8 Comparative Genomics

Asli Ismihan Ozen · Tammi Vesth · David W. Ussery
Department of Systems Biology, Center for Biological Sequence Analysis, Kemitorvet, The Technical University of Denmark, Lyngby, Denmark

Prokaryotic Classification

Classification covers the theory and practice of how to order characterized organisms into different groups based on their degree of relatedness. Together with identification and nomenclature, classification is a part of taxonomy, a science that deals with the relatedness of organisms. The goal of many taxonomists is to have a classification system that reflects the natural relationships among organisms. This natural system has been depicted mostly as phylogeny (Doolittle 1999)—or an evolutionary tree—which is a diagram that shows ancestor-descendant relationships of organisms based on their evolutionary history. However, inferring a true phylogeny for prokaryotic organisms is very challenging due to the diversity of these organisms, as well as frequent horizontal transfer of genes.

Prokaryotes, known as unicellular organisms with no nuclear membrane structure, have a history of more than 3.5 billion years on earth, yet humans have been aware of them for only the past few centuries, after first being described by Robert Hooke in the seventeenth century. Louis Pasteur and other scientists of the nineteenth century described microorganisms in detail, and began to categorize them. Their classification was dependent on the development of microbial techniques such as isolating and growing microorganisms in pure cultures, staining and microscope observations. In their early observations, lack of guidelines on naming inevitably led to a vast number of invalid names and synonyms. Ferdinand Cohn made the first classification system of bacteria in 1872; six genera of bacteria were classified based on their shape, cellular structures, pigmentation, and metabolic activities (Cohn 1872).

At the beginning of the twentieth century, besides morphology, the use of physiological and biochemical information could be incorporated. European scientists also proposed physiology, metabolism, pigments, and pathogenicity as new systems for classification. However, some of these methods were then criticized for being not important for assessing taxonomic ranks. Later, advances in biochemistry and molecular biology from the isolation of nucleic acids to elucidation of macromolecular structure of proteins and nucleic acids led to the foundation of genomic sciences. Development of computers in the 1950s was another important step in bacterial taxonomy, where they were first used for analysis of phenotypic and molecular data. Between the years 1960–1980, numerical taxonomy and chemotaxonomy were on the rise (Stackebrandt 2006; Schleifer 2009).

In late 1950s, scientists were able to identify the molecules conserved throughout history of life, such as proteins, DNA, or RNA molecules. The idea of using these molecules as blueprints of the evolutionary history of organisms emerged in the 1960s (Zuckerkandl et al. 1962). Tertiary structure and sequence analysis of molecules, such as cytochrome C, ferrodoxins, and fibrinopeptides, and also immunological approaches were being used afterward. However, the interest in these methods decreased as rapid sequencing techniques for DNA became more significant.

The first genotypic approach that allowed bacteriologists to classify prokaryotes on the basis of their phylogenetic relatedness was DNA-DNA hybridization (DDH) (Wayne et al. 1987). In the following years, more genotypic studies, including comparative analysis of Ribosomal RNA (ribonucleic acid) genes and protein-coding gene sequence, allowed more insight to the relationships of prokaryotes (Schleifer 2009). The small subunit rRNA (16S rRNA in prokaryotes) was shown to be one of the universally conserved molecules became the primary molecule of interest. Being ubiquitous, having functional consistency, genetic stability, appropriate size, and independently evolving domains caused this molecule to be chosen for phylogenetic analysis and this approach became a classical tool for taxonomy (Harayama and Kasai 2006). An important study by Carl Woese

E. Rosenberg et al. (eds.), *The Prokaryotes – Prokaryotic Biology and Symbiotic Associations*, DOI 10.1007/978-3-642-30194-0_11,

revolutionized bacterial taxonomy, proposing the new kingdom of *Archaebacteria* (Woese and Fox 1977). His later studies concluded in a phylogenetic scheme of three main branches of life (*Bacteria*, *Archaea*, and *Eukarya*) that he called Domains (Woese et al. 1990).

In other genotypic classifications, many protein-coding genes were used for phylogenetic relatedness, some of which are *recA*, *gyrB*, genes of some chaperonins, RNA polymerase subunits (i.e., *rpoB*) or sigma factors (*rpoD*), elongation factor G (*fus*). The most accepted criteria for selection of these proteins is such that, they should not be subjected to horizontal gene transfer (HGT), should be present in all bacteria, preferably in single copies and at least two highly conserved regions for the design of PCR primers (Yamamoto and Harayama 1996). These properties give them an advantage of being more appropriate for phylogenetic analysis of closely related bacteria than 16S rRNA analysis.

In addition to the single gene based methods, Multi Locus Sequence Typing (MLST) has been widely used for genotypic characterization and classification of prokaryotes by comparing multiple housekeeping gene sequences (Maiden et al. 1998). However, usually a different set of genes is useful for different set of organisms, and some difficulties occur in primer design for amplification of genes in all strains if the analysis is not conducted all in silico. A widely used website and database currently is mlst.net (Aanensen and Spratt 2005).

Current Taxonomy of Prokaryotes

Classification is done by comparing a newly identified organism with the collection of previously classified organisms and then assigning it with a previously described or new species. If a bacterial species is considered novel, the proper naming for the new or existing taxa are made by nomenclature that is based on the International Code of Nomenclature of Bacteria (Lapage et al. 1992), also named as the *Bacteriological Code*. Nomenclature is, however, subject to changes because classification is a dynamic process. The publication of names for novel prokaryotic taxa is made in the International Journal of Systematic and Evolutionary Microbiology (IJSEM), which is the official journal for this purpose. IJSEM also publishes "Validation Lists" which contain new names published in other journals (Tindall et al. 2006). An updated list of approved names for microorganisms based on the international rules can also be found at The DSMZ—Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (German Collection of Microorganisms and Cell Cultures) depository (http://www.dsmz.de).

In taxonomy, groups of organisms that are brought together based on shared properties are called "taxa" or "ranks," and prokaryotic taxonomy makes use of several ranks or levels. The current classification scheme has a hierarchical structure, where the higher taxonomic ranks consist of the lower ranked groups. In other words, higher taxa (e.g., genus) contain lower taxa (e.g., species). In an ideal classification system based on evolutionary history, clear clusters of taxonomic units in a phylogenetic tree are seen such that species that share a common ancestor would form the genus and genera that share a common ancestor form a family and so forth. Major sources for bacterial names and taxonomical order are Bergey's Manual of Systematic Bacteriology (Brenner et al. 2005a), Bergey's Taxonomic Outlines (http://www.bergeys.org/outlines.html), and the comprehensive list available at The Taxonomic Outline of Bacteria and Archaea (TOBA) journal (Garrity et al. 2007).

Taxonomy tools historically have been mainly based on laborious laboratory experiments trying to characterize bacteria based on their phenotypic and biochemical properties until molecular approaches and sequencing technologies were developed. Today, such research can be handled using robotic and computational techniques, where most of the knowledge gained from results rely on the data that is being handled.

The Explosion of Sequenced Bacterial Genomes

Biological data generated by researchers worldwide has been growing with a tremendous rate, especially with the advances in molecular biology techniques in the past 50 years. Much of this vast information can now be accessed through biological databases that hold records for experimental data, sequence data, classification schemes, literature, and some also provide computational analysis tools.

A part of this huge biological information is the genomic sequences. In modern molecular biology and genetics, a "genome" is the entirety of an organism's hereditary information. Therefore, genomics can be referred to as the science of genome analysis. As such the field of comparative microbial genomics (CMG) work with comparing the entire DNA material of a microbial organism to other organisms. The first two complete bacterial genome sequences were published in 1995. As the technologies advanced and the sequencing cost went down, many more sequences were being published and more databases were established to handle this information. One of the most used databases is based on GenBank, now located as part of the National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/). The NCBI GenBank holds the nucleotide sequence data from expression sequence tag (EST), genome survey sequences, other high-throughput sequences such as whole-genome sequences and genome annotations of thousands of organisms. Both prokaryotic and eukaryotic data is available (Benson et al. 2008). GenBank is a part of an international collaboration called International Sequence Database Collaboration, which also consists of DNA Data Base of Japan (DDBJ) and the European Molecular Biology laboratory (EMBL). Another part of NCBI that is highly related to this chapter is NCBI Taxonomy. Although claiming not to be a primary source, NCBI provides taxonomical information that is gathered from various sources.

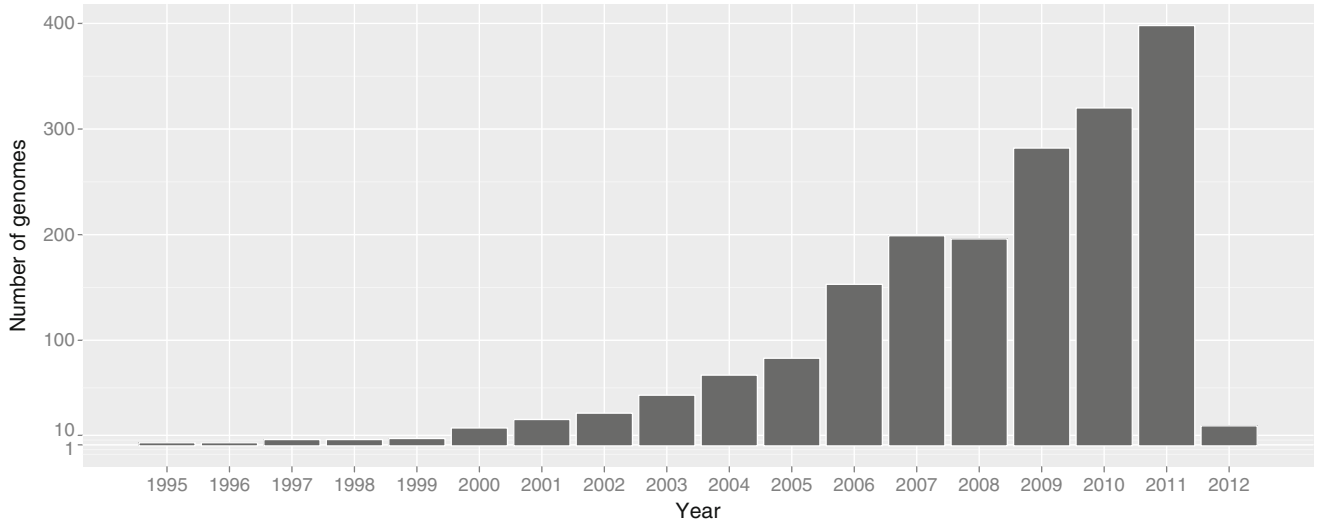

Fig. 8.1
Genomes published and deposited to public NCBI GenBank since 1995 (Data gathered from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, Jan. 2012))

In May 2011, NCBI GenBank contained around 1,500 genome sequences labeled as "finished." Six months later, at the time of writing (November 2011), this number has gone up to 1,790. Currently (Feb. 2012), the NCBI "Genome Projects" is changing to "BioProjects," in order to relate genomic information to other data types, such as the transcriptome, proteome, and metagenome.

In addition to NCBI (GenBank) and EMBL (Nucleotide Sequence Database), a source for genomic information is the Genomes Online database (GOLD). GOLD aims to provide an accurate and complete set of finished and ongoing genome projects with a broad range of information on each project. The sequence data itself is not stored in the database, however, external links to where the data can be found is given, most of which are to the NCBI Genome Project pages. GOLD also provides taxonomical information, though not the primary source (Bernal et al. 2001; Liolios et al. 2010).

