sample size for describing and comparing diversity.

In addition to sample size, the definition of an OTU differs from study to study. For 16S rRNA genes, the percent similarities usually employed are 97% for species, 95% for genus, 90% for family/class, and 80% for division (Schloss and Handelsman 2004). As mentioned above, these delimiters are not always equivalent to physiological uniqueness. Thus, while useful for analysis and comparison, these cutoffs are somewhat arbitrary. A recent analysis of 56,215 16S rRNA genes in the RDP-II demonstrated that the percent similarity chosen to cluster OTUs has a profound affect on diversity estimates (Schloss and Handelsman 2004). Based on this RDP-II dataset, statistical estimates of the total number of OTUs in the global environment ranged between 9,867 OTUs (for 90% similarity clustering) to 325,040 OTUs (for 100% similarity clustering). Such widely disparate estimates underline the sensitivity of OTU definition to the relative magnitude of reported microbial diversity. The practical significance of this problem may be highlighted by considering a hypothetical scenario in which a comparison of two communities would be biased by the similarity cutoff chosen for clustering OTUs. In one hypothetical community, each organism occurs in conjunction with another closely related organism with a 16S rRNA gene of x% similarity. In a second community with an equivalent number of representatives, organisms are distantly related to each other with 16S rRNA gene similarities all much lower than x%. If OTUs were clustered with x% or greater similarity, both communities would be considered equally diverse. However, if clustering was done with a similarity cutoff less than x%, the two communities would be considered to have widely different diversities. Approaches that consider multiple OTU definitions may help minimize bias associated with deriving meaningful estimates of diversity in the context of microbial ecology.

A further difficulty arises with algorithms used for grouping sequences. Consider a case in which 16S rRNA gene from a clone library are grouped if they differ by one or no bases. If sequence A differs by one base from B and B differs by one other base from C, then A and B could be grouped or B and C could be grouped. A program such as Fastgroup (Seguritan and Rohwer 2001), which was developed specifically for the clustering of 16S rRNA gene sequences, would group all sequences (A, B, C) together since all of the sequences are within one base of another member of the group, despite the fact that A and C differ by two bases. This is obviously a problematic approach for forming distinct OTU clusters, since widely disparate sequences may be considered to be the same phylotype as an artifact of clustering large datasets. Schloss and Handelsman (2005) recognized this clustering artifact in creating DOTUR, which would group A and B in one OTU and C in a second group. DOTUR has an added advantage that multiple similarity criteria for OTU grouping may be examined simultaneously for its impact on measured diversity. Since Fastgroup and possibly DOTUR

are sensitive to sequence input order, a future modification may involve the bootstrapping of OTU clustering by randomizing the sequence input order, providing "consensus" OTU clusters.

Relative estimations and comparisons of community diversity are commonly made from fingerprint profiles using molecular methods, but the use of fingerprints to estimate diversity is highly controversial. There are general biases and caveats that apply to diversity estimates based on fingerprints. Caution must precede diversity estimates from fingerprints since community complexity can attain a threshold in which the number of bands is too high for adequate resolution, generating a fingerprint "smear". Fingerprints from a community of thousands of distinct OTUs may have only tens of discernible bands. For such communities, estimates of diversity may be highly biased by accounting for only relatively abundant OTUs. In such cases, diversity estimates would not correlate with species richness, due to the difficulty of detecting individual bands above the background. Another issue related to resolution is that multiple differing sequences may migrate to the same location and obscure accurate estimations of diversity. Further, the presence of multiple and differing ribosomal operons may contribute multiple bands from individual organisms in a fingerprint; and this restricts the extent to which fingerprints may reflect the actual species diversity of environmental samples (Fromin et al. 2002).

It is important to recognize that bias is common to all experimental techniques used to obtain qualitative and quantitative information from natural systems. Biases inherent to DNA-based microbial community analyses are probably no worse than the biases that are an accepted component of well established macro-community analyses. There is perhaps a tendency for microbiologists, accustomed to laboratory science, to over-react to the uncertainties inherent in field ecology. The biases associated with culture-based analyses of microbial communities are clearly worse than those with nucleic acid-based analyses, probably by orders of magnitude. PCR bias is often singled out as a factor diminishing the reliability of certain DNA-based analyses. However, other biases that are common to all nucleic acid-based analyses may be of far more consequence than PCR bias. In particular, nucleic acid extraction is clearly an important factor affecting all nucleic acid-based analyses. The magnitude of bias encountered with the assessment of phylogenetic composition of environmental nucleic acids and its effect on the resulting data should be estimated as accurately as possible. Ecological conclusions and principles should be cautiously inferred from PCR-based approaches. Thus can we fully benefit from the power of nucleic acid-based methods without being paralyzed by the awareness of inherent biases.

7.6 Phylogenetic Assessment of Environmental Nucleic Acids

The majority of nucleic acid-based methods fall into two broad categories: whole-community and partial-community analysis (Ranjard et al. 2000b), perhaps better referred to as total-DNA and single-gene (or single-amplicon) analysis. Total-DNA analysis involves approaches in which DNA extracts are analyzed for specific properties characteristic of the entire community. These techniques include the analysis of DNA guanine and cytosine (G+C) content by density centrifugation (e.g. Holben and Harris 1995), genome diversity by measuring reassociation kinetics of denatured DNA (e.g. Torsvik et al. 1990), and broad-scale community similarities by cross-community DNA hybridizations (e.g. Xia et al. 1995). Even though total-DNA approaches provide comprehensive information about community genetic diversity and similarity, these approaches do not provide qualitative assessment of phylotypes within microbial populations. Additionally, total-DNA assessment techniques have requirements that complicate their routine use in the analysis of multiple samples in a relatively unbiased manner. A large amount (> 50 μg) of minimally sheared DNA is required for reassociation and hybridization analyses, which is prohibitive for many environmental studies (Ranjard et al. 2000b). Furthermore, obtaining high-molecular-weight DNA precludes the use of rigorous DNA extraction approaches and may bias DNA extraction toward organisms that lyse readily, such as gram-negative organisms. Additionally, DNA reassociation approaches are time-consuming. For environmental DNA reassociation analysis, such as that pioneered by Torsvik's group at the University of Bergen, each sample requires a processing time of up to several weeks (Torsvik et al. 1998).

Single-gene analyses focus on PCR amplification of specific DNA fragments from community DNA (or RNA) extracts; and the 16S rRNA gene is most frequently studied with these approaches. PCR primer sequences can be modified to target different taxonomic levels, either to provide "universal" amplification or narrow specificity for individual phylogenetic groups. PCR products are analyzed either by gel-based separation (fingerprinting) or by sequencing of PCR amplicons. As illustrated in Fig. 7.1 and discussed below, fingerprinting methods are advantaged by being rapid and amenable to relatively high sample throughput, but limited by the available amount of phylogenetic information. Sequencing-based methods offer greater phylogenetic information but labor and cost limitations preclude the analysis of more than a select few samples. The successful application of microarrays in microbial ecology has high appeal since this technology has the potential to bridge the gap between fingerprinting and sequencing, offering high sample throughput while monitoring the presence of many

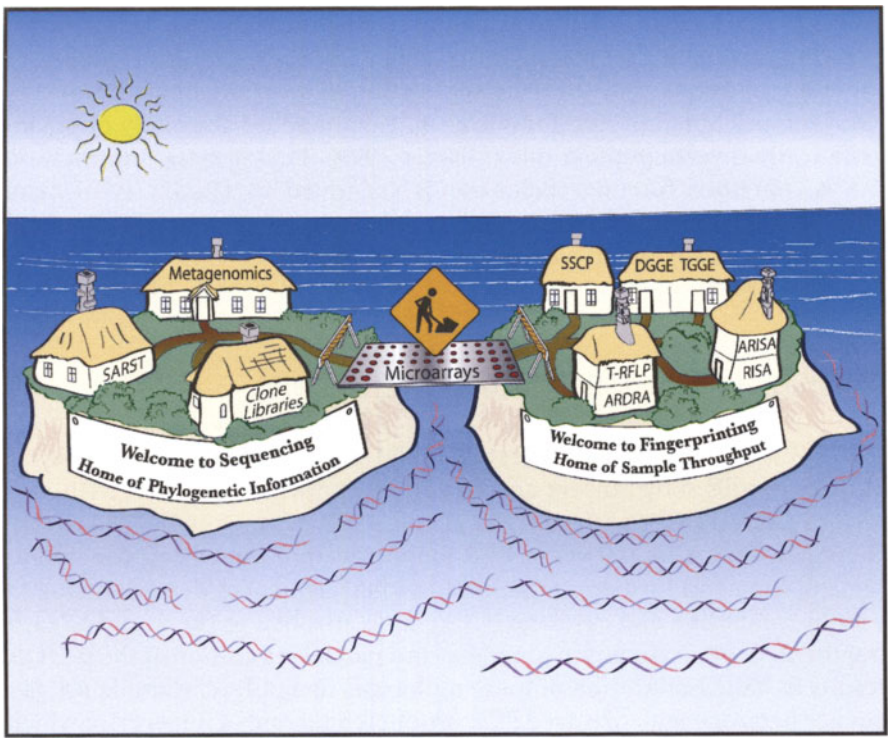


Fig. 7.1. Strengths and limitations associated with methods used for assessing the phylogenetic diversity of environmental nucleic acids. *SARST* Serial analysis of ribosomal sequence tags, *SSCP* single-stranded conformational polymorphism, *DGGE* denaturing gradient gel electrophoresis, *TGGE* temperature gradient gel electrophoresis, *T-RFLP* terminal fragment length polymorphism, *ARDRA* amplified ribosomal DNA restriction analysis, *RISA* ribosomal intergenic spacer analysis. The UBC Media Group is thanked for technical assistance in preparing this illustration

phylotypes (potentially 10^4) simultaneously. Below, we discuss the status of fingerprinting, sequencing, and microarray techniques with recent examples of their application in microbial ecology.

7.7 Fingerprinting

Fingerprinting methods provide advantages of being rapid, affordable, relatively easy to use and amenable to high sample throughput. Methods in microbial ecology that separate 16S rRNA gene PCR amplicons in a gel matrix based on sequence heterogeneity include denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE),

single-stranded conformational polymorphism (SSCP) and terminal restriction fragment length polymorphism (T-RFLP). The accessibility of these methods has enabled their application in a variety of environments (Muyzer and Smalla 1998; Torsvik et al. 1998) and for the study of phylogenetically diverse populations (Muyzer 1999). Phylogenetic information can be obtained for fingerprint bands generated by DGGE, TGGE, and SSCP by excising regions of the gel for subsequent PCR amplification and sequencing, as suggested by Muyzer and coworkers (1993).

7.7.1 Denaturing Gradient Gel Electrophoresis

DGGE was initially used for detecting mutations within the human genome, but was modified by Muyzer and coworkers for separating 16S rRNA amplicons (Muyzer 1999; Muyzer and Smalla 1998; Muyzer et al. 1993, 2004). DGGE involves a migration of PCR amplicons into increasing concentrations of urea and formamide until individual sequences denature. Due to a 40-bp guanidine and cytosine (GC) clamp attached to the 5′ end of each fragment, denaturation is incomplete, and partial separation of the strands results in halted migration of these molecules in a polyacrylamide gel. Sequence heterogeneity of mixed PCR products generates a fingerprint which is characteristic for each community. The complexity and resolution of 16S rRNA gene PCR products is sensitive to the variable regions selected for fingerprint analysis (Yu and Morrison 2004).

Short PCR products (< 500 bp) are commonly separated by DGGE, although short amplicons limit the phylogenetic information that may be obtained by sequencing fingerprint bands and restrict the choice of target for PCR amplification. A 500-bp size limit is cited (e.g. Muyzer et al. 2004) as a result of early computer-based predictions suggesting that polyacrylamide resolution of single-base substitutions in DNA is optimal for sequences between 25 bp and 500 bp (Myers et al. 1985). Indeed, comparisons of DGGE fragment lengths have demonstrated that shorter PCR products offer higher resolution, yielding greater numbers of fingerprint bands (Yu and Morrison 2004). However, studies examining marine viral diversity have consistently resolved longer PCR amplicons (550–700 bp) with DGGE (Short and Suttle 1999, 2000). Longer fragment lengths obviate the need for the addition of a GC clamp since amplicons do not completely denature. Since the detection of single-base differences is not usually required or desired for generating microbial community fingerprints, the use of longer PCR products should be appropriate for future experimental designs.

Other recently reported methods avoid artifacts encountered with DGGE. Consistent electrophoresis times and modified PCR reaction protocols in-

crease DGGE reproducibility and reduce problematic band profiles (Janse et al. 2004; Sigler et al. 2004). Another notable drawback with DGGE is the challenge of pouring consistent gradients. Many published studies that employ DGGE report the use of the DCode universal mutation detection system (BioRad). The gradient-forming wheel included with this system (affectionately referred to in our laboratory as the "wheel of fortune") requires manual manipulation, which leads to gel-to-gel inconsistencies. Careful gel normalization is critical if fingerprints are to be compared to those either from the same or from different gels (Ferrari and Hollibaugh 1999; Powell et al. 2003). Toward solving this problem, the application of fluorophore labels on DGGE primers (Bano and Hollibaugh 2000) provides a means by which intra-lane standards may be run with each sample. This simple modification has improved the sensitivity and normalization of fingerprints for the comparison of multiple samples (Neufeld and Mohn 2005a). Furthermore, fluorophore-labeled internal standards prove useful for measuring the magnitude of bias affecting DNA-handling steps that precede DGGE, such as DNA extraction and PCR (Petersen and Dahllöf 2005).

7.7.2 Temperature Gradient Gel Electrophoresis

Instead of a chemical gradient, TGGE involves a migration of PCR products into an increasing temperature gradient (Muyzer 1999; Muyzer and Smalla 1998). Like DGGE, PCR primers are synthesized with a GC clamp and PCR products are separated by their resistance to denaturation. TGGE was adapted from gene mutation analysis (Rosenbaum and Riesner 1987) for microbial ecology (Felske et al. 1998). A variation of TGGE is temporal temperature gradient gel electrophoresis (TTGE) which increases the electrophoresis buffer temperature at a defined ramp rate over the course of the gel run (Ogier et al. 2002). The use of TGGE/TTGE is simpler and likely more consistent than DGGE since gels are poured with a uniform concentration of denaturant. While one study demonstrated that DGGE provided higher resolution of highly similar DNA fragments than TGGE (Farnleitner et al. 2000), it seems intuitive that once optimized properly, TTGE should perform comparably to DGGE. For example, Ogier and coworkers (2002) used TTGE to successfully separate 16S rRNA gene amplicons from 48 closely related bacteria recognized for their involvement in cheese production. Furthermore, they demonstrated high resolution, even with a 700-bp 16S rRNA gene amplicon, which further supports the suggestion that > 500 bp fragments are amenable to efficient separation with denaturing gradient fingerprint techniques.

7.7.3 Single-stranded Conformational Polymorphism

SSCP analysis was also originally used for gene mutation analysis (Orita et al. 1989) and adapted for the analysis of microbial communities (Lee et al. 1996; Schwieger and Tebbe 1998). As the name implies, SSCP analysis involves the electrophoresis of single-stranded PCR amplicons. Denatured sequences assume unique single-stranded conformations that differentially inhibit electrophoresis. SSCP has been used to profile various phylogenetic groups from microbial communities, including bacteria (Delbes et al. 2000), fungi (Peters et al. 2000), and Archaea (Leclerc et al. 2004). Without requiring GC-clamped primers or the formation of a gradient gel and being amenable to automated DNA sequencer analysis (Zumstein et al. 2000), SSCP is an adaptable and versatile method for the rapid profiling of microbial communities.

7.7.4 Terminal Restriction Fragment Length Polymorphism

Other fingerprinting techniques separate PCR amplicons by size rather than sequence heterogeneity. One approach involves PCR amplification of 16S rRNA genes, followed by restriction endonuclease digestion with combinations of frequent-cutting enzymes (e.g. Porteous et al. 1997). This method is known as amplified ribosomal DNA restriction analysis (ARDRA) and is well suited to the characterization of environmental isolates (Nazaret et al. 2003). However, since each digestion yields at least two fragments for a given amplicon, the digested products generate complex and poorly resolved fingerprints, even for communities with low diversity. Toward solving this dilemma, T-RFLP uses fluorophore labels introduced on one PCR primer. Only the terminal fragment linked to the labeled primer fluoresces, greatly simplifying the restriction pattern (Avaniss-Aghajani et al. 1994). T-RFLP electrophoresis is typically performed on a sequencing gel, which provides high resolution, sensitivity, quantitation, sample throughput, and accurate sizing of individual fragments by the use of size standards. T-RFLP has proved particularly useful for comparing the similarity of multiple bacterial communities, with examples including aquifer sands (Liu et al. 1997), marine samples (Moeseneder et al. 1999), soils (Dunbar et al. 2000; Hackl et al. 2004), and infant fecal communities (Wang et al. 2004).

Despite obvious advantages, T-RFLP has some disadvantages. The use of a DNA sequencer for T-RFLP prevents the excision and sequencing of fingerprint bands, as is common for DGGE or TGGE. Partially circumventing this limitation, OTU information can be indirectly inferred by comparison

of T-RF lengths to 16S rRNA gene databases of theoretical T-RF lengths (Kent et al. 2003; Marsh et al. 2000). This is effective only if the community has a low relative complexity, since the phylogenetic specificity of T-RFs decreases as the number of T-RF peaks increases (Dunbar et al. 2001). An appropriate alternative involves the comparison of T-RF peaks with T-RF sizes calculated from clone library sequences generated from the same samples (Hackl et al. 2004). Another disadvantage is that T-RFLP is not well suited to estimating OTU richness and evenness, since digestion with different enzymes produces substantially different band patterns and complexity (Dunbar et al. 2000). Also, single-stranded DNA artifacts generated by PCR could potentially lead to additional bands and increased estimates of OTU diversity (Egert and Friedrich 2003). Finally, difficulties in loading consistent amounts of DNA in each lane led some researchers to report inconsistency in patterns from replicate samples (Dunbar et al. 2001; Osborn et al. 2000). Provided measures are taken to minimize (or account for) PCR-generated artifacts (Egert and Friedrich 2003) and careful fingerprint standardization is employed, T-RFLP shows great promise as a leading methodology for the rapid assessment of phylogenetic composition of environmental nucleic acid from multiple sites and replicate samples.

7.7.5 Ribosomal Intergenic Spacer Analysis

In the majority of characterized microorganisms, the 16S rRNA gene is adjacent to the 23S rRNA gene and is separated by an intervening region of variable length. Ribosomal intergenic spacer analysis (RISA) separates PCR products that span the 5′ end of the 16S rRNA gene, through the spacer, and into the 3′ end of the 23S rRNA gene. Since the spacer region is not as conserved evolutionarily as the ribosomal genes, RISA offers higher resolution than the analysis of the 16S rRNA gene. Fingerprints possibly reflect species or sub-species taxonomic distributions (Jensen et al. 1993). A recent study demonstrated that isolates of *Brevundimonas albus* differing substantially in morphology and physiology all contained identical 16S rRNA genes, but that the spacer region length polymorphism correlated with cell morphology polymorphisms (Jaspers and Overmann 2004). RISA has been applied to the study of soils (Borneman and Triplett 1997; Ranjard et al. 2000a), wastewater treatment systems (Smith et al. 2003; Yu and Mohn 2001), and other environments (Gonzalez et al. 2003). The majority of studies targeting the ribosomal intergenic spacer incorporate fluorophore labels for automated use of a DNA sequencer (Leckie et al. 2004; Yannarell et al. 2003). Automated ribosomal intergenic spacer analysis (ARISA) provides

increased resolution (Fisher and Triplett 1999), which enables rapid and reproducible comparisons of bacterial and fungal communities from multiple samples (Ranjard et al. 2001). However, unlike T-RFLP, ARISA does not permit phylogenetic identification of the myriad bands that represent most microbial communities.

7.7.6 Additional Considerations

Fingerprinting approaches have some common limitations. DGGE and SSCP fingerprints, for example, are estimated to focus on only the most abundant OTUs: those that comprise greater than 1% of a given community (Lee et al. 1996; Muyzer et al. 1993). Complex communities with thousands of phylotypes typically generate patterns with only 10 – 30 discernible bands and thus do not provide reasonable estimates of community diversity and composition for all but relatively simple communities (Fig. 7.2). Multiple ribosomal RNA operons contribute multiple bands for some organisms, particularly with RISA, while different sequences from distinct organisms commonly migrate to identical positions (Kirk et al. 2004), all of which further complicate pattern interpretations. Also, while bands can be excised from gels and sequenced, the technical difficulty of this process and the presence of multiple OTUs in single bands (Casamayor et al. 2000; Zhang et al. 2005) hinder the feasibility of this approach. The 16S rRNA gene

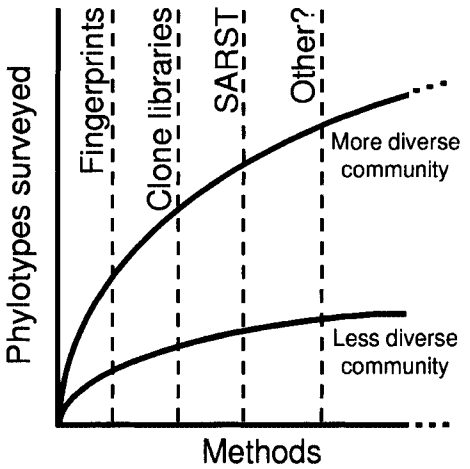


Fig. 7.2. Relationship between methodologies (*dotted lines*) and phylogenetic diversity coverage. Rarefaction curves are shown for two hypothetical communities, indicating that the depth of coverage (phylotypes surveyed) obtained with a given methodology differs, depending on the overall magnitude of the phylogenetic diversity

fingerprint methods as described above provide an indication of the phylogenetic diversity of the most predominant organisms within communities, overall measures of community similarity, a high sample throughput (Fig. 7.1), and an estimate of which of the relatively abundant OTUs respond to specific environmental factors.

7.8 Sequencing

There are several approaches for sequencing phylogenetic genes from environmental nucleic acids. These methods usually involve either collecting 16S rRNA genes from clone libraries, modifying clone library generation for greater sequence throughput, or analyzing metagenomic libraries (Fig. 7.1). Unlike fingerprinting methods, sequence-based analyses are labor-intensive and poorly amenable to the analysis of multiple samples with replication.

7.8.1 16S rRNA Gene Libraries

Cloning and sequencing PCR products that contain fragments of the 16S rRNA gene provides information for both the phylogenetic identity, and to some extent, the relative abundance of community OTUs. Ironically, while the sample-throughput of sequence-based methods is limited, comparisons between samples are greatly facilitated by the type of data obtained. The advantage of sequence-based techniques over gel fingerprints is that each sequence has associated phylogenetic information that may be directly related to other samples, and the depth of coverage of a community is correlated with the number of sequences obtained (Fig. 7.2). Such sequence data can easily be stored in databases and readily compared, with little ambiguity, to data from other studies. This is in sharp contrast to fingerprint data, which is challenging to compare even within studies. Cloned inserts from *Escherichia coli* transformants can either be sequenced directly or subjected to an initial screen to group (de-replicate) similar inserts. Grouping related inserts minimizes the number of sequencing reactions required for characterizing cloned 16S rRNA gene sequences. Fingerprinting methods have been used, such as ARDRA (to a large extent) and DGGE (to a lesser extent), to screen clone libraries prior to sequencing representative OTUs. Furthermore, a variety of hybridization-based screening methods (Liesack and Stackebrandt 1992; Ravenschlag et al. 1999; Schramm et al. 2002; Snaidr et al. 1997; Valinsky et al. 2002b) have been used to reduce the number of clones that require sequencing.

During the 1990s, clone libraries generated from environmental samples were predominantly aimed at describing diversity and discovering novel phylogenetic groups in a variety of environments (Morris et al. 2002). These experiments generated libraries from single samples from soils (Borneman et al. 1996; Liesack and Stackebrandt 1992; Stackebrandt et al. 1993; Zhou et al. 1997), activated sludge (Blackall et al. 1998), and landfill soils (Lloyd-Jones and Lau 1998). The value of comparing multiple samples has become evident; and 16S rRNA gene libraries are now more frequently generated from multiple samples that are related by location or treatment. Comparing the composition and diversity of samples from different treatments provides a more robust approach for elucidating the impact of environmental factors on community composition. For studies of microbial ecology focusing on soil environments, multiple 16S rRNA gene clone libraries have now gauged the impact of soil type (Girvan et al. 2003; Zhou et al. 2002), plant cover (Dunbar et al. 1999; Kuske et al. 2002), time (Lipson and Schmidt 2004), and human disturbance (Chow et al. 2002; McCaig et al. 1999) on bacterial community composition.

