UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉTUDE PHYLOGÉNÉTIQUE DES γ-PROTÉOBACTÉRIES BASÉE SUR LES GÈNES 16S ARNr ET DES GÈNES CODANT POUR DES PROTÉINES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN BIOLOGIE

PAR

HOON-YONG LEE

NOVEMBRE 2006

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

PHYLOGENETIC ANALYSIS OF γ-PROTEOBACTÉRIA BASED ON 16S rRNA GENES AND PROTEIN-CODING GENES

THESIS

PRESENTED

AS A PARTIAL REQUIREMENT

FOR THE MASTER IN BIOLOGY

BY

HOON-YONG LEE

NOVEMBER 2006

RÉSUMÉ

Les bactéries se prêtent très bien aux analyses phylogénétiques en raison du fait que plusieurs génomes ont été entièrement séquencés – et leur nombre continue d'augmenter – et leurs génomes sont petits et ne contiennent que peu d'ADN répétitif en contraste avec les génomes des eucaryotes.

La molécule d'ARNr 16S, l'ARN composant de la petite sous-unité du ribosome, a été établie comme la macromolécule de choix pour les analyses phylogénétiques et l'identification des espèces bactériennes. En dépit de sa position centrale, l'utilisation de l'ARNr 16S n'est pas sans inconvénients. En effet, l'évolution d'un seul gène n'est pas nécessairement représentative de l'évolution d'un génome entier, des hétérogénéités entre des copies du gène d'ARNr 16S - des allèles - ne sont pas exceptionnelles, et des copies paralogues peuvent suggérer des phylogénies différentes. De plus, au cours des dernières années, des analyses phylogénétiques basées sur des séquences codant pour des protéines conservées - telles des gènes de ménage (house-keeping genes) - ont donné des résultats discordants.

Parmi les bactéries, les γ-protéobactéries avec 33 génomes entièrement séquencés sont les plus étudiées, certainement en raison de leurs importances cliniques et biologiques.

L'objectif de mon travail a été, d'une part d'étudier l'hétérogénéité des séquences des allèles de l'ARNr 16S chez 33 γ-protéobactéries et de déterminer si un seul allèle par bactérie est suffisant pour établir une phylogénie, d'autre part de construire les arbres phylogénétiques de ces 33 γ-protéobactéries à partir de l'analyse comparée des séquences de certains gènes de ménage à savoir: l'adénylate kinase (*adk*), la shikimate déshydrogénase (*aroE*), la glucose-6-phosphate déshydrogénase (*gdh*) et de leurs séquences concaténées. Ces phylogénies seront comparées entre elles et aussi à celle établie à partir du gène de l'ARNr 16S.

Les analyses phylogénétiques ont démontré que la plupart des séquences alléliques d'une même souche sont identiques ou quasi identiques et sont regroupées à l'intérieur d'un même genre et d'une même espèce. À cause de leurs homogénéités, un seul allèle du gène de

ACKNOWLEDGEMENTS

I am very grateful to the many people who have assisted me in the completion of this project. Without their great assistance, I would not have been able to complete this project successfully.

First and foremost, I would like to thank Dr. Jean-Charles Côté for giving me the opportunity to pursue a master's degree in his laboratory and for being an outstanding director. I especially appreciated his patience, guidance and support during my graduate studies. His encouraging attitude helped me keep the objectives of this study clear in my mind and his decisive guidance has made possible the publication of two scientific papers.

I would like to thank Prof. Lucie Lamontagne for agreeing to be my co-director. I also wish to express my great appreciation to the other two members of my Committee, Prof. Cédric Chauve and Prof. Vladimir Makarenkov. All are thanked for their thoughtful suggestions during the course of this study. I would like to thank Prof. Anne Bergeron and Prof. Josée Harel for their thoughtful review of this work.

I wish to express my sincere gratitude to Prof. Young Sup Chung for his support and encouragement during my study in Graduate School. Whenever I encountered problems in the course of my study, or anything else, he was always there to give his best support.

I would like to thank Brahim Soufiane, Dr. Dong Xu, Mouna Cheikh-Rouhou, Sabarimatou Yakoubou and Suzanne Fréchette, my colleagues here in the Laboratory of Molecular and Applied Microbiology at the Agriculture and Agri-Food Canada, Horticulture Research and Development Centre, in Saint-Jean-sur-Richelieu for their scientific advices and friendship. I also wish to acknowledge that this research was conducted entirely at the Horticulture Research and Development Centre.

Finally, I would like to express my deep appreciation to my parents, brothers and sisters for their constant love, support and encouragement throughout my life. I also would like to thank my nephews and friends for easing my stress and for their encouragements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RÉSUMÉ

Les bactéries se prêtent très bien aux analyses phylogénétiques en raison du fait que plusieurs génomes ont été entièrement séquencés – et leur nombre continue d'augmenter – et leurs génomes sont petits et ne contiennent que peu d'ADN répétitif en contraste avec les génomes des eucaryotes.

La molécule d'ARNr 16S, l'ARN composant de la petite sous-unité du ribosome, a été établie comme la macromolécule de choix pour les analyses phylogénétiques et l'identification des espèces bactériennes. En dépit de sa position centrale, l'utilisation de l'ARNr 16S n'est pas sans inconvénients. En effet, l'évolution d'un seul gène n'est pas nécessairement représentative de l'évolution d'un génome entier, des hétérogénéités entre des copies du gène d'ARNr 16S - des allèles - ne sont pas exceptionnelles, et des copies paralogues peuvent suggérer des phylogénies différentes. De plus, au cours des dernières années, des analyses phylogénétiques basées sur des séquences codant pour des protéines conservées - telles des gènes de ménage (house-keeping genes) - ont donné des résultats discordants.

Parmi les bactéries, les γ-protéobactéries avec 33 génomes entièrement séquencés sont les plus étudiées, certainement en raison de leurs importances cliniques et biologiques.

L'objectif de mon travail a été, d'une part d'étudier l'hétérogénéité des séquences des allèles de l'ARNr 16S chez 33 γ-protéobactéries et de déterminer si un seul allèle par bactérie est suffisant pour établir une phylogénie, d'autre part de construire les arbres phylogénétiques de ces 33 γ-protéobactéries à partir de l'analyse comparée des séquences de certains gènes de ménage à savoir: l'adénylate kinase (*adk*), la shikimate déshydrogénase (*aroE*), la glucose-6-phosphate déshydrogénase (*gdh*) et de leurs séquences concaténées. Ces phylogénies seront comparées entre elles et aussi à celle établie à partir du gène de l'ARNr 16S.

Les analyses phylogénétiques ont démontré que la plupart des séquences alléliques d'une même souche sont identiques ou quasi identiques et sont regroupées à l'intérieur d'un même genre et d'une même espèce. À cause de leurs homogénéités, un seul allèle du gène de

l'ARNr 16S est suffisant pour construire la phylogénie des γ-protéobactéries au niveau du genre et de l'espèce.

Les arbres phylogénétiques construits à partir de chacun des trois gènes de ménage - *adk*, *aroE* et *gdh* - et des séquences concaténées des trois gènes sont, en général, similaires à l'arbre construit à partir du gène de l'ARNr 16S, aux niveaux de la famille, du genre, de l'espèce et de la souche, avec certaines exceptions. Les gènes de ménage, cependant, montrent un plus haut taux de substitutions nucléotidiques que le gène de l'ARNr 16S. De plus, parce que les trois gènes de ménage ont chacun un pourcentage plus bas de similarité de séquences que le gène de l'ARNr 16S, ils ont montré une meilleure différenciation des neuf souches de bactéries entériques *Escherichia-Shigella-Salmonella.* Ils peuvent potentiellement révéler des relations sur des branches plus profondes d'un arbre phylogénétique que ne le peut le gène de l'ARNr 16S. De plus, puisque ces gènes de ménage sont utilisés dans les typages de séquence multi-locus "multilocus sequence typing (MLST)", nous pouvons nous attendre à ce que le nombre de séquences disponibles dans le domaine public pour ces trois gènes de ménage augmente rapidement les rendant ainsi d'autant plus utiles pour complémenter le gène de l'ARNr 16S dans des études phylogénétiques, ou pour certaines autres analyses phylogénétiques très ciblées.

Mots clés- γ-protéobactéries, gènes de ménage, phylogénie, ARNr 16S

SUMMARY

Bacteria offer the most opportunities for phylogenetic analyses, because many genomes have been fully sequenced – and their number is rapidly increasing - and the genomes are small and contain little repetitive sequence.

The 16S rRNA, the RNA component of the ribosome small subunit (SSU rRNA), has been established as the macromolecule of choice for single-gene phylogenetic analyses and identifications of bacterial species. Despite its predominance in phylogenetic analyses, the use of 16S rRNA gene sequences is not without drawbacks. The evolution of a single gene may not represent the evolution of an entire genome, heterogeneities between 16S rRNA copies, alleles, are not a rare occurrence, and paralogous copies may possibly infer different phylogenies. In addition, in recent years, bacterial phylogenetic analyses inferred from various conserved proteins (such as house-keeping genes) have given incongruent phylogenetic results.