One part of comparative microbial genomics is to monitor the available microbial genomic data. Even though the sequences available may only be a small fraction of the real world, the information gathered is growing every day. It took 14 years to sequence the first thousand bacterial genomes (1995–2009), and already in 2012, less than 3 years later, the two thousandth genome sequence has been deposited to GenBank. Not only has the cost of genome sequencing decrease dramatically but also the time and effort put into the task has also gone down. Also the computational power and software to handle sequencing data is being revolutionized and fast assembly and interpretation is increasing the number of published genomes (Ansorge 2009). The increase in genome data has given rise to a whole new area of problems when it comes to publication and sharing of data. Databases usually have their own formatting of the raw data and though some are more used than others, a clear standard for genome publication has yet to be established (Médigue and Moszer 2007).

Statistics on Prokaryotic Genomes

With such a large amount of data, it is interesting to see the trends in the basic statistics of the genomic data and comparisons on different taxonomic levels and years of sequence publications. The data presented in this section is taken from the NCBI complete genomes list in Jan. 2012 and GenBank files for 1,500 sequenced genomes were downloaded (November 2011).

Data Growth over the Years

Figure 8.1 illustrates how many genomes were published each year, since 1995. The two first complete genomes to be sequenced and deposited was *Mycoplasma genitalium* G37 (Fraser et al. 1995) and *Haemophilus influenzae* Rd KW20 (Fleischmann et al. 1995). From 1995 until 1999, only 25 genomes were published as complete and they covered 14 phyla. Of these the Archaeal genomes constitute a large portion compared to the fraction today (around 31% of the 25 compared to 99 out of 1,500 (6.6%)). The genomes from this first period of genome sequencing cover a large span of the microbial landscape with no obvious medical bias. From 2000 to 2010 the number steadily increased from 26 to 1,423 with the major phyla covered being *Firmicutes*, *Gamma*, and *Alpha* subdivisions of *Proteobacteria*. It is possible that producing a complete genome sequence is becoming less popular due to the improvement in software that can work on draft genomes (Chain et al. 2009).

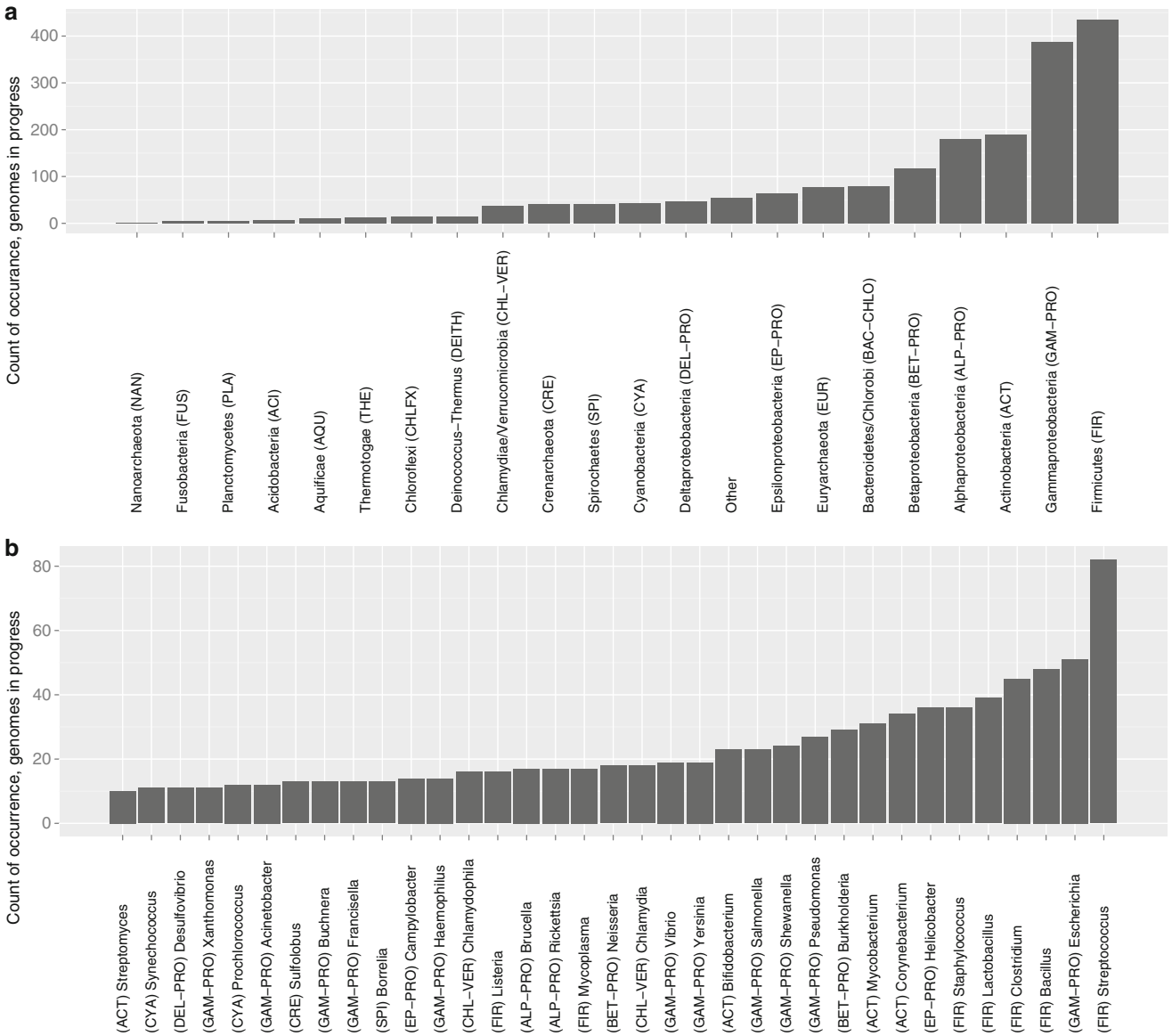

Fig. 8.2

Number of genomes sequences from each phyla and genera. Only genera with more than 10 representative genomes are shown

The cost of sequencing and the development of cheaper sequencing methods have most definitely had an impact on the rate of sequencing (Sboner et al. 2011).

Taxonomy Analysis, Most Sequenced Phyla and Genera

Figure 8.2a shows the number of genomes within each phyla; *Firmicutes* and *Gamma Proteobacteria* are the most highly represented. The plot in *Fig. 8.2b* shows genera with more than 10 sequenced genomes. The genus of *Streptococcus* is highly overrepresented (63 genomes) while the closest other group is *Escherichia* (45 genomes). According to supporting data from the GOLD database (http://genomesonline.org, March 2011) over 73% of the listed *Streptococcus* and more than 64% of *Escherichia* are labeled as pathogens.

It is likely that some organisms are sequenced because of their medical relevance. Organisms belonging to the *Escherichia* genus have a considerable role within the medical world, with *Escherichia coli* being the cause of serious food poisoning. The tendency of pathogens to be more often sequenced is, however, not as strong overall. The fraction of pathogens within each genus varies from 7% (*Lactobacillus*) to 98% (*Listeria*), for the six different genera belonging to the *Firmicutes*. For all the genera listed in *Fig. 8.2b* the range covers everything from 3% (*Synechococcus*) and 100% (*Borrelia* and *Rickettsia*, supporting data from GOLD). It should be noted that the

annotation of "pathogen" is not always accurate; for example, the first sequenced genome, *Haemophilus influenzae* is listed as a "pathogen," although the strain sequenced (Rd KW20) is a nonpathogenic strain.

In any event, it is clear that there is a strong sequencing bias, making the available data for certain phyla and genera considerably more than others. It could be expected that more pathogens than nonpathogens would be sequenced, as the immediate interest in these organisms is larger. However, the bias is not directly linked to pathogenicity, as some genera are sequenced more often though not being serious pathogens. For example, species like *Escherichia coli* (urinary tract infections, simple diarrhea, dysentery-like conditions) include pathogens but are not as severe as other species like *Borrelia* (Lyme disease) or *Listeria* (Listeriosis in newborn infants, elderly patients, and immunocompromised). *Escherichia coli* is, however, a significant player in the financial aspect of medical relevance. These organisms, though rarely lethal, can occur frequently in the population, and still require treatment; this is a burden on any healthcare system. Another factor could be the economic cost, as some organisms grow less easily or replicate very slowly making experiments long and expensive. Some pathogens require extreme safety procedures when cultured and this consumes time, space, and money. The historic factor could also be partly responsible for sequencing bias. Some organisms became model organisms from the early stage of microbiology and as such, many procedures are optimized for these organisms. Unfortunately, due to the large variety within the microbial world, many organisms will not respond well to procedures developed with *Escherichia coli* as the template. Taxonomical bias in sequencing data is, as stated, a complex and multifaceted discussion that will probably never end. However, these tendencies should be kept in mind when accessing the data available for analysis.

Basic Genome Statistics

The sequences of 1,500 genomes have been obtained from NCBI GenBank and analyzed according to basic statistical parameters. Here, basic genome statistics refers to certain DNA properties of the genome, such as genome size, frequencies of A and T bases in the DNA, and bias on the third codon positions for the open reading frames.

Figure 8.3a is a box and whisker plot showing the variation of genome sizes within each phylum. As seen, several phyla have a wide distribution of sizes. Phyla containing only a few genomes (less than five sequences) show very little size variation that could be the result of sequencing several closely related strains. For most phyla size is not a key feature, although *Chlamydia* and *Nanoarchaeota* are expected to be small genomes. The genomes within the *Firmicutes* are distributed over a broad spectrum (580–8,300 kb), while *Gamma Proteobacteria* size varies between 32−7,215 kb. Large genomes are often seen within *Planctomycetes*, *Beta Proteobacteria*, and *Actinobacteria* while small genomes are found within *Epsilon Proteobacteria*, *Chlamydiae*/*Verrucomicrobia*, and *Nanoarchaeota*. Of the largest genomes (more than 8,000 kb, 32 genomes), members of *Actinobacteria* are prevalent (14 genomes) ranging from 925 kb to 11,937 kb. The largest genome, as of May 2011, at 13,033.779 kb, was *Sorangium cellulosum* So ce 56 (soil-dwelling bacteria) (Schneiker et al. 2007).

An interesting perspective on genome size is the focus on the minimal genome for a free-living organism. Defining the minimal genome is a science in itself and has been heavily discussed in the scientific community (Galperin 2006). In 1995, the genome of *Mycoplasma genitalium* (a parasite) was published, and at that time this was thought to be the smallest genome of any free-living organism (Fraser et al. 1995). Of the 1,500 genomes in this study, *M. genitalium* is the 18th smallest genome. Upon closer inspection, the eight smallest "genomes" are described as phage, Integrating and conjugative elements (ICEs), pathogenicity island, or genomic island, so not free-living organisms. These nongenomic sequences have been reported to GenBank, and since have been deleted from the list of genomes during this work. Other genomes smaller than *M. genitalium* consist of *Buchnera* (an endosymbiant (Pérez-Brocal et al. 2006) and *Nanoarchaeum*) another symbiont (Waters et al. 2003). The remaining seven genomes are *Candidatus* species, from proposed genera, and all of these are described as symbionts (McCutcheon et al. 2009). It is worth mentioning that the smallest genome of a "true" free-living organism (as opposed to parasites like *M. genitalium*) is considerably larger, containing more than a thousand protein-encoding genes. Two proposed "minimal free-living organisms" are *Pelagibacter ubique* (heterotroph, 133th smallest in this study; DeWall and Cheng 2011) and *Prochlorococcus marinus* (autotroph, 209th smallest in this study; Moya et al. 2009). Note that these genomes are still smaller than the largest viral genomes (Arslan et al. 2011).