As discussed above, a major limitation in applying clone library methodology to the analysis of microbial communities is that the high diversity common to most environments precludes an adequate description of subdominant populations and statistically valid comparisons of diversity. Borneman and Triplett (1997) discovered the worst case scenario by sequencing 50 clones from each of two Amazonian soils and found not a single duplicate sequence. Zhou and coworkers (1997) also found maximum diversity among 43 clones screened from a Siberian tundra sample. The occurrence of 16S rRNA gene libraries replete with only singletons prompted Lunn and coworkers (2004) to generate statistical diversity predictions suitable for libraries without duplicates. Most clone libraries surveyed by a recent report were considered insufficiently sampled for adequate coverage of the environmental sample being characterized. The majority of libraries that were judged to be "sufficiently sampled" for generating stable richness estimates were derived from aquatic environments (Kemp and Aller 2004), which are predicted to harbor low relative diversity (Hagstrom et al. 2002). Cost and labor limitations often preclude sufficient sampling of clone libraries to enable the detection of significant differences in diversity. While most studies limit sampling to several hundred clones per sample at most, perhaps thousands (Tiedje et al. 1999) or tens of thousands (Dunbar et al. 2002; Schloss and Handelsman 2005) of clones must be sequenced for reliable comparisons of complex communities, such as those found in soils and sediments.

7.8.2 Serial Analysis of Ribosomal Sequence Tags

Recent molecular innovations have increased the throughput with which complex mixtures of nucleic acids are characterized. For example, serial analysis of gene expression (SAGE) was developed to profile the expression of mRNA in eukaryotic cells (Velculescu et al. 1995). By concatenating short 14-bp portions of mRNA transcripts known as expressed sequence tags (ESTs), each sequencing reaction generates data from many transcripts. SAGE methodology was recently modified to enable the concatenation of short and variable regions of the 16S rRNA gene from bacterial communities (Neufeld et al. 2004a, b). Serial analysis of ribosomal sequence tags (SARST) generated an average of between five and ten ribosomal sequence tags (RSTs) from each sequencing reaction from soils and with high reproducibility (Neufeld et al. 2004b). Large datasets of RSTs recently enabled the comparison of soil diversity, providing evidence that arctic tundra can harbor higher phylogenetic diversity than forest soils (Neufeld and Mohn 2005b). Such short variable regions restrict the phylogenetic specificity of RSTs and also prevent phylogenetic and statistical analyses that rely on sequence alignments (Schloss and Handelsman 2005; Singleton et al. 2001). However, RST sequence surveys have facilitated direct comparisons of OTU composition and diversity within complex communities to an extent unparalleled by the traditional analysis of clone libraries (Neufeld and Mohn 2005b). Further, sequencing of longer portions of the 16S rRNA gene is possible by developing primers specific to RSTs of interest and coupling them with other universal primers further downstream, as was recently demonstrated for SARST (Neufeld et al. 2004b) and DGGE (Höfle et al. 2005).

The quantity of phylogenetic information assessed by SARST increases the coverage efficiency with which microbial communities are sequenced (Fig. 7.2) and in some cases may enable complete sampling of phylogenetic diversity from environmental DNA extracts. However, SARST is limited in the number of samples that may be processed simultaneously, since SARST methodology requires multiple time-consuming steps. Recent variations of SARST, such as SARST-V6 and iSARST, have modified the protocol to help reduce the time and effort leading to sequencing of concatamer inserts. These two modifications have further demonstrated the utility of this approach by generating large RST libraries from bacterial communities in hydrothermal vents (Kysela et al. 2005) and from bovine rumens (Yu et al. 2005), respectively. As with SAGE (Ruijter et al. 2002), lack of replication unfortunately results in unanswered questions about experimental and biological variability. Toward circumventing limitations of sample throughput, future improvements to SARST should focus on facili-

tating the simultaneous preparation of environmental samples, from PCR to sequencing, perhaps in 96-well microtiter plates.

7.9 Metagenomics

Metagenome analysis is an increasingly popular approach for studying nucleic acids from the environment. By cloning DNA directly, without prior amplification of particular genes, the bias associated with PCR (but not that associated with DNA extraction or cloning) is avoided, and phylogenetic information may be coupled with genomic information from uncultured organisms. Metagenomics provides an unparalleled ability to link physiological roles with the myriad organisms previously recognized only by 16S rRNA gene sequences. Recent approaches have screened large cloned inserts in cosmid, fosmid or BAC libraries for phylogenetic markers prior to sequencing the genes flanking these markers (Leveau et al. 2004; Liles et al. 2003; Sebat et al. 2003). However, we do not summarize these studies here since they do not offer efficient phylogenetic assessment, and excellent reviews have recently been published by Jo Handelsman (2004) and elsewhere in this book (Chap. 8). An alternative metagenomic methodology involves the random shotgun sequencing of extracted environmental DNA. High-throughput sequencing of random cloned inserts (ca. 3 kb) from each library generates large numbers of insert sequences for the analysis of metabolic and phylogenetic diversity of environmental samples. Presumably, the smaller clone insert size greatly reduces DNA extraction and cloning biases. This approach was recently applied to acid mine drainage biofilms (Tyson et al. 2004) and planktonic organisms from the Sargasso Sea (Venter et al. 2004). Together, these two ambitious studies performed over 2×10^6 sequencing reactions and collected more than 10^9 bases of non-redundant DNA sequence. Venter and coworkers (2004) demonstrated the use of several phylogenetic markers in addition to 16S rRNA for gauging the phylogenetic diversity of their samples. Using these markers, they estimated that their sequence dataset contained genes from approximately 450 unique bacterial species. Further, by applying multiple richness estimators, the marine samples they collected were predicted to contain potentially more than 1,000 bacterial species. By microbial community standards, this is modest diversity, especially given that over half of the sequence data appeared to originate from a single organism. However, even for this "simple" community, a strong and unparalleled advantage of the metagenomic approach is that information about metabolic diversity, encoded by many megabases of microbial genomes, is available in the resulting sequence data. Bioinformatic approaches enabled the association of specific phy-

logenetic groups with physiological capabilities, offering insight into the community structure (phylogenetic and physiological diversity) of the marine environment. For example, Venter and coworkers (2004) confirmed a wide distribution and diversity among the marine bacteria of rhodopsins for harvesting solar energy. Using a similar approach, Tyson and coworkers (2004) helped uncover key nutrient-cycling genes associated with specific biofilm inhabitants.

Unfortunately, enormous sequencing efforts and costs involved in metagenomic approaches are prohibitive for most laboratories, even for generating data from a single sample. The lack of amenability to the analysis of multiple samples will continue to limit widespread adoption of this approach and will likely inhibit hypothesis testing of the ecological role of individual organisms within ecosystems and the interactions between organisms, both of which are important for microbial ecology. Despite these limitations, as the cost of sequencing continues to decrease, metagenomics provides an increasingly practical means by which phylogenetic and functional diversity is explored in the environment. There is arguably no better means by which environmental community structure may be studied than with methods that generate sequence data linking phylogeny and physiology.

7.10 Array Technology

Sequencing-based approaches provide phylogenetic information about individual communities but are not readily amenable to the comparison of multiple samples with replication. Gel fingerprints rapidly compare multiple samples and provide a rough similarity measure, but do not readily provide phylogenetic information related to community composition. Array technology shows great promise in microbial ecology, since it potentially offers the advantages of both sequencing and fingerprinting methodologies: simultaneous quantitation and comparison of many phylotypes and functional guilds in multiple samples (Fig. 7.1). In principle, tens of thousands of targets (probes) can be simultaneously assayed. Arrays share with fingerprinting the advantage of high sample throughput but offer the added advantages of more quantitative data on phylotypes, resolution of more phylotypes, more readily manageable and comparable data, and potentially greater sensitivity. Greater reliability may be obtained by including multiple (redundant) probes for individual OTUs and using differentially labeled reference samples as internal standards for hybridization. The reference samples may be used to facilitate the normalized comparison of multiple samples. A major advantage of the microarray method over traditional fingerprinting methods is the ease of unambiguous comparisons

between samples (i. e. no need to determine if bands match). The potential advantages of array technology provide an enticing goal for many investigators of microbial ecology.

The use of arrays in microbial ecology is still in development and has not progressed far beyond the level of initial testing on defined mixtures of known targets. Microarrays have been optimized for eventual environmental studies of catabolic genes (Bodrossy et al. 2003; Dennis et al. 2003; Rhee et al. 2004; Wu et al. 2001) and phylogenetic markers (Castiglioni et al. 2004; Loy et al. 2002; Peplies et al. 2003, 2004; Small et al. 2001; Wilson et al. 2002). However, few studies have successfully monitored microbial communities with microarrays. Initial surveys have included the analyses of methane and ammonia-oxidizing populations in landfill covers (Stralis-Pavese et al. 2004), nitrogen-cycling communities in river sediment (Taroncher-Oldenburg et al. 2003), and ribosomal RNA from estuarine sediment (El Fantroussi et al. 2003). Poor sensitivity and unknown specificity of probe sets are current limitations for studying environmental mixtures of nucleic acids, and recent reviews have covered these limitations in detail (Cook and Sayler 2003; Kelly 2003; Zhou and Thompson 2002). One way to increase sensitivity and decrease non-specific hybridization is to probe PCR-amplified targets. With present technology, amplified targets are probably necessary for the microarray analysis of most complex communities (Cook and Sayler 2003), in which total DNA exacerbates cross-hybridization and most individual populations are undetectable. The tradeoff for this approach is the potential introduction of PCR bias.

A further limitation of array methods is that one must know *a priori* which phylotypes will be examined. Further, a hybridization probe is only useful if its desired specificity matches its actual specificity. For environmental samples, the unknown and diverse assemblage of organisms may preclude the application of arrays containing probes developed using available database sequences. Typically, the probes must either be produced by amplifying DNA from the environment of interest (e. g. cloned rRNA gene fragments) or must be designed (e. g. on the basis of rRNA sequences from the environment of interest) and synthesized. Synthetic probes offer the advantage of enabling bioinformatic approaches to maximize the specificity of probes, essentially as is currently done for the development of arrays for transcriptomic analysis of single genomes. Arrays are not appropriate for exploring the diversity of a new environment, but they are ideal for characterizing the spatial and temporal patterns of diversity within an environment, following some preliminary assessment, such as rRNA gene clone library analysis. We recently used SARST to collect a large number of 16S rRNA gene sequences from a composite of eight soil samples related by location and treatment (Neufeld et al. 2005). The collected sequences were used for designing a series of habitat-specific probes for

comparing the set of individual soil samples. The microarray successfully functioned as a high-throughput fingerprinting technique, since the microarray clustering of samples agreed with DGGE fingerprint clustering. The probe signal diversity decreased with increasing pollutant concentration, and the microarray enabled rapid identification of the probe OTUs for which hybridization signals correlated with pollutant contamination.

Other novel array applications provide alternative means by which community diversity may be assessed. By modifying the typical probe and target hybridization strategy, Valinsky and coworkers (2002b, 2004) combined array technology with clone library approaches to generate a method known as oligonucleotide fingerprinting of rRNA genes (OFRG). This approach involved placing individual ribosomal DNA clones in an array format for subsequent screening with a strategic set of oligonucleotide probes. Bioinformatics-based approaches identified a set of probes for which the signal intensities generated by the set of all hybridizations would provide phylogenetic identifications for all clones on the array. Over two dozen probes were designed for the strategic identification of both bacteria (Valinsky et al. 2002b) and fungi (Valinsky et al. 2002a). The set of signal intensities for hybridization of each clone to each probe are theoretically capable of high phylogenetic resolution, providing a means by which the diversity represented by clone libraries may be rapidly assessed. OFRG was used to identify specific bacterial (Yin et al. 2003a) and fungal (Yin et al. 2003b) phylotypes associated with soils that suppressed infection by a parasitic nematode. Currently, OFRG involves using Nylon membranes for printing 16S rRNA clone arrays, which may limit the number of phylotypes that may be surveyed with this approach. However, many thousands of phylotypes may be screened rapidly by combining the OFRG technique with higher density spotting (Borneman, personal communication), or possibly, with a glass slide microarray format.

7.11 Composite Methodologies

Just as replication is critical for robust experimental design, multiple methodologies produce a more conclusive and complete picture of community diversity and population dynamics. Methodologies can be combined to produce a more powerful analysis, strengthen observations, or provide complementary observations for the same experiment. A few recent examples follow. Sigler and Zeyer (2002) examined the diversity of bacteria along the forefields of receding glaciers. They demonstrated similar band OTU diversities for samples analyzed with both DGGE and RISA, thus confirming their observations with these two fingerprinting methods. Holben et al.

(2004) used GC fractionation of PCR products prior to DGGE and demonstrated that this additional methodology could reveal minority band OTUs otherwise obscured and facilitate band sequencing. This novel combination of methods improved the phylogenetic depth with which DGGE could profile chicken digesta community diversity. As an example of combined fingerprinting and sequencing methodologies, Noll and coworkers (2005) analyzed bacterial succession in a flooded rice paddy soil. They profiled 16S rRNA (RNA) and 16S rRNA genes (DNA) using T-RFLP of all samples and prepared small clone libraries from a representative subset of samples. Statistical analyses of T-RFLP patterns demonstrated an impact of time and oxygen on community structure after flooding and the establishment of stable communities after three weeks. Clone libraries helped provide phylogenetic information for predominant phylotypes from each stage of succession, suggesting *r*- and then *K*-selected populations following paddy flooding. Finally, community changes were more clearly resolved with the RNA analysis, demonstrating increased sensitivity of this dynamic ribosomal RNA molecule.

Combined methodologies are clearly the most powerful approach toward understanding microbial community diversity and ecology. The unique advantages and disadvantages inherent in sequence- and fingerprint-based methods (Fig. 7.1) make combinations of these approaches particularly helpful. Much as combining colors from opposite sides of the color-wheel provides clear contrasts, combining sequencing approaches with fingerprint analysis generates datasets with complementary information. Future approaches in microbial ecology should continue to employ multiple existing methods in conjunction with sample replication for more "holistic" insight into microbial diversity in the environment.

7.12 Conclusion

The critical importance of microbial diversity and the terrific challenges in understanding this topic have driven researchers for more than two decades to develop creative approaches to analyze environmental nucleic acids. Of the many nucleic acid-based approaches now available, most can either be characterized by the ability either to rapidly generate fingerprint profiles from multiple samples or to efficiently determine multiple sequences from a select few samples. Array technology potentially bridges the gap between the two approaches, providing the benefits of both. Arrays have yet to be demonstrated as widely applicable and practical, but this is an area to watch for emerging, powerful new methods. Each of the particular nucleic acid-based methods has unique characteristics (e. g. biases, level of phylogenetic

resolution) that tailor it for particular applications. Thus, we are now at a point where researchers can select from many options, using one or more methods to address specific research aims. As in other fields, studies of microbial ecology are strengthened by replication and the use of multiple approaches. It is becoming routine to use greater replication than in the past and to employ multiple methods in a single study. We have barely begun to appreciate the basic phylogenetic and physiological diversity of most microbial communities, and there remains a vast amount to learn about the relationships between organisms and their environments. We are indeed witness to the Golden Age of microbial ecology, continually supported by novel molecular approaches and fueled by the thrill of discovery.

References

Amann RI, Ludwig W, Schleifer K-H (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Avaniss-Aghajani E, Jones K, Chapman D, Brunk C (1994) A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. Biotechniques 17:144–146

Bano N, Hollibaugh JT (2000) Diversity and distribution of DNA sequences with affinity to ammonia-oxidizing bacteria of the β subdivision of the class *Proteobacteria* in the Arctic Ocean. Appl Environ Microbiol 66:1960–1969

Becker S, Boger P, Oehlmann R, Ernst A (2000) PCR bias in ecological analysis: a case study for quantitative *Taq* nuclease assays in analyses of microbial communities. Appl Environ Microbiol 66:4945–4953

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) GenBank. Nucleic Acids Res 28:15–18

Besemer K, Moeseneder MM, Arrieta JM, Herndl GJ, Peduzzi P (2005) Complexity of bacterial communities in a river-floodplain system (Danube, Austria). Appl Environ Microbiol 71:609–620

Blackall LL, Burrell PC, Gwilliam H, Bradford D, Bond PL, Hugenholtz P (1998) The use of 16S rDNA clone libraries to describe the microbial diversity of activated sludge communities. Water Sci Technol 37:451–454

Blank CE, Cady SL, Pace NR (2002) Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone national park. Appl Environ Microbiol 68:5123–5135

Bodrossy L, Stralis-Pavese N, Murrell JC, Radajewski S, Weilharter A, Sessitsch A (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. Environ Microbiol 5:566–582

Bohannan B, Hughes J (2003) New approaches to analyzing microbial biodiversity data. Curr Opin Microbiol 6:282–287

Borneman J, Triplett EW (1997) Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. Appl Environ Microbiol 63:2647–2653

Borneman J, Skroch PW, O'Sullivan KM, Palus JA, Rumjanek NG, Jansen JL, Nienhuis J, Triplett EW (1996) Molecular microbial diversity of an agricultural soil in Wisconsin. Appl Environ Microbiol 62:1935–1943

Casamayor EO, Schäfer H, Bañeras L, Pedrós-Alió C, Muyzer G (2000) Identification of and spatio-temporal differences between microbial assemblages from two neighboring sulfurous lakes: Comparison by microscopy and denaturing gradient gel electrophoresis. Appl Environ Microbiol 66:499–508

Castiglioni B, Rizzi E, Frosini A, Sivonen K, Rajaniemi P, Rantala A, Mugnai MA, Ventura S, Wilmotte A, Boutte C, Grubisic S, Balthasart P, Consolandi C, Bordoni R, Mezzelani A, Battaglia C, De Bellis G (2004) Development of a universal microarray based on the ligation detection reaction and 16S rRNA gene polymorphism to target diversity of cyanobacteria. Appl Environ Microbiol 70:7161–7172

Chao A (1984) Non-parametric estimation of the number of classes in a population. Scand J Stat 11:265–270

Chao A, Lee S-M (1992) Estimating the number of classes via sample coverage. J Am Stat Assoc 87:210–217

Chase JM, Leibold MA (2002) Spatial scale dictates the productivity–biodiversity relationship. Nature 416:427–430

Chow ML, Radomski CC, McDermott JM, Davies J, Axelrood PE (2002) Molecular characterization of bacterial diversity in Lodgepole pine (*Pinus contorta*) rhizosphere soils from British Columbia forest soils differing in disturbance and geographic source. FEMS Microbiol Ecol 42:347–357

Coeyne T, Vandamme P (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett 228:45–49

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31:442–443

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. Phil Trans R Soc Lond B 345:101–118

Connon SA, Giovannoni SJ (2002) High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol 68:3878–3885

Cook KL, Sayler GS (2003) Environmental application of array technology: promise, problems and practicalities. Curr Opin Biotechnol 14:311–318

Curtis TP, Sloan WT (2004) Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. Curr Opin Microbiol 7:221–226

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99:10494–10499

Dahllöf I, Baillie H, Kjelleberg S (2000) *rpoB*-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. Appl Environ Microbiol 66:3376–3380

Delbes C, Moletta R, Godon JJ (2000) Monitoring of activity dynamics of an anaerobic digester bacterial community using 16S rRNA polymerase chain reaction-single-strand conformation polymorphism analysis. Environ Microbiol 2:506–515

Dennis P, Edwards EA, Liss SN, Fulthorpe R (2003) Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl Environ Microbiol 69:769–778

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Dunbar J, Takala S, Barns SM, Davis JA, Kuske CR (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. Appl Environ Microbiol 65:1662–1669

Dunbar J, Ticknor LO, Kuske CR (2000) Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis. Appl Environ Microbiol 66:2943–2950

Dunbar J, Ticknor LO, Kuske CR (2001) Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities. Appl Environ Microbiol 67:190–197

Dunbar J, Barns SM, Ticknor LO, Kuske CR (2002) Empirical and theoretical bacterial diversity in four Arizona soils. Appl Environ Microbiol 68:3035–3045

Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie van Leeuwenhoek 73:25–33

Egert M, Friedrich MW (2003) Formation of pseudo-terminal restriction fragments, a PCR-related bias affecting terminal restriction fragment length polymorphism analysis of microbial community structure. Appl Environ Microbiol 69:2555–2562

El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, Yershov GM, Stahl DA (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. Appl Environ Microbiol 69:2377–2382

Farnleitner AH, Kreuzinger N, Kavka GG, Grillenberger S, Rath J, Mach RL (2000) Comparative analysis of denaturing gradient gel electrophoresis and temporal temperature gradient gel electrophoresis in separating *Escherichia coli uidA* amplicons differing in single base substitutions. Lett Appl Microbiol 30:427–431

Farrelly V, Rainey F, Stackebrandt E (1995) Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl Environ Microbiol 61:2798–2801

Felske A, Wolterink A, van Lis R, Akkermans AD (1998) Phylogeny of the main bacterial 16S rRNA sequences in Drentse A grassland soils (The Netherlands). Appl Environ Microbiol 64:871–879

Ferrari VC, Hollibaugh JT (1999) Distribution of microbial assemblages in the Central Arctic Ocean basin studied by PCR/DGGE: analysis of a large data set. Hydrobiologia 401:55–68

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science 296:1061–1063

Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65:4630–4636

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27:4218–4222

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner R, Magrum L, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457–463

Fox GE, Wisotzkey J, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170

Fromin N, Hamelin J, Tarnawski S, Roesti D, Jourdain-Miserez K, Forestier N, Teyssier-Cuvelle S, Gillet F, Aragno M, Rossi P (2002) Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns. Environ Microbiol 4:634–643

Giovannoni SJ, DeLong EF, Olsen GJ, Pace NR (1988) Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells. J Bacteriol 170:720–726

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60–63

Girvan MS, Bullimore J, Pretty JN, Osborn AM, Ball AS (2003) Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. Appl Environ Microbiol 69:1800–1809

Girvan MS, Campbell CD, Killham K, Prosser JI, Glover LA (2005) Bacterial diversity promotes community stability and functional resilience after perturbation. Environ Microbiol 7:301–313

Gonzalez N, Romero J, Espejo RT (2003) Comprehensive detection of bacterial populations by PCR amplification of the 16S-23S rRNA spacer region. J Microbiol Methods 55:91–97

Graham DE, Overbeek R, Olsen GJ, Woese CR (2000) An archaeal genomic signature. Proc Natl Acad Sci USA 97:3304–3308

Green JL, Holmes AJ, Westoby M, Oliver I, Briscoe D, Dangerfield M, Gillings M, Beattie AJ (2004) Spatial scaling of microbial eukaryote diversity. Nature 432:747–750

Griffiths BS, Kuan HL, Ritz K, Glover LA, McCaig AE, Fenwick C (2004) The relationship between microbial community structure and functional stability, tested experimentally in an upland pasture soil. Microb Ecol 47:104–113

Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ (2003) Influence of depth and sampling time on bacterial community structure in an upland grassland soil. FEMS Microbiol Ecol 43:35–43

Grundmann GL (2004) Spatial scales of soil bacterial diversity – the size of a clone. FEMS Microbiol Ecol 48:119–127

Gutell RR, Larsen N, Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol Rev 58:10–26

Hackl E, Zechmeister-Boltenstern S, Bodrossy L, Sessitsch A (2004) Comparison of diversities and compositions of bacterial populations inhabiting natural forest soils. Appl Environ Microbiol 70:5057–5065

Hagstrom A, Pommier T, Rohwer F, Simu K, Stolte W, Svensson D, Zweifel UL (2002) Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. Appl Environ Microbiol 68:3628–3633

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685

Hill JE, Seipp RP, Betts M, Hawkins L, van Kessel AG, Crosby WL, Hemmingsen SM (2002) Extensive profiling of a complex microbial community by high-throughput sequencing. Appl Environ Microbiol 68:3055–3066

Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM (2004) cpnDB: a chaperonin sequence database. Genome Res 14:1669–1675

Hill TCJ, Walsh KA, Harris JA, Moffett BF (2003) Using ecological diversity measures with bacterial communities. FEMS Microbiol Ecol 43:1–11

Höfle MG, Flavier S, Christen R, Bötel J, Labrenz M, Brettar I (2005) Retrieval of nearly complete 16S rRNA gene sequences from environmental DNA following 16S rRNA-based community fingerprinting. Environ Microbiol 7:670–675