Among bacteria, the γ-proteobacteria are the most extensively studied and contain the highest number of fully sequenced genomes – 33 –, certainly because most of them are bacteria of clinical or biological importance.

The purpose of my study was two-fold: first, to study the heterogeneity of 16S rRNA allelic sequences in 33 γ-proteobacteria and determine whether a single allele could be sufficient for inferring phylogenies; second, to construct phylogenies in these 33 γ-proteobacteria inferred from nucleotide sequence comparisons of the house-keeping genes adenylate kinase (*adk*), shikimate dehydrogenase (*aroE*), glucose-6-phosphate dehydrogenase (*gdh*) and their concatenated sequences. These phylogenies were compared to each other and further compared to a 16S rRNA gene-inferred phylogeny.

The phylogenetic analysis reveals that most 16S rRNA allelic sequences from same strain are identical or nearly identical and 16S rRNA allelic sequences are clustered within genera and species. Because of their homogeneity, a single 16S rRNA allelic sequence is sufficient to reconstruct the phylogeny of γ-proteobacteria at the genera and species level.

Phylogenetic trees inferred from each of the three house-keeping genes *adk*, *aroE* and *gdh*, and of the concatenated sequences of all three genes, are, in general, similar to a 16S rRNA gene-inferred tree, at the family, genus, species and strain levels. The house-keeping genes, however, show a higher rate of nucleotide sequence substitutions than 16S rRNA gene. In addition, because the three house-keeping genes have a lower percentage of gene sequence similarity than the 16S rRNA gene, they showed a better resolution for the nine core *Escherichia-Shigella-Salmonella* enterics. They can possibly probe deeper branches of a phylogenetic tree than the 16S rRNA gene. In addition, since these house-keeping genes are used in multilocus sequence typing (MLST), it is expected that the number of sequences publicly available for these three house-keeping genes will be getting bigger over time proving them very useful either at complementing 16S rRNA-inferred phylogenies or for specific, targeted, phylogenetic analysis.

INTRODUCTION

Over the last few years, the amount of bacterial DNA and protein sequence information available has grown as a result of the increasing development in sequencing and cloning techniques. Several bacterial genomes have been sequenced and a large number of DNA sequences have been recorded. These data are freely available, and most can be downloaded and analysed using the Internet. Gene sequences are collected into the International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/), the European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/), and the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) in the USA.

In the last 20 years, the rapid advances in DNA sequencing technology have led to major changes in the way that prokaryotes are classified. Sequence analyses of highly conserved regions in the bacterial genome, such as the small subunit 16S ribosomal RNA genes are now widely used for phylogenetic analysis and species identification. This has also led to the discovery of new prokaryotic species.

The bacteria offer the most opportunities for phylogenetic analyses, because many genomes are available – and the number is rapidly increasing - and the genomes are small and contain little repetitive sequence (Lerat *et al.*, 2003). Currently, of all bacterial groups, the γ-proteobacteria are the most intensively studied and contain the highest number of fully

sequenced genomes, certainly because most of them are bacteria of clinical or biological importance. By mid-2004, the genomes had been fully sequenced for at least 33 γ-proteobacteria. In this study, therefore, these 33 γ-proteobacterial species and strains were selected for phylogenetic analysis.

First, the phylogenetic analysis of the 33 γ-proteobacteria was inferred through a series of steps as summarized in Fig. 1.1. Here, we used the nucleotide sequences of the 16S rRNA gene. We analysed the heterogeneity in 16S ribosomal RNA (16S rRNA) gene sequences of the 33 γ-proteobacteria.

Second, the phylogenetic analysis of the 33 γ-proteobacteria was also inferred through a series of steps as summarized in Fig. 1.2. Here, we used the nucleotide sequences of three house-keeping genes, namely: adenylate kinase (*adk*); shikimate dehydrogenase (*aroE*); glucose-6-phosphate dehydrogenase (*gdh*); and the concatenated *adk*, *aroE* and *gdh* gene sequences to reconstruct the phylogenies for the 33 γ-proteobacteria. These phylogenies were compared with the phylogeny inferred from the 16S rRNA gene sequences.

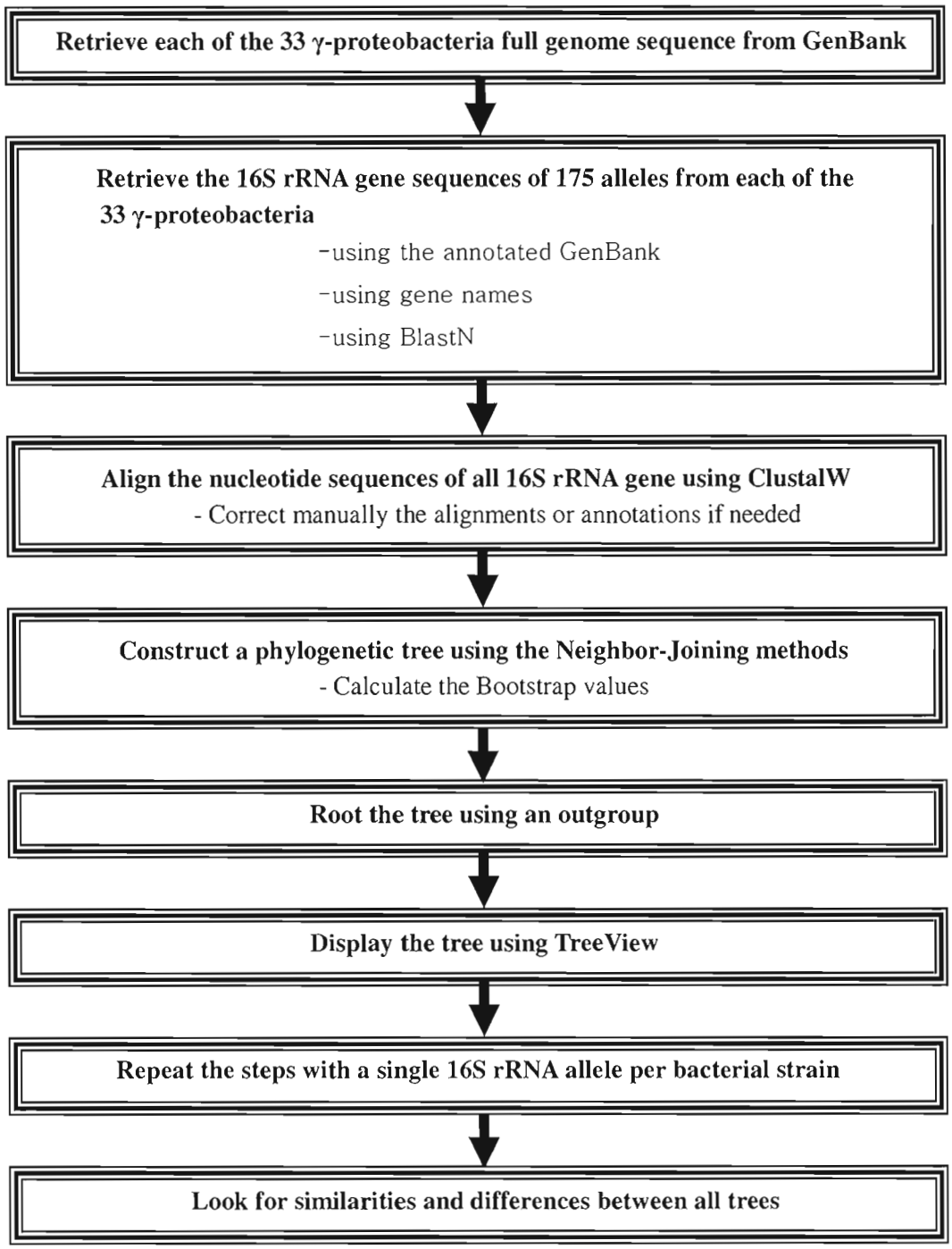


Figure 1.1 -Flow chart of the steps to follow for phylogenetic analysis based on 16S rRNA