Another genome statistics commonly used is the percentage of AT (*Fig. 8.3b*), which is calculated as the average AT content of all the DNA sequence. Genomes with high AT content include *Candidatus Zinderia insecticola* CARI (86%, *Beta Proteobacteria*), *Candidatus Carsonella ruddii* PV DNA (83%, *Gamma Proteobacteria*), and *Buchnera aphidicola* str. Cc (*Cinara cedri*; 80%, *Gamma Proteobacteria*). These are all extremely small genomes. They are also all symbiotic organisms living inside insects, the spittlebug *Clastoptera arizonana*, jumping plant lice and plant lice, respectively. Genomes with low AT (high GC) content include *Anaeromyxobacter dehalogenans* (*Delta Proteobacteria;* Sanford et al. 2002) and *Cellulomonas flavigena* (*Actinobacteria*; Abt et al. 2010), both with around 25% AT. These genomes consist of an anaerobic and aerobic soil-bacteria, respectively. The AT content within each phyla shows some specificity with phyla like *Acidobacteria*, *Actinobacteria*, and *Deinococcus*/*Thermus* having a significant skew toward low AT and *Fusobacteria*, *Epsilon Proteobacteria*, and *Aquificae* having a skew toward high AT (*Fig. 8.3b*).

Can the AT content of an organism be related to its size? The answer can be both "yes" and "no." The numbers from 1,500 genomes show that these two properties are not always proportional to each other. However, for very large and small genomes,

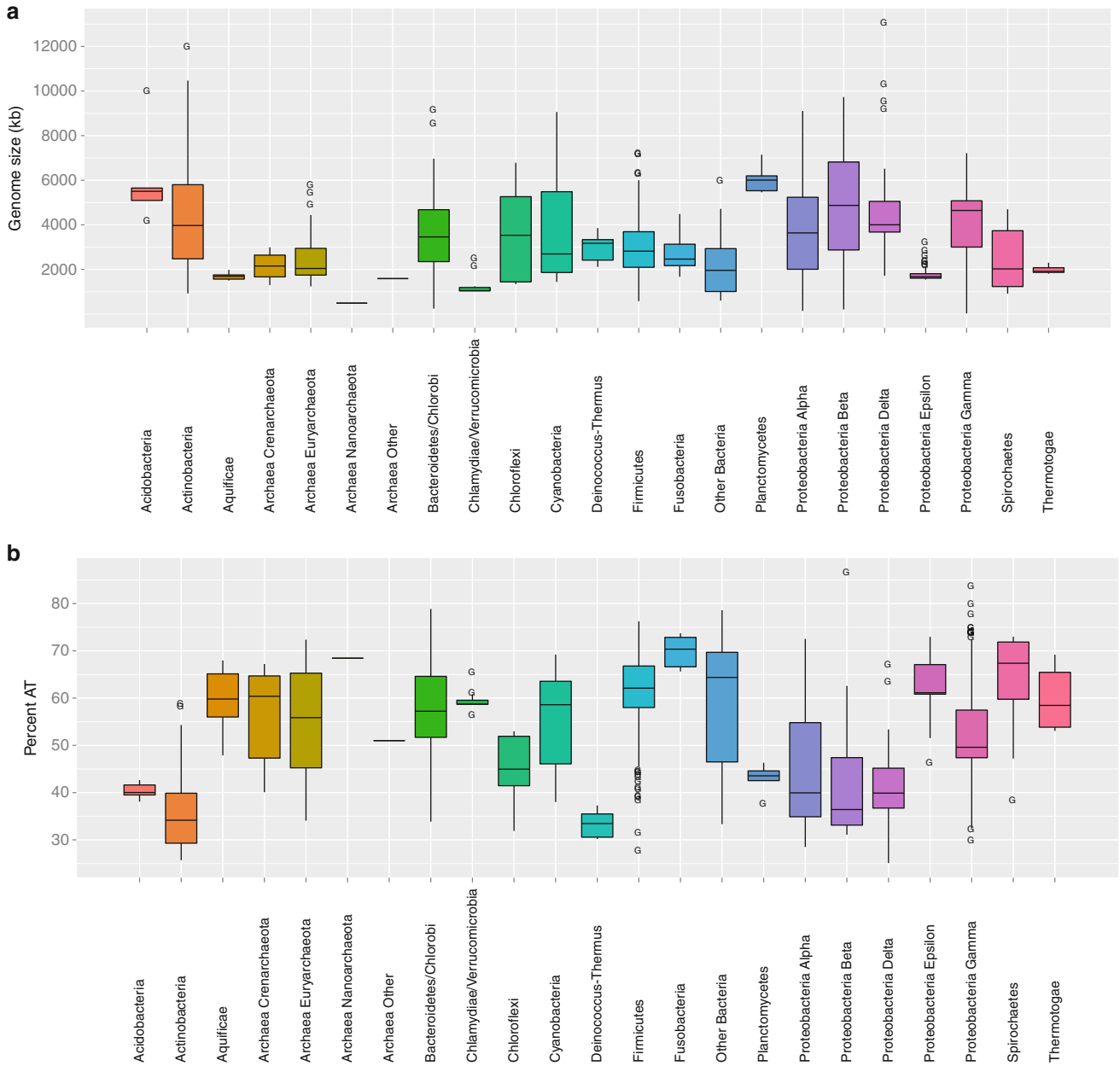

Fig. 8.3

Boxplots showing the distribution of genome size (in kilo base-pairs, *panel a*) and AT content (in percentage, *panel b*) for each phyla (as described by NCBI Taxonomy). The *middle bar* is the 50 % percentile, the *bottom* and *top* of the *box* are the 25 % and 75 % percentiles (Q1 and Q3, respectively). *Whisker bars* extend to the most extreme data point which is no more than +/− 1.58IQR/sqrt (n), where IQR is the interquartile range (IQR = Q3 − Q1). Any data point that exceeds this limit is plotted as an individual data point (outlier). The genome size was calculated as the sum of lengths of all contigs

the answer can be "yes." *Figure 8.4* shows a scatterplot of genome size and AT content (Pearson correlation coefficient of −0.48), showing that small genomes have high AT content and large genomes have low AT content. The analysis also shows a cloud around the middle values, indicating that average size corresponds to an AT content with high fluctuations.

On the other hand, an interesting relation is seen between AT content and bias in third codon position. *Figure 8.5* illustrates a strong correlation between these two properties of a genome (Pearson correlation coefficient of −0.94). The third codon position is the most variable position for the codon and this is where the largest variation in base use would be expected. The correlation was therefore expected and shows that high AT content in a genome correlates with a bias close to −1 (which is 100% AT in the third codon) and low AT content correlates with a bias close to +1 (which is 100% GC in the third codon).

Fig. 8.4

Scatter plot showing percentage AT compared to total genome size (kb) for 1,500 genome sequences. The Pearson Correlation Coefficient (PCC) for this data is −0.48, which shows a medium correlation. PCC is often used to measure the linear dependence between two variables, and takes a value between +1 and −1, where 0 reflects no linear correlation

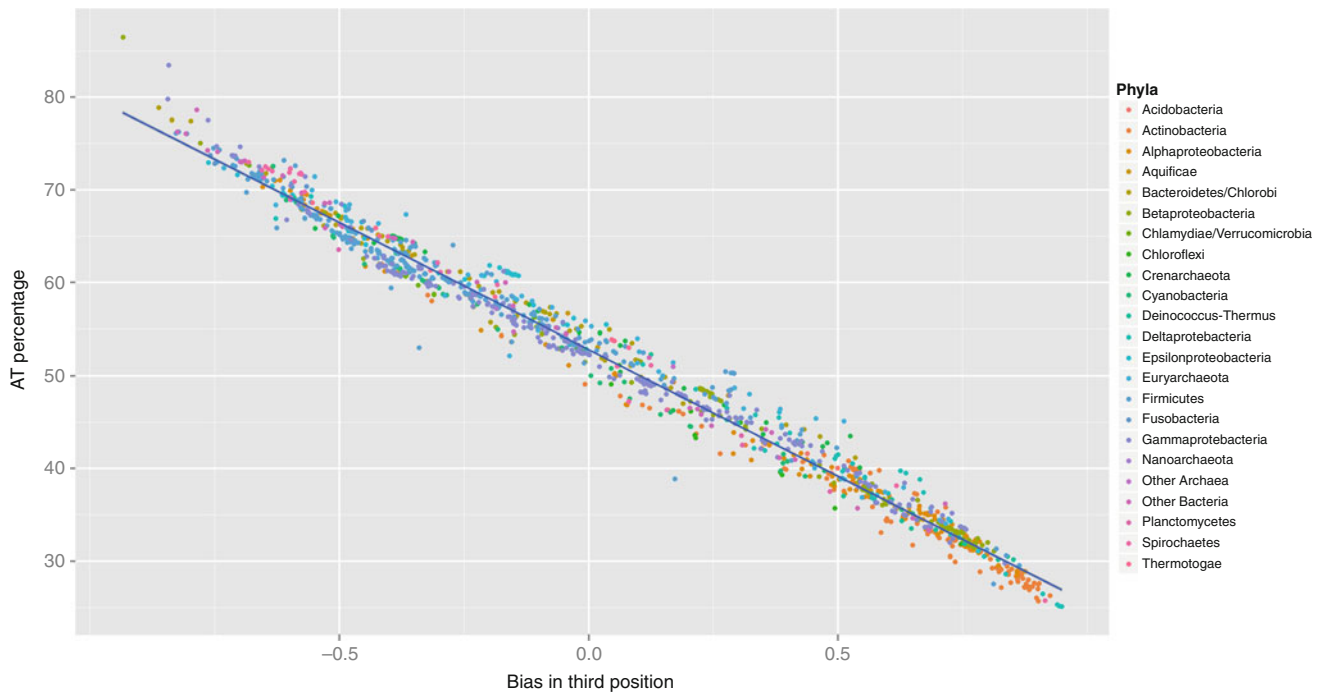

Fig. 8.5

Scatter plot showing percentage AT compared to base bias in third codon position for 1,500 genome sequences. Bias is calculated so that 100 % A or T in third codon position gives a score of −1, 100 % G or C in third position gives a score of +1. The Pearson Correlation Coefficient for this data is −0.94, which shows a strong correlation. PCC is often used to measure the linear dependence between two variables, and takes a value between +1 and −1, where 0 reflects no linear correlation

Thousands of Genome Sequences

Availability of thousands of genomes makes it possible to investigate phylogenies based on genomic information and see how current taxonomy is affected. The development of many computational tools and increasing computational power makes it possible to compare whole genomes in a reasonable time, yet comparison of thousands of whole genomes is still a tedious process. Therefore, three data sets were selected that represent different taxonomic levels of prokaryotes. The first data set is chosen to cover a wide coverage of all the prokaryotic organisms (126 genomes and 23 phyla). The second data set is a representative of a well-defined prokaryotic family (*Enterobactericiae* family, 50 genomes). The third one is chosen as an example of a prokaryotic species and close relatives (*Escherichia coli, Escherichia fergusonii, Shigella*). Different computational methods that we have encountered to be fitting in the current taxonomy of prokaryotes were shown for each data set in the following sections.