Holben WE, Feris KP, Kettunen A, Apajalahti JHA (2004) GC fractionation enhances microbial community diversity assessment and detection of minority populations of bacteria by denaturing gradient gel electrophoresis. Appl Environ Microbiol 70:2263–2270

Holben WE, Harris D (1995) DNA-based monitoring of total bacterial community structure in environmental samples. Mol Ecol 4:627–631

Horner-Devine MC, Leibold MA, Smith VH, Bohannan BJM (2003) Bacterial diversity patterns along a gradient of primary productivity. Ecol Lett 6:613–622

Horner-Devine MC, Carney KM, Bohannan BJM (2004a) An ecological perspective on bacterial biodiversity. Proc R Soc Lond Ser B Biol Sci 271:113–122

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM (2004b) A taxa–area relationship for bacteria. Nature 432:750–753

Hugenholtz P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int J Syst Evol Microbiol 53:289–293

Hugenholtz P, Pace NR (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol 14:190–197

Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998) Novel division level bacterial diversity in a Yellowstone hot spring. J Bacteriol 180:366–376

Hughes AR, Stachowicz JJ (2004) Genetic diversity enhances the resistance of a seagrass ecosystem to disturbance. Proc Natl Acad Sci USA 101:8998–9002

Hughes JB, Bohannan BJM (2004) Application of ecological diversity statistics in microbial ecology. In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds) Molecular microbial ecology manual, 2nd edn. Kluwer, London, pp 1321–1344

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 67:4399–4406

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Janse I, Bok J, Zwart G (2004) A simple remedy against artifactual double bands in denaturing gradient gel electrophoresis. J Microbiol Methods 57:279–281

Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M (2002) Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions *Acidobacteria*, *Actinobacteria*, *Proteobacteria*, and *Verrucomicrobia*. Appl Environ Microbiol 68:2391–2396

Jaspers E, Overmann J (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol 70:4831–4839

Jensen MA, Webster JA, Straus N (1993) Rapid identification of bacteria on the basis of polymerase chain reaction-amplified ribosomal DNA spacer polymorphisms. Appl Environ Microbiol 59:945–952

Joseph SJ, Hugenholtz P, Sangwan P, Osborne CA, Janssen PH (2003) Laboratory cultivation of widespread and previously uncultured soil bacteria. Appl Environ Microbiol 69:7210–7215

Kelly JJ (2003) Molecular techniques for the analysis of soil microbial processes: functional gene analysis and the utility of DNA microarrays. Soil Sci 168:597–605

Kemp PF, Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. FEMS Microbiol Ecol 47:161–177

Kent AD, Smith DJ, Benson BJ, Triplett EW (2003) Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities. Appl Environ Microbiol 69:6768–6776

Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H, Trevors JT (2004) Methods of studying soil microbial diversity. J Microbiol Methods 58:169–188

Klappenbach JA, Dunbar JM, Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328–1333

Kopczynski ED, Bateson MM, Ward DM (1994) Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. Appl Environ Microbiol 60:746–748

Kroes I, Lepp PW, Relman DA (1999) Bacterial diversity within the human subgingival crevice. Proc Natl Acad Sci USA 96:14547–14552

Kropf S, Heuer H, Gruning M, Smalla K (2004) Significance test for comparing complex microbial community fingerprints using pairwise similarity measures. J Microbiol Methods 57:187–195

Kuske CR, Ticknor LO, Miller ME, Dunbar JM, Davis JA, Barns SM, Belnap J (2002) Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in an arid grassland. Appl Environ Microbiol 68:1854–1863

Kysela DT, Palacios C, Sogin ML (2005) Serial analysis of V6 ribosomal sequence tags (SARST-V6). Environ Microbiol 7:356–364

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82:6955–6959

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 95:9413–9417

Leckie SE, Prescott CE, Grayston SJ, Neufeld JD, Mohn WW (2004) Characterization of humus microbial communities in adjacent forest types that differ in nitrogen availability. Microb Ecol 48:29–40

Leclerc M, Delgenes JP, Godon JJ (2004) Diversity of the archaeal community in 44 anaerobic digesters as determined by single strand conformation polymorphism analysis and 16S rDNA sequencing. Environ Microbiol 6:809–819

Lee D, Zo Y, Kim S (1996) Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. Appl Environ Microbiol 62:3112–3120

Leveau JHJ, Gerards S, de Boer W, van Veen JA (2004) Phylogeny-function analysis of (meta)genomic libraries: screening for expression of ribosomal RNA genes by large-insert library fluorescent in situ hybridization (LIL-FISH). Environ Microbiol 6:990–998

Liesack W, Stackebrandt E (1992) Occurrence of novel groups of the domain Bacteria as revealed by analysis of genetic material isolated from an Australian terrestrial environment. J Bacteriol 174:5072–5078

Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl Environ Microbiol 69:2684–2691

Lipson DA, Schmidt SK (2004) Seasonal changes in an alpine soil bacterial community in the colorado rocky mountains. Appl Environ Microbiol 70:2867–2879

Liu W, Marsh T, Cheng H, Forney L (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol 63:4516–4522

Lloyd-Jones G, Lau PCK (1998) A molecular view of microbial diversity in a dynamic landfill in Québec. FEMS Microbiol Lett 162:219–226

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer KH, Wagner M (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl Environ Microbiol 68:5064–5081

Lunn M, Sloan WT, Curtis TP (2004) Estimating bacterial diversity from clone libraries with flat rank abundance distributions. Environ Microbiol 6:1081–1085

Marsh TL, Saxman P, Cole J, Tiedje J (2000) Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. Appl Environ Microbiol 66:3616–3620

Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl Environ Microbiol 68:3673–3682

Martinez-Murcia A, Benlloch S, Collins M (1992) Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. Int J Syst Bacteriol 42:412–421

McCaig AE, Glover LA, Prosser JI (1999) Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. Appl Environ Microbiol 65:1721–1730

McCann KS (2000) The diversity–stability debate. Nature 405:228–233

Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Appl Environ Microbiol 65:3518–3525

Morris CE, Bardin M, Berge O, Frey-Klett P, Fromin N, Girardin H, Guinebretière M-H, Lebaron P, Thiéry JM, Troussellier M (2002) Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999. Microbiol Mol Biol Rev 66:592–616

Mulder CP, Uliassi DD, Doak DF (2001) Physical stress and diversity–productivity relationships: the role of positive interactions. Proc Natl Acad Sci USA 98:6704–6708

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. Curr Opin Microbiol 2:317–322

Muyzer G, Smalla K (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie Van Leeuwenhoek 73:127–141

Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol 59:695–700

Muyzer G, Brinkhoff T, Nübel U, Santegoeds C, Schäfer H, Wawer C (2004) Denaturing gradient gel electrophoresis (DGGE) in microbial ecology. In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds) Molecular microbial ecology manual, 2nd edn. Kluwer, London

Myers R, Fischer S, Lerman L, Maniatis T (1985) Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis. Nucleic Acids Res 13:3131–3145

Nazaret S, Brothier E, Ranjard L (2003) Shifts in diversity and microscale distribution of the adapted bacterial phenotypes due to Hg(II) spiking in soil. Microb Ecol 45:259–269

Nee S (2003) Unveiling prokaryotic diversity. Trends Ecol Evol 18:62–63

Neufeld JD, Mohn WW (2005a) Fluorophore-labeled primers improve the sensitivity, versatility and normalization of denaturing gradient gel electrophoresis (DGGE). Appl Environ Microbiol 71:4893–4896

Neufeld JD, Mohn WW (2005b) Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils revealed with serial analysis of ribosomal sequence tags (SARST). Appl Environ Microbiol 71:5710–5718

Neufeld JD, Driscoll BT, Knowles R, Archibald FS (2001) Quantifying functional gene populations: comparing gene abundance and corresponding enzymatic activity using denitrification and nitrogen fixation in pulp and paper mill effluent treatment systems. Can J Microbiol 47:925–934

Neufeld JD, Yu Z, Lam W, Mohn WW (2004a) SARST, serial analysis of ribosomal sequence tags. In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds) Molecular microbial ecology manual, 2nd edn. Kluwer, London, pp 543–568

Neufeld JD, Yu Z, Lam W, Mohn WW (2004b) Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities. Environ Microbiol 6:131–144

Neufeld JD, Mohn WW, de Lorenzo V(2005) Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. Environ Microbiol (in press)

Noll M, Matthies D, Frenzel P, Derakshani M, Liesack W (2005) Succession of bacterial community structure and diversity in a paddy soil oxygen gradient. Environ Microbiol 7:382–395

Nübel U, Engelen B, Felske A, Snaidr J, Wieshuber A, Amann R, Ludwig W, Backhaus H (1996) Sequence heterogeneities of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J Bacteriol 178:5636–5643

Nübel U, Garcia-Pichel F, Kühl M, Muyzer G (1999a) Quantifying microbial diversity: morphotyoes, 16S rRNA genes, and carotenoids of oxygenic phototrophs in microbial mats. Appl Environ Microbiol 65:422–430

Nübel U, Garcia-Pichel F, Kühl M, Muyzer G (1999b) Spatial scale and the diversity of benthic cyanobacteria and diatoms in a salina. Hydrobiologia 401:199–206

Ogier J-C, Son O, Gruss A, Tailliez P, Delacroix-Buchet A (2002) Identification of the bacterial microflora in dairy products by temporal temperature gradient gel electrophoresis. Appl Environ Microbiol 68:3691–3701

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40:337–365

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86:2766–2770

Osborn AM, Moore ERB, Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ Microbiol 2:39–50

Palleroni NJ (1997) Prokaryotic diversity and the importance of culturing. Antonie van Leeuwenhoek 72:3–19

Palys T, Nakamura LK, Cohan FM (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol 47:1145–1156

Peixoto RS, da Costa Coutinho HL, Rumjanek NG, Macrae A, Rosado AS (2002) Use of *rpo*B and 16S rRNA genes to analyse bacterial diversity of a tropical soil using PCR and DGGE. Lett Appl Microbiol 35:316–320

Peplies J, Glockner FO, Amann R (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl Environ Microbiol 69:1397–1407

Peplies J, Lau SC, Pernthaler J, Amann R, Glockner FO (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. Environ Microbiol 6:638–645

Peters S, Koschinsky S, Schwieger F, Tebbe CC (2000) Succession of microbial communities during hot composting as detected by PCR-single-strand-conformation polymorphism-based genetic profiles of small-subunit rRNA genes. Appl Environ Microbiol 66:930–936

Petersen DG, Dahllöf I (2005) Improvements for comparative analysis of changes in diversity of microbial communities using internal standards in PCR-DGGE. FEMS Microbiol Ecol 53:339–348

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64:3724–3730

Porteous LA, Seidler RJ, Watrud LS (1997) An improved method for purifying DNA from soil for polymerase chain reaction amplification and molecular ecology applications. Mol Ecol 6:787–791

Powell SM, Bowman JP, Snape I, Stark JS (2003) Microbial community variation in pristine and polluted nearshore Antarctic sediments. FEMS Microbiol Ecol 45:135–145

Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, Zhou J (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol 67:880–887

Ragan MA (2001) Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev 11:620–626

Rainey FA, Ward N, Sly LI, Stackebrandt E (1994) Dependence on the taxon composition of clone libraries for PCR amplified, naturally occuring 16S rRNA, on the primer pair and the cloning system. Experientia 50:796–797

Ranjard L, Nazaret S, Gourbière F, Thioulouse J, Linet P, Richaume A (2000a) A soil microscale study to reveal the heterogeneity of Hg(II) impact on indigenous bacteria by quantification of adapted phenotypes and analysis of community DNA fingerprints. FEMS Microbiol Ecol 31:107–115

Ranjard L, Poly F, Nazaret S (2000b) Monitoring complex bacterial communities using culture-independent molecular techniques: application to soil environment. Res Microbiol 151:167–177

Ranjard L, Poly F, Lata JC, Mougel C, Thioulouse J, Nazaret S (2001) Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. Appl Environ Microbiol 67:4479–4487

Ranjard L, Lejon DP, Mougel C, Schehrer L, Merdinoglu D, Chaussod R (2003) Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA fingerprinting analysis of fungal and bacterial communities. Environ Microbiol 5:1111–1120

Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394

Ravenschlag K, Sahm K, Pernthaler J, Amann R (1999) High bacterial diversity in permanently cold marine sediments. Appl Environ Microbiol 65:3982–3989

Reysenbach A-L, Giver LJ, Wickham GS, Pace NR (1992) Differential amplification of rRNA genes by polymerase chain reaction. Appl Environ Microbiol 58:3417–3418

Rhee SK, Liu X, Wu L, Chong SC, Wan X, Zhou J (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. Appl Environ Microbiol 70:4303–4317

Rochelle PA, Cragg BA, Fry JC, Parkes RJ, Weightman AJ (1994) Effect of sample handling on estimation of bacterial diversity in marine sediments by 16S rRNA gene sequence analysis. FEMS Microbiol Ecol 15:215–226

Rosenbaum V, Riesner D (1987) Temperature-gradient gel electrophoresis. Thermodynamic analysis of nucleic acids and proteins in purified form and in cellular extracts. Biophys Chem 26:235–246

Rosenzweig M (1995) Species diversity in space and time. Cambridge University, Cambridge

Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67

Rosswall T, Kvillner E (1978) Principal-components and factor analysis for the description of microbial populations. Adv Microb Ecol 2:1–48

Ruijter JM, van Kampen AH, Baas F (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. Physiol Genomics 11:37–44

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Amheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxanomic units and estimating species richness. Appl Environ Microbiol 71:1501–1506

Schloss PD, Larget BR, Handelsman J (2004) Integration of microbial ecology and statistics: a test to compare gene libraries. Appl Environ Microbiol 70:5485–5492

Schmalenberger A, Schwieger F, Tebbe CC (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl Environ Microbiol 67:3557–3563

Schmidt TM (1997) Multiplicity of ribosomal RNA operons in prokaryotes. In: De Bruijn JF, Lupiski JR, Weinstock G (eds) Bacterial genomes: physical structure and analysis. Chapman and Hall, London, pp 221–229

Schramm A, Fuchs BM, Nielsen JL, Tonolla M, Stahl DA (2002) Fluorescence in situ hybridization of 16S rRNA gene clones (clone-FISH) for probe validation and screening of clone libraries. Environ Microbiol 4:713–720

Schwieger F, Tebbe CC (1998) A new approach to utilize PCR-single-strand-conformation polymorphism for 16S rRNA gene-based microbial community analysis. Appl Environ Microbiol 64:4870–4876

Sebat JL, Colwell FS, Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. Appl Environ Microbiol 69:4927–4934

Seguritan V, Rohwer F (2001) FastGroup: a program to dereplicate libraries of 16S rDNA sequences. BMC Bioinform 2:9

Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E (2001) Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Appl Environ Microbiol 67:4215–4224

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5:335–344

Short SM, Suttle CA (1999) Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. Hydrobiologia 401:19–33

Short SM, Suttle CA (2000) Denaturing gradient gel electrophoresis resolves virus sequences amplified with degenerate primers. Biotechniques 28:20–26

Sigler WV, Zeyer J (2002) Microbial diversity and activity along the forefields of two receding glaciers. Microb Ecol 43:397–407

Sigler WV, Miniaci C, Zeyer J (2004) Electrophoresis time impacts the denaturing gradient gel electrophoresis-based assessment of bacterial community structure. J Microbiol Methods 57:17–22

Singleton DR, Furlong MA, Rathbun SL, Whitman WB (2001) Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. Appl Environ Microbiol 67:4374–4376

Skirnisdottir S, Hreggvidsson GO, Hjorleifsdottir S, Marteinsson VT, Petursdottir SK, Holst O, Kristjansson JK (2000) Influence of sulfide and temperature on species composition and community structure of hot spring microbial mats. Appl Environ Microbiol 66:2835–2841

Small J, Call DR, Brockman FJ, Straub TM, Chandler DP (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl Environ Microbiol 67:4708–4716

Smith NR, Yu Z, Mohn WW (2003) Stability of the bacterial community in a pulp mill effluent treatment system during normal operation and a system shutdown. Water Res 37:4873–4884

Snaidr J, Amann R, Huber I, Ludwig W, Schleifer K-H (1997) Phylogenetic analysis and in situ identification of bacteria in activated sludge. Appl Environ Microbiol 63:2884–2896

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21:108–110

Stackebrandt E, Goebel B (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44:846–849

Stackebrandt E, Liesack W, Goebel BM (1993) Bacterial diversity in a soil sample from a subtropical Australian environment as determined by 16S rDNA analysis. FASEB J 7:232–236

Stackebrandt E, Frederiksen W, Garrity GM, Grimont P, Kampfer P, Maiden M, Nesme X, Rossello-Mora R, Swings J, Truper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047

Stahl DA, Lane DJ, Olsen GJ, Pace NR (1985) Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. Appl Environ Microbiol 49:1379–1384

Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346

Stralis-Pavese N, Sessitsch A, Weilharter A, Reichenauer T, Riesing J, Csontos J, Murrell JC, Bodrossy L (2004) Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. Environ Microbiol 6:347–363

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol 69:1159–1171

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9:550–557

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res 30:2083–2088

Tiedje JM, Asuming-Brempong S, Nüsslein K, Marsh TL, Flynn SJ (1999) Opening the black box of soil microbial diversity. Appl Soil Ecol 13:109–122

Tilman D, Knops J, Wedin D, Reich P, Ritchie M, Siemann E (1997) The influence of functional diversity and composition on ecosystem processes. Science 277:1300–1302

Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Appl Environ Microbiol 56:782–787

Torsvik V, Daae FL, Sandaa RA, Øvreås L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. J Biotechnol 64:53–62

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Valinsky L, Della Vedova G, Jiang T, Borneman J (2002a) Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. Appl Environ Microbiol 68:5999–6004

Valinsky L, Della Vedova G, Scupham AJ, Alvey S, Figueroa A, Yin B, Hartin RJ, Chrobak M, Crowley DE, Jiang T, Borneman J (2002b) Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. Appl Environ Microbiol 68:3243–3250

Valinsky L, Scupham AJ, Vedova GD, Liu Z, Figueroa A, Jampachaisri K, Yin B, Bent E, Mancini-Jones R, Press J, Jiang T, Borneman J (2004) Oligonucleotide fingerprinting of ribosomal RNA genes (OFRG). In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds) Molecular microbial ecology manual, 2nd edn. Kluwer, London, pp 569–585

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso sea. Science 304:66–74

Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. Appl Environ Microbiol 64:3075–3078

von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213–229

Wang G, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology 142:1107–1114

Wang G, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl Environ Microbiol 63:4645–4650

Wang M, Ahrne S, Antonsson M, Molin G (2004) T-RFLP combined with principal component analysis and 16S rRNA gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages. J Microbiol Methods 59:53–69

Wang Y, Zhang Z, Ramanan N (1997) The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. J Bacteriol 179:3270–3276

Ward DM, Ferris MJ, Nold SC, Bateson MM (1998) A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol Mol Biol Rev 62:1353–1370

Watanabe K, Yamamoto S, Hino S, Harayama S (1998) Population dynamics of phenol-degrading bacteria in activated sludge determined by *gyrB*-targeted quantitative PCR. Appl Environ Microbiol 64:1203–1209

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583

Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuczmarski TA, Andersen GL (2002) High-density microarray of small-subunit ribosomal DNA probes. Appl Environ Microbiol 68:2535–2541

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Stackebrandt E, Macke TJ, Fox GE (1985) A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol 6:143–151

Woese CR, Kandler O, Wheeler ML (1990) Towards a natural system of organsims: Proposal for the domains *Archaea*, *Bacteria*, and *Eucarya*. Proc Natl Acad Sci USA 87:4576–4579

Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl Environ Microbiol 67:5780–5790

Xia X, Bollinger J, Ogram A (1995) Molecular genetic analysis of the response of three soil microbial communities to the application of 2,4-D. Mol Ecol 4:17–28

Yannarell AC, Kent AD, Lauster GH, Kratz TK, Triplett EW (2003) Temporal patterns in bacterial communities in three temperate lakes of different trophic status. Microbial Ecol 46:391–405

Yap WH, Zhang Z, Wang Y (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol 181:5201–5209

Yin B, Valinsky L, Gao X, Becker JO, Borneman J (2003a) Bacterial rRNA genes associated with soil suppressiveness against the plant-parasitic nematode *Heterodera schachtii*. Appl Environ Microbiol 69:1573–1580

Yin B, Valinsky L, Gao X, Becker JO, Borneman J (2003b) Identification of fungal rDNA associated with soil suppressiveness against *Heterodera schachtii* using oligonucleotide fingerprinting of ribosomal RNA genes. Phytopathology 93:1006–1013

Yu Z, Mohn WW (2001) Bacterial diversity and community structure in an aerated lagoon revealed by ribosomal intergenic spacer analyses and 16S ribosomal DNA sequencing. Appl Environ Microbiol 67:1565–1574

Yu Z, Morrison M (2004) Comparisons of different hypervariable regions of *rrs* genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. Appl Environ Microbiol 70:4800–4806

Yu Z, Yu M, Morrison M (2005) Improved serial analysis of V1 ribosomal sequence tags (SARST-V1) provides a rapid, comprehensive, sequence-based characterisation of bacterial diversity and community composition. Environ Microbiol (in press)

Zhang X, Yan X, Gao P, Wang L, Zhou Z, Zhao L (2005) Optimized sequence retrieval from single bands of temperature gradient gel electrophoresis profiles of the amplified 16S rDNA fragments from an activated sludge system. J Microbiol Methods 60:1–11

Zheng D, Alm EW, Stahl DA, Raskin L (1996) Characterization of universal small-subunit rRNA hybridization probes for quantitative molecular microbial ecology studies. Appl Environ Microbiol 62:4504–4513

Zhou J, Davey ME, Figueras JB, Rivkina E, Gilichinsky D, Tiedje JM (1997) Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. Microbiology 143:3913–3919

Zhou J, Thompson DK (2002) Challenges in applying microarrays to environmental studies. Curr Opin Biotechnol 13:204–207

Zhou J, Xia B, Treves DS, Wu LY, Marsh TL, O'Neill RV, Palumbo AV, Tiedje JM (2002) Spatial and resource factors influencing high microbial diversity in soil. Appl Environ Microbiol 68:326–334

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

Zumstein E, Moletta R, Godon JJ (2000) Examination of two years of community dynamics in an anaerobic bioreactor using fluorescence polymerase chain reaction (PCR) single-strand conformation polymorphism analysis. Environ Microbiol 2:69–78

8 Metagenome Analyses

Frank Oliver Glöckner, Anke Meyerdierks

8.1 Introduction

Robert Koch's invention of pure culture techniques at the end of the nineteenth century focused microbiology on the isolation of bacteria for laboratory studies. Even today, in clinical diagnostics and foodstuff biotechnology, cultivation remains the gold standard because full characterisation of metabolic capabilities, resistance and pathogenesis can still only be achieved with pure cultures. "Winds of change" (Olsen et al. 1994) blew in the field of microbiology when the first cultivation-independent investigations reported an immense array of completely unexpected microbial diversity in the environment (Torsvik et al. 1990). Today it is estimated that only 1% of the microbial diversity in the biosphere can be assessed by means of standard cultivation techniques (Amann et al. 1995; Curtis et al. 2002). Although new approaches have recently been introduced to gain access to the "not currently cultureable majority" (Connon and Giovannoni 2002; Rappe et al. 2002; Zengler et al. 2002), they are not keeping pace with the substantial set of molecular tools to address the diversity and structure of microbial communities. Examples of these molecular tools are the powerful PCR-based methods that have been established for direct amplification, cloning and analysis of ribosomal RNA (rRNA) genes from the environment (Pace et al. 1985; Olsen et al. 1986; Giovannoni et al. 1990; Ward et al. 1990). Beyond diversity, the design and application of specific rRNA-targeted oligonucleotide probes allows insights into the structure of microbial communities *in situ* (Stahl and Amann 1991). Over the past 10 years, this has become a standard method in molecular ecology (Amann et al. 1995; Pace 1997). The impact of the new methods can even be monitored by noting the exponential increase in the number of 16S rRNA sequences in public databases like RDP II and ARB (Cole et al. 2003; Ludwig et al. 2004). Currently (October 2005), more than 184,990 16S rRNA sequences are publicly available, with the vast majority originating in thus-far uncultured bacteria. Taken together, it is clear now that the vast

Frank Oliver Glöckner, Anke Meyerdierks: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany, E-mail: fog@mpi-bremen.de

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

majority of prokaryotic diversity is not represented in culture collections and therefore the abilities of most prokaryotes are largely unknown.