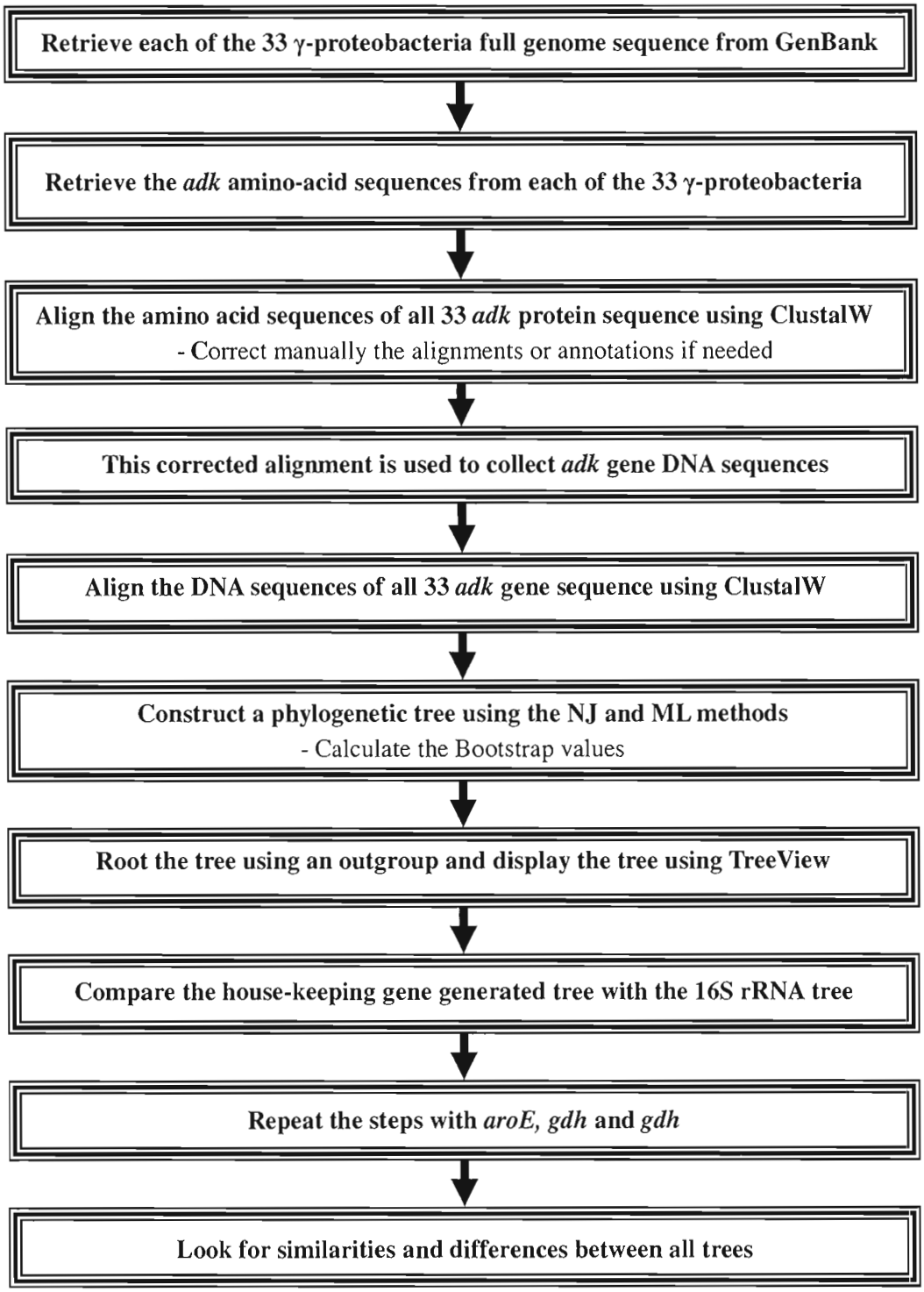


Figure 1.2 -Flow chart of the steps to follow for phylogenetic analysis based on house-keeping genes

We present here the phylogenetic analysis of 33 γ-proteobacteria inferred from nucleotide sequences comparison of the 16S rRNA gene. Two phylogenies are presented, a first one inferred from nucleotide sequence comparisons of all 175 16S rRNA alleles from all 33 γ-proteobacteria, a second one inferred from nucleotide sequence comparisons of a single 16S rRNA allele from each of the 33 γ-proteobacteria. Both phylogenies are compared. Congruencies and differences are discussed (Chapter I).

We also present here the phylogenetic analysis of 33 γ-proteobacteria inferred from nucleotide sequences comparison of three house-keeping genes, *adk*, *aroE* and *gdh*. Five phylogenies are presented, inferred from nucleotide sequence comparisons of the 33 *adk*; 32 *aroE*; 31 *gdh*; the concatenated 31 *adk*, *aroE* and *gdh* sequences; and 33 16S rRNA alleles, respectively. All phylogenies are compared. Congruencies and differences are discussed (Chapter II).

16S ribosomal RNA gene (16S rRNA)

Most prokaryotes have three different ribosomal RNAs, the 5S, 16S and 23S ribosomal RNA, present in multiple copies. The 5S rRNA and 23S rRNA are the RNA component of the ribosome large subunit (LSU rRNA), and 16S rRNA is the RNA component of the ribosome small subunit (SSU rRNA). The 5S rRNA has been extensively studied, but it is usually too small for reliable phylogenetic inference. The 16S rRNA and 23S rRNA are sufficiently large to be useful.

The 16S rRNA, the RNA component of the ribosome small subunit, has been established as the macromolecule of choice for single-gene phylogenetic analyses and identifications of species (Lane *et al.*, 1985; Woese, 1987; Woese *et al.*, 1990). The 16S rRNA is an essential component of protein synthesis and is present in all organisms. For phylogeny purposes, 16S rRNA sequences have proven useful because the 16S rRNA gene is highly conserved throughout bacteria and even very distant bacterial species can be compared, and the gene is easy to amplify and sequence using universal primers (Stackebrandt *et al.*, 1991). It is assumed that the homology between 16S rRNA sequences from different bacteria reflects the phylogenetic relationship between these organisms.

However, the use of 16S rRNA sequences for single-gene phylogenetic analyses has some limits. The copy number of 16S rRNA genes per bacterial genome ranges between 1 and 15 (Klappenbach *et al.*, 2001). For example, there are seven *rrn* operons in *Escherichia coli* and *Salmonella typhimurium* (Hill and Harnish, 1981); and 9 or 10 *rrn* operons in *Bacillus subtilis* (Loughney *et al.*, 1983). Heterogeneities between *rrn* operons are not a rare occurrence and paralogous copies may infer different phylogenies. In addition, 16S rRNA gene may undergo occasional lateral gene transfer (LGT) or recombination (Ueda *et al.*, 1999; Yap *et al.*, 1999). Finally, 16S rRNA gene sequences may not be adequate to analyse phylogenetic relationships between closely related species because its gene is highly conserved (Gürtler and Stanisich, 1996; Kolbert and Persing, 1999).

In this study, we have studied the heterogeneity in 16S rRNA gene sequences of 175 alleles from the 33 γ-proteobacteria to address these problems (*see* Chapter I). A phylogenetic tree with all 175 alleles is presented and clustering at the genus, species and strain levels are discussed.

Molecular phylogenetic studies with house-keeping genes

Despite the success of rRNA microbial taxonomy, the evolutionary relationships between major groups of prokaryotes is still unclear because phylogenetic analysis of single gene sequences is insufficient to resolve deep branches (Fitz-Gibbon and House, 1999).

Several genes have been proposed to complement rRNA genes in bacterial phylogenetic analysis. They include *atpD* (beta subunit of the membrane ATP synthase; Ludwig *et al.*, 1993), *gyrB* (subunit B protein of DNA gyrase, topoisomerase type II; Yamamoto and Harayama, 1995), *infB* (translation initiation factor 2; Hedegaard *et al.*, 1999), *recA* (RecA protein; Gaunt *et al.*, 2001), and *rpoB* (RNA polymerase beta subunit; Mollet *et al.*, 1997) to name a few. These genes are not transmitted horizontally, their evolutionary rate is higher than the one of 16S rRNA, and they are present in most bacteria.

In recent years, multilocus sequence typing (MLST) (Maiden *et al.*, 1998) has been developed as a molecular typing method. It is similar to multilocus enzyme electrophoresis (MLEE), but characterizes the alleles present at multiple house-keeping genes. Owing to the

development of MLST within species, large amounts of sequence data have become available for a number of pathogens. Maiden and colleagues (1998) have selected six house-keeping genes for MLST of *Neisseria*. The genes selected are *abcZ* (encoding the putative ABC transporter), *adk* (adenylate kinase), *aroE* (shikimate dehydrogenase), *gdh* (glucose-6-phosphate dehydrogenase), *pdhC* (pyruvate dehydrogenase subunit), and *pgm* (phosphoglucomutase), on the basis that they are unlinked, variations accumulate slowly and are likely to be selectively neutral, and they are present in many bacteria. With the increasing number of research groups using MLST as a typing method, the number of sequences for these six house-keeping genes is rapidly increasing for a wide variety of bacteria. It is worth addressing whether these house-keeping genes could be used to complement 16S rRNA gene to infer phylogenies.

House-keeping genes are a class of highly expressed, highly conversed protein encoding genes that show a high degree of codon bias. They evolve more slowly than typical protein encoding genes, but more rapidly than rRNA genes (Bruns *et al.* 1991). Thus, house-keeping genes are often used to construct gene trees of closely related taxa (Lawrence *et al.*, 1991). Each house-keeping gene can be analysed separately and the resulting phylogenies can be compared to see if they support or conflict with each other.

Of the six house-keeping genes used in MLST, only three, *adk*, *aroE* and *gdh*, have been studied here. The three other house-keeping genes used in MLST, *abcZ*, *pdhC* and *pgm* were not universally distributed in our 33 γ-proteobacteria and thus proved unfit for our

specific phylogenetic analyses. (*see* Chapter II).