Whole-Genome-Based Tools for Taxonomy

The previous section showed the growth in available sequence data as well as the bias and diversity in this data. This large diversity and coverage opens the doors to large-scale phylogenetic analysis of genome sequences. As a result, great insight into bacterial evolution and diversity has come from comparison of many microbial genome sequences in the last decade. The differences, even between strains of a distinct taxonomic cluster, show that bacteria represent a great diversity, which led to the formation of the hypothetical concepts of "pan-genomes" and "core-genomes." The pan-genome contains the total number of genes found in the gene pool of a set of genomes (Ussery et al. 2009) and can be viewed in three separate parts. The part that consists of conserved essential genes common to all genomes compared (core-genome). It has been seen that core-genomes of phylogenetically coherent groups contain genes that are less prone to horizontal gene transfer and are more stable such as housekeeping genes. The genes essential for colonization, survival, or adaptation to a specific environment are thought to form the lifestyle genes, which can also be named as the "shell" for frequently occurring genes. The third part is called "accessory" or "cloud" genes, as these are rarely found, often strain specific and nonessential (Lapierre and Gogarten 2009). Though hypothetical, these terms can serve use for defining and classifying bacteria. These different "genomes" can be used to explain the differences and similarities between species or genera, and visualized by pan-genome trees (Snipen and Ussery 2010). An elaborate work on the comparisons of genomic DNA using oligonucleotide-based methods and proteomes with a pan-genome approach was presented in a study by Bohlin et al., where *Brucella* species were classified using 32 genomes (Bohlin et al. 2010).

Other genome-based methods include measures for replacement of the DDH (DNA-DNA-Hybridization) analysis, such as the Average Nucleotide Identity (ANI) between pairs of genomes or the Average Amino acid Identity (AAI) of the shared genes between two genomes. The study of Goris et al. on pairwise comparison of complete sequenced genomes showed the ANI of the core genes show results similar to analysis of 16S rRNA sequence identity and DDH similarity values, concluding that a 70% DDH value corresponds to 95% ANI. Hence ANI has been shown to be an alternative to the tedious DDH method (Goris et al. 2007). Another genome-based method, AAI, has been shown to result in strong correlation between 16S rRNA gene identity and AAI-based phylogenetic trees congruent with core genome-based trees (Konstantinidis and Tiedje 2005).

There are many methods and types of data used to build evolutionary trees. The results of the whole-genome-based tools are usually values representing similarity between organisms, which can then be converted to distance-based phylogenies. Distance methods constitute a major part of phylogenetic analysis. "Least squares" is one of these methods where the sum of squares of difference between the observed and the predicted distances of a tree should be minimized. Unweighted (Cavalli-Sforza and Edwards 1967) and weighted (Fitch and Margoliash 1967; Beyer et al. 1974) algorithms are suggested for least squares. Minimum Evolution, Neighbor joining, and UPGMA are all distance-based methods. There are also methods that rely on probabilities of evolutionary change. Maximum likelihood is one of them, where different evolutionary rates can be taken into account and several models can be implemented (Felsenstein 2004). The evolutionary models and the distance methods should be chosen carefully when phylogenies are generated, as they might result in different results even for a small set.

rRNA Phylogenetic Trees

In this section the 16S rRNA and 23S rRNA comparison of 126 various organisms from all bacterial and archaeal phyla is presented (*Fig. 8.6*). This data set represents a collection of distantly related prokaryotic organisms. The first criteria for the selection of organisms for this dataset, was to get the largest and smallest genomes from each phyla (taxonomy reference is Genome metadata from NCBI and GOLD). More organisms from each phyla were selected from different environments or host associations, in order to get a less biased data in total.

Ribosomal RNA sequences of all 126 genomes were predicted using RNAmmer program (Lagesen et al. 2007). Foreach genome one 16S and 23S rRNA sequence was selected based on the highest RNAmmer score and appropriate length (Lagesen et al. 2010). The length requirements were between 1,400 and 1,700 bp for 16s rRNA sequences, 2,500 and 3,800 bp for 23S rRNA sequences. Once the RNA sequences were gathered they were aligned using CLUSTALW with default parameters (10 for gap opening penalty, 0.20 gap extension penalty, 30% Delay divergent sequences, 0.5 for DNA transitions weight, IUB for DNA weight matrix) (Larkin et al. 2007). After obtaining the alignments, the phylogenetic trees were constructed using MEGA5 (Tamura et al. 2011) and

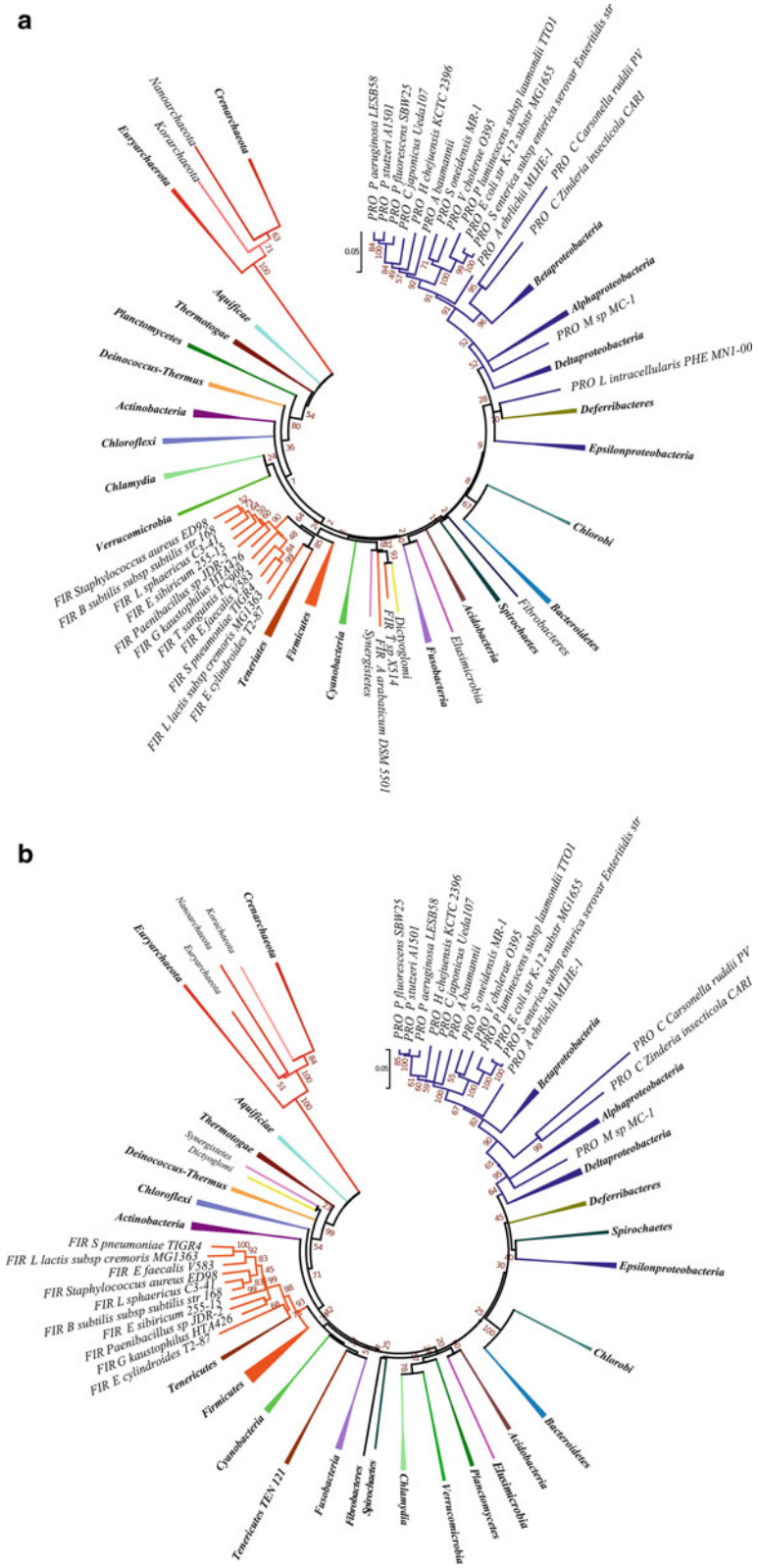

Fig. 8.6

(a)16S rRNA and (b) 23S rRNA tree with NJ method and 1,000 bootstrap resamplings from ClustalW alignment. The trees are viewed and colored with MEGA5. Branch lengths are measured in the number of substitutions per site. Each phylum is collapsed when possible, except classes of *Proteobacteria* were collapsed instead of phyla

Neighbor-Joining (NJ) with 1,000 bootstrap resamplings. The bootstrap values in these phylogenies were transformed to percentages. They give a statistical measure for how reliable a branch separation is. Therefore, higher percentages support a stronger evidence of grouping, meaning a more prominent common ancestor, whereas lower percentages mean the separation on that branch is statistically insignificant.

Ribosomal RNA phylogenies are usually able to distinguish the domains, phyla, and genera in a given set. The distances on this type of phylogeny show the divergence in the rRNA sequences. According to the bootstrap values in the 16S rRNA phylogenetic tree (*Fig. 8.6a*), the phyla level clusters are significant with higher than 80% bootstrap value on their roots. To better illustrate this, the two different clades were left uncollapsed while the remaining phyla clades were collapsed. The correspondence of the significant clades to phyla in prokaryotes is, however, an expected result. Phylum, as a taxonomic unit is not defined by the official nomenclature (International Code of Nomenclature of Bacteria (Lapage et al. 1992)). The highest rank according to the official nomenclature is a class; however, the rank phylum is also being used in prokaryotic taxonomy quite often and seems to serve practical use for the taxonomists. Historically, phyla were referred as divisions and Prokaryotes, as one of the superkingdoms proposed by Whittaker and Margulis, were divided into three divisions based on cell wall structure or absence (Whittaker and Margulis 1978; Gibbons and Murray 1978). Although the classification largely changed since the division of archaeal phyla (Murray 1989) were discovered, most of these names are still in use today. In the 2nd edition of Bergey's Manual, phylum was defined as the major prokaryotic lineages, based on the 16S rDNA sequence data and used as main organizational unit (Brenner et al. 2005b).

Ribosomal RNA based phylogenies usually involves the 16S rRNA subunit comparisons. Here we show that 23S rRNA phylogeny can also be useful. When the same dataset is analyzed using 23S rRNA genes, the bootstrap values are generally higher than 16S rRNA phylogenies (*Fig. 8.6b*). There are exceptions to this, for example, the bootstrap value on the roots of the *Gamma Proteobacteria* clade that is higher on the 16S rRNA tree. The generally higher values for the 23S rRNA analysis might be due to the size or information content of the sequences and maybe due to the different mutation rates of the genes. The separation of phyla on *Fig. 8.6b* is significant, but the order is different, which might lead to the idea of having different relationships among different phyla. However, since the bootstrap values are very low at that level, it is still not relevant to conclude how close *Firmicutes* is to *Proteobacteria* based on rRNA phylogeny.