The main drawback of the rRNA-targeted methodology is that usually there is no way to infer the physiology, biochemistry, or ecological function of an uncultivated micro-organism from phylogenetic information alone. Ribosomal RNAs, like other highly conserved phylogenetic markers, belong to a rather static set of genes that is needed to maintain the basic functionality of the cell. Such housekeeping genes are not subject to direct selective pressure when organisms have to adapt to changing environmental conditions. To obtain insights into their ecophysiology, the 'adaptive' pool of metabolic, resistance and defence genes has to be investigated because their successful deployment ensures continued survival in the new environment. The logical extension of the single gene approach is to analyse larger genomic fragments directly extracted from microbial assemblages. This was implemented by Schmidt et al. (1991), who constructed the first environmental genomic library with the bacteriophage λ from DNA directly extracted from microbial biomass of the north central Pacific Ocean and screened it for 16S rRNA genes. A few years later, Stein and DeLong significantly improved the method, applied it to Oregon coastal waters and produced the first environmental clone library that contained large genomic fragments of about 40 kbp (Stein et al. 1996). Subsequent complete sequencing of one of the 40-kbp fragments gave new insights into the genetic capabilities and genomic organisation of an uncultured marine archaeon. The true significance of this new technique only emerged four years later, with the publication of bacterial artificial chromosome (BAC) libraries from soil and marine picoplankton (Beja et al. 2000b; Rondon et al. 2000). It was Handelsman (1998) who introduced the name 'metagenomics', which is now defined as the "functional and sequenced-based analysis of the collective microbial genomes contained in an environmental sample" (Riesenfeld et al. 2004). Other terms that have been used to describe the same method are 'environmental genomics' (Stahl and Tiedje 2002), 'ecogenomics' (Stein et al. 1996), and many others (for an overview, see Riesenfeld et al. 2004). The expectation that the method might bridge the gaps between diversity, structure and function seems to have been fulfilled, as can be illustrated in two recent examples. A novel light-driven proton pump (proteorhodopsin) was identified on a 150-kbp genome fragment assigned to the uncultured SAR86 group (Beja et al. 2000a, 2001). Finding this new kind of photosynthetic energy generation by a marine γ-proteobacterium was unexpected; and the implications for the global energy balance of our world's oceans are far-reaching. This is also true for a recently published study that combined biochemistry and metagenomics to analyse an enzyme probably involved in the anaerobic oxidation of methane (AOM). The results indicate that a conspicuous nickel protein might be the key enzyme to allow the reversal of

methanogenesis (Krüger et al. 2003). Since microbially mediated AOM is the major biological sink of the greenhouse gas methane in marine sediments, understanding the underlying metabolic process is of pivotal importance.

In biotechnologically oriented bio-prospecting, metagenomics has already become a standard tool. The screening of metagenomic libraries has identified novel antibiotics, antibiotic resistance genes and several genes for biopolymer degradation and biosynthesis (for reviews, see Handelsman et al. 2002; Schloss and Handelsman 2003; Riesenfeld et al. 2004). The recent explosion of interest and activity in the field of metagenomics is mainly driven by accelerated sequencing techniques. This overwhelming power of high-throughput sequencing has been demonstrated by Tyson et al. (2004) and Venter et al. (2004), who sequenced the metagenome of microbial communities, applying a shotgun approach. Thereby, Tyson's results were especially interesting because they demonstrated for the first time that it is possible to reconstruct nearly complete genomes from a mixed microbial community. The Sargasso Sea approach of Venter and co-workers, with an estimated 1,800 genomic species and containing over a million previously unknown genes, underlined once more that we have only scratched the surface of microbial diversity and function.

Genomics has introduced a new dimension in biology, guiding it to a massively parallel and high-throughput endeavour. This quantum leap in production has surpassed our current ability to interpret and use the resultant deluge of data. To best proceed, we must find a way to ensure that not only computer scientists can store, analyse and integrate all the data coming from genomics, metagenomics and post-genomics, but biologists as well. This means creating interfaces so that researchers in the laboratory are able to access and work with the data on a routine basis. This can only be achieved by the expansion and continuous support of the emerging field of bioinformatics. Nevertheless, bioinformatic approaches with homology-based functional predictions currently only provide us with hints on gene function for about 50 – 60% of the sequences retrieved (Nelson 2003). Improving this situation demands that high-throughput post-genomic methodologies are invented or further developed. To this end, microarrays are now a standard laboratory tool for gene expression and genotyping; and it is encouraging to see that their application to complex microbial assemblages also seems to be possible (Dennis et al. 2003; Peplies et al. 2004). Nevertheless, to nail down the function of a protein, laborious and time-consuming expression, protein–protein interaction and functional studies are required. Bioinformatics will back up this process with predictive in silico models in order to obtain priorities for the set of genes that has to be analysed in detail. Without further support in this field, it can be safely assumed that the 'gold-mine of sequences' which has now been opened will not be fully exploited in the near future.

8.2 Construction and Screening of Metagenome Libraries

The construction of metagenomic libraries follows a general scheme, independent of the vector system chosen (Fig. 8.1). Genomic DNA is isolated either directly from an environmental sample, or from an enriched culture of target organisms. The DNA is further purified to remove contaminants, e. g. polyphenolic substances or metal ions, which could interfere with the subsequent enzymatic manipulation of the DNA. In the next steps, the genomic DNA fragments are trimmed either by end-repair or restriction enzyme digestion, properly size-selected and concentrated. The DNA is ligated to a cloning vector and transferred into *Escherichia coli* host cells. Arraying of individual clones is indicated in most cases. After the initial characterization of the library with respect to, e. g. the number of recombinant clones, the average insert size and the number of clones without insert, screening for selected DNA sequences or expressed proteins is performed. This often results in the full sequencing of interesting clones from the metagenomic library (here called metagenomic clones) and their subsequent detailed bioinformatic analysis.

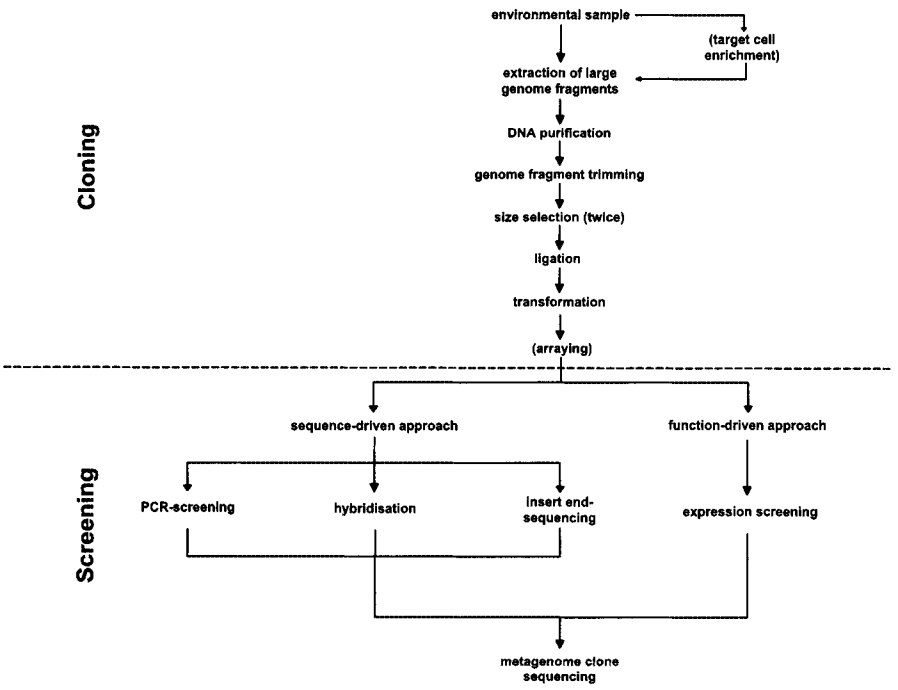


Fig. 8.1. Overview on the construction of metagenomic libraries and the downstream screening methods

8.2.1 Small and Large Insert Libraries

Several metagenomic libraries with low molecular weight (LMW) DNA inserts (< 30 kbp), i.e. plasmid and bacteriophage λ libraries, have been constructed in the past. Early studies of such small insert libraries, e.g. by Cottrell et al. (1999) and Henne et al. (1999), focused on the analysis of single metabolic genes of uncultured micro-organisms. Recently, it was shown that small insert libraries in a reporter gene construct are also an intelligent tool to identify novel catabolic operons in substrate-induced gene expression screening (Uchiyama et al. 2005). Additionally, high-throughput sequencing of LMW DNA insert libraries can be a powerful strategy to get access to drafts or nearly complete assemblies of genomes of uncultured micro-organisms (Tyson et al. 2004; Venter et al. 2004).

Although the success and impact of these studies are far-reaching, it has to be considered that phylogenetic markers, which allow the reliable assignment of fragments to certain phylogenetic groups, are often missing on small inserts. In the case of limited sequencing power or a high microbial diversity within the sample, e.g. in soil samples (Handelsman et al. 2002), the assignment of genetic capabilities present on a small continuous DNA region (contig) is nearly impossible in small insert libraries. Moreover, the *in silico* assembly of larger fragments implies the risk of creating chimeras, especially when the overlaps of the reassembled fragments are too short.

In contrast to small insert libraries, metagenomic libraries with high molecular weight (HMW) DNA inserts (>30 kbp; large insert libraries) give access to DNA contigs which often carry phylogenetic markers, such as the 16S rRNA gene (e.g. Schleper et al. 1997; Beja et al. 2002). This reduces the risk of creating chimeras and also the sequencing effort, which is particularly adventurous when specific questions, e.g. about defined metabolic pathways in a certain microbial group, are to be answered. Large insert libraries have therefore been favoured in most of the reported metagenomic studies.

8.2.2 High-capacity Vectors: Cosmids, Fosmids or BACs?

Before the construction of a large insert metagenomic library starts, an appropriate vector system has to be chosen. Large DNA fragments can generally be cloned into a variety of different vectors (Green et al. 1997; Tao and Zhang 1998). For the construction of large insert metagenomic libraries, three of these vector types are commonly used: cosmid, fosmid and BAC vectors.

Cosmid vectors (Collins and Hohn 1978) represent the oldest type of high-capacity vector and have been used in the construction of several metagenome libraries (Entcheva et al. 2001; Piel 2002; Courtois et al. 2003; Schmeisser et al. 2003; Sebat et al. 2003; Lopez-Garcia et al. 2004). These vectors are composed of conventional plasmids in which one or two bacteriophage λ *cos* sites have been integrated, allowing the utilization of the in vitro packaging system of bacteriophage λ that accounts for high-cloning efficiencies. Consequently, the average capacity of cosmid vectors is 30–45 kbp, due to the size limitation of DNA that can be packaged into bacteriophage λ phage heads. Once in the host cell, cosmids are present in high copy number. The construction of cosmid libraries is straightforward; and the subsequent analysis of the libraries is facilitated due to their natural amplification (Collins and Hohn 1978; Sambrook and Russel 2001). However, the system has two major drawbacks. First, chimeras, rearrangements and deletions have been observed (Monaco and Larin 1994) and, second, the insert sizes are more or less uniform, but relatively small compared to a complete genome.

To overcome the inherent instability, fosmid vectors (Kim et al. 1992) have been developed and used for metagenomic library construction (Stein et al. 1996; Schleper et al. 1997, 1998; Beja et al. 2002; Quaiser et al. 2002, 2003). The fosmid cloning system also uses the *in vitro* packaging of bacteriophage λ, but the vector is derived from the *E. coli* F(ertility)-factor and carries replication and partition sequences of the F-factor plasmid. This accounts for its low copy number (1 - 2 copies per cell) and prevents two different fosmids from being maintained in a single cell. Additionally, the expression of toxic gene products that could be lethal for the host is reduced. However, the screening of libraries present in low copy vectors is more laborious than screening those in high copy vectors. Therefore, fosmid vectors have been constructed which carry a second inducible replicon. This allows maintenance of the library in the low copy state and the selective induction of 10 - 50 copies per cell for screening and further analysis.

BAC vectors (Shizuya et al. 1992) have been developed to solve the insert size limitation of cosmids and fosmids. BAC vectors, like fosmid vectors, carry the F-factor replication and partition sequences from *E. coli* and are therefore low copy vectors exhibiting the advantages already mentioned. To circumvent the size limitation of bacteriophage λ packaging, in BAC library construction the DNA is introduced into the host cell by electroporation. Particular *E. coli* host strains that have proved to transform well with large insert constructs are used for this purpose (Sheng et al. 1995). It has been shown that DNA fragments larger than 300 kbp can be cloned into BAC vectors and stably maintained in *E. coli* (Shizuya et al. 1992; Kim et al. 1996; Zimmer and Verrinder 1997). Recently, BAC vectors with a second inducible replicon became available (Handelsman et al. 2002; Wild et al. 2002). Never-

theless, BAC cloning is up to 100 - 1,000 times less efficient than cosmid and fosmid cloning due to the diminished efficiency of electroporation as compared to *in vitro* packaging and transition. Additionally, the average insert size of BAC libraries is negatively correlated with the number of BACs generated (Leonardo and Sedivy 1990; Woo et al. 1994; Sheng et al. 1995; Zimmer and Verrinder 1997).

Since fosmid and BAC vectors with an inducible copy number are available, cosmids seem to have gone more or less out of fashion because of the already mentioned drawbacks. Generally, the BAC cloning system is more widely used in genome analysis, compared to the fosmid system. However, most of the groups utilizing the BAC cloning system work on eukaryotic cells or cultured micro-organisms and take advantage of the large insert sizes. For groups working on metagenomics, especially for those focusing on soil or marine sediments, BAC cloning is still a challenge; and often the average insert size is not much higher than those of fosmids (Hughes et al. 1997; Beja et al. 2000b; Rondon et al. 2000; MacNeil et al. 2001; De la Torre et al. 2003). Metagenome BAC libraries with an average insert size of more than 80–100 kbp seem to be currently out of reach, even when using 'clean' environmental samples like plankton (Beja et al. 2000b) or worm symbionts (Blazejak et al., in preparation). However, it is comfortable to have a metagenomic clone with a large insert, because genome walking can be laborious, especially in a library from a highly diverse sample.

In summary, the authors suggest to give BAC cloning a try. Thus, contacting experienced groups prior to cloning is highly recommended. If BAC cloning does not work out or if a library is urgently needed, it is recommended to go for fosmids. Fosmid cloning is straightforward and has been successful with all samples we have investigated. Even with a small amount of DNA, one can easily get a library containing at least a few thousand clones.

Finally, it has to be taken into account that genes encoded on an HMW DNA insert might be heterologously expressed. These products can be toxic for the host and therefore cannot be cloned and stably maintained in *E. coli* (Beja et al. 2000a). Therefore, a combination of HMW and LMW DNA insert libraries might sometimes be indicated. Additionally, high-capacity shuttle vectors have already been developed for different purposes (Handelsman 1998; Courtois et al. 2003).

8.2.3 Library Size

The general equation for calculating the library size needed to cover the genome of an organism in pure culture with a given probability is defined

as follows: $N = [\ln(1 - P)]/[\ln(1 - f)]$, where P is the desired probability, f is the fractional proportion of the genome in a single recombinant (e. g. average insert size/genome size) and N is the necessary number of clones in the library (Sambrook and Russel 2001).

In practice, the size of a metagenomic library that is needed to be representative of the environmental sample, or statistically contains one clone that carries the marker gene of interest is heavily dependent on, e. g. the diversity of the sample, the abundance of the target organisms, their genome size and the clonability of the genomic fragments (e. g. toxic gene products).

8.2.4 Isolation and Purification of HMW DNA

To reduce the library size and therefore the effort involved in library construction and screening, enrichment of target organisms is highly recommended, at least when the scientific inquiry is targeted to microbial populations of low abundance. Possible enrichment strategies applied prior to metagenomic library construction span from concentration with filters of different pore sizes (Beja et al. 2000a), to density gradient centrifugation (Schleper et al. 1998; Hallam et al. 2003), enrichment cultures (Entcheva et al. 2001) and selective lysis of cells. Alternatives to the enrichment of target cells include the enrichment of genomic DNA with respect to the average G+C content, the incorporation of bromodeoxyuridine and the incorporation of stable isotopes (Schloss and Handelsman 2003).

HMW genomic DNA fragments can subsequently be isolated from enrichments or directly from the environmental sample following one of two distinct methods: 'liquid phase-based' and 'solid phase-based'. The liquid phase-based method is the most straightforward DNA isolation technique and is applicable to nearly every environmental sample. A popular liquid phase-based protocol for the extraction of large amounts of genomic DNA from environmental samples is the lysis protocol published by Zhou et al. (1996). Cells are lysed in a high-salt buffer containing proteinase K, CTAB (hexadecylmethylammonium bromide) and SDS (sodium dodecyl sulfate). The lysis of gram-positive bacteria is thereby supported by several freeze-thaw cycles included in the protocol. If it is necessary, the obtained DNA can be further purified, by gel electrophoresis (Rondon et al. 2000; Quaiser et al. 2003), anion exchange chromatography (Krüger et al. 2003), density gradient centrifugation (MacNeil et al. 2001; Courtois et al. 2003), or by using commercially available kits to remove polyphenolic compounds and other contaminating substances that would interfere with the subsequent cloning steps. After purification, the DNA is suitable for the construction of metagenomic libraries. Nonetheless, even when DNA extraction is done very care-

fully, avoiding excessive shaking of the extraction tubes and using wide bore tips, the maximal size of the isolated DNA does generally not exceed 150–200 kbp. This also seems to be the cut-off when using commercially available kits for the preparation of genomic DNA. Therefore, this preparation method is suitable for the construction of metagenome cosmid and fosmid libraries, but hardly for the construction of BAC libraries with average insert sizes exceeding 50 kbp (Rondon et al. 2000; MacNeil et al. 2001).

The isolation of high-quality genomic DNA fragments of more than 200 kbp from environmental samples, especially from soil or sediments, is more difficult, because HMW DNA of more than 100 kbp is prone to shearing forces generated by standard pipetting or sample mixing. Therefore, the environmental sample or enriched cells are embedded in agarose plugs and dialysed against different buffers supplemented with enzymes and detergents to remove proteins and lipids from the embedded cells, leaving naked HMW DNA behind, which is thereafter ready for enzymatic manipulations (Green et al. 1997; Sambrook and Russel 2001). This was effective for a plankton sample (Beja et al. 2000a), but its application to other environmental samples, especially sediment and soil samples, resulted in agarose plugs with partly degraded genomes and inhibitors, which could not be removed by the lysis procedure or dialysis. In these cases, further purification of the DNA is necessary, e.g. by electrophoresis through conventional agarose gels, or two-phase agarose gels containing polyvinylpyrrolidone (Quaiser et al. 2002). This results in a diminishment and shearing of the genomic DNA and, in the worst case, the purity and size of the HMW DNA still might be inadequate for producing large-scale BAC libraries with high average insert sizes.

8.2.5 Construction of Large Insert Metagenomic Libraries

The construction of large insert libraries in general is well described, e.g. by Sambrook and Russel (2001) and Green et al. (1997). Therefore, only critical steps will be discussed here.

The extraction and purification of HMW DNA is followed by either end-repair or restriction digestion, to make the fragments compatible to the ends of the chosen cloning vector. Sticky-end cloning of DNA is more efficient than blunt-end cloning. Therefore, trimming the ends of HMW DNA by partial restriction digestion seems to be the method of choice for metagenomic library construction. However, its drawbacks are:

1. The average size of the genomic DNA prior to partial digestion has to be at least three- to five-fold larger than the size of the DNA fragments that are supposed to be ligated to the vector.

2. A cloning bias can occur if recognition sequences for the restriction enzyme are under-represented in some parts of the genome.

Blunt-ending, in contrast, does not lead to a marked reduction of the DNA fragment size, and blunt-end cloning is less biased.

The trimming of the HMW inserts is followed by a size selection step. This is another crucial cloning step for two reasons. First, during the introduction of constructs into the host cell by electroporation, constructs with smaller insert sizes are favoured. Therefore, insufficient size selection will lead to a smaller average insert size of the resulting library (Osoegawa et al. 1998). Second, proper size selection is important to avoid chimera formation resulting from the ligation of two genome fragments to one vector molecule. Therefore, pulsed field gel electrophoresis (PFGE) is preferred over conventional agarose gel electrophoresis. During conventional agarose gel electrophoresis, in which a constant electric field induces DNA to migrate in a single direction, all DNA molecules larger than approx. 15–25 kbp migrate with nearly identical mobility. In contrast, PFGE produces size-dependent mobility of large DNA up to > 5 Mbp in agarose gels (Sambrook and Russel 2001). Consequently, access to a PFGE is a prerequisite for BAC cloning. It is also highly recommended for fosmid cloning, although conventional gel electrophoresis performed at low voltage is also possible (Hallam et al. 2003). Two rounds of size selection might be necessary if large amounts of DNA are separated on the gel (Osoegawa et al. 1998; Rondon et al. 2000). The obtained size-selected DNA is either electroeluted from the gel slice, or, alternatively, the gel slice is enzymatically digested using GELase® (Epicentre) or β-agarase. In both cases, the DNA is subsequently concentrated, e. g. (a) by filter dialysis on VSWP filters against polyethyleneglycol containing dialysis buffer, (b) by centrifugation through filter devices, or (c) by precipitation (only fosmid or cosmid cloning). The ligation is ideally done overnight at 16 °C, or for two consecutive days at 4 °C, in a volume of 20–100 µl, testing different vector:insert ratios. Subsequent *in vitro* packaging in phage heads is usually done following the instructions of the manufacturers of bacteriophage λ packaging extracts. Filter dialysis of the ligation mixture is best done prior to electroporation.

8.2.6 Storage of Metagenomic Libraries

The storage of metagenomic libraries mainly depends on the equipment available in the laboratory, the sparseness of the sample, the library size and the screening method. Large (> 20,000 clones) cosmid and fosmid libraries are sometimes stored in pools of recombinant clones that have been washed off their agar plates with a freezer medium containing cryoprotectants

(e.g. 5 – 7% glycerine) and stored at 80 °C. Storage of a library in pools is suitable and less time-consuming than the picking and arraying of clones, especially when the subsequent screening is done by hybridisation or expression screening. However, it is useless when end-sequencing is intended and it complicates PCR screening, because a second hybridisation step has to be carried out in order to identify the individual positive clones. It also must be noted that slowly growing clones might be overgrown in pools. At least in cases where the environmental sample is sparse, the library size is relatively small, or when PCR screening or end-sequencing is planned, an effort should be made to array the library and store it in several copies.

8.2.7 Screening of Metagenomic Libraries

Two different screening strategies, the DNA-driven and the protein-driven approaches, have to be differentiated. The DNA-driven approach is based on specific oligonucleotides or probes that are used in: (a) PCR screening, (b) hybridisation, or (c) insert end-sequencing.

PCR screening is the fastest method to screen a whole library, especially when DNA pools have been prepared. The scheme finally used for pooling is dependent on the number of positive clones theoretically present in the library. Thereby, it is advisable to have one statistically positive clone in a given DNA superpool. Examples for complex pooling schemes are given by Kim et al. (1996) and Asakawa et al. (1997). A particular drawback of PCR screening can be the cross-hybridisation of PCR primers with chromosomal DNA of the host. Different solutions for overcoming this problem have been published. One is the selective hydrolysis of chromosomal DNA by using an ATP-dependent DNase prior to PCR screening (Beja et al. 2000a; Liles et al. 2003). Another method is to include host-specific, terminally modified oligonucleotides in the PCR reaction (Goodman and Liles 2001; Liles et al. 2003). Finally, a subsequent RFLP analysis step can be added to the PCR protocol in order to identify positive clones (Liles et al. 2003).

For hybridisation, colony blots (e.g. Asakawa et al. 1997; Osoegawa et al. 2000) and spotted DNA (Rondon et al. 1999) have been successfully used for the screening of libraries present in low and high copy vectors. However, colony blots with libraries that are cloned in uninducible low copy vectors might sometimes be a bit tricky with respect to the signal to noise ratio, depending on the length of the probe, the labelling efficiency, the hybridisation efficiency and the signal detection system.

The extraction of large insert constructs from host cells for insert end-sequencing can be done by simple alkaline lysis of the recombinant clones with subsequent alcohol precipitation of the DNA, or by using commercially

available lysis and purification kits. The precise DNA extraction methods and the sequencing conditions applied vary greatly. At any rate, the sequences often have lower quality than those usually obtained from the sequencing of conventional plasmids.