Proteobacteria

Members in the Division Proteobacteria, within the Domain Bacteria, comprise at present the largest and phenotypically most diverse phylogenetic lineage. The proteobacteria constitute one of the largest divisions within prokaryotes and form the vast majority of the Gram-negative bacteria. In 1988, Stackebrandt *et al.* named the proteobacteria after the Greek god Proteus, who could have many different shapes, because of the great diversity of forms found in it. This group of organisms, often referred to as 'purple bacteria and relatives', encompasses bacteria with a great diversity of phenotypes, physiological attributes, and habitats (Stackebrandt *et al.*, 1988; Gupta, 2000). The proteobacteria contain more than 460 genera and encompass a major proportion of the known Gram-negative bacteria. As the proteobacteria include a large number of known human, animal and plant pathogens, the group is of great biological significance. Proteobacteria have been classified based on homology of 16S ribosomal RNA or by hybridization of ribosomal DNA with 16S and 23S ribosomal RNA (Fox *et al.*, 1980; Woese *et al.*, 1985a; Woese, 1987; De Ley, 1992). The division of proteobacteria has been subdivided into five major classes, α- (Woese *et al.*, 1984a), β- (Woese *et al.*, 1984b), γ- (Woese *et al.*, 1985b), δ- and ε- (Fig. 2)

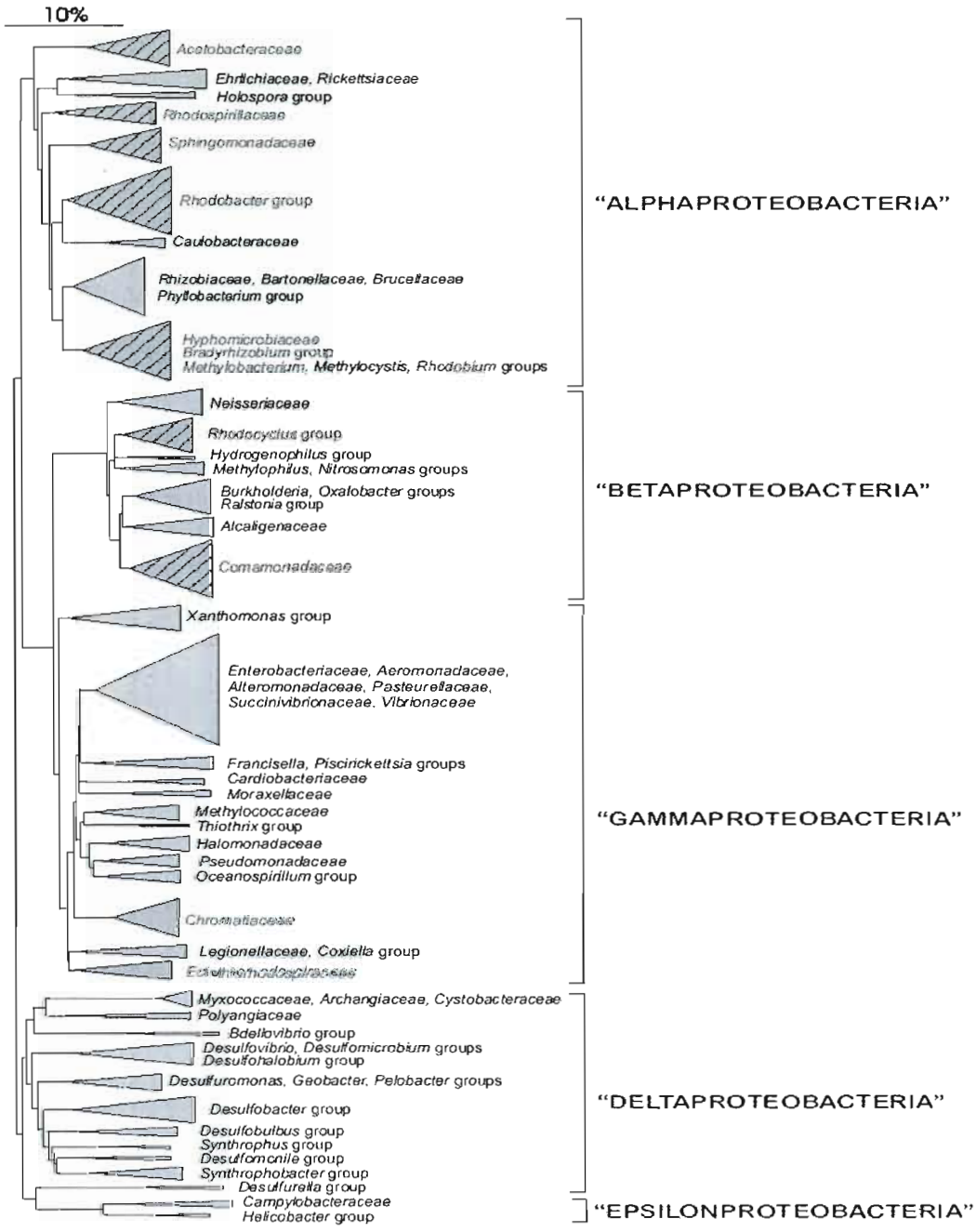


Figure 2 -Phylogenetic tree of the proteobacteria based on 16S rDNA sequences of the type strains of the proteobacterial genera

Fig. 2 is a simplified phylogenetic tree of the proteobacteria based on the nearly complete 16S rDNA sequences of the type strains of the type species of the majority of proteobacterial genera. The names of the families and major groups are also noted. The δ- and ε-proteobacteria classes form the deeper branches of the division; the α-proteobacteria are also clearly separated, whereas the closer relationship between the β- and γ-proteobacteria lineages may indicate the common origin of the latter groups (Ludwig and Klenk, 2001); (adapted from http://141.150.157.117:8080/prokPUB/chaphtm/379/02_00.htm).

γ (Gamma)-proteobacteria

The γ-proteobacteria, classified on the basis of sequence signatures structural differences in the SSU rRNA (Woese, 1987), include free-living and commensal species, intracellular symbionts, and human-, animal- and plant pathogens.

Most 16S rDNA trees show that the members of the γ-proteobacteria represent a monophyletic group which includes in fact also the β -proteobacteria as a major line of descent. The γ-proteobacteria are the largest proteobacterial group which comprise at least 200 genera and 750 species, including several families of utmost biological importance, most notably the *Coxiellaceae*, *Enterobacteriaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae*, *Shewanellaceae* and *Xanthomonadaceae*.

The γ-proteobacteria contain the photosynthetic purple sulfur bacteria together with a great number of familiar chemoorganotrophic bacterial groups, such as the *Enterobacteriaceae*, *Legionellaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae*. The class includes some important human and animal pathogens. Note that the *Enterobacteriaceae* family has been known since 1937, as a classical phenotypic group (Rahn, 1937). The *Enterobacteriaceae* family is fully supported by modern molecular taxonomy. However, the *Pseudomonadaceae* family differs in that it turned out to be phylogenetically extremely heterogeneous, because its members are scattered over the α-proteobacteria, β-proteobacteria and γ-proteobacteria. The genus *Pseudomonas* is presently restricted to all species phylogenetically which are related to its type species, *Pseudomonas aeruginosa*, a member of the γ-proteobacteria. All of the other *pseudomonads*, however, which belong to the α- and β-classes, have been allocated to new genera such as *Brevundimonas*, *Sphingomonas*, *Comamonas*, *Burkholderia*, *Ralstonia*, etc.

In this study, the 33 γ-proteobacterial species and strains which have been analysed here are listed in Tables 1 and 2. These include one ***Coxiellaceae*** species, 17 ***Enterobacteriaceae*** strains (encompassing 9 species), three ***Pasteurellaceae*** species, three ***Pseudomonadaceae*** species, one ***Shewanellaceae*** species, four ***Vibrionaceae*** strains (three species), and four ***Xanthomonadaceae*** strains (three species). An ε-proteobacterium, *Helicobacter pylori*, was also included as an outgroup.

Coxiellaceae. A family of Gram-negative bacteria in the order *Legionellales*, includes the genus *Coxiella*. The clinically important *Legionellae* occurs in surface water, mud, thermally polluted lakes and streams. In addition, it may enter the human respiratory tract when water is aerosolised in showers and through air-conditioning systems. *Legionella pneumophila* is a pathogen for humans causing pneumonia. An obligate intracellular bacterial parasite of small free-living amoebae (previously classified as *Sarcobium lyticum*; Drozanski, 1991) belongs also to the genus *Legionella* (Hookey *et al.*, 1996).

The genus *Coxiella* belongs to the same phylogenetic lineage as the *Legionellae*. Although phylogenetically distinct, *Coxiella* shares similarities in their parasitic lifestyle. *Coxiella burnetii*, an obligate parasitic bacterium grows preferentially in the vacuoles of the host cells, causes Q-fever, a pneumonia-like infection that is transmitted among animals by insect bites (e.g., ticks). *Coxiella burnetii* occasionally causes disease in humans.

Enterobacteriaceae. The *Enterobacteriaceae* family is the best studied group of microorganisms. Their popularity is of medical and economic importance, because of the ease of their isolation and cultivation, rapid generation time, and the ease with which they can be genetically manipulated. *Enterobacteriaceae* are distributed worldwide. They are found in water and soil and as normal intestinal flora in humans and many animals. They live saprophytically, as symbionts, epiphytes, and parasites. Their hosts include animals ranging from insects to humans, and fruits, vegetables, grains, flowering plants, and trees.