Average Nucleotide Identities (ANI) and Tetra Nucleotide Frequency Calculations

Average nucleotide identity was developed as an alternative to the DDH values, and was initially based on comparison of all shared genes among two genomes. Later on an advance in the method, to make it more similar to DDH, was made by randomly chopping up the genome sequences in 1,020 nucleotide fragments regardless of whether or not they correspond to any ORFs. The fragments from two genomes are aligned using a BLAST (Altschul et al. 1990) algorithm or a fast alignment tool such as MUMmer without fragmentation.

In this section, 50 genomes from different genera of the *Enterobacteriaceae* family (data gathered from NCBI GenBank) are compared based on their ANI values. ANI calculations were performed as explained in the paper by Richter and Rosello-Mora's using Jspecies (Richter and Rosselló-Móra 2009). The genome sequence comparison was based on nucleotide MUMmer (NUCmer) which is a fast DNA alignment tool for large-scale comparisons (Delcher et al. 1999). MUMmer aligns two given genome sequences based on maximal unique matches (MUMs) between the sequences. A "MUM" is an exact string match that occurs once in each genome. Once the MUMs are identified, they are sorted in ascending order according to their positions in the genomes. After the global MUM-alignment, the gaps between them are closed based on the properties of the gaps. A gap can be a single nucleotide polymorphism, an insertion or deletion where a large sequence is found in one but not the other genome, tandem repeats, or polymorphic regions. If gaps are found, they are aligned using the Smith-Waterman algorithm (Smith and Waterman 1981).

Comparisons based on the tetranucleotide frequencies were calculated using Jspecies and the algorithms from Teeling et al. (2004). In this method, all possible combinations of tetranucleotide frequencies (256 frequencies) for each sequence is calculated and their z-scores are computed based on the difference between the observed and the expected frequencies for a genomic fragment. The similarity between the two sequences (or genomic fragments) in terms of having similar patterns of tetranucleotides is addressed by calculating the Pearson correlation coefficient for their z-scores. Similar patterns are expected to correlate and therefore have higher correlation coefficients, whereas the distant patterns would have lower correlation coefficients. Oligonucleotide frequencies are thought to carry species-specific signals, where longer signatures carry more signals. Thus, closely related organisms are expected to show similar distribution of the usage of these signatures.

Figure 8.7 shows a pairwise genome comparison of ANI value (heatmap). The genomes are manually ordered based on 16S rRNA similarities. It is seen that DNA similarity within a genus is higher compared to the similarity between genera. It is therefore possible to distinguish groups of genus and species based on their DNA similarity. For comparison, Tetra Nucleotide frequencies were calculated for the same data and ordered based on 16S rRNA similarity. The two heatmaps are expected to show similar results, with ANI values above 96% identity would correspond to very high Tetra Nucleotide frequencies correlation coefficients of $\geq$0.99 (Richter and Rosselló-Móra 2009). It is seen that within genera, sequences are highly correlated based on tetranucleotide signature usage. Changing the order of the matrix based on hierarchical clustering might give a better resolution using Tetra Nucleotide frequencies.

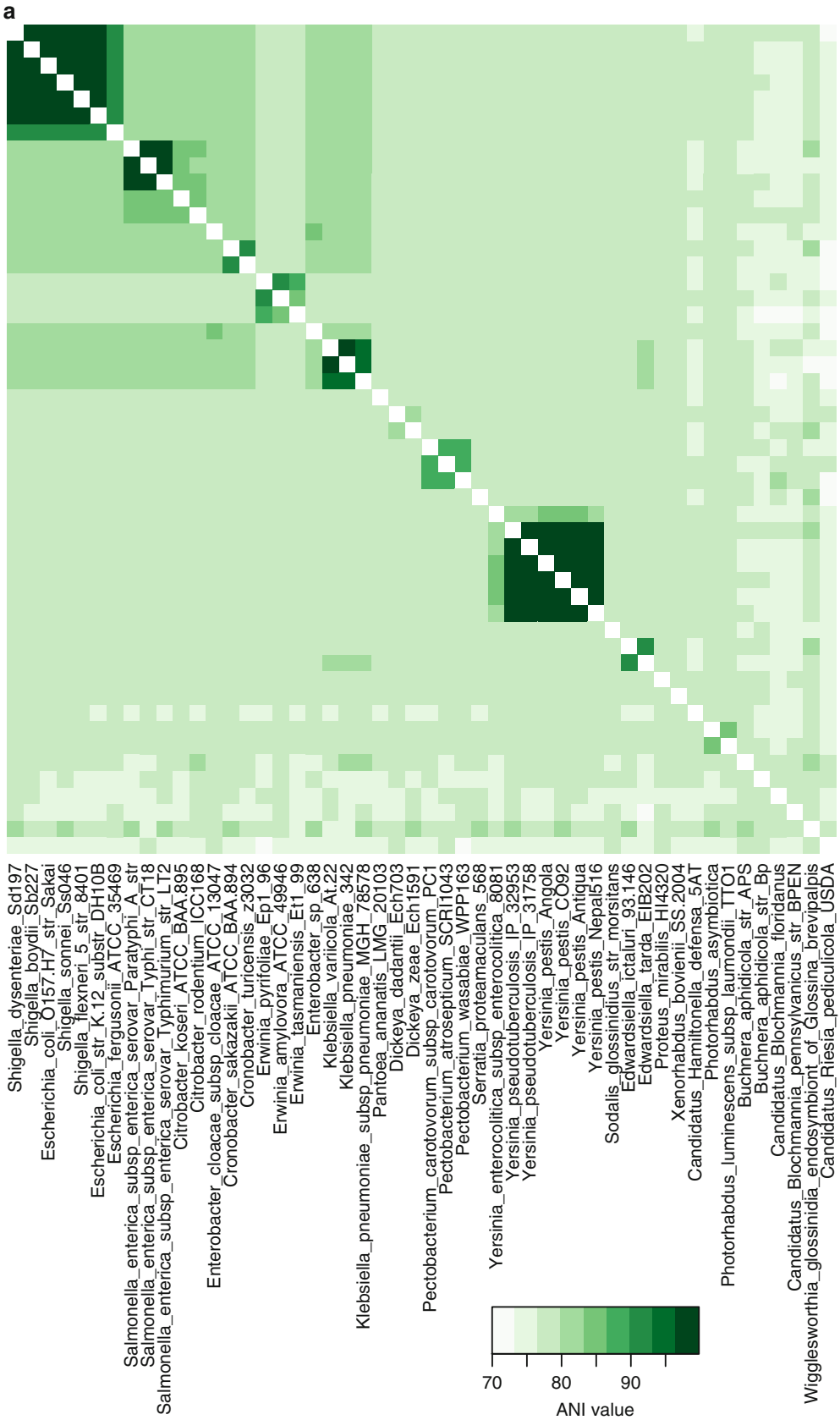

Fig. 8.7 (continued)

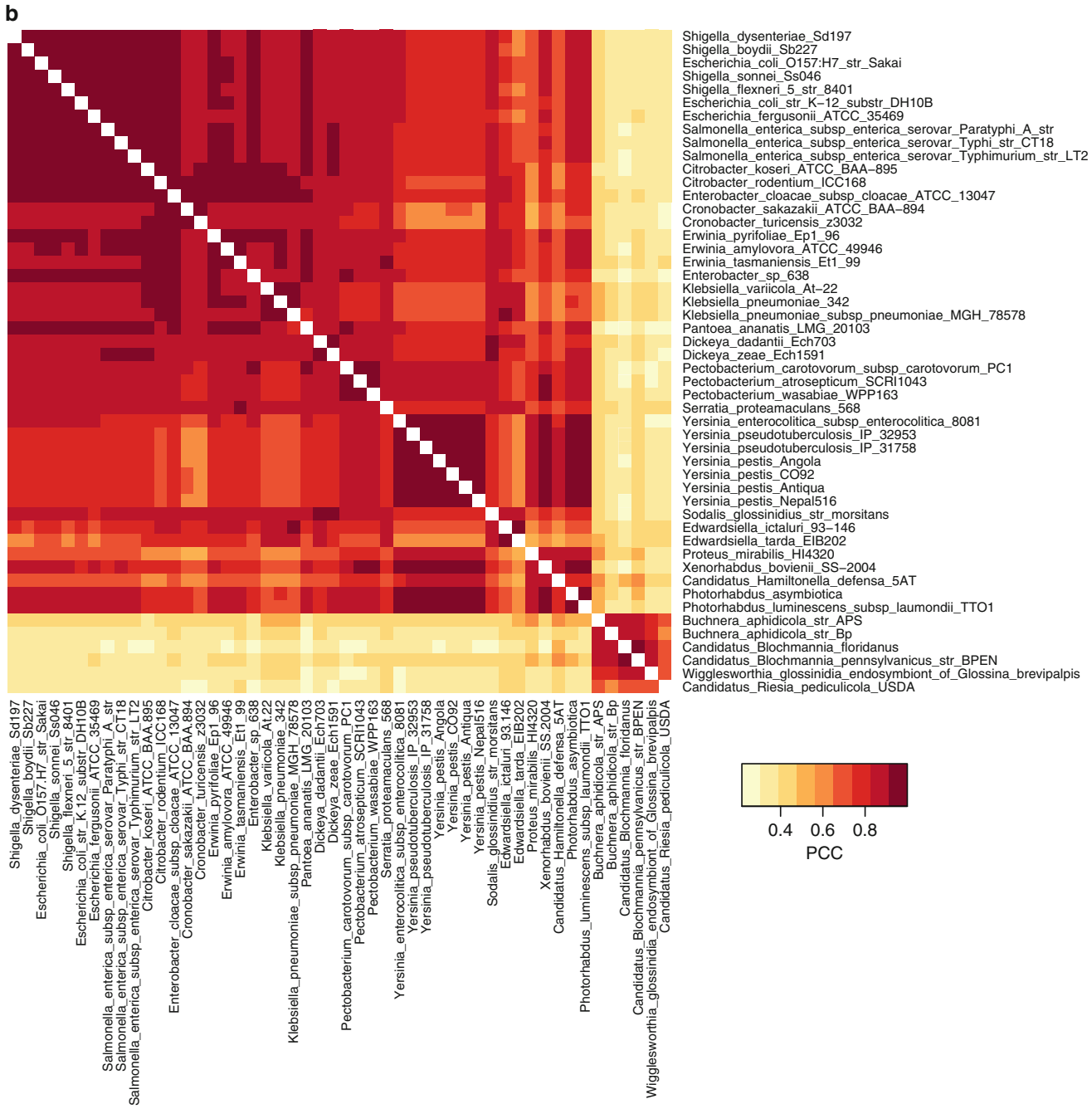

Fig. 8.7

(a) Shows heatmap for ANI values between each genome retrieved by pairwise alignments with MUMmer. Darker colors indicate higher percentages. The columns and rows are ordered based on the clusters in the 16S rRNA NJ tree. (b) Shows Pearson Correlation coefficients yielded by tetranucleotide frequency calculations between genomes of Genera set. The comparison based on Tetra calculations shows higher similarity between organisms as the colors get darker