An example of the successful combined application of all three different methods is the recent analysis of a fosmid library from a methanotrophic microbial mat. Using established primer sets, PCR screening was carried out to reveal the diversity of archaeal 16S rRNA genes in the library. Published primer sets as well as probes (labelled selected PCR products) were used to analyse the library by PCR screening and hybridisation for the presence of a key gene of a metabolic process of interest. Whole fosmid sequencing of the identified clones gave access to an apparent operon putatively encoding the key enzyme. Further analysis of the operon by bioinformatic tools revealed that the deduced amino acid sequence of the key enzyme was slightly different from those already known. The determination of insert end-sequences led to the identification of other key genes of the metabolic process investigated. After full-length sequencing, the fosmid insert was assigned to a certain species by using bioinformatics tools to compare its sequence characteristics with a fosmid insert that carried a phylogenetic marker (see below). The insert end-sequences were valuable for further genome-walking (Krüger et al. 2003; Meyerdierks, unpublished data).

Finally, the protein-driven approach takes advantage of the fact that genes present on the cloned inserts can be heterologously expressed in the host cell, as long as the transcription and translation machinery of the host and donor strain are compatible. This approach is predominantly applied to identify enzymes for biotechnological purposes. An overview of the possibilities and various screening strategies is given in the review articles of Handelsman et al. (2002) and Riesenfeld et al. (2004).

8.2.8 Sequencing of Large Insert Constructs

Large insert constructs are generally sequenced in a shotgun approach as described, e.g. by Sambrook and Russel (2001). A crucial step in the preparation of a shotgun library from low copy constructs is thereby the removal of contaminating residual chromosomal DNA of the host. This is accomplished either by caesium chloride density gradient centrifugation, or enzymatic digestion with ATP-dependent exonuclease. After shotgun library construction, vector-specific primers are used for the sequencing of insert ends. About 400 sequencing reactions are approximately required to sequence a fosmid in eight-fold coverage. Sequences are assembled using specific assembly software, such as Phrap (http://www.phrap.com/).

Remaining gaps are usually closed by primer-walking. Since the extraction methods for fosmids and BACs have improved and vectors with inducible copy number are available, transposon-mediated sequencing, e.g. performed by Courtois et al. (2003), is becoming more popular.

8.3 Sequence Analysis

After finishing the sequencing and assembly phase by external companies or in house facilities, raw sequence information is generally stored and shipped in simple flat file formats like the FASTA format. It consists of a sequence name and description on a single line starting with the 'greater than' symbol (>) and followed by the sequence itself. The advantage of this format is that it can be handled by nearly all currently available bioinformatic tools. The disadvantage is that there is no standard for the order or content of information in the description line. This will immediately cause consistency problems when complex information has to be exchanged between programs. If this is needed, structured standards like the EMBL or GenBank formats have to be given preference so as to circumvent the loss, mixing, or truncation of data (Mount 2001).

8.3.1 Marker Genes

The first step in revealing the affiliation of genes amplified from a metagenomic library is either to generate a database of orthologues for multiple sequence comparison and phylogenetic affiliation, or to add them to an existing one. If screening has been performed for ribosomal RNA genes, this is rather simple, since several databases like RDP II and ARB exist (Cole et al. 2003; Ludwig et al. 2004). The easiest way to assign the sequence of interest to the currently emerging "taxonomic outline of the prokaryotes" (Garrity et al. 2002) is to use the web-based classifier program on the RDP II homepage (http://rdp.cme.msu.edu/). For eubacterial sequences, this will provide a first indication of the taxonomic affiliation, based on octamer representation. A more thorough phylogenetic investigation can be obtained with, e.g. the phylogenetic software suite ARB (Ludwig et al. 2004). Since ARB only runs as a local installation, both the software and the database have to be downloaded and installed on a workstation or PC from www.arb-home.de. The advantage is that the small subunit database of ARB has a manually curated comprehensive alignment comprising not only the domain *Bacteria*, but also *Archaea* and *Eucarya*. After importing,

the automatic aligner implemented in ARB aligns the sequences according to the closest relatives in the database. To assign the sequences to the general phylogenetic tree delivered with each database, a quick add procedure can be initiated. This feature is unique among currently available programs for phylogenetic tree reconstructions, since a parsimony algorithm (Swofford et al. 1996) is used for the assignment of new sequences to an existing tree without changing the overall topology. This is especially useful when working with partial sequences, because the results are much more reliable than those obtained via the alternative method of truncating all sequences to the shortest one for phylogenetic reconstructions. Furthermore, once the sequences are stored in the local database, ARB offers a plethora of distance matrix, parsimony and maximum likelihood methods (Swofford et al. 1996) for subsequent in-depth phylogenetic analysis.

A first indication of the affiliations between functional markers can be obtained by pairwise sequence alignments against public databases at the European Bioinformatics Institute (EBI; www.ebi.ac.uk) or the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) with, e.g. a BLAST implementation (Altschul et al. 1990). Nevertheless, when using this approach, it has to be kept in mind that these kinds of program rely on heuristics to reduce the search space and speed up the search process. Furthermore, the algorithm involves no model of evolution. In a worst case scenario, especially when sequences share less than 30% identity on the protein level, the results obtained might be misleading. To really investigate the relationship of functional genes, it is in most cases necessary to build up a local database of homologues for multiple sequence alignment. Depending on the gene of interest, the seed or full alignments provided by knowledge databases of protein families, such as Pfam (Bateman et al. 2004), can be used as a starting point. Common multiple sequence alignment tools include ClustalW (Chenna et al. 2003) and MAFFT (Katoh et al. 2002). After manual refinement of the alignment, phylogenetic reconstructions can be performed using the Phylip package (http://evolution.genetics.washington.edu/phylip.html), for example, to clearly establish the phylogenetic relationship of the marker of interest. Corresponding to the analysis of ribosomal RNA genes, all these individual tasks (and many more) can also be handled within the ARB system with the advantage of a common graphical user interface. A collection of currently available databases for ARB can be found at http://arb-db-central.swiki.net/.

8.3.2 End-Sequences

The sequencing power available today offers the possibility to end-sequence 500 – 800 bases of many fosmid or BAC clones within a short time. This provides first fascinating insights into the metabolic capabilities and the taxonomic groups available in the metagenomic library under investigation. When sequencing is finished, all reads can be checked for similarity against a comprehensive nucleotide sequence database like the EMBL database provided by EBI with BLASTn (http://www.ebi.ac.uk/blast2/nucleotide.html). The objective of this approach is to find significant hits ($< 10^{-3}$) to ribosomal RNA genes for extraction and subsequent processing, as described for the marker genes. When the phylogenetic reconstructions are stable, the corresponding clone(s) can be directly assigned to distinct taxonomic groups. To get reliable results for protein-coding genes, the sequences have to be translated into all six possible reading frames before performing a BLAST search against a comprehensive protein database like UniProt (Apweiler et al. 2004), or GenBank from NCBI. The reason is that, for enzyme function maintenance, the evolutionary pressure is on the amino acid rather than on the nucleotide sequence, because of the degenerated genetic code. Furthermore, in many circumstances a given amino acid could be replaced by an isofunctional one while still retaining its operational integrity. This second condition requires searching for similarities among proteins themselves, taking into account appropriate amino acid substitution matrices like PAM or BLOSUM (Korf et al. 2003). The easiest way to perform a combination of six-frame translation and BLAST search is to use the BLASTx algorithm (http://www.ebi.ac.uk/blast2/). This increases the number of searches by a factor of six, leading to the problem that most web-based systems restrict batch jobs to a maximum of some hundreds, due to performance problems. To circumvent this, a local installation of the BLAST programs and their corresponding databases is highly recommended. This is also helpful for parsing the results of a BLAST search, which can easily exceed 10,000 hits for several hundred end-sequences. To structure all these results for data mining, command line tools like MSPcrunch (Sonnhammer and Durbin 1994) are available. To get an overview of the taxonomic distribution within the BLAST hits, a taxonomic breakdown can be performed, with, e.g. the SEALS system (Walker and Koonin 1997). The principle is to take the best BLAST hit – please do not forget to do vector clipping before BLASTing – for every sequence searched against GenBank and to store the corresponding general identifier (gi). This information is used to extract the taxonomic information provided by the taxonomy browser of NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/) on the level of domains, phyla, genera and species. The results give a first

impression of the diversity covered by the clones in the metagenomic library. Nevertheless, one must remember that BLAST is not a phylogenetic program and its results might not always reflect true evolutionary relationships. As a further step, a functional classification of the proteins can be performed by BLASTing all reads (BLASTx) against the clusters of orthologous groups of proteins (COGs) database (Tatusov et al. 2003). If a significant hit (*e*-value $< 10^{-5}$) to one of the COGs stored in the database is found, the output contains not only a potential function, but additionally a single-letter code for functional classification. The full list of 18 (initial version) or 25 (updated version) functional categories provided by COG is available on the website (http://www.ncbi.nlm.nih.gov/COG/). The functional classification thereby obtained will provide a general overview of the metabolic capabilities found within the library. This will assist in the selection of specific BACs or fosmids for complete sequencing and in-depth analysis.

8.3.3 Cosmids, Fosmids or BACs

Correlation of Metagenomic Fragments

The two most common problems with metagenomic libraries are: (1) the BAC or fosmid fragments that carry the metabolic genes of interest lack suitable markers for their phylogenetic classification and (2) BAC or fosmid fragments from the same organism cannot be reliably identified as such, unless they overlap. In both cases, measures such as the average G+C content of the fragments, the best BLAST hits and the codon usage of the corresponding coding regions are commonly used to provide further hints. These measures, however, can produce ambiguous or even misleading results and should be supplemented by tools taking intrinsic genomic signatures into account (Teeling et al. 2004a, 2004b). Numerous studies have shown that oligonucleotide frequencies within DNA sequences exhibit species-specific patterns (Karlin et al. 1998); and for tetranucleotides it has even been demonstrated that their frequencies carry an innate but weak phylogenetic signal (Pride et al. 2003). This technology has already been shown to be a valuable tool for the analysis of metagenomic libraries created from samples where the anaerobic oxidation of methane is the prevailing process (Krüger et al. 2003; Meyerdierks, unpublished data). To facilitate the analysis of tetranucleotide frequencies and make them easily applicable to users, a web server and stand-alone programs are available at www.megx.net/tetra. A limitation of statistical analyses of oligonucleotide distributions is that small sequence lengths hamper the underlying statistics and thus overall reliability. While the method works quite well for fosmid-sized fragments (ca. 40 kbp), it is currently not well suited for the

analysis of single-read end-sequences, which are typically shorter than 1 kbp. With neuronal networks or naïve Bayesian classifiers, however, substantial species-specific information can be inferred even from sequences shorter than 10 kbp (Sandberg et al. 2001; Abe et al. 2003). This indicates that, in the future with more sophisticated or combinatory methods, intrinsic DNA signatures can support the process of assembling short sequences. Work is in progress and there is a clear chance that intrinsic DNA signatures can help to cluster the short (2–3 kbp) sequences that are generated in huge amounts by present-day environmental shotgun approaches.

Functional Annotation

Functional annotation can be regarded as the final step in the process of analysing genomic fragments obtained from metagenomic studies. At this level, the investigator gets a substantial insight into the wealth of the genetic potential available in the environment. Annotation should be handled with care since frowsy annotations will – like the proverbial first ice crystal – start a snowball effect by continuous error propagation. In general, errors can be introduced by inconsistencies in functional assignments between and even within a single genome and by a simplistic procedure to assign potential functions to the genes found. Unfortunately there is currently no 'gold standard' for consistent genome annotation available and no binding rules exist which have to be used by all annotators. A first step to address the problem is to provide a controlled vocabulary with unique identifiers and a clear hierarchy. This has been recently introduced by the Gene Ontology consortium (Ashburner et al. 2000) and will hopefully get the standard for genome annotations in the future.

To exploit the currently available data sources for functional predictions, comprehensive software systems are needed to store, analyse and visualize data and support the decision process by providing information from various sequence-based analysis tools. Performing a simple BLAST search against the UniProt or COGs database and taking the best hit for gene annotation is definitely not adequate! Storing tool results in simple spreadsheets will lead to redundancy and hamper the necessary integration and correlation of data. Relational database management systems with a consistent internal data representation and a defined data model, including an applications programmer's interface (API), are a prerequisite for data management. The API allows customized data mining and the implementation of self-written tools that fit your personal needs. A 'state of the art' analysis pipeline for genomic data includes gene-finding and standard bioinformatic tools for similarity-, pattern- and profile-based searches as well as prediction of signal peptides, transmembrane helices, transfer and other stable RNAs. Additionally, the analysis of global and local G+C content and

skews as well as codon usage and further statistical parameters can help in distinguishing coding from non-coding regions. Adequate annotation systems include automatic annotation of the protein coding regions as well as web-based and user-friendly annotation facilities for manual refinement in annotation jamborees. Since the advent of genomics at the beginning of the 1990s, several annotation systems have been made available. The most prominent are MAGPIE (Gaasterland and Sensen 1996), PEDANT Pro (Frishman et al. 2001), WIT/ERGO (Overbeek et al. 1999, 2000) and ARTEMIS (Rutherford et al. 2000). Currently, the most advanced is the recently developed GenDB system, which is furthermore free for academic use (Meyer et al. 2003).

Without going into detail about the pros and cons of the different annotation systems, the authors of this chapter would like to state that the data model and versatility of GenDB seems to be the most appropriate for the emerging demands of metagenomics. Therefore, we will restrict our description of the annotation process to this system. For a local installation, it can be obtained on several DVDs (see www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/appl.htm). Thereby, it should be noted that, for good performance, at least a small cluster of Unix-based servers (5 - 10), appropriate network connections and additional tools like the Sun Grid Engine (http://gridengine.sunsource.net/) and MySQL (http://www.mysql.com/) are needed. Starting with gene prediction, several options can currently be chosen in GenDB: (1) run single gene finders like Glimmer (Delcher et al. 1999), Critica (Badger and Olsen 1999) or Getorf (http://emboss.sourceforge.net/apps/getorf.html), (2) use a combination of two gene finders united together in the tool Reganor (McHardy et al. 2004). In practice it seems that, for large genome fragments (over several hundred kilobasepairs), the Reganor system gives significantly improved specificity compared to, e.g. Glimmer alone, but for BAC- and fosmid-sized sequences, it is usually better to go with Glimmer. The reason is that Reganor has a slight tendency for gene underprediction (McHardy et al. 2004). This might cause gene loss, which is especially dangerous when working on small fragments with only a limited set of genes originally present. The inherent tendency for 20% gene overprediction (Guo et al. 2003) by using Glimmer alone can be taken into consideration since, for BAC and Fosmid sized fragments, this will only cause between eight and 20 additional ORFs to be spuriously identified. The standard tools and databases for providing functional observations for the predicted genes can be found in Table 8.1. Information about the location of a protein can be procured by the prediction of signal peptides with signalP (Bendtsen et al. 2004) and transmembrane helices prediction with TMHMM (Krogh et al. 2001). tRNAscan-SE (Lowe and Eddy 1997) can be used to find and assign transfer RNAs within the sequence.

Table 8.1. Standard tools and databases providing functional observations

Tool	Database	Reference
BLASTn	GenBank	http://www.ncbi.nlm.nih.gov
	EMBL	http://www.ebi.ac.uk/
BLASTp or BLASTx	GenBank	http://www.ncbi.nlm.nih.gov
	UniProt	Apweiler et al. (2004)
	Swiss-Prot	Boeckmann et al. (2003)
HMMER	Pfam	Bateman et al. (2004)
InterProscan[a]	InterPro	Mulder et al. (2003)

[a] InterPro itself is a metadatabase that provides access to commonly used signature database, like Prosite, Prints, Pfam, ProDom, SMART, TIGR-fams, SCOP, Cath and MSD (see http://www.ebi.ac.uk/interpro/)

Once the calculations are finished, automatic annotation systems like Metanor (provided by GenDB) or MicHanThi (currently being developed at the Max Planck Institute for Marine Microbiology, Bremen), try to automatically generate annotations for all predicted genes, based on the observations returned by the individual tools. This supports the manual annotation process by providing additional information for decision-making. The MicHanThi system is currently able to separate and annotate *hypothetical* and *conserved hypothetical* genes in a nearly quantitative manner. For genes with significant hits in primary or secondary databases, like UniProt or Pfam, the system is consistently able to assign the correct functional category. In the subsequent manual annotation process, every predicted gene has to be investigated for significant hits to entries in the databases. Starting with hits to Swiss-Prot and taking into account Pfam and InterPro results, the annotators have to integrate the information, read additional literature and finally assign a certain function to the gene. GenDB supports this process by providing graphical representations of the coverage of BLAST hits, the relative location of Pfam and InterPro hits, as well as signal peptides and transmembrane helices. After a gene function has been assigned, the annotation should be supplemented by information on the gene name, Enzyme Commission (EC) number and Gene Ontology (Ashburner et al. 2000) classifications.

With the history system implemented in GenDB, all annotation changes are tracked and thus parallel annotations by different experts can be handled for every gene. To make assignments consistent, it is highly recommended to give a list of stringent annotation rules to all people involved in the process. Initial training of the annotators, continuous monitoring and a final crosscheck of the annotations are also needed to achieve high quality functional assignments. The experience gained from several genome and metagenome projects processed so far demonstrates that there is no way of

circumventing the manual inspection of each predicted ORF. After finishing the annotation process, metabolic reconstructions can be performed as far as possible. The easiest way to do so is to automatically map the EC numbers to the corresponding KEGG pathway maps (Kanehisa et al. 2004) provided by the GenDB system.

Limitations of the GenDB system are that it cannot currently handle thousands, or even several hundreds, of metagenome fragments in a single project and treat them like a single 'meta-organism' for, e.g. metabolic reconstruction. Furthermore, comparison between or the assembly of overlapping fragments is not implemented in the current GenDB version (2.0.1). This situation will be significantly enhanced with the next release of GenDB, which will be merged with the comparative genomics tool 'The Seed' (http://theseed.uchicago.edu/FIG/; Meyer, personal communication).

8.4 Summary, Pitfalls and Outlook

Metagenome analysis is the method of choice to access the untapped functional diversity beyond cultivation-driven approaches. Sequence- and function-driven explorations have already greatly impacted basic research and biotechnological applications (for reviews, see DeLong 2004; Riesenfeld et al. 2004). Nevertheless, there are some limitations that have to be addressed, the first being the size of the metagenomic library itself. Assuming a marine sample with a mid-range diversity of 100 species/ml, equalling about 500 Mbp of unique DNA, about 58,000 fosmid-sized clones are needed in theory to clone every part of the metagenome with a probability of at least 99% (see formula in Sect. 8.2.3). An unequal density of community members in most cases means that only the more abundant species will be represented in the library. To address minor populations (< 1%) 100 – 1,000 times more clones are needed, which still exceeds currently available technological resources. Furthermore, although BAC- and fosmid-sized cloning approaches have clear advantages in providing more information about gene context and in reducing the number of clones needed, the cell-lysing procedure used with these methods is gentle compared to that of small-insert libraries. Therefore, the corresponding metagenomic library will often be shifted in favour of organisms with 'easy to open' cell walls. As with genomic studies, gene product toxicity is also a concern in metagenomic analysis. As mentioned above, vectors with inducible copy number help to minimize this problem. The assembly procedure, especially for shotgun approaches, is an additional issue only sparsely addressed so far. The danger of assembling 'virtual chimeric organisms' is obvious and can only be avoided with more sophisticated methods coming from genome

linguistics. To guide the assembly process, substantial efforts have to be invested in order to sequence as many cultivated environmental organisms as possible (Tyson et al. 2004). The ongoing genome sequencing projects underway at Genoscope, the Joint Genome Institute and those financed by the The Betty and Gordon Moore Foundation are sorely needed. To solve the aforementioned problems, high-throughput automation is needed in library construction, screening, sequencing and analysis. Bioinformatics will be the key discipline for storing, analysing and classifying the flood of data, by keeping pace with the exponentially growing sequencing efforts. The bottlenecks that can already be identified are appropriate post-metagenomic techniques, like expression analysis at the gene and protein levels, along with functional characterisation of key enzymes. "The big picture" (Rodriguez-Valera 2004) will only emerge if we can manage to generate, join and integrate all these data together with information on the biogeochemistry of the habitats (www.megx.net). If this happens, we may be able to understand some additional basic principles about the development and future of life on earth.

Acknowledgements. We thank Thierry Lombardot, Hanno Teeling, Katrin Knittel, Marc Mußmann, Nicole Dubilier, Marisano James and Alexander Goesmann for improving the manuscript by critical reading and helpful suggestions.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13:693–702

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Apweiler R, et al (2004) UniProt: the universal protein knowledgebase. Nucleic Acid Res 32:D115–D119

Asakawa S, et al (1997) Human BAC library: construction and rapid screening. Gene 191:69–79

Ashburner M, et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16:512–524

Bateman A, et al (2004) The Pfam protein families database. Nucleic Acid Res 32:D138–D141

Beja O, et al (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906

Beja O, et al (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ Microbiol 2:516–529

Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. Nature 411:786–789

Beja O, et al (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl Environ Microbiol 68:335–345

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Boeckmann B, et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acid Res 31:365–370

Chenna R, et al (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acid Res 31:3497–3500

Cole JR, et al (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acid Res 31:442–443

Collins J, Hohn B (1978) Cosmids – type of plasmid gene-cloning vector that is packageable invitro in bacteriophage lambda-heads. Proc Natl Acad Sci USA 75:4242–4246

Connon SA, Giovannoni SJ (2002) High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol 68:3878–3885

Cottrell MT, Moore JA, Kirchman DL (1999) Chitinases from uncultured marine microorganisms. Appl Environ Microbiol 65:2553–2557

Courtois S, et al (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl Environ Microbiol 69:49–55

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99:10494–10499

De la Torre JR, et al (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci USA 100:12830–12835

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acid Res 27:4636–4641

DeLong EF (2004) Microbial population genomics and ecology. In: Fraser CM, Read TD, Nelson KE (eds) Microbial genomics. Humana, Totowa, pp 419–442

Dennis P, Edwards EA, Liss SN, Fulthorpe R (2003) Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl Environ Microbiol 69:769–778

Entcheva P, Liebl W, Johann A, Hartsch T, Streit WR (2001) Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl Environ Microbiol 67:89–99

Frishman D, et al (2001) Functional and structural genomics using PEDANT. Bioinformatics 17:44–57

Gaasterland T, Sensen CW (1996) MAGPIE: automated genome interpretation. Trends Genet 12:76–78

Garrity GM, Jonson KL, Bell J, Searles DB (2002) Taxonomic outline of the prokaryotes. In: Bergey's manual of systematic bacteriology, 2nd edn. Springer, Berlin Heidelberg New York

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso sea bacterioplankton. Nature 345:60–63

Goodman RM, Liles M (2001) Template specific termination in a polymerase chain reaction. US Patent

Green ED, Birren B, Klapholz S, Myers RM, Hieter P (1997) Genome analysis: a laboratory manual, 1st edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Guo FB, Ou HY, Zhang CT (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acid Res 31:1780–1789

Hallam SJ, Girguis PR, Preston CM, Richardson PM, DeLong EF (2003) Identification of methyl coenzyme M reductase A (*mcrA*) genes associated with methane-oxidizing archaea. Appl Environ Microbiol 69:5483–5491

Handelsman JRM (1998) Molecular biological access to the chemistry of unknown soil microbes – a new frontier for natural products. Chem Biol 5:R245–R249

Handelsman J, Liles M, Mann D, Riesenfeld C, Goodman RM (2002) Cloning the metagenome: culture-independent access to the diversity and functions of the uncultivated microbial world. In: Brendan W, Dorrell N (eds) Functional microbial genomics. Academic, San Diego, pp 241–255

Henne A, Daniel R, Schmitz RA, Gottschalk G (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl Environ Microbiol 65:3901–3907

Hughes DS, Felbeck H, Stein JL (1997) A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. Appl Environ Microbiol 63:3494–3498

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acid Res 32:D277–D280

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185–225

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acid Res 30:3059–3066

Kim UJ, Shizuya H, Dejong PJ, Birren B, Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an *F*-factor based vector. Nucleic Acid Res 20:1083–1085

Kim UJ, et al (1996) Construction and characterization of a human bacterial artificial chromosome library. Genomics 34:213–218

Korf I, Yandell M, Bedell J (2003) BLAST. O'Reilly & Associates, Cambridge

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Krüger M, et al (2003) A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. Nature 426:878–881