Escherichia coli (Fig. 3) is considered the most thoroughly studied bacterial species. The *Enterobacteriaceae* family includes *Escherichia coli*, and a multiple of inhabitants of the intestinal tract of warm-blooded animals (e.g., *Salmonella* and *Shigella*). The *Enterobacteriaceae* comprise a relatively homogeneous phylogenetic group within this large cluster. They are mostly facultative anaerobic carbohydrate-degrading microorganisms; and some perform a mixed acid fermentation, whereas others carry out the butanediol fermentation. The enterics also comprise plant pathogenic bacteria, the plague-causing *Yersinia pestis*, and *Photorhabdus*, symbionts of entomopathogenic nematodes.

Enterobacteriaceae live in the intestinal lumen. An interesting and somewhat distant member of the enterics is *Buchnera aphidicola*, an endosymbiont of aphids (Baumann *et al.*, 1995). The symbiotic association between aphids and *Buchnera* seems to be obligate and mutualistic: *Buchnera* synthesises tryptophan, cysteine and methionine. *Buchnera* supplies these essential amino acids to the aphid host. A parallel evolution of *Buchnera* and aphids seems to have occurred, and Baumann *et al.* (1998) estimated the origin of this symbiotic association at 200–250 million years ago. The correlation of sequence diversity of 16S rDNA of symbionts with the age of their hosts has led to the calibration of the molecular clock of 16S rDNA in recent organisms (Moran *et al.*, 1993), and is assumed to generate 1% sequence divergence within 25–50 million years (Stackebrandt, 1995). In addition, the endosymbionts of carpenter ants constitute a distinct taxonomic group within the γ-proteobacteria and are phylogenetically closely related to *Buchnera* and symbionts of tsetse flies. Comparison of the phylogenetic trees of the bacterial

endosymbionts and their host species suggests a highly synchronous cospeciation process of both partners (Sauer *et al.*, 2000).

Pasteurellaceae. The family *Pasteurellaceae* contain several pathogens of vertebrates. They include the following major genera: *Pasteurella*, *Haemophilus* and *Actinobacillus*. Organisms of the genus *Pasteurella* are Gram-negative, non-motile, facultatively anaerobic coccobacilli. The *Pasteurella* is the oldest recognized genus of the family *Pasteurellaceae* consisting of several species. Species of this genus are found in both animals and humans. The *Haemophilus* genus represents a large group of Gram-negative rods that can grow on blood agar. *Haemophilus influenzae*, because of its small genome, became the first free-living organism whose entire genome was sequenced (Fleischmann *et al.*, 1995).

Pseudomonadaceae. Two decades ago, the family *Pseudomonadaceae* contained an extremely heterogeneous group of aerobic, rod-shaped and mostly polarly flagellated bacteria (Palleroni, 1984). It phylogenetically spread over the α-proteobacteria, β-proteobacteria and γ-proteobacteria. At present, the family is restricted to close 90 *Pseudomonas* species. A separate lineage within the γ-proteobacteria is formed around the type species *Pseudomonas aeruginosa*. The *Pseudomonads* of the α-proteobacteria and β-proteobacteria have been transferred to other genera (Willems *et al.*, 1991; Kersters *et al.*, 1996; Anzai *et al.*, 2000; *Caulobacteraceae*, *Alcaligenaceae*, *Comamonadaceae*, *Burkholderia*, *Oxalobacter* and Related Groups). Some authentic *Pseudomonas* species

Legend

Figure 3: *Escherichia coli* (Colorized low-temperature electron micrograph of a cluster of bacteria. Individual bacteria in this photo are oblong and colored brown).

Taken from the "Tree of Life Web Project. 2006. Proteobacteria. Version 10 March 2006 (temporary)". http://tolweb.org/Proteobacteria/2302/2006.03.10 in The Tree of Life Web Project, http://tolweb.org

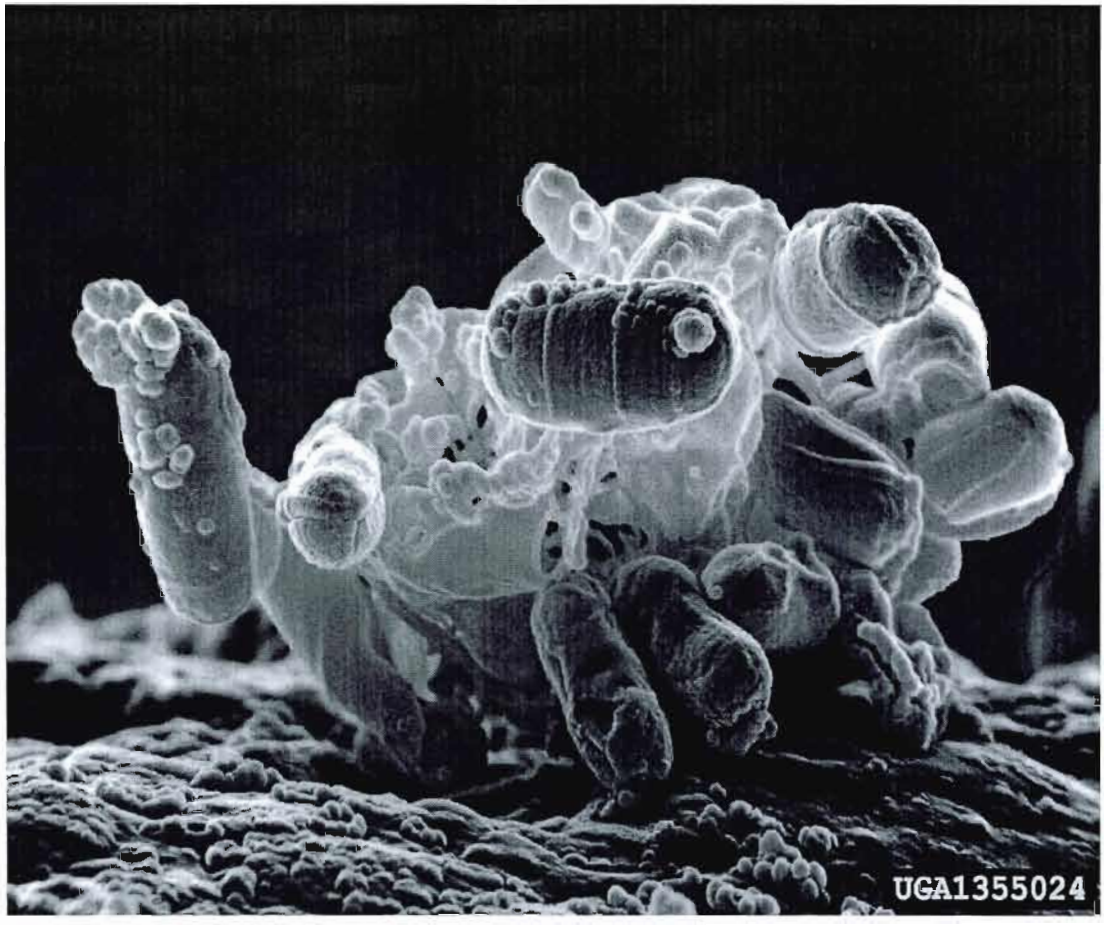


Figure 3 -*Escherichia coli*

(e.g., *P. fluorescens*) produce fluorescent pigments. *Pseudomonads* can grow on a great variety of organic substrates, including aromatic hydrocarbons, because some of them possess efficient oxygenases. Such bacteria play key roles in the purification of wastewater and clean-up of oil spills. Some of the *Pseudomonas* species (e.g., *P. syringae*) are plant pathogens, whereas *P. aeruginosa* and some other fluorescent *pseudomonads* can be involved in serious nosocomial infections. The free-living nitrogen fixers of the genera *Azotobacter* and *Azomonas* belong also to this phylogenetic lineage, together with the cellulose-degrading *Cellvibrio* and a few other related genera.

Shewanellaceae. The family *Shewanellaceae*, Gram-negative, rod-shaped bacteria, is a member of the γ-proteobacteria. The genus *Shewanella* (MacDonell and Colwell, 1985), facultatively anaerobic proteobacteria, associated with aquatic habitats. During the last decade, the organisms of this genus have received a significant amount of attention. This is because of the important roles in co-metabolic bioremediation of halogenated organic pollutants (Petrovskis *et al.*, 1994), destructive souring of crude petroleum (Semple and Westlake, 1987) and the dissimilatory reduction of magnesium and iron oxides (Myers and Nealson, 1988).