BLASTMatrix Using Reciprocal Best Hits

The dataset of 50 genomes was used to illustrate the BLASTMatrix method. Each proteome was pairwise compared using BLASTP using a reciprocal best-hit criterion. The method is based on an all-against-all BLASTP analysis where all proteins are compared to all other proteins in the dataset. Then a reciprocal best-hit criterion is implemented. According to this, a BLAST hit will be considered significant only if the length of the alignment is at least 95% of the longest protein and has

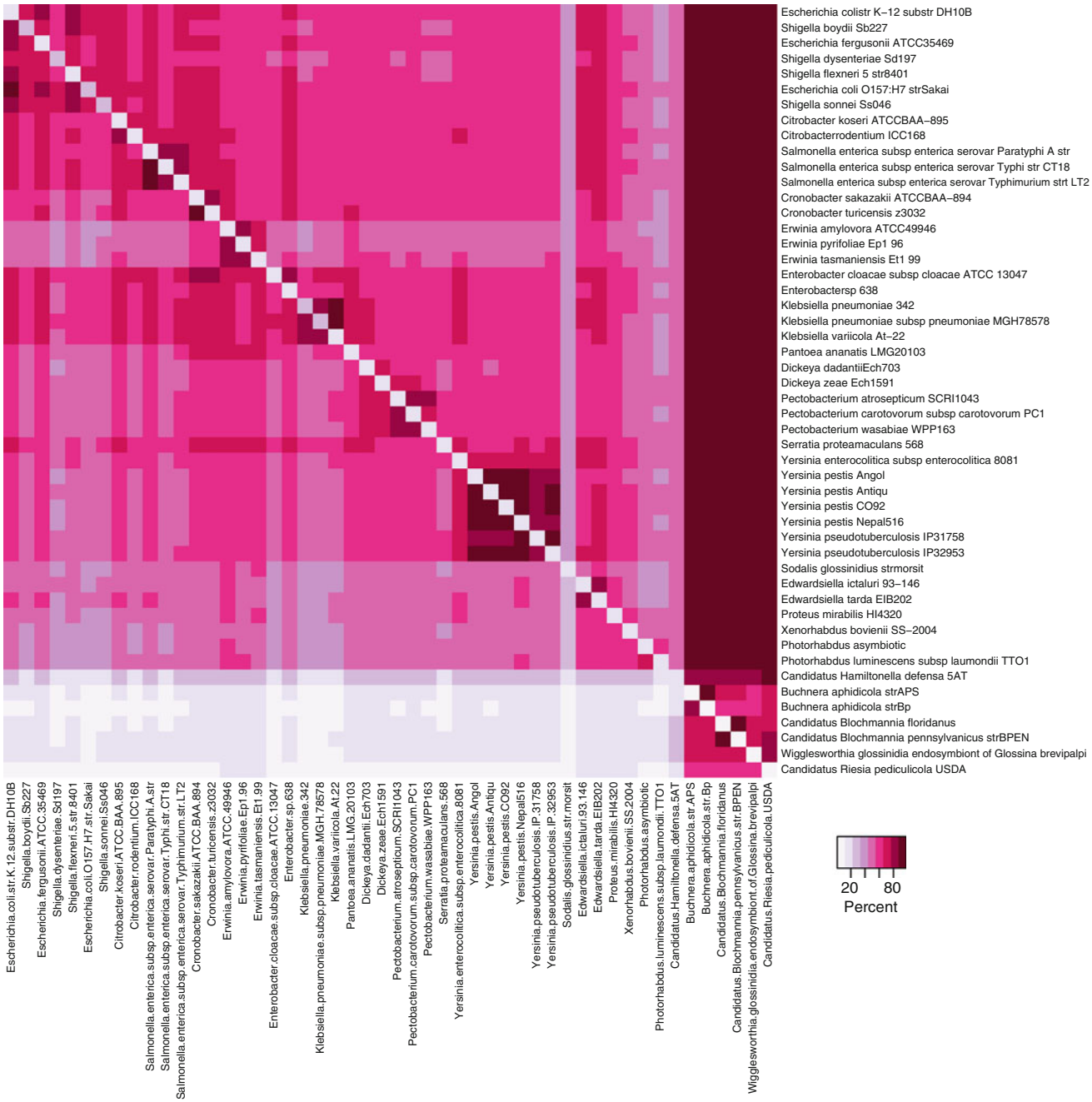

Fig. 8.8

Proteome comparison between the genomes of Enterobacteriaceae based on BLASTP searches

95% sequence identity. There are three possible outcomes for each protein: (1) The protein does not have any significant hits to any other protein, (2) the protein has one or more hits where the hit with the highest bitscore is chosen as best hit, and (3) the protein has more than one hit where there can be several best hits. Once identified based on this three possibilities, best hits for each protein are stored. For each proteome, a hit will be counted if the best hit of protein X in proteome A is also the best hit of the corresponding protein Y on proteome B. In the case where there are several best hits, the hit is counted if protein Y is one of the best hits for protein X and vice-a-versa. When a proteome is BLAST searched against itself, a protein will have a hit to itself and it might have another hit to another protein in the same genome. The results are then put through homology reduction by the HOBOHM algorithm, in order to reduce the self-hits (Hobohm et al. 1992).

The output of the BLASTMatrix analysis is a matrix of numbers. *Figure 8.8* shows a heatmap of these values though not the actual value. In the discussion below the actual values are presented. In the matrix for the *Enterobacteriaceae* family, the

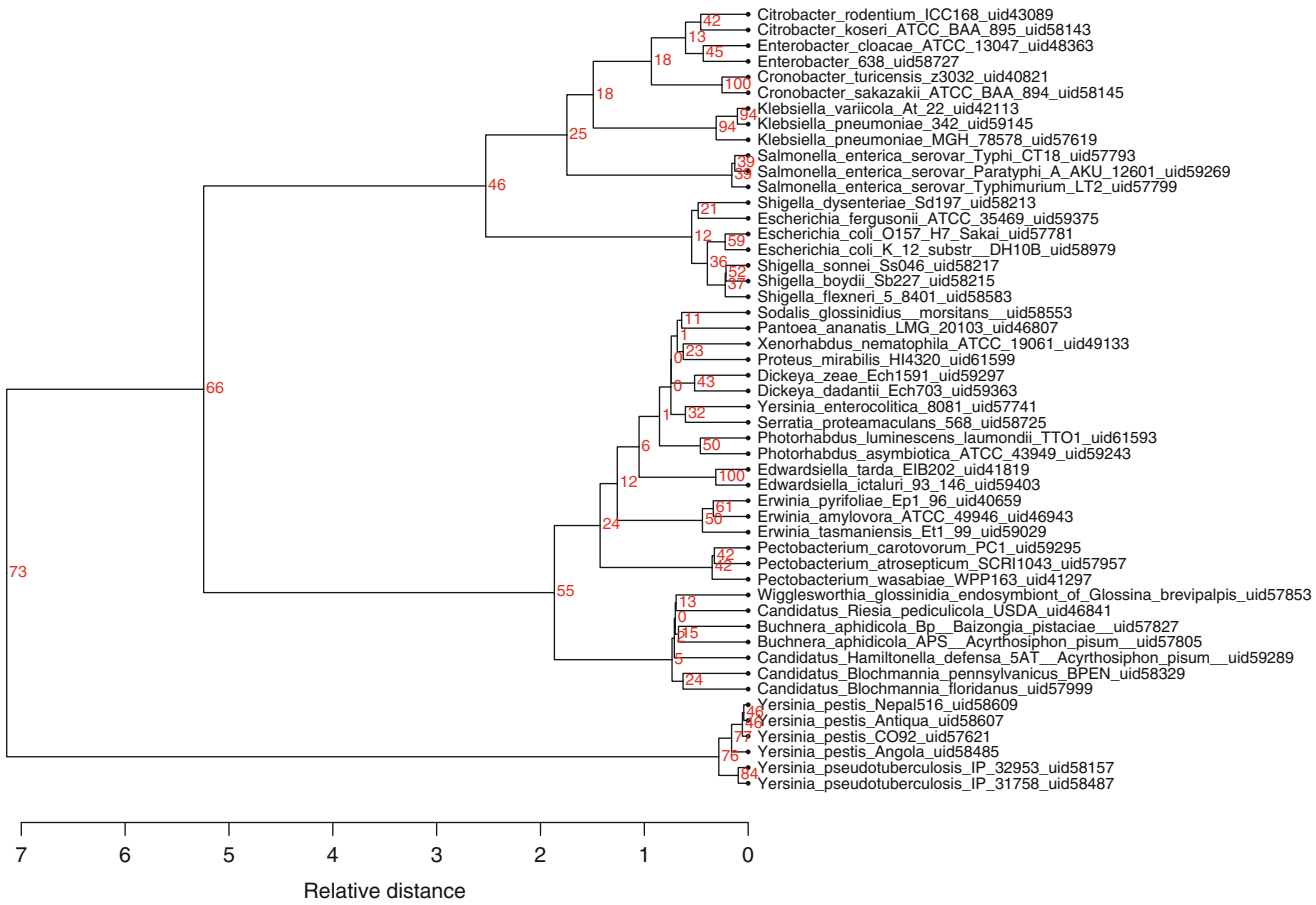

Fig. 8.9

CVTree for the Genera set, generated by using "euclidian" distances with "ward" linkage hierarchical clustering. The values on branches indicate bootstrap values

order is based on the rRNA phylogenies, because a clear separation of each genus was possible. According to this matrix, proteome similarity within some genera might seem to be higher than others; however, the darkest colors around the diagonal (≥90%) are due to the comparison within a species level, rather than different species in the same genus. Average proteome similarity between *Escherichia* and *Shigella* clusters is 73%. Homology between *Salmonella* strains are 80–96% while *Citrobacter* species show an average of 72% similarity. The similarity values for *Cronobacter* are 88% on average, 80% for *Erwinia*, around 70% for *Enterobacter*. *Klebsiella* species have up to 94% similarity, *Dickeya* species around 74% and *Pectobacterium* species have 75–83% similarity (Fig. 8.8). In the *Yersinia* cluster, similarity values between the strains of same species are 92–98% for *Y. pestis* and 88% for *Y. pseudotuberculosis*. Homology between *Yersinia* species is around 88%. *Edwardsiella* species have 81%, *Buchnera* strains have around 89%, *Candidatus Blochmannia* strains have around 96% similarity.

It is not clear if there should be a proteome similarity cut-off to specify genera, however, the values within a genus are generally between 70% and 80% and within the species it is higher, 80–98%. For example, *Klebsiella variicola* can actually be a *Klebsiella pneumoniae* because it has 94% similarity to the *K. pneumoniae* 342. Another issue is the presence of the reduced genomes in this family. In this plot, the reduced genomes are seen in the upper right corner (Fig. 8.8). The percentage of proteins that these organisms share with the others is very high because of the small size of their proteomes, which creates a dark band on the top of the matrix and a lighter one on the right hand side. This might be due to these genomes containing generally conserved core genes of the whole family, and they have actually very few accessory genes compared to the other genomes. They also have very low internal homology. A *Shigella* genome, on the other hand, has up to 30% internal homology. The largest proteome, *Sodalis*, shares 32–40% with the genomes that are not among the reduced genomes group. Except the reduced and expanded genomes, the similarity levels among different genera range between 40% and 72%.