Leonardo ED, Sedivy JM (1990) A new vector for cloning large eukaryotic DNA segments in *Escherichia coli*. Bio-Technology 8:841–844

Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl Environ Microbiol 69:2684–2691

Lopez-Garcia P, Brochier C, Moreira D, Rodriguez-Valera F (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. Environ Microbiol 6:19–34

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acid Res 25:955–964

Ludwig W, et al (2004) ARB: a software environment for sequence data. Nucleic Acid Res 32:1363–1371

MacNeil IA, et al (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. J Mol Microbiol Biotechnol 3:301–308

McHardy AC, Goesmann A, Pühler A, Meyer F (2004) Development of joint application strategies for two microbial gene finders. Bioinformatics 20:1622–1631

Meyer F, et al (2003) GenDB – an open source genome annotation system for prokaryote genomes. Nucleic Acid Res 31:2187–2195

Monaco AP, Larin Z (1994) Yacs, Bacs, Pacs and Macs – artificial chromosomes as research tools. Trends Biotechnol 12:280–286

Mount DW (2001) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Mulder NJ, et al (2003) The InterPro database, 2003 brings increased coverage and new features. Nucleic Acid Res 31:315–318

Nelson KE (2003) The future of microbial genomics. Environ Microbiol 5:1223–1225

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40:337–365

Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6

Osoegawa K, et al (1998) An improved approach for construction of bacterial artificial chromosome libraries. Genomics 52:1–8

Osoegawa K, et al (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. Genome Res 10:116–128

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896–2901

Overbeek R, et al (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acid Res 28:123–125

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740

Pace NR, Stahl DA, Olsen GJ, Lane DJ (1985) Analyzing natural microbial populations by rRNA sequences. ASM News 51:4–12

Peplies J, Lau SCK, Pernthaler A, Amann R, Glöckner FO (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. Environ Microbiol 6:638–645

Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. Proc Natl Acad Sci USA 99:14002–14007

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 13:145–158

Quaiser A, et al (2002) First insight into the genome of an uncultivated crenarchaeote from soil. Environ Microbiol 4:603–611

Quaiser A, et al (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. Mol Microbiol 50:563–575

Rappe MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418:630–633

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525–552

Rodriguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiol Lett 231:153–158

Rondon MR, Raffel SJ, Goodman RM, Handelsman J (1999) Toward functional genomics in bacteria: analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. Proc Natl Acad Sci USA 96:6451–6455

Rondon MR, et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541–2547

Rutherford K, et al (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Sambrook J, Russel DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res 11:1404–1409

Schleper C, Swanson RV, Mathur EJ, DeLong EF (1997) Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. 179:7803–7811

Schleper C, Delong EF, Preston CM, Feldman RA, Wu KY, Swanson RV (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 180:5003–5009

Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14:303–310

Schmeisser C, et al (2003) Metagenome survey of biofilms in drinking-water networks. Appl Environ Microbiol 69:7298–7309

Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173:4371–4378

Sebat JL, Colwell FS, Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. Appl Environ Microbiol 69:4927–4934

Sheng Y, Mancino V, Birren B (1995) Transformation of *Escherichia coli* with large DNA molecules by electroporation. Nucleic Acid Res 23:1990–1996

Shizuya H, et al (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an *F*-factor-based vector. Proc Natl Acad Sci USA 89:8794–8797

Sonnhammer ELL, Durbin R (1994) A workbench for large-scale sequence homology analysis. Comput Appl Biosci 10:301–307

Stahl DA, Amann R (1991) Development and application of nucleic acid probes. In: Stackebrandt E, Goodfellow M (eds) Nucleic acid techniques in bacterial systematics. Wiley, Chichester, pp 205–248

Stahl DA, Tiedje JM (2002) Microbial ecology and genomics: a crossroad of opportunity. American Academy of Microbiology, Washington, D.C.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. J Bacteriol 178:591–599

Swofford DL, Olsen GJ, Waddel PJ, Hillis DM (1996) Phylogenetic Inference. In: Hillis DM, Moritz C, Marble BK (eds) Molecular systematics, 2nd edn. Sinauer Associates, Sunderland, Mass., pp 407–514

Tao Q, Zhang HB (1998) Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res 26:4901–4909

Tatusov RL, et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO (2004a) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6:938–947

Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004b) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences sequences. BMC Bioinform 5:163

Torsvik V, Goksoyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Appl Environ Microbiol 56:782–787

Tyson GW, et al (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Uchiyama T, Abe T, Ikemura T, Watanabe K (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nat Biotechnol 23:88–93

Venter JC, et al (2004) Environmental genome shotgun sequencing of the Sargasso sea. Science 304:66–74

Walker DR, Koonin EV (1997) SEALS: a system for easy analysis of lots of sequences. Intell Sys Mol Biol 5:333–339

Ward DM, Weller R, Bateson MM (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345:63–65

Wild J, Hradecna Z, Szybalski W (2002) Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. Genome Res 12:1434–1444

Woo SS, Jiang JM, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of sorghum-bicolor. Nucleic Acid Res 22:4922–4931

Zengler K, et al (2002) Cultivating the uncultured. Proc Natl Acad Sci USA 99:15681–15686

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl Environ Microbiol 62:316–322

Zimmer R, Verrinder GA (1997) Construction and characterization of a large-fragment chicken bacterial artificial chromosome library. Genomics 42:217–226

9 DNA Microarrays for Bacterial Genotyping

Ulrich Nübel, Markus Antwerpen, Birgit Strommenger, Wolfgang Witte

9.1 Introduction

DNA microarrays provide a means for the detection of thousands of discrete nucleic acid sequences in a single experiment. They are widely used tools in molecular biology research for profiling differential gene expression and studying DNA variation. Since DNA microarrays were invented in the early 1990s, most of the technological development has focused on analyses of mammalian gene expression. However, recent years have witnessed a rapid increase in microarray usage for bacterial genotyping (Fig. 9.1). The majority of these studies has been concerned with the identification of genomic differences among related bacterial isolates or the detection of diagnostic marker sequences in clinical or environmental samples.

Microarrays designed for screening complete sets of predicted open reading frames from given bacterial chromosomes are direct offsprings of bacterial genome sequencing projects, numerous of which have been completed recently (for updated overviews, see: http://www.genomesonline.org; http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Such whole-genome microarrays enable comprehensive inventories of all the genes of a bacterium in overnight experiments. In contrast, diagnostic microarrays, for economic reasons, commonly apply significantly lower numbers of diagnostic features. They are being developed for the detection of key virulence and antibiotic resistance genes for a growing number of pathogenic bacteria. Other microarrays achieve phylogenetic bacterial identification, mutation detection, or hybridization-based multi-locus sequence typing by interrogating suitable nucleic acid molecules. In environmental microbiology, several research groups have developed specialized DNA microarrays to profile microbial communities in samples from diverse settings, for example, soil, sea water, or human feces. Many of these tools currently are in the "proof of principle" state and may be introduced to routine and automated clinical diagnostics or environmental monitoring in the future.

Ulrich Nübel: Robert Koch Institut, Burgstrasse 37, 38855 Wernigerode, Germany, E-mail: nuebelu@rki.de

Molecular Identification, Systematics, and Population Structure of Prokaryotes
E. Stackebrandt (Ed.)

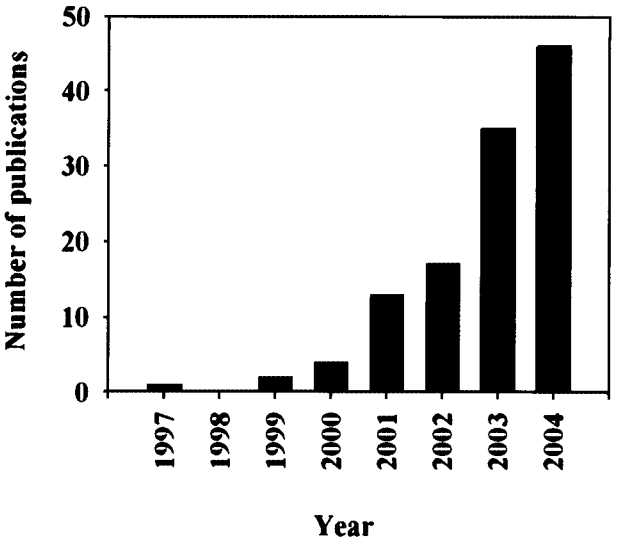


Fig. 9.1. Numbers of scientific publications on the use of microarrays for bacterial genotyping, in the years 1997–2004

As applications and requirements for microarrays for bacterial genotyping are diverse, so are the concomitant concepts for technical realization. DNA microarrays differ broadly with respect to their manufacturing process, probe type and density, support material, and methods for DNA hybridization and detection. Both commercial and "homebrew", or self-printed, arrays are being used. In this chapter, we review conceptually different applications of DNA microarrays for bacterial genotyping. As an introduction to the field, we start with a concise overview on the technical principals being employed.

9.2 Technical Principles

DNA microarrays apply nucleic acid molecules (commonly referred to as probes) that are immobilized on solid supports, to interrogate nucleic acid molecules (referred to as targets) from a sample. They make use of the unique feature of nucleic acids to form duplex structures among complementary molecules and have evolved from membrane-based "blotting" methods (Southern et al. 1999). In contrast to the use of Nylon or nitrocellulose membranes as support materials, the introduction of rigid, impermeable, flat substrates, such as glass, enabled the drastic miniaturization of DNA arrays and facilitated fluorescence-based detection and automated array manufacture and handling (Schena et al. 1995; Southern et al. 1999).

Probes usually are DNA, but may also consist of nucleic acid analogues, such as peptide nucleic acids (Brandt et al. 2003). DNA probes can be PCR amplification products, cDNA, plasmid DNA, or synthetic oligonucleotides of variable length. Even entire bacterial genomes are conceivable as microarray probes, analogous to genome dot blots on membranes (Shen et al. 1998; Zhang et al. 2004). While arrays equipped with cDNAs or PCR amplicons for probes remain popular for research application, oligonucleotides offer a number of advantages. They get designed on the basis of sequence databases and are then chemically synthesized, with lengths of up to approximately 100 nucleotides. Quality control is performed cost-effectively by HPLC or mass spectrometry. In contrast, reference DNAs are needed for the enzymatic production of PCR amplicons and sequence analyses are essential to exclude erroneous probes. Short oligonucleotide probes on microarrays enable highly specific hybridizations, discriminating target DNAs which differ at single nucleotide positions only (Bodrossy et al. 2003; Nübel et al. 2004; Urakawa et al. 2003). Long oligonucleotides, in turn, have been reported to allow for low detection limits, comparable to PCR amplicon probes (Kane et al. 2000; Letowski et al. 2004).

Probes may be synthesized and subsequently gridded onto the substrates using robotic printers. In this way, glass slides of the standard format (25 × 75 mm) may be assembled with several thousand discrete probes. Considerable effort has been invested in the development of chemical methods for attaching probes to derivatized or unmodified glass surfaces (Lindroos et al. 2001; Zammatteo et al. 2000). The methodology used may strongly influence the amount of probe immobilized per slide area, its steric availability for hybridization, the signal-to-noise ratio achieved, and array costs. Specialized slides are available commercially, together with recommendations for appropriate immobilization protocols.

As an alternative to printing pre-fabricated probes onto slides, oligonucleotides may be directly synthesized onto the array support. Three competing technologies have been developed. The company Affymetrix uses photolithographic masking techniques similar to those applied for computer chip manufacturing (Fodor et al. 1991). In this way, extremely high probe densities of 10^6 probes cm^{-2} are achieved. High set-up and manufacturing costs, however, limit this technology to large-volume applications. In contrast, bench-top machines are available for maskless light-directed oligonucleotide syntheses, which apply micromirror arrays to photo-deprotect the ends of oligonucleotide chains to be grown (Singh Gasson et al. 1999). Another approach recently introduced ink-jet printing technology to deliver synthesis reagents to individual probe locations on glass slides (Hughes et al. 2001).

The most common approach to screen for sequence complementation between probes on the microarray and target nucleic acids in a sample

involves labeling the target DNA with fluorescent dyes and detecting hybridization events by use of a specialized fluorimetric scanner. High image resolution is required to cope with the small feature sizes achieved on microarrays. Nucleic acid labeling may be achieved enzymatically, for example through DNA polymerase-mediated insertion of modified nucleotides (Vora et al. 2004), or chemically (Kelly et al. 2002; Zhang et al. 2001). Reagent kits for nucleic acid labeling with a number of popular fluorescent dyes are available from various manufacturers. Aside from fluorescence-based assays, non-fluorescent dyes, radioisotopes, and nanometer-sized gold particles have all been successfully used to label and detect DNA on microarrays (Taton et al. 2000). In addition, several companies are experimenting with systems that provide direct electrical readouts with spatial resolution to avoid scanometric, optical detection methods altogether (Gabig-Ciminska et al. 2004; McKendry et al. 2002). These technologies may allow a reduction in the size and cost of analytical devices. Hybridization of target DNA to arrayed oligonucleotide probes may also be combined with enzymatic polymerase or ligase reaction assays to improve the specificity of the reaction (Tönnisson et al. 2000) or to enhance the fluorescence signal based on DNA amplification (Adessi et al. 2000; Westin et al. 2001).

9.3 Applications

9.3.1 Comparative Genome Hybridization

In 1995, the sequencing of the first entire bacterial chromosome – from the facultative human pathogen *Haemophilus influenzae* – was successfully accomplished (Fleischmann et al. 1995). Since that breakthrough, sequencing technology has been constantly improved and accelerated, accompanied by a significant reduction in costs. As a consequence, at the time of writing this chapter (December 2004), genome sequences from 239 prokaryotic isolates have been published and another 534 are in progress, according to the Genomes online database (www.genomesonline.org). The wealth of sequence data generated has provided unprecedented insights into the evolution of prokaryotes, including most of the bacteria pathogenic to humans. Perhaps the most remarkable finding is the highly dynamic and mosaic structure of many bacterial chromosomes due to frequent gene acquisition and loss (Daubin et al. 2003; Gogarten et al. 2002).

Despite indisputably enormous progress, the sequence analysis and concomitant bioinformatic data management of a single bacterial genome still requires several months; and the accompanying costs are prohibitive for

epidemiological studies on large numbers of isolates. Once the genome sequence of a bacterium is established, however, it can be compared to those from related isolates through "comparative genome hybridization" analyses applying DNA microarrays. These microarrays are usually equipped with probes against all open reading frames predicted in silico on the basis of genome sequence data. Pairwise comparisons are then performed by simultaneously hybridizing to the microarray genomic DNAs that have been extracted from two bacterial isolates and labeled with different fluorescent dyes. The fluorescence signals at each probe indicate whether the interrogated genomic region is present in both genomes or is absent in a tested isolate. Naturally, by using arrays based on a single genome, only genes present in the previously sequenced isolate can be detected and unknown genes unique to newly tested isolates will be missed. To reduce this limitation, more recently microarrays have been introduced that are based on complete sets of genes from several isolates of a species (Dunman et al. 2004; Lindsay et al. 2004; Porwollik et al. 2003).

Microarray-based comparative genome hybridizations have accompanied many of the more recent genome sequencing projects and a comprehensive review of the literature would fill an entire chapter. In fact, microarrays representing the genomes from several bacterial species have become commercially available and are on the way to become standard tools in the molecular biology laboratory (Table 9.1). We will therefore focus on a few selected examples to illustrate the potential of this technique. Due to their social and economic relevance, research on pathogenic bacteria is most advanced. Comparisons of genomes from related bacterial pathogens with different disease phenotypes have elucidated the origins and evolution of infectious diseases (Hacker et al. 2003), discovered previously unknown virulence-associated genes, suggested novel strategies for therapy and containment, and indicated targets suitable for nucleic acid-based diagnostics.

The first whole-genome DNA microarray reported was derived from the genome sequence from *Mycobacterium tuberculosis* and was equipped with 4,896 PCR-generated probes, representing 99.4% of the predicted open reading frames (Behr et al. 1999). It was used to investigate the genetic differences between *M. tuberculosis*, *M. bovis*, and attenuated strains of the latter species, which are regularly administered as vaccines against tuberculosis. Behr et al. (1999) detected 16 large genomic regions that were deleted in the vaccine strains, encompassing 129 open reading frames. These data enabled these authors to reconstruct the genealogy of 13 variants of the tuberculosis vaccine and to determine when historically each of the deletions in the genomes of the attenuated strains had occurred. Moreover, correlations of genomic composition and associated phenotypes suggested rational approaches for the design of improved vaccines (Behr et al. 1999).

Table 9.1. Whole bacterial genome microarrays from commercial suppliers: Affymetrix (Santa Clara, Calif.; www.affymetrix.com), DNAMicroarray (San Diego, Calif.; www.dnamicroarray.com), TIGR (the Institute for genomic research, Pathogen functional genomics resource center, Rockville, Md.; www.pfgrc.tigr.org/), Eurogentec (Seraing, Belgium; www.eurogentec.com), Scienion (Berlin, Germany; www.scienion.de), MWG Biotech (Ebersberg, Germany; www.mwg-biotech.com), IFR (IFR Microarray Facility, Norwich, UK; www.ifr.bbsrc.ac.uk), Sigma-Genosys (The Woodlands, Tex.; www.sigma-genosys.com), Cambrex (Cambrex BioScience, East Rutherford, N.J.; www.cambrex.com), Operon (Cologne, Germany; www.operon.com)

Organism	Strain (if specified)	Probe type (if specified)	No. of probes	Supplier
Bacillus anthracis	Ames, A2012	70-mer	5,823	TIGR
B. anthracis, *B. cereus*[a]	Ames, A2012, ATCC14579	70-mer	5,309	Operon
B. subtilis		25-mer	[b]	Affymetrix
B. subtilis		Amplicon	4,096	Eurogentec
Bordetella pertussis[a]		70-mer		Operon
Campylobacter jejuni[a]	NCTC11168	70-mer	1,601	Operon
C. jejuni		50-mer	1,632	MWG Biotech
C. jejuni		70-mer		DNAMicroarray
C. jejuni			1,700	IFR
Chlamydophila pneumoniae	D/UW-3/Cx, AR39, CWL029, J138	70-mer	2,065	TIGR
C. pneumoniae[a]	AR39, CWL029, TW-183	70-mer	1,350	Operon
Clostridium botulinum	Hall strain A	70-mer	3,618	TIGR
Escherichia coli, *Shigella flexneri*	K12, O157, S.f. 2a str 301		6,379	IFR
E. coli[a]	K12, O157:H7 (EDL 933), O157:H7 (RIMD)	70-mer	5,978	Operon
E. coli		25-mer	[b]	Affymetrix
E. coli	K12, O157:H7 (EDL 933), O157:H7 (RIMD)	50-mer	6,176	MWG Biotech
E. coli	K12	50-mer	4,288	MWG Biotech.
E. coli	K12	25-mer	[b]	Affymetrix
E. coli		70-mer		DNAMicroarray
E. coli	K12	Amplicon	4,155	Cambrex
E. coli	K12	Amplicon	4,000	Scienion

Table 9.1. (continued)

Organism	Strain (if specified)	Probe type (if specified)	No. of probes	Supplier
Haemophilus influenzae		70-mer		DNAMicroarray
H. influenzae[a]	Rd	70-mer	1,714	Operon
Helicobacter pylori		Amplicon	1,681	Sigma-Genosys
H. pylori	J99, 26695	50-mer	1,877	MWG Biotech
H. pylori		Amplicons	1,621	Eurogentec
H. pylori	26695, j991	70-mer	2,572	TIGR
Lactococcus lactis		Amplicon		Eurogentec
Listeria monocytogenes	EGD-e, F2365 (4b) F6854 (1/2a). H7858 (4b)	70-mer	6,347	TIGR
L. monocytogenes[a]	EGD	70-mer	2,857	Operon
Mycobacterium smegmatis	MC2155	70-mer	6,746	TIGR
M. tuberculosis		Amplicon	3,875	Sigma-Genosys
M. tuberculosis	H37Rv, CDC1551	70-mer	4,127	TIGR
Neisseria gonorrhoeae[a]	MC58, Z2491, FAM18, ALPHA14	70-mer	2,872	Operon
N. gonorrhoeae; *N. meningitidis*	FA1090, ATCC700825, Z2491 (A), MC58 (B)	70-mer	6,389	TIGR
Pseudomonas aeruginosa		25-mer	[b]	Affymetrix
Salmonella enterica	LT2a		4,414	IFR
Salmonella spp.	LT2a, DT104, SL1344, PT4, 287/91		5,080	IFR
S. enterica[a]		70-mer	5,578	Operon
S. enterica	LT2, CT18	Amplicon	5,405	TIGR
Staphylococcus aureus		Amplicon	2,334	Scienion
S. aureus	N315, Mu50, NCTC8325, Col	25-mer	[b]	Affymetrix.
S. aureus	COL, Mu50, MW2, N315	Amplicon	2,480	TIGR
Streptococcus agalactiae	NEM316, 2603 V/R, A909	70-mer	2,850	TIGR
S. mutans	UA159	70-mer	1,948	TIGR
S. pneumoniae			2,200	IFR
S. pneumoniae		Amplicons	2,085	Eurogentec
S. pneumoniae	R6	50-mer	2,043	MWG Biotech
S. pneumoniae	TIGR4, R6, G54	Amplicon	2,131	TIGR
Vibrio cholerae	N16961 biotype ElTor	70-mer	3,811	TIGR
Yersinia pestis	CO92, KIM	70-mer	4,829	TIGR

[a] oligonucleotide probes are supplied

[b] approx. 20 probe pairs per putative open reading frame

Staphylococcus aureus is the most common cause of infections in hospitals, which may result in life-threatening endocarditis, septicemia, or toxic-shock syndrome. Most alarmingly, this species is becoming increasingly resistant to antibiotics. Genomes from seven strains have been sequenced to date. Comparative hybridizations of genomic DNAs from 36 strains to a microarray equipped with amplicon probes derived from the genome sequence of strain COL established that 22% of the *S. aureus* genome is strain-specific (Fitzgerald et al. 2001). Eighteen large genomic difference regions were detected that consisted mostly of mobile (or once mobile) genetic elements, including bacteriophages, pathogenicity islands, plasmids, and transposons. Many of the genes on these elements have virulence and resistance functions; and their apparently frequent horizontal transfer among strains may have important clinical implications (Fitzgerald et al. 2001; Lindsay and Holden 2004). For example, it was discovered that strains of methicillin-resistant *S. aureus* (MRSA), which are a serious public health threat, must have evolved multiple times independently through horizontal transfer of the SCC*mec* element, since it was found in the genomes of five distinct phylogenetic groupings (Fitzgerald et al. 2001). More recently, microarrays have been constructed to incorporate every open reading frame from seven sequenced *S. aureus* genomes (Dunman et al. 2004; Lindsay et al. 2004). These tools enable even more comprehensive surveys of the genomic make-up of novel isolates than any array based on a single genome and have been demonstrated to provide superior discriminative power for strain typing when compared to pulsed-field gel electrophoresis and ribotyping (Dunman et al. 2004).

Bacillus anthracis is a highly pathogenic bacterium that may cause anthrax. Phylogenetically, this species is an offshoot of the closely related *B. cereus*, which is a comparatively harmless, opportunistic pathogen causing food poisoning. By comparative hybridization of genomic DNA from a number of *B. cereus* strains to a microarray based on the genome sequence from *B. anthracis*, it was established that genomes of the two species had up to 92% of genes in common and that most of the putative chromosomal virulence genes known from *B. anthracis* had homologues in *B. cereus*, including genes encoding hemolysins, phospholipases, and iron acquisition factors (Read et al. 2003). Even though several smaller genomic regions were detected that apparently were unique to *B. anthracis*, the drastically different pathogenicity was attributed to genes on species-specific plasmids and nonsense mutations in an important positive regulator of gene expression (Read et al. 2003).

9.3.2 Diagnostic Detection of Virulence Genes

Their capabilities for highly parallel gene detection make DNA microarrays attractive for diagnostic applications in the fields of medical microbiology, food microbiology, and environmental monitoring. Specialized microarrays may be used for typing cultivated bacterial isolates or for the detection of bacterial nucleic acids in clinical or environmental samples. Assays for the latter type of application usually include gene-specific PCR amplification of target DNA prior to microarray hybridization. Compared to whole-genome arrays applied for research, the number of probes on diagnostic arrays is usually low for economic reasons.