Vibrionaceae. *Vibrionaceae* are facultative anaerobic inhabitants of brackish, estuarine and pelagic waters and sediments, and form the dominant culturable microflora in the gut of molluscs, shrimps and fish. This family harbors several pathogens (e.g., *Vibrio cholerae*, the causal agent of cholera) as well as luminous bacteria (e.g., *Photobacterium* and

several *Vibrio* species), occurring free-living in seawater, as well as symbionts in the light organs of many fish and invertebrates (Dunlap and Kita-Tsukamoto, 2001). Bioluminescent bacteria occur also among the genera *Photorhabdus* and *Shewanella.*

Xanthomonadaceae. The plant pathogenic *Xanthomonas* species, *Xylella* (a phytopathogen living in the xylem of various plants) and *Stenotrophomonas*, together with some yellow-pigmented N_2O-producing bacteria isolated from ammonia-supplied biofilters (Finkmann *et al.*, 2000) constitute a clearly separated phylogenetic lineage among the *Chromatibacteria* sensu (Cavalier-Smith, 2002). i.e., the large complex formed by the β-proteobacteria and the γ-proteobacteria. Depending on the treeing algorithms used and the number of rDNA-sequences included, the *Xanthomonads* cluster peripherally linked either to the β-proteobacteria or to the γ-proteobacteria (Fig. 2). The first complete genome sequence published of a plant pathogenic bacterium (Simpson *et al.*, 2000) was that of *Xylella fastidiosa*, a pathogen causing important diseases in citrus trees, grapevines and other plants.

Elaboration of the problematic

Although the 16S rRNA gene has been most used in phylogenetic studies, the evolution of a single gene may not represent the evolution of an entire genome. Because not only 16S rRNA genes may undergo lateral gene transfer or recombination (Reischl *et al.*, 1998; Ueda *et al.*, 1999; Yap *et al.*, 1999; Schouls *et al.*, 2003) but also because 16S rRNA genes are highly conserved, the classification of closely related bacterial species may be

problematic (Gürtler and Stanisich, 1996; Kolbert and Persing, 1999). Another problem derives from the copy number of 16S rRNA genes in bacteria. Heterogeneities between copies are not a rare occurrence and paralogous copies may infer different phylogenies.

In addition, in recent years, careful alignments and phylogenetic analyses of large numbers of conserved proteins (such as house-keeping genes) have given incongruent phylogenetic results (Turner and Young, 2000; Parker *et al.*, 2002). The extent of these incongruences has led to consideration that there may not be a single tree that can be used to represent the history of life. A robust phylogeny based on more genes could then be used to reconstruct genome-scale events, including LGT and rearrangements. Thus, gene trees reconstructed from a single gene may not infer robust phylogenetic relationships among taxa (Li and Graur, 1991).

The study of two or more housekeeping gene sequences has already been recommended for improving the reliability of phylogenetic inference (Yamamoto and Harayama, 1998; Stackebrandt *et al.*, 2002) and DNA sequence comparison of house-keepings has been used for phylogenetic analysis of γ-proteobacteria (Hedegaard *et al.*, 1999; Angen *et al.*, 2003).

In summary, the purpose of my study was two-fold: first, to study the heterogeneity of 16S rRNA allelic sequences in 33 γ-proteobacteria and determine whether a single allele could be sufficient for inferring phylogenies; second, to construct phylogenies in these 33 γ-

proteobacteria inferred from nucleotide sequence comparisons of the house-keeping genes adenylate kinase (*adk*), shikimate dehydrogenase (*aroE*), glucose-6-phosphate dehydrogenase (*gdh*) and their concatenated sequences. These phylogenies were compared to each other and further compared to a 16S rRNA gene-inferred phylogeny.

The following Chapters I and II present first, a "study of the heterogeneity of 16S rRNA genes in γ-proteobacteria - implications for phylogenetic analysis" and second, a "phylogenetic analysis of γ-proteobacteria inferred from nucleotide sequence comparisons of the house-keeping genes *adk*, *aroE* and *gdh* - comparisons with phylogeny inferred from 16S rRNA gene sequences", respectively. Key points are addressed as follows:

- Problems associated with incorrect annotations in GenBank.
- The variation in the number of 16S rRNA alleles in γ-proteobacteria.
- The analysis of the heterogeneity in 16S rRNA gene sequences in γ-proteobacteria.
- The comparison between a phylogenetic tree inferred from 175 16S rDNA allelic sequences and 33 16S rDNA allelic sequences in 33 γ-proteobacteria.
- The distribution of house-keeping genes in γ-proteobacteria, and the absence of some house-keeping genes in some γ-proteobacteria.
- The phylogenetic analysis based on nucleotide sequences (and to a lesser extent, phylogenetic analysis based on amino acid sequences).
- The discriminatory power of the house-keeping genes among the "core" enterics, the percentage sequence similarities, and quantitative data.

- The different percentage of nucleotide sequence divergence for all three house-keeping genes, the concatenated sequences and the 16S rRNA gene and quantitative data.
- Comparisons between phylogenetic tree inferred from house-keeping genes with phylogenetic tree inferred from 16S rRNA genes.

CHAPTER I

STUDY OF THE HETEROGENEITY OF 16S rRNA GENES IN γ-PROTEOBACTERIA: IMPLICATIONS FOR PHYLOGENETIC ANALYSIS

Audrey Olivier, Hoon-Yong Lee and Jean-Charles Côté

JOURNAL OF GENERAL AND APPLIED MICROBIOLOGY, vol. 51, p. 395-405, 2005

Summary

We have analysed the heterogeneity in 16S rRNA gene sequences of 175 alleles from 33 strains covering 23 species and 16 genera of the γ-proteobacteria deposited in GenBank. We show that most allelic sequences from same strain are identical or nearly identical. A phylogenetic analysis reveals that allelic sequences are clustered within genera and species, except for *Escherichia coli* and *Shigella flexneri*, where they can be intertwined. For some γ-proteobacteria, allelic sequences are even sub-clustered at the strain level. We conclude that for the proteobacteria studied, a single 16S rRNA allelic sequence is sufficient to reconstruct the phylogeny of proteobacteria at the genera and species level. Because of their homogeneity, different alleles from same strains would yield similar phylogenies.

1. 1. Introduction

The proteobacteria, formerly known as "purple bacteria and their relatives", are a major class of bacteria (Stackebrandt *et al.*, 1988; Holt *et al.*, 1994; Zinder, 1998; Gupta, 2000). This group of predominantly Gram-negative bacteria contains more than 200 genera with a great diversity of phenotypic and physiological attributes. Proteobacteria have been classified based on homology of 16S ribosomal RNA or by hybridization of ribosomal RNA or DNA with 16S and 23S ribosomal RNA (Fox *et al.*, 1980; Woese, *et al.*, 1985; Woese, 1987). The class has been divided into five major groups, α-, β-, γ-, δ- and ε-. The γ-proteobacteria include several families of utmost biological significance, several human, animal and plant pathogens, most notably the *Enterobacteriaceae*, *Legionellaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae*, and *Xanthomonadaceae*. Because of their biological importance, the genomes of more than 33 γ-proteobacteria have been fully sequenced and are freely available in GenBank.

The 16S ribosomal RNA, the RNA component of the ribosome small subunit, has been established as the macromolecule of choice for single-gene phylogenetic analyses (Lane *et al.*, 1985; Woese, 1987; Woese *et al.*, 1990). The 16S rRNA is an essential component of protein synthesis and is present in all bacteria. Because it is under highly constrainted function, its gene is highly conserved throughout bacteria and even very distant bacterial species can be compared. The gene is easy to amplify and sequence using universal primers (Stackebrandt *et al.*, 1991). It is assumed that the homology between 16S rRNA

sequences from different bacteria reflects the phylogenetic relationship between these organims. However, the copy number of 16S rRNA genes per bacterial genome ranges between 1 and 15 (Klappenbach *et al.*, 2001). Consequently, one of the premise in 16S rRNA sequences inferred phylogenies is that a single 16S rRNA allele, anyone, is representative of its taxon. It has long been assumed that the 16S rRNA allelic sequences within a bacterial isolate were nearly identical and the homology was governed by concerted evolution (Hillis *et al.*, 1991). In the last few years, however, 16S rRNA sequence heterogeneities have been reported for some bacteria at the intraspecies or intrastrain levels. This is the case for *Phormium* yellow leaf phytoplasma (Liefting *et al.*, 1996); *Mycobacterium* strain "X" (Ninet *et al.*, 1996); *Paenibacillus polymyxa* (Nubel *et al.*, 1996) ; *Mycoplasma capricolum* (Pettersson *et al.*, 1998); *Mycobacterium celatum* (Reischl *et al.*, 1998); *Escherichia coli* (Martínez-Murcia *et al.*, 1999); *Paenibacillus turicensis* (Bosshard *et al.*, 2002); and *Veillonella* spp. (Marchandin *et al.*, 2003). In addition, some extensive studies on *Escherichia* spp. and *Salmonella* spp. (Cilia *et al.*, 1996) and in GenBank (Clayton *et al.*, 1995; Coenye and Vandamme, 2003; Acinas, *et al.*, 2004) have clearly shown that intraspecies and intrastrain heterogeneities were more widespread than initially thought, that the level of heterogeneity varies among taxa, and that in some cases, 16S rRNA allelic sequences could differ up to several percents. These raise serious questions regarding the bacterial phylogenies inferred from single 16S rRNA allele per taxon. Should some paralogous alleles of 16S rRNA gene in a given bacterial strain be more divergent than orthologous alleles in another bacterial strain, comparison of 16S rRNA gene nucleotide sequences between different bacterial strains, species or genus could yield erroneous

positioning of the taxon on the phylogenetic tree. We report here the analysis of the heterogeneity in 16S rRNA gene sequences of 175 alleles from 33 strains covering 23 species and 16 genera of the γ-proteobacteria deposited in GenBank. A phylogenetic tree with all 175 alleles is presented and clustering at the genus, species and strain levels is discussed.