Composition Vector Trees (CVTree)

In this section, a CVTree for the 50 *Enterobacteriaceae* genomes is presented. In this method, frequencies of overlapping

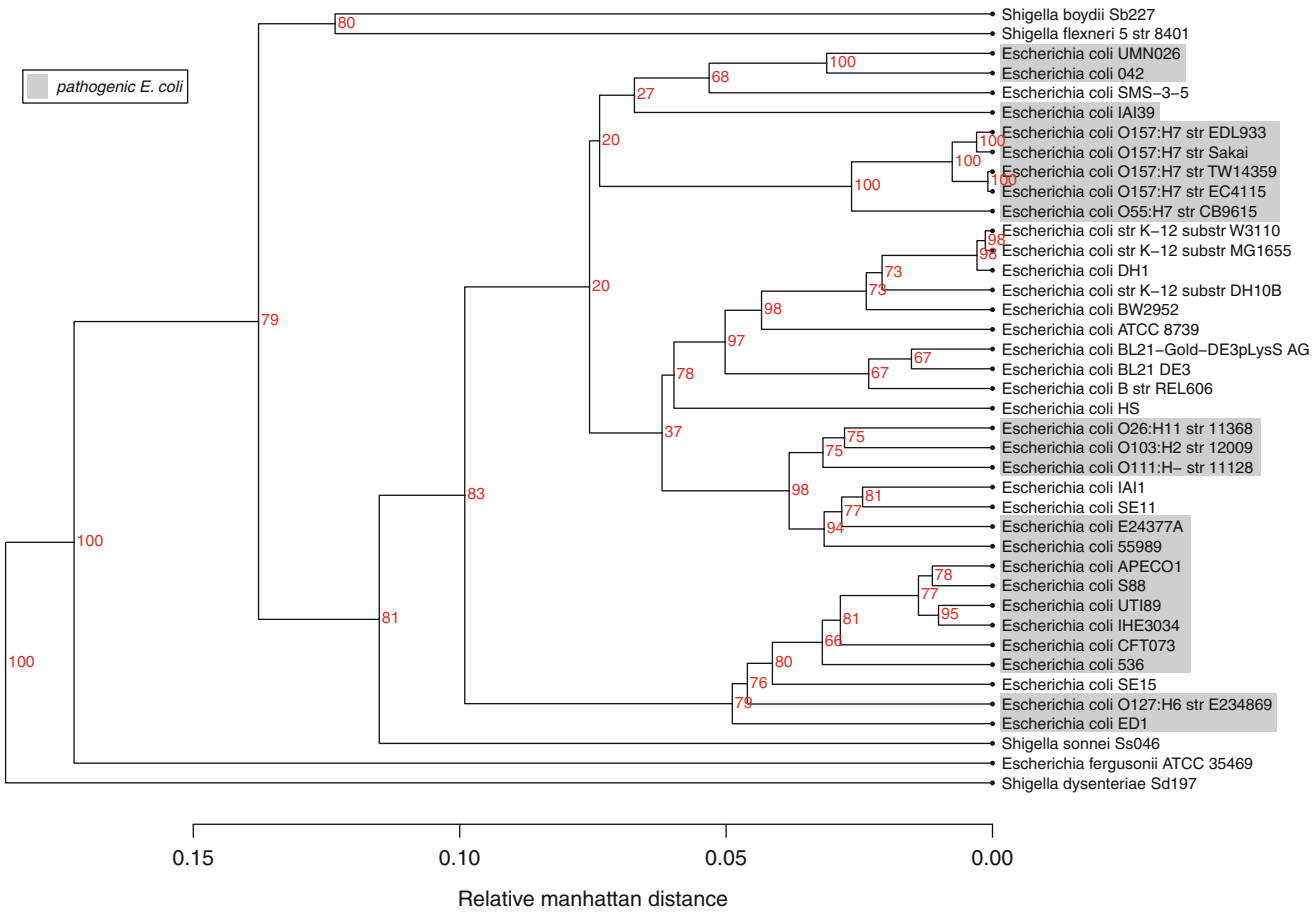

Fig. 8.10
Pan-genome tree for the *Escherichia* and *Shigella* set based on the "shell" genes of the data set. The *gray boxes* indicate pathogenic strains of *E. coli*

oligopeptides of length K are calculated. The random background is subtracted from these frequencies with a (K-2) order Markov model to avoid the bias from neutral mutations, highlighting the selective evolution. After this procedure, the pairwise distance is calculated using the correlation between two organisms. The method describes the resulting distances as $D = (1 - C)/2$, where C is the correlation between genomes and D is the distance (Qi et al. 2004). The updated CVTree method with a faster and more stable web server performance is described in Xu and Hao (2009). The computation was performed on the web server using proteomes for all genomes and a K-parameter set to 6.

The outcome from the analysis is a distance matrix based on amino acid sequence comparison, which is then used to generate phylogenetic trees using neighbor joining. The web server provides an NJtree, made using the PHYLIP package, with a Newick format output. However, the NJtree did not provide bootstrap values so the data was instead analyzed using R, version 2.11. With this approach, the distance matrix from the CVTrees output was used to generate phylogenetic trees with bootstrap value. In *Fig. 8.9*, new distances were calculated from the distance matrix using the "Euclidian" distance function and hierarchical clustering using the "ward" method (Ward 1963). For statistical significance 100 bootstrap samplings were made on these trees, giving values between 0 and 100, showing the reliability for each branches.

The CVTree for the *Enterobactericiae* family can be seen in *Fig. 8.9*. In this tree, *Yersinia pestis* and *pseudotuberculosis* species are clearly separated from the rest of the taxa with 73% bootstrap value, whereas *Yersinia enterocolitica* remains clustered with *Serratia*. Two main clusters are seen after this, where the first consists of *E. coli*, *Shigella*, *Salmonella*, *Klebsiella*, *Cronobacter*, *Citrobacter*, and *Enterobacter*. The branching order is very similar to rRNA phylogeny as well.

To summarize, the CVTree is an alignment-free method, where the sequence similarities are investigated by the use of hexamer frequencies (K = 6); therefore, it is also quite fast. It is seen that the distance matrix generated by the algorithm does not give a clear picture for classification when very diverse organisms are analyzed. On a genera level, all the genera are intact and the clustering is very similar to the rRNA trees. On the species level, *Shigella* and *E. fergusonii ATCC 35469* are separated from *E. coli*.

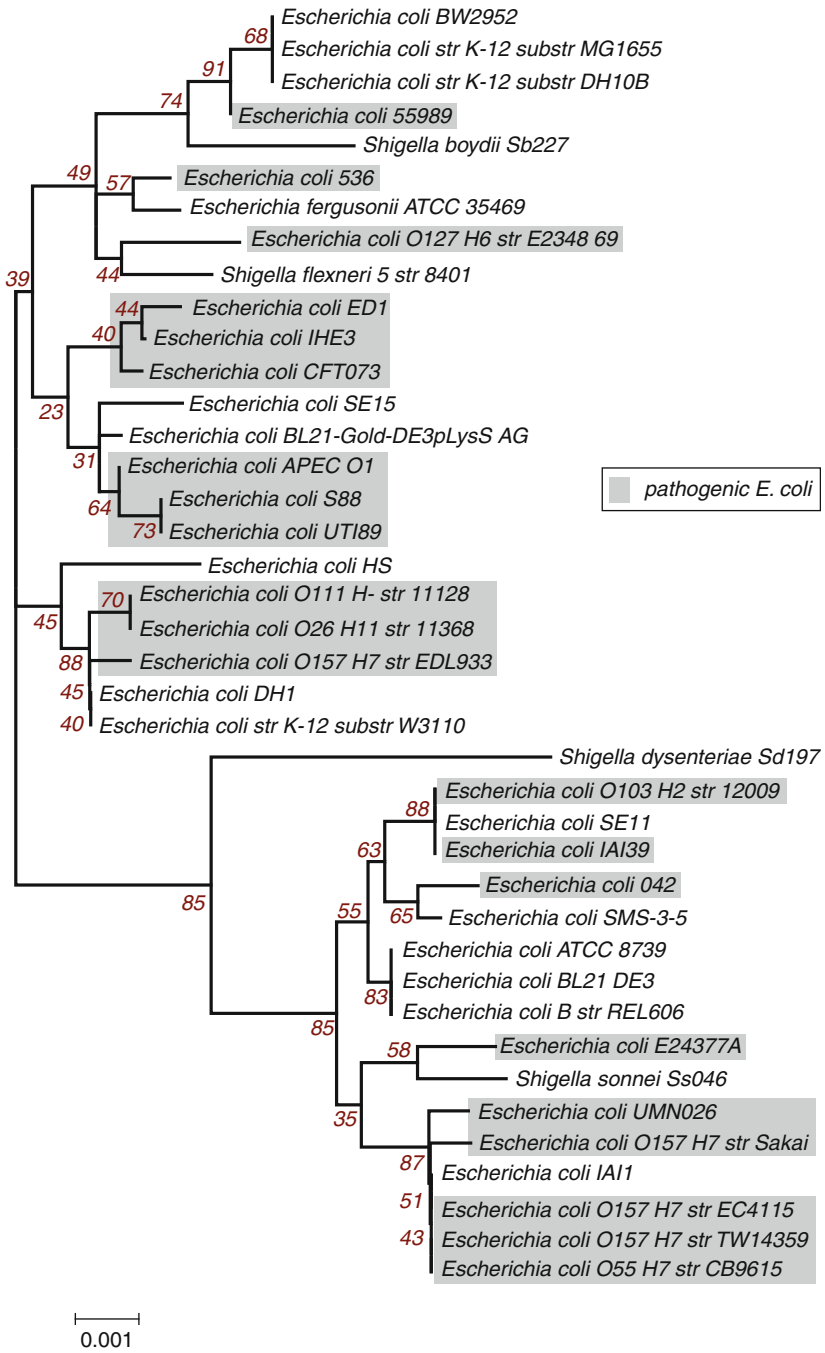

Fig. 8.11

16S rRNA phylogenetic tree for *Escherichia* and *Shigella* genomes. Sequences are obtained as explained in rRNA phylogenies previously. The tree was generated using ClustalW alignments with neighbor-joining method and statistically tested with 1,000 bootstrap resamplings

Pan-genome Trees

Pan-genome family trees were generated using BLASTP (Altschul et al. 1990) similarity between each proteome. According to this, genes that have a significant hit to each other are considered to be in one gene family, where the significance cut-off is chosen for each BLAST hit (50% identity over an alignment with a length of at least 50% of the longest gene). Once the gene families are assigned, a matrix is constructed containing the gene families in columns and the genomes in rows, having 1 for the presence of that gene family in the corresponding genome, 0 otherwise. The tree is constructed by calculating Manhattan distances from this matrix and making hierarchical clustering using the Unweighted Pair-Group Method with Arithmetic mean (UPGMA) algorithm. For the stabilome view, the gene families that are represented in only one

genome (ORFans) are weighted down and the tree is based on "shell" genes between the genomes (Snipen and Ussery 2010).

In this section, a pan-genome tree for genomes of 36 *Escherichia coli* and 4 *Shigella* species was constructed (data gathered from NCBI GenBank) (Fig. 8.10). This tree, compared to a 16S rRNA neighbor-joining phylogeny (Fig. 8.11), shows more clear separation on the pathogenicity of the *E. coli* strains. The pan-genome tree shows relationships among different strains of a family with a higher resolution. *Shigella* species and *E. fergusonii ATCC 35469* are also clearly separated from the *E. coli*, showing that they show clear differences in their proteome composition, therefore, they can be separated from *E. coli*.