Bacterial virulence factors include adhesins, invasins, capsules, toxins, siderophores, and secretion systems. The genes encoding virulence factors (virulence genes) determine the potential pathogenic properties of a bacterium and can have strong impact on clinical symptoms of the infectious disease it may cause. Since virulence genes are often located on mobile genetic elements, horizontal transfer of genes between strains may cause even closely related bacteria to differ significantly with respect to their pathogenic potential (Bekal et al. 2003; Hacker et al. 2003).

Most diagnostic microarrays for detection of bacterial virulence genes have applied oligonucleotide probes. Call et al. (2001) described an assay for detecting and genotyping enterohemorrhagic *Escherichia coli* directly from chicken rinsate, which included a cultivation enrichment step, immunomagnetic cell capture, PCR amplification of four virulence-associated target genes, and subsequent hybridization to an oligonucleotide microarray. A group at the United States Federal Drug Administration published a series of papers describing oligonucleotide microarrays for detecting subsets of virulence genes from *E. coli, Shigella* spp., and *Salmonella enterica* (Chizhikov et al. 2001), *Listeria* spp. (Volokhov et al. 2002), *Campylobacter* spp. (Volokhov et al. 2003a), *B. anthracis* (Volokhov et al. 2004), and *Clostridium perfringens* (Al-Khaldi et al. 2004). All these assays included gene-specific PCR amplification and subsequent hybridization to short oligonucleotide probes (usually, 20 – 25 nucleotides in length). In most cases, target genes were chosen to enable the identification of the respective bacterial species. Sergeev and coworkers (2004) from the same laboratory described the use of an oligonucleotide microarray for detection and discrimination between 17 major serological types of staphylococcal heat-stable enterotoxins that are among the leading causes of gastroenteritis following the consumption of contaminated food.

A microarray with 383 oligonucleotide probes based on putative virulence-associated genes present in sequenced genomes of four *Staphylococcus aureus* strains was used for typing 12 *S. aureus* strains of different

geographic origin. Differences in the topologies of the resulting gene difference distance tree and a phylogenetic tree based on nucleotide sequences from housekeeping genes were accounted to sporadic horizontal genetic transfer (Saunders et al. 2004). Another microarray was invented for the simultaneous detection of five important marine fish pathogens (Gonzalez et al. 2004).

Two microarrays equipped with PCR amplicon probes (25 and 105 probes, respectively) targeting virulence genes from *E. coli* were reported. Hybridizations with fluorescently labeled bacterial genomic DNAs enabled a genotypic differentiation of *E. coli* isolates; and it was predicted that these tools may facilitate the identification of newly emerging pathotypes in the future (Bekal et al. 2003; Van Ijperen et al. 2002). Similarly, Marokhazi et al. (2003) applied a microarray with 96 amplicon probes to study the distribution of toxin genes in insect-pathogenic *Photorhabdus* spp.

9.3.3 Diagnostic Detection of Resistance Determinants

The past decade has seen a steady increase in the incidence of antibiotic-resistant bacteria. This also includes the emergence of multidrug-resistant isolates which constitute a major problem, especially in the nosocomial setting (Witte 1999). Rapid determination of the antimicrobial susceptibility of a clinical isolate is therefore crucial to prevent treatment failures. Besides that, the monitoring of antibiotic resistant organisms or resistance genes is essential epidemiologically to monitor and prevent the spread of multi-resistant organisms inside a hospital and between hospitals and the community (Fluit et al. 2001). Genetic causes of resistance may be horizontally acquired resistance genes, often located on mobile genetic elements including plasmids, or chromosomal mutations in genes encoding target proteins for the respective antimicrobial agents. Rapid detection of these genetic determinants can predict resistances and hence may assist drug prescribers, especially when dealing with infections caused by bacteria that – under laboratory conditions – grow very slowly and require several days for culture-based susceptibility testing (Bergeron and Ouellette 1998). Microarray technology offers the tools for parallel screening for a multitude of relevant resistance traits.

Detection of Resistance Genes

In 2001, Hamels et al. (2001) published the first report on a diagnostic microarray for the identification of nosocomially important MRSA isolates. This array enabled the detection of the species-specific marker *fem*A, together with *mec*A, the gene for penicillin binding protein 2a, which is

responsible for methicillin resistance. An expanded array for the detection of various antibiotic resistance genes and some relevant toxin genes of *S. aureus* was described (Monecke et al. 2003). Similar arrays focusing on the same pathogen are now commercially available (Chipron, Berlin, Germany; Clondiag, Jena, Germany). Volokhov et al. (2003b) designed a microarray for the specific detection of six different genes (*erm*A, *erm*B, *erm*C, *ere*A, *ere*B, *msr*A/B) leading to resistance to macrolide, lincosamide, and streptogramin B compounds in *S. aureus* and *Streptococcus pyogenes*. All microarrays mentioned so far are based on oligonucleotide probes and utilize a PCR amplification step to reach a detection limit appropriate for diagnostic applications. Lee et al. (2002) described a microarray equipped with PCR amplicon probes for the detection of different types of β-lactam antibiotic resistance genes in Gram-negative bacteria (including PSE, OXA, FOX, MEN, CMY, TEM, SHV, OXY, *Amp*C). Multiplex PCR amplification of target genes enabled their detection from a single bacterium. In contrast, Call et al. (2003) developed a system for the detection of 17 different tetracycline resistance genes in a variety of Gram-negative bacteria, using bacterial genomic DNA as target. One microgram of DNA was necessary for a single hybridization experiment, which makes this assay more useful for epidemiological studies than for diagnostic purposes.

Analysis of Resistance-mediating Point Mutations

Mutations are likely the most relevant mechanism of resistance development in *Mycobacterium tuberculosis* (Musser 1995). Several DNA microarrays have been developed in the recent past that can be used to detect relevant mutations. Wade et al. (2004) presented a microarray able to detect nucleotide substitutions, deletions, and insertions in the *pncA* gene that cause resistance to the important tuberculosis drug pyrazinamide. Target DNA from 57 mycobacterial isolates was PCR-amplified and transcribed into RNA in vitro, which was subsequently hybridized to a set of 79 short oligonucleotide probes (14 – 20 nucleotides in length) on the microarray. In this way, each nucleotide position of the *pncA* gene was interrogated by two overlapping oligonucleotides, enabling the detection of all but one resistance-mediating mutations in this gene.

Rifampicin is another potent drug against tuberculosis and other bacterial infections. It binds and inhibits the bacterial RNA polymerase (Campbell et al. 2001). Resistance is caused by mutations in the gene *rpoB*, encoding the β-subunit of the RNA polymerase, due to concomitant structural changes in the rifampicin-binding site. Several microarrays have been described for detecting these mutations in *M. tuberculosis*. A high-density Affymetrix array used some 65,000 oligonucleotides to probe mutations in the *rpoB* gene and, simultaneously, determine the sequence of a short

stretch in the 16S ribosomal RNA gene to identify the mycobacterial species. Both investigated genes were PCR-amplified and fluorescently labeled prior to microarray hybridization (Troesch et al. 1999). Similarly, Sougakoff et al. (2004) applied another Affymetrix microarray with an unknown number of probes, PCR amplification of *rpoB*, and a proprietary DNA labeling system. In addition, several low-density microarrays for detection of the same mutations in *M. tuberculosis* have been described (Mikhailovich et al. 2001; Strizhkov et al. 2000). Rifampicin may be used for the treatment of infections caused by other bacteria, for example *Staphylococcus aureus* and *Bacillus anthracis*, and similar resistance-mediating mutations have been found in these species (Aubry-Damon et al. 1998; Vogler et al. 2002). However, microarray detection of these mutations will require the design of adjusted probes, since *rpoB* sequences differ significantly in the different species.

Fluoroquinolones are broad-spectrum antibiotics targeting two bacterial enzymes, DNA gyrase and DNA topoisomerase IV. Resistance may be caused by mutations in the encoding genes *gyrA/B* and *parC/E*. DNA microarrays have been described for the detection of these mutations in *S. aureus* (Couzinet et al. 2005), *Neisseria gonorrhoeae* (Booth et al. 2003), and *E. coli* (Yu et al. 2004). For *S. aureus*, an Affymetrix microarray was used and information about the number and characteristics of the probes was not provided (Couzinet et al. 2005). The four genes of interest were PCR-amplified prior to hybridization analysis; and between 90% (*gyrA*) and 95% (*parC*, termed *grlA* in *S. aureus*) of mutations were recognized correctly. In contrast, Yu et al. (2004) presented a low-density microarray with 52 oligonucleotide probes to screen for silent and resistance-mediating mutations at two amino acid residues in GyrA protein from *E. coli* that were considered most important. Probes were designed in such a way that interrogated nucleotides were positioned in the center of the probes and each nucleotide position of interest was represented by four different probes with either one of the four possible nucleotides at the position in question (Yu et al. 2004). Booth et al. (2003) used a microarray with 21 oligonucleotide probes to interrogate four mutations in *gyrA* and four mutations in *parC* from *N. gonorrhoeae*.

9.3.4 Multi-locus Sequence Typing by Hybridization

Identification of clonal lineages can unambiguously be achieved by multilocus sequence typing (MLST), which was first introduced to molecular population studies on *N. meningitidis* (Maiden et al. 1998) and has now been developed for a number of bacterial species. MLST analyses house-

keeping genes which are thought to be selectively neutral. Most MLST systems are based on sequence analyses of fragments of about 450–500 bp generated by PCR. As sequencing in both directions for seven to eight loci is rather laborious, MLST so far has been restricted to more basic population studies. However, oligonucleotide arrays are a promising tool for the use of MLST for epidemiological typing on a broader scale. Such a microarray has been developed for *S. aureus* (Van Leeuwen et al. 2003). A database of allele reference sequences of the seven polymorphic housekeeping genes (www.mlst.net) was utilized to design the array. For every base interrogated within the reference sequence, four probes of equal length had been synthesized on the chip that differed at the interrogated position. Analysis of two sets of reference strain collections revealed that chip-defined MLST was concordant with "conventional" MLST and highly reproducible (Van Leeuwen et al. 2003). However, the set of probes applied represents only a limited selection of the sequences (alleles) currently known from *S. aureus*.

9.3.5 Composite Gene Detection for Epidemiological Typing

Basic diagnostics in clinical microbiology covers species identification and antimicrobial susceptibility testing. Epidemiology of infectious diseases not only includes the monitoring of the incidence and prevalence among the human population but also the discovery of reservoirs and routes of transmission of the pathogens under surveillance. This can be achieved by typing. Besides the species characteristics, organisms exhibit a number of additional properties which can be used for further differentiation into "types" below the species level.

Many of the bacteria capable of causing infections in macroorganisms are conditional pathogens, with *E. coli* and *S. aureus* as the most prominent examples. They are widely disseminated among the human population as colonizers of skin and the mucosas, intestinal flora included. They cause disease when they are at the wrong place at the wrong time, but nevertheless, this is associated with particular virulence-associated characteristics that are encoded by genes often carried on discrete genetic elements, similar to genes conferring acquired antibiotic resistance. Advanced epidemiological typing has to identify clonal lineages (strains) with particular epidemic and pathogenic potential. This is especially important for the detection and tracing of strains that are of epidemic virulence and contain particular virulence-associated genes, as shown in Table 9.2 for staphylococci, enterococci and *E. coli*. Microarrays containing probes for antibiotic resistance genes, for sequences relevant to MLST, and for virulence genes are very powerful

Table 9.2. Examples for the demonstration of virulence-associated characteristics for confirmation of clinical diagnostics, prediction of the course of an infection and early recognition of particular virulent clonal lineages

Species	Kind of infection	Genetic determinant
Staphylococcus aureus (Dinges et al. 2000; Lina et al. 1999; Vandenesch et al. 2003)	Toxic-shock syndrome	*tst* (toxic-shock syndrome toxin)
		seb, *sec* (enterotoxins B and C)
		eta, *etb* (exfoliative toxins A, B)
	Deep-seated infections of skin and soft tissue, necrotizing pneumoniae	*luk*S-*luk*F (Panton–Valentine leukocidin)
Enterococcus faecium (Homan et al. 2002; Rice et al. 2003; Willems et al. 2001)	Septicemia, endocarditis	*esp* (enterococcal surface protein)
		hyl (hyluronidase)
Escherichia coli (Bingen-Bidois et al. 2002; Dobrindt et al. 2003; Fratamico et al. 1995; Friedrich et al. 2002; Hilali et al. 2000 ; Johnson et al. 2002)	Septicemia, meningitis	*cnf-1* (necrotizing cytotoxin)
	Urosepticemia, pyelonephritis	*papC* (pilus as adhesion)
		hyl (α-hemolysin)
		aero (aerobactin)
		afa (adhesion, nonfimbrial)
	Hemolytic uremic syndrome	*stx* (shiga toxin)
		eae (intimin)

tools for genotyping which not only allow efficient local, national and international tracking of epidemic clones but also provide predictions on important kinds of disease in case of infection. This will be illustrated by two examples.

S. aureus has a rather clonal population structure with particularly successful clonal lineages prevailing in colonization and infection (Enright et al. 2002). The wide dissemination of MRSA is mainly due to the acquisition of antibiotic resistance genes, including the SCC*mec* elements (coding for methicillin resistance) by successful clonal lineages (Enright et al. 2002; Robinson and Enright 2003), which can be disseminated worldwide (Witte 2004). This gets even more problematic when particularly virulent strains acquire *mec*A, for example those capable of causing invasive infections and containing the *luk*S-*luk*F determinant (for Panton–Valentine leukocidin), or vice versa, when epidemic virulent MRSA acquire *luk*S-*luk*F (Vandenesch et al. 2003). Rapid identification is a prerequisite for efficient infection control. A PCR amplicon-based microarray containing probes for 3,623 *S. aureus* genes known from seven MRSA sequenced so far was just recently presented (Lindsay et al. 2004). The comparison of 61

invasive versus 101 carriage isolates clearly demonstrated an association of invasive potential with virulence-associated genes on mobile genetic elements. Thirty well conserved genes showed a strong correlation with MLST typing.

The treatment of enterococcal infections gets difficult in the case of glycopeptide resistance, especially in *Enterococcus faecium*. Although glycopeptide-resistant *E. faecium* (GREF) is already widely disseminated among humans and other animals (Klare et al. 2003), there is a particular nosocomial population – termed C17 – which is characterized by its unique MLST profile, especially the *pur*K allele (Homan et al. 2002). Most of these isolates also possess the *esp* gene, carried on a pathogenicity island and coding for the enterococcal surface protein. A particularly virulent subpopulation additionally possesses *hyl*, coding for hyaluronidase (Rice et al. 2003). A composite microarray containing capture probes for resistance genes *pur*K, *esp*, and *hyl* will provide sufficient information about the options for antibiotic chemotherapy, epidemic potential in the nosocomial setting, and pathogenicity.

9.3.6 Detection of Genes Associated with Metabolic Functions

Many metabolic processes in nature are performed by mixed microbial communities rather than individual species. However, very little is known about the complexity of the composition of communities and their associated interactions, due to the technical limitations of the experimental tools available. Microarrays hold considerable promise for more thorough investigations into the functioning of microbial communities because they may enable the parallel monitoring of many community members. Extensive data on community gene composition may reflect the diverse metabolic capacity of the microorganisms present in a sample. Several studies applying DNA microarrays have targeted bacterial genes that are directly linked to metabolic functions. Detection sensitivities achieved and abilities to generate meaningful quantitative data are important issues in this field.

Wu et al. (2001) described a microarray based on PCR-amplicon probes to monitor genes involved in nitrogen cycling (*nirS*, *nirK*, *amoA*, *pmoA*). Probes were generated from bacterial cultures and from genes cloned from environmental samples. These authors concluded that the measurement of relative abundances of target genes in environmental samples (marine sediments) was complicated through potential cross-hybridizations of unknown genes with divergent sequences (Wu et al. 2001; Zhou 2003). Rhee et al. (2004) recently presented a DNA microarray equipped with 1,662 oligoncleotide probes, targeting diverse bacterial genes involved in the

biodegradation of xenobiotics. Probes were approximately 50 nucleotides in length and non-amplified target DNA was fluorescently labeled using random primers. Another study from the same group applied similar technology and probes against genes involved in nitrogen cycling (*nirS*, *nirK*, *amoA*, *nifH*, *pmoA*) and sulfite reductase (*dsrAB*; Tiquia et al. 2004). Both assays achieved discrimination of sequences with less than 88% similarity; and detection limits were determined at 10 ng of bacterial DNA. However, the presence of excess environmental non-target DNAs affected detection sensitivities unfavourably (Rhee et al. 2004; Tiquia et al. 2004). For the four genes tested, microarray-based quantification was consistent with results obtained by real-time PCR (Rhee et al. 2004). Furthermore, linear correlations of hybridization signal intensities and amounts of DNA were observed. However, Denef et al. (2003), using a microarray for the detection of enzymes catalyzing the oxygenation of polychlorinated biphenyls, observed that the slopes and intercepts of regression lines depended on the particular gene detected and concluded that thorough quantitative analyses would require the determination of standard curves for each individual probe, or at the least, for a representative subset of the probes. Denef et al. (2003) used an approach applying a probe against lambda phage DNA in each probe spot on the microarray and spiking target DNA with known amounts of lambda DNA, which had previously been introduced to enable normalization of hybridization signals for variation in spot quality and hybridization efficiency across the microarray slide (Cho and Tiedje 2002).

A DNA microarray for the detection of particulate methane monooxygenase (*pmoA*) from methanotrophic bacteria was equipped with 59 short oligonucleotide probes (17 – 26 nucleotides). Considerable effort was invested to establish the sequence specificity of the hybridization assay. Based on hybridizations with PCR products derived from bacterial cultures, environmental DNA clones, and soil samples, the detection limit was determined as 5% of the total cells containing *pmoA* (Bodrossy et al. 2003). Cross-reactivity of the probes with ammonia monooxygenase genes (*amoA*) from ammonia-oxidizing bacteria was named as a limitation of this approach. In a follow-up study, this microarray was modified and used to investigate soil samples from landfill lysimeter sites (Stralis-Pavese et al. 2004). Sample-specific differences were found that correlated with lysimeter plant cover, methane supply, and depth in the soil. Certain methanotrophs known to prefer elevated oxygen concentrations were exclusively found in a lysimeter with an air leakage. Significant variation among replicate samples was attributed to sample heterogeneities (Stralis-Pavese et al. 2004).

In another report on the application of an oligonucleotide microarray for monitoring nitrogen cycle genes, Taroncher-Oldenburg et al. (2003) concluded that hybridization patterns differed between two river sediment samples, although these results were not validated with any other method.

9.3.7 Phylogenetic Identification

Small subunit ribosomal RNA (rRNA) is universal to all living beings. Comparative analyses of rRNA gene sequences have elucidated the phylogenetic relationships among all kinds of organisms and, in particular, have promoted great progress in the systematics of prokaryotic life (Woese 1992). The retrieval of bacterial rRNA gene sequences from natural ecosystems has led to the discovery of many heretofore unknown phylogenetic lineages (Pace 1997). Public databases of – at the time of writing – more than 70,000 rRNA gene sequences exist, that have provided the phylogenetic framework for studies directed towards the understanding of microbial diversity and community composition. In this tradition, it is obvious to develop DNA microarrays for the detection of these genes to identify prokaryotes.

Significant baseline work has been performed in the laboratory of D. A. Stahl at the University of Washington (Seattle, Wash., USA). In a series of papers, the abilities of DNA microarrays equipped with short oligonucleotides (< 20 nucleotides) to discriminate single-nucleotide differences among 16S rRNA gene sequences was systematically evaluated (El Fantroussi et al. 2003; Koizumi et al. 2002; Liu et al. 2001; Urakawa et al. 2002, 2003). For these studies, a system immobilizing oligonucleotide probes in polyacrylamide gel pads attached to a glass surface was used (Guschin et al. 1997). In that way, it was possible to determine curves of "non-equilibrium dissociation" at increasing temperature for each of the probes in parallel (Liu et al. 2001). The position of the mismatches between probe and target DNA, the type of the mismatch, and the hybridization conditions strongly influenced the dissociation curves and signal intensities. It was demonstrated that target DNAs with single-basepair mismatches to the probes could be discriminated from perfectly matching targets, even if the mismatches were located close to either end of the probe (Urakawa et al. 2003). This is an important feature, especially if environmental samples are to be investigated that may contain nucleic acids with unknown sequences. A disadvantage of this technology is that devices and slides are custom-built and not generally available. More recently, similar non-equilibrium dissociation profiles were successfully measured on more conventional glass microarrays, with no need for gel pads (Li et al. 2004).

Studies on the detection of PCR-amplified 16S rRNA genes using conventional, planar glass microarrays equipped with oligonucleotide probes have been reported from several additional laboratories. Loy et al. (2002) presented an array with 132 probes for the detection of diverse sulfate-reducing bacteria in samples from periodontal tooth pockets and from a cyanobacterial mat. Peplies et al. (2003) developed an array with 20 probes specialized for the detection of some groups of planktonic bacteria from sea water sam-

ples. Other papers reported on the use of oligonucleotide microarrays for 16S rRNA gene identification of intestinal bacteria in human fecal samples (Wang et al. 2002, 2004), diverse bacteria in a Siberian high-temperature oil reservoir (Bonch-Osmolovskaya et al. 2003), *Campylobacter* spp (Keramas et al. 2003, 2004), and 15 different fish pathogens (Warsen et al. 2004). Wilson et al. (2002) used a photolithographic Affymetrix microarray with 31,179 oligonucleotide probes to determine through hybridization the sequences of a 83-basepair segment of PCR-amplified 16S rRNA genes from cultivated bacteria and from an air sample. It was possible to obtain sequence information from previously unknown organisms (Wilson et al. 2002). An attractive feature of 16S rRNA gene-directed microarrays is that they may also be used to investigate native ribosomal RNA instead of the encoding genes, as has been demonstrated in several cases (Adamczyk et al. 2003; Chandler and Jarrell 2004; El Fantroussi et al. 2003; Peplies et al. 2004; Small et al. 2001). RNA molecules may be more abundant in actively growing bacteria, and hence enzymatic amplification steps such as PCR, which otherwise may bias any subsequent analysis, have usually been avoided. When rRNA was extracted from activated sludge samples which previously had been incubated with [^{14}C]-bicarbonate for 26 h, radioactivity on the microarray surface indicated which bacterial populations had incorporated the radiolabeled carbon due to metabolic activity (Adamczyk et al. 2003).

Other housekeeping loci universally distributed among bacteria that have been probed by oligonucleotide microarrays for bacterial identification include *gyrB* and *parE* genes (Roth et al. 2004), 23S rRNA genes (Anthony et al. 2000; Hong et al. 2004), and the internal transcribed spacer region between genes for 16S rRNA and 23S rRNA (Nübel et al. 2004). The latter target was found to be particularly well suited for discrimination of *Bacillus anthracis* and closely related species, albeit it was universally amplifiable through PCR from a wide range of bacteria due to highly conserved flanking regions encoding rRNA genes.

9.3.8 Random Hybridization Fingerprinting

Kingsley and coworkers (2002) suggested the use of microarrays to generating organism-specific fingerprints by hybridization of genomic DNA to short oligonucleotide probes with random sequences. They presented an array equipped with 47 nonamer oligonucleotides that was able to distinguish 14 strains of *Xanthomonas* spp, *E. coli*, and *Pseudomonas putida*. Hybridization fingerprints were reproducibly different, even between two very closely related strains of *X. oryzae* (Kingsley et al. 2002). Similarly, Belosludtsev and coworkers (2004) recently introduced a high-density microarray with

14,283 different 12- and 13-mer oligonucleotide probes generated in situ by applying light-directed synthesis methodology, which could be used to differentiate several organisms including *B. anthracis*, *Yersinia pestis*, *Streptococcus pneumoniae*, and *Homo sapiens*. The discriminative power of this tool for closely related bacterial strains was not established. These microarray applications are conceptually similar to PCR-based randomly amplified polymorphic DNA (RAPD) analyses, which apply short oligonucleotides with random sequences to generate electrophoretic band patterns. The approach is universally applicable and requires no a priori knowledge about DNA sequences to be detected. On the downside, however, as for other random hybridization-based methods (for example, RAPD), no specific information is obtained about the genetic make-up of the bacterium under investigation. Data are not easily storable in databases and the transfer and comparison of results between laboratories is likely to be difficult. In a related application, Cho and Tiedje (2001) used a microarray with 96 randomly chosen genome fragments of approximately 1 kb for probes to generate hybridization patterns from 12 strains of *Pseudomonas* spp and suggested this method to replace the laborious DNA–DNA hybridization experiments that are commonly used for bacterial species delignation.