1. 2. Materials and Methods

The 33 γ-proteobacteria species and strains analysed in this study are listed in Table 1. They were selected on the basis that their complete genome sequences were freely available through GenBank at the National Center for Biotechnology Information (NCBI) completed microbial genomes database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html; May 2004). These include one *Alteromonadaceae* species, 17 *Enterobacteriaceae* strains (nine species), one *Legionellaceae* species, three *Pasteurellaceae* species, three *Pseudomonadaceae* species, four *Vibrionaceae* strains (three species), four *Xanthomonadaceae* strains (three species). An ε-proteobacterium, *Helicobacter pylori*, was included as an outgroup.

The GenBank accession numbers of the fully sequenced bacterial genomes are given along with the locations of the 16S ribosomal RNA (16S rRNA) genes as annotated in GenBank in Table 1. In the absence of annotation, BLASTn was used to identify in a target bacterial genome the genes and alleles orthologous to a query 16S rRNA allelic sequence

Table 1- List of γ-proteobacteria species and strains used in this study. Complete bacterial genome GenBank accession number, 16S rRNA gene name indicated in GenBank, suggested 16S rRNA gene name for discriminating between alleles when necessary, locations of the start and end points of the allele as indicated in GenBank, and locations after corrections when necessary are presented.

Species	Accession number	Gene name	Gene name in Genebank	Gene name suggested	Locations as annotated	Locations after corrections
Gammaproteobacteria		**16S rRNA**				
Alteromonadaceae						
Shewanella oneidensis MR-1	NC_004347	16S rRNA	*Sp16SA*		269814-271342 (+)	269805-271347
			Sp16SB		216788-218316 (+)	216779-218321
			Sp16SC		46116-47644 (+)	46107-47649
			Sp16SD		4945901-4947429 (-)	4945896-4947438
			Sp16SE		4691979-4693507 (-)	4691974-4693516
			Sp16SF		4682869-4684397 (-)	4682864-4684406
			Sp16SG		4380831-4382359 (-)	4380826-4382368
			Sp16SH		3738378-3739906 (-)	3738373-3739915
			Sp16SI		2973049-2974577 (-)	2973044-2974586
Enterobacteriaceae						
Blochmannia floridanus	NC_005061	16S rRNA	*16S rRNA*	*Bflo*	616478-618033 (-)	616527-618111
Buchnera aphidicola str. Bp	NC_004545	16S rRNA		*bP*	266932-268450 (+)	266925-268485
Buchnera aphidicola str. Sg	NC_000061	16S rRNA	*rrs*	*bSg*	275522-277031 (+)	275515-277066
Buchnera aphidicola str. APS	NC_002528	16S rRNA	*rrs*	*bAPS*	274065-275524 (+)	274037-275584
Escherichia coli K12	NC_000913	16S rRNA	*rrsA*		4033120-4034661 (+)	
			rrsB		4164238-4165779 (+)	
			rrsC		3939431-3940971 (+)	
			rrsD		3424858-3426399 (-)	
			rrsE		4205725-4207266 (+)	
			rrsG		2727636-2729178 (-)	
			rrsH		223771-225312 (+)	
Escherichia coli O157:H7 EDL933	NC_002655	16S rRNA	*rrsA*		4900418-4901959 (+)	
			rrsB		5044690-5046231 (+)	
			rrsC		4804233-4805773 (+)	
			rrsD		4229456-4230997 (-)	
			rrsE		5085699-5087240 (+)	
			rrsG		3519578-3521124 (-)	
			rrsH		227103-228644 (+)	

Species	Accession number	Gene name	Gene name in Genebank	Gene name suggested	Locations as annotated	Locations after corrections
Gammaproteobacteria		**16S rRNA**				
Escherichia coli O157:H7 VT2-Sakai	NC_002695	16S rRNA	*rrsA*		4831654-4833195 (+)	
			rrsB		4975927-4977468 (+)	
			rrsC		4735252-4736793 (+)	
			rrsD		4162234-4163775 (-)	
			rrsE		5016953-5018494 (+)	
			rrsG		3449735-3451276 (-)	
			rrsH		227102-228643 (+)	
Escherichia coli CFT073	NC_004431	16S rRNA	*rrsA*		4561194-4562736 (+)	
			rrsB		4699063-4700613 (+)	
			rrsC		4442840-4444382 (+)	
			rrsD		3858590-3860131 (-)	
			rrsE		4739329-4740870 (+)	
			rrsG		2992309-2993859 (-)	
			rrsH		235186-236727 (+)	
Photorhabdus luminescens TTO1	NC_005126	16S rRNA	*16s_rRNA*	*1*	58438-59982 (+)	
			16s_rRNA	*2*	536233-537777 (+)	
			16s_rRNA	*3*	801957-803501 (+)	
			16s_rRNA	*4*	1472782-1474326 (+)	
			16s_rRNA	*5*	5143181-5144725 (-)	
			16s_rRNA	*6*	5473370-5474914 (-)	
			16s_rRNA	*7*	5509393-5510937 (-)	
Salmonella typhimurium LT2 SGSC1412	NC_003197	16S rRNA	*rrsA*		4196059-4197600 (+)	
			rrsB		4351130-4352673 (+)	4351130-4352671
			rrsC		4100132-4101675 (+)	4100132-4101673
			rrsD		3570463-3572006 (-)	3570466-3572007
			rrsE		4394675-4396219 (+)	4394675-4396217
			rrsG		2800118-2801660 (-)	2800119-2801660
			rrsH		289189-290732 (+)	289179-290720
Salmonella enterica serovar Typhi CT18	NC_003198	16S rRNA	*16s_rRNA*	*A*	3598527-3600068 (-)	
			16s_rRNA	*B*	3747528-3749069 (-)	
			16s_rRNA	*C*	3556311-3557853 (-)	
			16s_rRNA	*D*	3421900-3423441 (-)	
			16s_rRNA	*E*	4257492-4259033 (+)	
			16s_rRNA	*G*	2715999-2717540 (-)	
			16s_rRNA	*H*	287479-289020 (+)	
Salmonella enterica serovar Typhi Ty2	NC_004631	16S rRNA		*1*	287477-289010 (+)	287470-289011
				2	2691223-2692757 (-)	2691222-2692764
				3	3407559-3409092 (-)	2407558-3409099
				4	3541972-3543505 (-)	3541971-3543512
				5	3584188-3585721 (-)	3584187-3585728
				6	3733020-3734553 (-)	3733019-3734560
				7	4242149-4243682 (+)	4242142-4243683

Species	Accession number	Gene name	Gene name in Genebank	Gene name suggested	Locations as annotated	Locations after corrections
Gammaproteobacteria		**16S rRNA**				
Shigella flexneri 2a str. 2457T	NC_004741	16S rRNA	*rrsA*		3723289-3724830 (-)	
			rrsB		3585007-3586548 (-)	
			rrsC		3820462-3822003 (-)	
			rrsD		3400184-3401730 (-)	
			rrsE		3544571-3546112 (-)	
			rrsG		2720026-2721566 (-)	
			rrsH		214156-215696 (+)	
Shigella flexneri 2a str. 301	NC_004337	16S rRNA	*rrsA*		4048482-4050023 (+)	
			rrsB		4186671-4188212 (+)	
			rrsC		3951291-3952832 (+)	
			rrsD		3410015-3411556 (-)	
			rrsE		4227109-4228650 (+)	
			rrsG		2726674-2728214 (-)	
			rrsH		214669-216209 (+)	
Wigglesworthia glossinidia brevipalpis	NC_004344	16S rRNA	*rrsH*	*rrsH+*	136582-138132 (+)	
			rrsH	*rrsH-*	684192-685742 (-)	
Yersinia pestis CO92	NC_003143	16S rRNA	*16s_rRNA*	*001*	12292-13763 (+)	12265-12807
			16s_rRNA	*002*	1217505-1218993 (+)	1217478-1219020
			16s_rRNA	*003*	3652660-3654148 (-)	3652633-3654175
			16s_rRNA	*004*	4178944-4180432 (-)	4178917-4180459
			16s_rRNA	*005*	4221808-4223296 (-)	4221781-4223323
			16s_rRNA	*006*	4388245-4389733 (-)	4388218-4389760
Yersinia pestis KIM	NC_004088	16S rRNA		*yr001*	12016-13600 (+)	12015-13557
				yr005	355930-357514 (+)	355929-357471
				yr008	522393-523977 (+)	522392-523934
				yr011	565257-566841 (+)	565256-566798
				yr014	1024389-1025973 (+)	1024388-1025930
				yr020	3415745-3417329 (-)	3415788-3417330
				yr024	4244973-4246557 (-)	4245016-4246558
Legionellaceae						
Coxiella burnetii RSA 493	NC_002971	16S rRNA		*Cb1*	165579-167035 (+)	165577-167116
Pasteurellaceae						
Haemophilus influenzae KW20 Rd	NC_000907	16S rRNA	*HIrrnA16S*		623825-625364 (+)	
			HIrrnB16S		657107-658646 (+)	
			HIrrnC16S		771212-772750 (+)	771210-772749
			HIrrnD16S		1820465-1822003 (-)	1820460-1821999
			HIrrnE16S		127176-128715 (-)	
			HIrrnF16S		246013-247552 (-)	
Haemophilus ducreyi 35000HP	NC_002940	16S rRNA		*1*	10643-12179 (+)	10642-12182
				2	240156-241692 (+)	240155-241695
				3	489319-490855 (+)	489318-490858
				4	616741-618277 (+)	616740-618280
				5	1539720-1541256 (-)	1539717-1541257
				6	1604540-1606076 (-)	1604537-1606077

