# 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 example, 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 $4 \times 10^5$ 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 $4 \times 10^5$ (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 taxonomic 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 $\sigma$70 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 reevaluation 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 ($\alpha$), beta ($\beta$), beta′ ($\beta'$), and omega ($\omega$) in the configuration $\alpha 2\beta\beta'\omega$. The structural gene for the $\beta$-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 $\beta'$-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

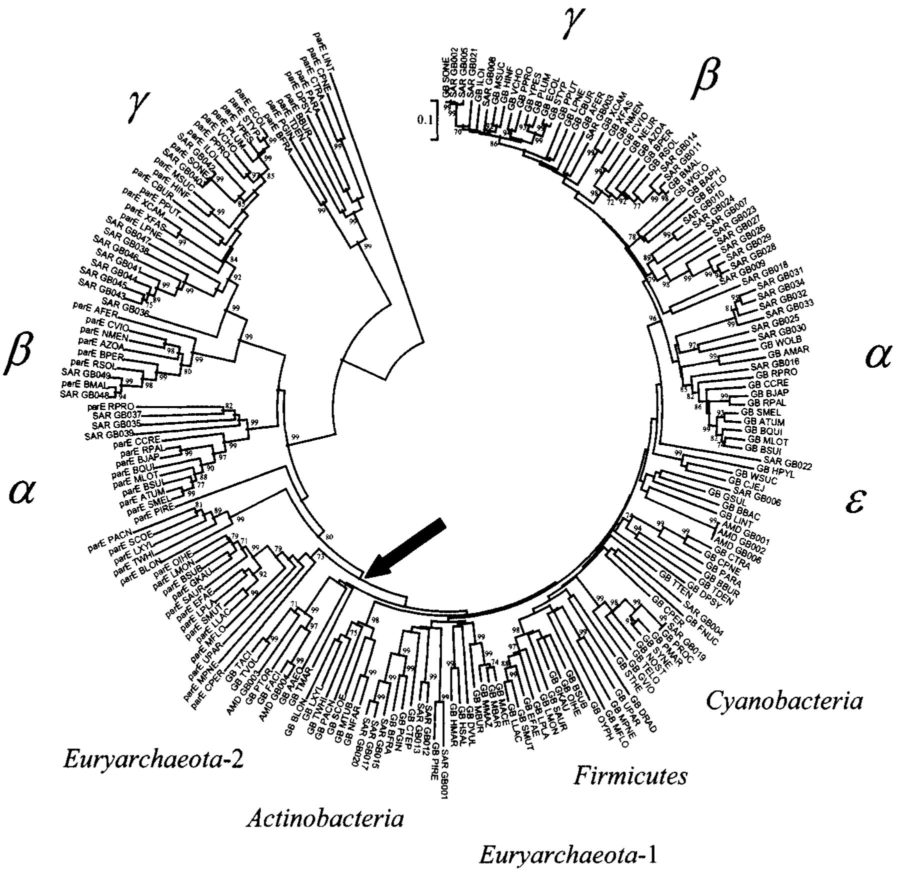

**Fig. 5.2.** Unrooted tree based on the amino acid sequences of GyrB and toposiomerase IV subunit B. Multiple alignment of the amino acid sequences was created using ClustalX (Thompson et al. 1997). The BLOSUM matrix was used for weight matrix parameters. The gap-open penalty was set to 20 and the gap-extension 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 $\alpha$-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 provision 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 $\gamma$-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