9.4 Present Limitations and Future Prospects

Present limitations to more widespread usage of DNA microarrays for bacterial genotyping relate to issues of detection sensitivity, detection specificity, and – particularly in the diagnostic field – their associated costs.

By using assays based on bacterial genomic DNA randomly labeled with fluorescent dyes, lower detection limits below 10 ng DNA are repeatedly reported (Wu et al. 2001; Zhou 2003). This amount of genomic DNA corresponds to approximately 10^6 bacterial genomes. For cultivation-independent detection of bacteria in environmental or clinical samples, this sensitivity will not be sufficient in most cases. Therefore, DNA microarrays designed for applications in these fields usually rely on prior PCR amplification of target genes. However, PCR is inherently gene-specific, and hence constitutes a narrow bottleneck for exploiting the multiplexing potential of microarrays. Several solutions have been suggested to circumvent this limitation, including improved protocols for random DNA labeling (Vora et al. 2004; Wang et al. 2003), the employment of techniques for post-hybridization signal enhancement borrowed from immunohistochemistry (Denef et al. 2003), and enzymatic amplification of the target DNA on the microarray. The latter approach applies oligonucleotide primers attached to the solid support, which enables a reduction of primer–primer interac-

tions through their spatial separation and has been reported to allow for multiplex DNA amplification (Adessi et al. 2000; Nallur et al. 2001; Westin et al. 2000). Applying rolling circle amplification on a microarray, 150 target DNA molecules could be detected (Nallur et al. 2001), which would be sufficient for many applications in diagnostic microbiology.

The sequence specificity of microarray hybridizations may be of concern because, commonly, global reaction conditions are applied to hybridize diverse target DNAs to many probes with different nucleotide sequences and different thermodynamic characteristics. Especially when analyzing very complex environmental samples, cross-hybridization events of imperfectly matching targets may obscur experimental results (El Fantroussi et al. 2003; Zhou 2003). Real-time hybridization monitoring may be a very valuable approach to improve the specificity of the detection of hybridization events. If the binding reaction between probe and target can be observed online, dissociation rate constants can be measured that reflect thermodynamic characteristics of probe–target duplexes and may differ significantly for perfectly matching and slightly mis-matching targets (Bier and Kleinjung 2001; Li et al. 2004; Liu et al. 2001). These measurements will also likely enable a significant acceleration of measurement speed, since reaction equilibrium will not be required. Unfortunately, devices appropriate for such parallel kinetic measurements are not yet readily available. Instead, nearly all microarray scanners measure dry microarrays, requiring hybridization reactions to be performed outside the measurement unit. Example systems reported to be suitable for monitoring hybridizations on microarrays online employ a fluidic system (Bier and Kleinjung 2001) or a temperature control unit and gel pad slides (Li et al. 2004; Liu et al. 2001). Both of these devices are custom-built.

Costs and inconveniences in handling DNA microarrays have hampered their introduction into the diagnostic field. However, prices for microarrays and devices have dropped significantly in the recent past due to a diversification of industrial suppliers; and several companies doing research in the field have announced that they plan to provide systems integrating several processing steps into single, affordable devices in the near future. Hence, there is reason for optimism.

References

Adamczyk J, Hesselsoe M, Iversen N, et al (2003) The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. Appl Environ Microbiol 69:6875–6887

Adessi C, Matton G, Ayala G, Turcatti G, et al (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28:E87

Al-Khaldi SF, Myers KM, Rasooly A, Chizhikov V (2004) Genotyping of *Clostridium perfringens* toxins using multiple oligonucleotide microarray hybridization. Mol Cell Probes 18:359–367

Anthony RM, Brown TJ, French GL (2000) Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. J Clin Microbiol 38:781–788

Aubry-Damon H, Soussy CJ, Courvalin P (1998) Characterization of mutations in the *rpoB* gene that confer rifampin resistance in *Staphylococcus aureus*. Antimicrob Agents Chemother 42:2590–2594

Behr MA, Wilson MA, Gill WP, et al (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284:1520–1523

Bekal S, Brousseau R, Masson L, et al (2003) Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays. J Clin Microbiol 41:2113–2125

Belosludtsev YY, Bowerman D, Weil R, et al (2004) Organism identification using a genome sequence-independent universal microarray probe set. Biotechniques 37:654–658, 660

Bergeron MG, Ouellette M (1998) Preventing antibiotic resistance through rapid genotypic identification of bacteria and of their antibiotic resistance genes in the clinical microbiology laboratory. J Clin Microbiol 36:2169–2172

Bier FF, Kleinjung F (2001) Feature-size limitations of microarray technology – a critical review. Fresenius Z Anal Chem 371:151–156

Bingen-Bidois M, Clermont O, Bonacorsi S, Terki M, Brahimi N, Loukil C, Barraud D, Bingen E (2002) Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* bearing pathogenicity island-like domains. Infect Immun 70:3216–3226

Bodrossy L, Stralis-Pavese N, Murrell JC, Radajewski S, Weilharter A, Sessitsch A (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. Environ Microbiol 5:566–582

Bonch-Osmolovskaya EA, Miroshnichenko ML, Lebedinsky AV, et al (2003) Radioisotopic, culture-based, and oligonucleotide microchip analyses of thermophilic microbial communities in a continental high-temperature petroleum reservoir. Appl Environ Microbiol 69:6143–6151

Booth SA, Drebot MA, Martin IE, Ng LK (2003) Design of oligonucleotide arrays to detect point mutations: molecular typing of antibiotic resistant strains of *Neisseria gonorrhoeae* and hantavirus infected deer mice. Mol Cell Probes 17:77–84

Brandt O, Feldner J, Stephan A, Schroder M, Schnolzer M, Arlinghaus HF, Hoheisel JD, Jacob A (2003) PNA microarrays for hybridisation of unlabelled DNA samples. Nucleic Acids Res 31:E119

Call DR, Brockman FJ, Chandler DP (2001) Detecting and genotyping *Escherichia coli* O157:H7 using multiplexed PCR and nucleic acid microarrays. Int J Food Microbiol 67:71–80

Call DR, Bakko MK, Krug MJ, Roberts MC (2003) Identifying antimicrobial resistance genes with DNA microarrays. Antimicrob Agents Chemother 47:3290–3295

Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, Darst SA (2001) Structural mechanisms for rifampicin inhibition of bacterial RNA polymerase. Cell 104:901–912

Chandler DP, Jarrell AE (2004) Automated purification and suspension array detection of 16S rRNA from soil and sediment extracts by using tunable surface microparticles. Appl Environ Microbiol 70 :2621–2631

Chizhikov V, Rasooly A, Chumakov K, Levy DD (2001) Microarray analysis of microbial virulence factors. Appl Environ Microbiol 67 :3258–3263

Cho JC, Tiedje JM (2001) Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol 67:3677–3682

Cho JC, Tiedje JM (2002) Quantitative detection of microbial genes by using DNA microarrays. Appl Environ Microbiol 68 :1425–1430

Couzinet S, Yugueros J, Barras C, et al (2005) Evaluation of a high-density oligonucleotide array for characterization of *grlA*, *grlB*, *gyrA* and *gyrB* mutations in fluoroquinolone resistant *Staphylococcus aureus* isolates. J Microbiol Methods 60:275–279

Daubin V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. Science 301:829–832

Denef VJ, Park J, Rodrigues JL, Tsoi TV, Hashsham SA, Tiedje JM (2003) Validation of a more sensitive method for using spotted oligonucleotide DNA microarrays for functional genomics studies on bacterial communities. Environ Microbiol 5:933–943

Dinges MM, Orwin PM, Schlievert PM (2000) Exotoxins of *Staphylococcus aureus*. Clin Microbiol Rev 13:16–34

Dobrindt U, Agerer F, Michaelis K, et al (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. J Bacteriol 185:1831–1840

Dunman PM, Mounts W, McAleese F, et al (2004) Uses of *Staphylococcus aureus* GeneChips in genotyping and genetic composition analysis. J Clin Microbiol 42:4275–4283

El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, Yershov GM, Stahl DA (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. Appl Environ Microbiol 69:2377–2382

Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG (2002) The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA) Proc Natl Acad Sci USA 99:7687–7692

Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. Proc Natl Acad Sci USA 98:8821–8826

Fleischmann RD, Adams MD, White O, et al (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fluit AC, Visser MR, Schmitz FJ (2001) Molecular detection of antimicrobial resistance. Clin Microbiol Rev 14:836–871

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251:767–773

Fratamico PM, Sackitey SK, Wiedmann M, Deng MY (1995) Detection of *Escherichia coli* O157:H7 by multiplex PCR. J Clin Microbiol 33:2188–2191

Friedrich AW, Bielaszewska M, Zhang WL, Pulz M, Kuczius T, Ammon A, Karch H (2002) *Escherichia coli* harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms. J Infect Dis 185:74–84

Gabig-Ciminska M, Andresen H, Albers J, Hintsche R, Enfors SO (2004) Identification of pathogenic microbial cells and spores by electrochemical detection on a biochip. Microb Cell Fact 3:2

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238

Gonzalez SF, Krug MJ, Nielsen ME, Santos Y, Call DR (2004) Simultaneous detection of marine fish pathogens by using multiplex PCR and a DNA microarray. J Clin Microbiol 42:1414–1419

Guschin DY, Mobarry BK, Proudnikov D, Stahl DA, Rittmann BE, Mirzabekov AD (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl Environ Microbiol 63:2397–2402

Hacker J, Hentschel U, Dobrindt U (2003) Prokaryotic chromosomes and disease. Science 301:790–793

Hamels S, Gala JL, Dufour S, Vannuffel P, Zammatteo N, Remacle J (2001) Consensus PCR and microarray for diagnosis of the genus *Staphylococcus*, species, and methicillin resistance. Biotechniques 31:1364–1362

Hilali F, Ruimy R, Saulnier P, Barnabe C, Lebouguenec C, Tibayrenc M, Andremont A (2000) Prevalence of virulence genes and clonality in *Escherichia coli* strains that cause bacteremia in cancer patients. Infect Immun 68:3983–3989

Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, Van Embden JD, Willems RJ (2002) Multilocus sequence typing scheme for *Enterococcus faecium*. J Clin Microbiol 40:1963–1971

Hong BX, Jiang LF, Hu YS, Fang DY, Guo HY (2004) Application of oligonucleotide array technology for the rapid detection of pathogenic bacteria of foodborne infections. J Microbiol Methods 58:403–411

Hughes TR, Mao M, Jones AR, et al (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 19:342–347

Johnson JR, Oswald E, O'Bryan TT, Kuskowski MA, Spanjaard L (2002) Phylogenetic distribution of virulence-associated genes among *Escherichia coli* isolates associated with neonatal bacterial meningitis in the Netherlands. J Infect Dis 185:774–784

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas J, Madore SJ (2000) Assessment of the sensistivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res 28:4552–4557

Kelly JJ, Chernov BK, Tovstanovsky I, Mirzabekov AD, Bavykin SG (2002) Radical-generating coordination complexes as tools for rapid and effective fragmentation and fluorescent labeling of nucleic acids for microchip hybridization. Anal Biochem 311:103–118

Keramas G, Bang DD, Lund M, Madsen M, Rasmussen SE, Bunkenborg H, Telleman P, Christensen CB (2003) Development of a sensitive DNA microarray suitable for rapid detection of *Campylobacter* spp. Mol Cell Probes 17:187–196

Keramas G, Bang DD, Lund M, Madsen M, Bunkenborg H, Telleman P, Christensen CB (2004) Use of culture, PCR analysis, and DNA microarrays for detection of *Campylobacter jejuni* and *Campylobacter coli* from chicken feces. J Clin Microbiol 42:3985–3991

Kingsley MT, Straub TM, Call DR, Daly DS, Wunschel SC, Chandler DP (2002) Fingerprinting closely related *Xanthomonas* pathovars with random nonamer oligonucleotide microarrays. Appl Environ Microbiol 68:6361–6370

Klare I, Konstabel C, Badstubner D, Werner G, Witte W (2003) Occurrence and spread of antibiotic resistances in *Enterococcus faecium*. Int J Food Microbiol 88:269–290

Koizumi Y, Kelly JJ, Nakagawa T, Urakawa H, El-Fantroussi S, Al-Muzaini S, Fukui M, Urushigawa Y, Stahl DA (2002) Parallel characterization of anaerobic toluene- and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA–DNA membrane hybridization, and DNA microarray technology. Appl Environ Microbiol 68:3215–3225

Lee Y, Lee CS, Kim YJ, Chun S, Park S, Kim YS, Han BD (2002) Development of DNA chip for the simultaneous detection of various beta-lactam antibiotic-resistant genes. Mol Cells 14:192–197

Letowski J, Brousseau R, Masson L (2004) Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. J Microbiol Methods 57:269–278

Li ES, Ng JK, Wu JH, Liu WT (2004) Evaluating single-base-pair discriminating capability of planar oligonucleotide microchips using a non-equilibrium dissociation approach. Environ Microbiol 6:1197–1202

Lina G, Piemont Y, Godail-Gamot F, Bes M, Peter MO, Gauduchon V, Vandenesch F, Etienne J (1999) Involvement of Panton–Valentine leukocidin-producing *Staphylococcus aureus* in primary skin infections and pneumonia. Clin Infect Dis 29:1128–1132

Lindroos K, Liljedahl U, Raitio M, Syvanen A-C (2001) Minisequencing on oligonucleotide microarrays: comparison of immobilisation chemistries. Nucleic Acids Res 29:E69

Lindsay JA, Holden MT (2004) *Staphylococcus aureus*: superbug, super genome? Trends Microbiol 12 :378–385

Lindsay JA, Witney A, Holden M, et al (2004)Comparative genomics using a seven strain *S. aureus* microarray: the ultimate typing tool and identification of genes associated with invasive strains. Int Symp Staphylococci Staphylococcal Infect 11:ME17

Liu WT, Mirzabekov AD, Stahl DA (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. Environ Microbiol 3:619–629

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer KH, Wagner M (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl Environ Microbiol 68:5064–5081

Maiden MC, Bygraves JA, Feil EJ, et al (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145

Marokhazi J, Waterfield N, LeGoff G, Feil E, Stabler R, Hinds J, Fodor A, Ffrench-Constant RH (2003) Using a DNA microarray to investigate the distribution of insect virulence factors in strains of *Photorhabdus* bacteria. J Bacteriol 185:4648–4656

McKendry R, Zhang J, Arntz Y, et al (2002) Multiple label-free biodetection and quantitative DNA-binding assays on a nanomechanical cantilever array. Proc Natl Acad Sci USA 99:9783–9788

Mikhailovich V, Lapa S, Gryadunov D, et al (2001) Identification of rifampin-resistant *Mycobacterium tuberculosis* strains by hybridization, PCR, and ligase detection reaction on oligonucleotide microchips. J Clin Microbiol 39:2531–2540

Monecke S, Leube I, Ehricht R (2003) Simple and robust array-based methods for the parallel detection of resistance genes of *Staphylococcus aureus*. Genome Lett:106–118

Musser JM (1995) Antimicrobial agent resistance in mycobacteria: molecular genetic insights. Clin Microbiol Rev 8:496–514

Nallur G, Luo C, Fang L, et al (2001) Signal amplification by rolling circle amplification on DNA microarrays. Nucleic Acids Res 29:E118

Nübel U, Schmidt PM, Reiß E, Bier F, Beyer W, Naumann D (2004) Oligonucleotide microarray for identification of *Bacillus anthracis* based on intergenic transcribed spacers in ribosomal DNA. FEMS Microbiol Lett 240:215–223

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740

Peplies J, Glöckner FO, Amann R (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl Environ Microbiol 69:1397–1407

Peplies J, Lau SC, Pernthaler J, Amann R, Glöckner FO (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. Environ Microbiol 6:638–645

Porwollik S, Frye J, Florea LD, Blackmer F, McClelland M (2003) A non-redundant microarray of genes for two related bacteria. Nucleic Acids Res 31:1869–1876

Read TD, Peterson SN, Tourasse N, et al (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423:81–86

Rhee SK, Liu X, Wu L, Chong SC, WanX, Zhou J (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. Appl Environ Microbiol 70:4303–4317

Rice LB, Carias L, Rudin S, et al (2003) A potential virulence gene, hyl_{Efm}, predominates in *Enterococcus faecium* of clinical origin. J Infect Dis 187:508–512

Robinson DA, Enright MC (2003) Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 47:3926–3934

Roth SB, Jalava J, Ruuskanen O, Ruohola A, Nikkari S (2004) Use of an oligonucleotide array for laboratory diagnosis of bacteria responsible for acute upper respiratory infections. J Clin Microbiol 42:4268–4274

Saunders NA, Underwood A, Kearns AM, Hallas G (2004) A virulence-associated gene microarray: a tool for investigation of the evolution and pathogenic potential of *Staphylococcus aureus*. Microbiology 150:3763–3671

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Sergeev N, Volokhov D, Chizhikov V, Rasooly A (2004) Simultaneous analysis of multiple staphylococcal enterotoxin genes by an oligonucleotide microarray assay. J Clin Microbiol 42:2134–2143

Shen Y, Stehmeier LG, Voordouw G (1998) Identification of hydrocarbon-degrading bacteria in soil by reverse sample genome probing. Appl Environ Microbiol 64:637–645

Singh Gasson S, Green RD, Yue YJ, Nelson C, Blattner F, Sussman MR, Cerrina F (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 17:974–978

Small J, Call DR, Brockman FJ, Straub TM, Chandler DP (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl Environ Microbiol 67:4708–4716

Sougakoff W, Rodrigue M, Truffot-Pernot C, Renard M, Durin N, Szpytma M, Vachon R, Troesch A, Jarlier V (2004) Use of a high-density DNA probe array for detecting mutations involved in rifampicin resistance in *Mycobacterium* tuberculosis. Clin Microbiol Infect 10:289–294

Southern E, Mir K, Shchepinov M (1999) Molecular interactions on microarrays. Nat Genet 21:5–9

Stralis-Pavese N, Sessitsch A, Weilharter A, Reichenauer T, Riesing J, Csontos J, Murrell JC, Bodrossy L (2004) Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. Environ Microbiol 6:347–363

Strizhkov BN, Drobyshev AL, Mikhailovich VM, Mirzabekov AD (2000) PCR amplification on a microarray of gel-immobilized oligonucleotides: detection of bacterial toxin- and drug-resistant genes and their mutations. Biotechniques 29:844–848

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol 69:1159–1171

Taton TA, Mirkin CA, Letsinger RL (2000) Scanometric DNA array detection with nanoparticle probes. Science 289:1757–1760

Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y, Zhou J (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. Biotechniques 36:664–670, 672, 674–665

Tönnisson N, Krug A, Kaasik K, Löhmussar E, Metspalu A (2000) Unravelling genetic data by arrayed primer extension. Clin Chem Lab Med 38:165–170

Troesch A, Nguyen H, Miyada CG, Desvarenne S, Gingeras TR, Kaplan PM, Cros P, Mabilat C (1999) *Mycobacterium* species identification and rifampin resistance testing with high-density DNA probe arrays. J Clin Microbiol 37:49–55

Urakawa H, Noble PA, El Fantroussi S, Kelly JJ, Stahl DA (2002) Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. Appl Environ Microbiol 68:235–244

Urakawa H, El Fantroussi S, Smidt H, Smoot JC, Tribou EH, Kelly JJ, Noble PA, Stahl DA (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. Appl Environ Microbiol 69:2848–2856

Van Ijperen C, Kuhnert P, Frey J, Clewley JP (2002) Virulence typing of *Escherichia coli* using microarrays. Mol Cell Probes 16:371–378

Van Leeuwen WB, Jay C, Snijders S, Durin N, Lacroix B, Verbrugh HA, Enright MC, Troesch A, Van Belkum A (2003) Multilocus sequence typing of *Staphylococcus aureus* with DNA array technology. J Clin Microbiol 41:3323–3326

Vandenesch F, Naimi T, Enright MC, et al (2003) Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton–Valentine leukocidin genes: worldwide emergence. Emerg Infect Dis 9:978–984

Vogler AJ, Busch JD, Percy-Fine S, Tipton-Hunton C, Smith KL, Keim P (2002) Molecular analysis of rifampin resistance in *Bacillus anthracis* and *Bacillus cereus*. Antimicrob Agents Chemother 46:511–513

Volokhov D, Rasooly A, Chumakov K, Chizhikov V (2002) Identification of *Listeria* species by microarray-based assay. J Clin Microbiol 40:4720–4728

Volokhov D, Chizhikov V, Chumakov K, Rasooly A (2003a) Microarray-based identification of thermophilic *Campylobacter jejuni*, *C. coli*, *C. lari*, and *C. upsaliensis*. J Clin Microbiol 41:4071–4080

Volokhov D, Chizhikov V, Chumakov K, Rasooly A (2003b) Microarray analysis of erythromycin resistance determinants. J Appl Microbiol 95:787–798

Volokhov D, Pomerantsev A, Kivovich V, Rasooly A, Chizhikov V (2004) Identification of *Bacillus anthracis* by multiprobe microarray hybridization. Diagn Microbiol Infect Dis 49:163–171

Vora GJ, Meador CE, Stenger DA, Andreadis JD (2004) Nucleic acid amplification strategies for DNA microarray-based pathogen detection. Appl Environ Microbiol 70:3047–3054

Wade MM, Volokhov D, Peredelchuk M, Chizhikov V, Zhang Y (2004) Accurate mapping of mutations of pyrazinamide-resistant *Mycobacterium tuberculosis* strains with a scanning-frame oligonucleotide microarray. Diagn Microbiol Infect Dis 49:89–97

Wang D, Urisman A, Liu YT, et al (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1:257–260

Wang RF, Beggs ML, Robertson LH, Cerniglia CE (2002) Design and evaluation of oligonucleotide-microarray method for the detection of human intestinal bacteria in fecal samples. FEMS Microbiol Lett 213:175–182

Wang RF, Beggs ML, Erickson BD, Cerniglia CE (2004) DNA microarray analysis of predominant human intestinal bacteria in fecal samples. Mol Cell Probes 18:223–234

Warsen AE, Krug MJ, LaFrentz S, Stanek DR, Loge FJ, Call DR (2004) Simultaneous discrimination between 15 fish pathogens by using 16S ribosomal DNA PCR and DNA microarrays. Appl Environ Microbiol 70:4216–4221

Westin L, Xu X, Miller C, Wang L, Edman CF, Nerenberg M (2000) Anchored multiplex amplification on a microelectronic chip array. Nat Biotechnol 18:199–204

Westin L, Miller C, Vollmer D, Canter D, Radtkey R, Nerenberg M, O'Connell JP (2001) Antimicrobial resistance and bacterial identification utilizing a microelectronic chip array. J Clin Microbiol 39:1097–1104

Willems RJ, Homan W, Top J, et al (2001) Variant *esp* gene as a marker of a distinct genetic lineage of vancomycin-resistant *Enterococcus faecium* spreading in hospitals. Lancet 357:853–855

Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuczmarski TA, Andersen GL (2002) High-density microarray of small-subunit ribosomal DNA probes. Appl Environ Microbiol 68:2535–2541

Witte W (1999) Antibiotic resistance in gram-positive bacteria: epidemiological aspects. J Antimicrob Chemother 44 [Suppl A]:1–9

Witte W (2004) International dissemination of antibiotic resistant strains of bacterial pathogens. Infect Genet Evol 4:187–191

Woese CR (1992) Prokaryote systematics: the evolution of a science. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH (ed) The prokaryotes, 2nd edn. Springer, Berlin Heidelberg New York, pp 3–18

Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl Environ Microbiol 67:5780–5790

Yu X, Susa M, Knabbe C, Schmid RD, Bachmann TT (2004) Development and validation of a diagnostic DNA microarray to detect quinolone-resistant *Escherichia coli* among clinical isolates. J Clin Microbiol 42:4083–4091

Zammatteo N, Jeanmart L, Hamels S, Courtois S, Louette P, Hevesi I, Remacle J (2000) Comparison between different strategies of covalent attachment of DNA to glass surfaces to uild DNA microarrays. Anal Biochem 280:143–150

Zhang L, Srinivasan U, Marrs CF, Ghosh D, Gilsdorf JR, Foxman B (2004) Library on a slide for bacterial comparative genomics. BMC Microbiol 4:12

Zhang Y, Price BD, Tetradis S, Chakrabarti S, Maulik G, Makrigiorgos GM (2001) Reproducible and inexpensive probe preparation for oligonucleotide arrays. Nucleic Acids Res 29:E66

Zhou J (2003) Microarrays for bacterial detection and microbial community analysis. Curr Opin Microbiol 6:288–294

Subject Index