Summary

This chapter presents analysis on genomic data that is in public databases and comparative genomics approaches to taxonomy based on rRNA, DNA, and protein molecular sequences.

It is clear that the available genome data is biased which cannot be attributed to any one reason. However, the monoculture approach to genome sequencing is causing a significant skew in sequencing data. From experiments, it is known that the diversity in the microbial world is tremendous, but in the statistical results, this diversity is not covered in the sequenced data. The advances in metagenomic sequencing and the sequencing of noncultivable cells will in time result in a much more realistic view of the bacterial world than is seen now. In the mean time, scientists should be aware of the bias in sequence data and not believe that what is sequenced so far is representative, even with roughly 2,000 genomes finished, and another 3,000 "draft" genomes available; there are currently about 30,000 bacterial genomes available in the "short-read archives," which in principle could be assembled, sometimes into less than 100 pieces—this means that there are ~35,000 genomes available now, and within a year or two, the number is likely to be in the hundreds of thousands. It seems likely that in the near future, draft genomes will become more common; if done properly and assembled well into only a few contigs, these draft genomes can provide useful information for core- and pan-genomes of a given taxa. However, one can hope that in the not too distant future, emerging third-generation sequencing technology will allow for the economical production of high quality full-length genomes for more reliable and robust information.

The amount of available data makes sequence-based taxonomy inevitable. In this chapter, organisms with different levels of taxonomic relations were selected. Since the reference taxonomy used is the current taxonomy, the results shown have been selected to be as close as possible to current taxonomy. rRNA phylogenies were used since it is a classical approach. Although the method is based on a single gene, the more conservative nature of the rRNA genes gives them a unique advantage of identifying more distantly related organisms. Whole-genome approaches, on the other hand, are more precise for understanding relations among closely related organisms. There are also differences between all the methods in the sense of using the sequence information; some use the sequence directly and make use of alignments and some reduces this information content into vectors of numbers. The latter can be thought more as numerical taxonomy, where several properties of organisms are measured and statistical significance tests and clustering methods are used to analyze relationships. As a result of all this analysis, it is suggested that there might not actually be one unified theory on taxonomy of living things, but several which classify well in different taxonomic levels. These methods are shown in Table 8.1, where a method explaining the relationships among different taxa are referred as inter-taxa, and methods that can delineate specific taxa from others regardless of their relations with others are referred as intra-taxa.

Table 8.1
Methods that can be used for investigating inter- and intra-taxa relationships. The highest level is the Three Domains of life, and the lowest level is within the Strains

Taxonomic levels	Inter-taxa	Intra-taxa
Superkingdom		16S and 23S rRNA phylogeny
Phyla	None	16S and 23S rRNA phylogeny
Genus	16S and 23S rRNA phylogeny, BLASTMatrix	16S and 23S rRNA phylogeny ANIm and Tetra, CVTree, BLASTMatrix, Pan-genome tree
Species	16S and 23S rRNA phylogeny, BLASTMatrix, Pan-genome tree, CVTtee, ANI and Tetra	16S and 23S rRNA phylogeny, BLASTMatrix, Pan-genome tree, CVTtee, ANI and Tetra
Strain	BLASTMatrix, Pan-genome tree, CVTree	Pan-genome tree

As seen from the results, the largest phylum in terms of having the highest bacterial genome sequence projects, *Proteobacteria*, seems to be a well-defined taxa, where most of the methods catches the clustering, and the classes of *Alpha*, *Beta*, *Gamma*, *Delta*, and *Epsilon Proteobacteria* are usually coherent within themselves based on rRNA base taxonomy. Second largest phyla, Firmicutes are usually clearly separated into two groups of different classes. Another group, *Cyanobacteria* is actually a subdivision, because they do not have any classes or orders defined. Its members usually cluster together in many methods. All the other phyla generally have their members clustered together. This makes sense in classification, if looking for clearly separated groups. However, from an evolutionary point of view, this result is not enough to understand the ancestral relations of different phyla.

In the levels of genera, relationships for *Enterobacteriaceae* family can be seen in many methods. These families include some clinically and industrially important bacteria. Genera inside *Enterobacteriaceae* are clearly separated into different clusters, although sometimes the *Enterobacter* genus clusters separately. The relations of reduced genomes vary in different methods but generally they are separated from the rest of the genera. They share the core genes with the rest of the family and have very few accessory genes.

On a species level it is seen that even though in the same species, bacteria can be very diverse in terms of proteome content. Distinguishing between different types of *E. coli* strains are not possible with classical methods, but with proteome comparisons. It is also clear that although they are historically known to be similar, *Shigella* and *E. coli* can be distinguished from each other with proteome comparison.

In theory, classification simply depends on how one defines the relatedness among the entities of the system. However, in reality the choice of the characters when building a system is not always simple. Whole-genome-based tools do not consistently agree with current taxonomy. In order to make them match the current taxonomical system, different methods should be used for the different levels of taxonomy. The emergence of large amount of molecular and genomic data made it evident that there is not one universal method to naturally classify prokaryotes. Taxonomists should therefore keep the skepticism when using genomic data and using the common methodologies.

References

Aanensen D, Spratt B (2005) The multilocus sequence typing network: mlst. net. Nucleic Acids Res 33:W728–W733

Abt B, Foster B, Others (2010) Complete genome sequence of cellulomonas flavigena type strain (134). Stand Genomic Sci 3:15–25

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25:195–203

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci 108:17486–17491

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36:D25–D30

Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database GOLD: a monitor of genome projects world-wide. Nucleic Acids Res 29:126–127

Beyer WA, Stein M, Smith TF, Ulam S (1974) A molecular sequence metric and evolutionary trees. Math Biosci 19:9–25

Bohlin J, Snipen L, Cloeckaert A, Lagesen K, Ussery D, Kristoffersen AB, Godfroid J (2010) Genomic comparisons of *Brucella* spp and closely related bacteria using base compositional and proteome based methods. BMC Evol Biol 10:249

Brenner DJ, Krieg NR, Staley JT, Garrity GM, Boone DR, Vos P, Goodfellow M, Rainey FA, Schleifer KH (2005a) Bergeys manual® of systematic bacteriology, vol 2. Springer, Boston

Brenner DJ, Staley JT, Krieg NR (2005b) Classification of procaryotic organisms and the concept of bacterial speciation. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual® of systematic bacteriology. Springer, Boston, pp 27–32

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: Models and estimation procedures. Evolution 32:550–570

Chain P, Grafham D, Fulton R, Fitzgerald M (2009) Genome project standards in a new era of sequencing. Science 326:4–5

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

Delcher A, Kasif S, Fleischmann R, Peterson J, White O, Salzberg S (1999) Alignment of whole genomes. Nucleic Acids Res 27:2369–2376

DeWall MT, Cheng DW (2011) The minimal genome–a metabolic and environmental comparison. Brief Funct Genomic Proteomic 105:312–315

Doolittle W (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

Fraser CM, Gocayne JD, Others (1995) The minimal gene complement of Mycoplasma genitalium. Science 270:397–403

Galperin MY (2006) The minimal genome keeps growing. Environ Microbiol 84:569–573

Garrity GM, Lilburn GT, Cole JR, Harrison SH, Euzeby J, Tindall BJ (2007) Introduction to the taxonomic outline of bacteria and archaea (TOBA) Release 7.7. The Taxonomic Outline of Bacteria and Archaea. http://www.taxonomicoutline.org/index.php/toba/article/download/190/223. Accessed 23 Feb 2012

Gibbons N, Murray R (1978) Proposals concerning the higher taxa of bacteria. Int J Syst Evol Microbiol 28:1–6

Goris J, Konstantinidis K, Klappenbach J, Coenye T, Vandamme P, Tiedje J (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91

Harayama S, Kasai H (2006) Bacterial phylogeny reconstruction from molecular sequences. In: Stackebrandt E (ed) Molecular identification, systematics, and population structure of prokaryotes. Springer, Berlin/New York, pp 105–140

Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. Protein Sci 1:409–417

Konstantinidis K, Tiedje J (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA 102:2567–2572

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108

Lagesen K, Ussery DW, Wassenaar TM (2010) Genome update: the 1000th genome–a cautionary tale. Microbiology 156:603–608

Lapage S, Sneath P, Lessel E, Skerman V, Seeliger H, Clark W (1992) International code of nomenclature of bacteria: bacteriological code, 1990 revision. American Society of Microbiology, Washington, DC

Lapierre P, Gogarten J (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25:107–110

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Liolios K, Chen I-MA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC (2010) The genomes on line database GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 38:D346–D354

Maiden M, Bygraves J, Feil E, Morelli G, Russell J, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant D, Others (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

McCutcheon JP, McDonald BR, Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet 57:e1000565

Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158:24–36

Moya A, Gil R, Latorre A, Peretó J, Pilar Garcillán-Barcia M, de la Cruz F (2009) Toward minimal bacterial cells: evolution vs. design. FEMS Microbiol Rev 331:225–235

Murray RGE (1989) The higher taxa, or, a place for everything. . .? In: Williams ST, Sharpe ME, Holt JG (eds) Bergey's manual of systematic bacteriology, vol 4, 1st edn. Williams & Wilkins, Baltimore, pp 2329–2332

Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A (2006) A small microbial genome: the end of a long symbiotic relationship? Science 314:312–313

Qi J, Luo H, Hao B (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 32:W45–W47

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci 106:19126–19131

Sanford RA, Cole JR, Tiedje JM, Al SET, Icrobiol APPLENM (2002) Characterization and description of *Anaeromyxobacter dehalogenans* gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. Appl Environ Microbiol 68:893–900

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think! Genome Biol 12:125

Schleifer KH (2009) Classification of Bacteria and Archaea: past, present and future. Syst Appl Microbiol 32:533–542

Schneiker S, Perlova O, Others (2007) Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol 25:1281–1289

Smith T, Waterman M (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Snipen L, Ussery DW (2010) Standard operating procedure for computing pangenome trees. Stand Genomic Sci 2:135–141

Stackebrandt E (2006) Exciting times: the challenge to be a bacterial systematist. Molecular Identification, Systematics, and Population Structure of Prokaryotes 1–21

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6:938–947

Tindall BJ, Kämpfer P, Euzéby JP, Oren A (2001) Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. Int J Syst Evol Microbiol 56:2715–2720

Ussery D, Wassenaar T, Borini S (2009) Computing for comparative microbial genomics: bioinformatics for microbiologists. Springer, London

Ward JH, Jr. (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 48:236–244

Waters E, Hohn MJ, Others (2003) The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci USA 100:12984–12988

Wayne L, Brenner D, Colwell R, Grimont P, Kandler O, Krichevsky M, Moore L, Moore W, Murray R, Stackebrandt E, Others (1987) International committee on systematic bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol 37:463–464

Whittaker R, Margulis L (1978) Protist classification and the kingdoms of organisms. Biosystems 10:3–18

Woese C, Fox G (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese C, Kandler O, Wheelis M (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Xu Z, Hao B (2009) CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 37:174–178

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 Evol Microbiol 46: 506–511

Zuckerkandl E, Pauling LB, Kasha M, Pullman B (1962) Molecular disease, evolution, and genetic heterogeneity. Academic, New York, pp 189–225