Species	Accession number	Gene name	Gene name in Genebank	Gene name suggested	Locations as annotated	Locations after corrections
Gammaproteobacteria		**16S rRNA**				
Pasteurella multocida PM70	NC_002663	16S rRNA		*Pm1*	341429-342970 (+)	
				Pm2	541958-543499 (+)	
				Pm3	1080340-1081882 (-)	
				Pm4	1690576-1692118 (+)	
				Pm5	1761353-1762894(+)	
				Pm6	1941231-1942772 (-)	
Pseudomonadaceae						
Pseudomonas aeruginosa PAO1	NC_002516	16S rRNA		*1*	722096-723631 (+)	
				2	4792196-4793731 (-)	
				3	5267724-5269259 (-)	
				4	6043208-6044743 (-)	
Pseudomonas putida KT2440	NC_002947	16S rRNA	*Pp16SA*		171387-172904 (+)	171372-172908
			Pp16SB		176817-178334 (+)	176802-178338
			Pp16SC		524947-526464 (+)	524932-526468
			Pp16SD		697822-699339 (+)	697807-699343
			Pp16SE		1325501-1327018 (+)	1325486-1327022
			Pp16SF		2548687-2550204 (+)	2548672-2550208
			Pp16SG		5311162-5312679 (-)	5311158-5312694
Pseudomonas syringae DC3000	NC_004578	16S rRNA	*Ps16SA*		666741-668258 (+)	666727-668265
			Ps16SB		827555-829072 (+)	827541-829079
			Ps16SC		1072092-1073609 (+)	1072078-1073616
			Ps16SD		3873151-3874668 (-)	3873144-3874682
			Ps16SE		6217622-6219139 (-)	6217615-6219153
Vibrionaceae						
Vibrio cholerae El Tor N16961	NC_002505	16S rRNA	*16Sa*		53823-55357 (+)	53816-55358
			16Sb		151059-152593 (+)	151052-152594
			16Sc		324147-325181 (+)	324140-325682
			16Sd		401751-403286 (+)	401744-403287
			16Se		762775-764309 (+)	762768-764310
			16Sf		2679884-2681418 (-)	2679883-2681425
			16Sg		2931745-2933279 (-)	2931744-2933286
			16Sh		2937466-2939001 (-)	2937465-2939008
Vibrio parahaemolyticus RIMD 2210633	NC_004603	16S rRNA	*16Sa*		33633-35103 (+)	33631-35183
			16Sb		*581626-583096 (+)*	*581624-583176*
			16Sd		2779817-2781287 (-)	2779737-2781289
			16Se		2885630-2887100 (-)	2885550-2887102
			16Sf		3066302-3067772 (-)	3066222-3067774
			16Sg		3071548-3073018 (-)	3071468-3073020
			16Sh		3131721-3133191 (-)	3131641-3133193
			16Si		3136967-3138437 (-)	3136887-3138439
			16Sj		3196944-3198414 (-)	3196865-3198416
			16Sk		3238281-3239751 (-)	3238201-3239753

Species	Accession number	Gene name	Gene name in Genebank	Gene name suggested	Locations as annotated	Locations after corrections
Gammaproteobacteria		**16S rRNA**				
Vibrio vulnificus CMCP6	NC_004459	16S rRNA		*16Sa*	475699-477241 (-)	
				16sb	932194-933736 (-)	
				16Sc	937697-939239 (-)	
				16Sd	978401-979943 (-)	
				16Se	1060773-1062317 (+)	
				16sf	1170287-1171829 (+)	
				16Sg	1389748-1391290 (+)	
				16Sk	1478333-1479875 (+)	
Vibrio vulnificus YJ016	NC_004459	16S rRNA	*rRNA-16Sa*		32728-34262 (+)	32721-34263
			rRNA-16Sb		177895-179429 (+)	177888-179430
			rRNA-16Sc		717240-718774 (+)	717233-718775
			rRNA-16Se		2946932-2948466 (-)	2946931-2948473
			rRNA-16Sf		2952670-2954200 (-)	2952669-2954207
			rRNA-16Sg		3041406-3042940 (-)	3041405-3042947
			rRNA-16Sh		3261628-3263162 (-)	3261627-3263169
			rRNA-16Si		3303433-3304967 (-)	3303432-3304974
Xanthomonadaceae						
Xylella fastidiosa Temecula 1	NC_004556	16S rRNA		*PD0048*	66041-67585 (+)	
				PD0133	171124-172668 (+)	
Xylella fastidiosa 9a5c	NC_002488	16S rRNA	*XFrrnaA-1*		66558-68102 (+)	
			XFrrnaA-2		172274-173818 (+)	
Xanthomonas campestris pv.campestris ATCC33913		NC_003902	16S rRNA	*SSU1*	4561295-4562841 (-)	
				SSU2	4949163-4950709 (-)	
Xanthomonas axonopodis pv. citri 306	NC_003919	16S rRNA		*SSU1*	4580055-4581601 (-)	
				SSU2	5069297-5070843 (-)	
Epsilonproteobacteria (outgroup)		**16S rRNA**				
Campylobacteraceae						
Helicobacter pylori 26695	NC_000915	16S rRNA		*HPrrnA16s*	1207583-1209081 (-)	1207581-1209081
				HPrrnB16s	1511137-1512634 (-)	1511135-1512634

from a phylogenetically-related species. In the presence of annotations, BLASTn was used to confirm the annotations and confirm that no allele had been missed. When necessary, letters or numbers, or in some cases names, were added to distinguish each 16S rRNA allele from same strain.

All 16S rRNA genes nucleotide sequences and bacterial names were collected in FASTA format. The multiple alignment of the nucleotide sequences of the 16S rRNA genes was done using ClustalW (Thompson *et al.*, 1994), version 1.83 (http://www.ddbj.nig.ac.jp/search/clustalw-e.html). The slow pairwise alignment parameter was selected. Kimura's correction for multiple substitutions was selected (Kimura, 1980). The Neighbor-Joining (Saitou and Nei, 1987) trees were bootstrapped using 1000 random samples of sites from the alignment. TreeView (Page, 1996), version 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html), was used to display and print the phylogenetic tree.

1. 3. Results and Discussion

The locations of the 16S rRNA alleles were annotated in GenBank for most bacterial genomes under study except *Vibrio vulnificus* CMCP6 and *Pasteurella multicoda* PM70. A total of 175 16S rRNA genes nucleotide sequences were retrieved from GenBank for the 33 γ-proteobacteria species and strains under study. Two more sequences were retrieved for *Helicobacter pylori*, the outgroup. The number of 16S rRNA alleles varies from 1 (*Blochmannia floridanus*, each of the three *Buchnera aphidicola* strains, and *Coxiella*

burnetii RSA 493) to 10 (*Vibrio parahaemolyticus* RIMD 2210633). The 16S rRNA alleles ranged in length from 1536 nucleotides for each of the four alleles in *Pseudomonas aeruginosa* PAO1, to 1585 nucleotides for the single allele in *Blochmannia floridanus*. A first multiple alignment revealed that the start and end points of the 16S rRNA alleles from *Shewanella oneidensis* MR-1, *Buchnera aphidicola, Buchnera aphidicola* Sg, *Buchnera sp.* APS, *Blochmannia floridanus, Salmonella typhimurium* LT2, *Salmonella enterica* subsp. *enterica ser. Typhi* Ty2, *Yersinia pestis* CO92, *Yersinia pestis* KIM, *Coxiella burnetii* RSA 493, *Haemophilus influenzae* KW20 Rd, *Haemophilus ducreyi* 35000HP, *Pseudomonas putida* KT2440, *Pseudomonas syringae* DC3000, *Vibrio cholerae* El Tor N16961, *Vibrio parahaemolyticus* RIMD 2210633, *Vibrio vulnificus* YJ016 and *Helicobacter pylori* 26695 had been improperly annotated in GenBank and needed corrections. The corrected locations of the start and end points are indicated in Table 1. The matrix generated after the corrected multiple alignment was 177 16S rRNA alleles X 1647 nucleotides in size, including the two 16S rRNA alleles from the outgroup. The rooted neighbor-joining tree based on the 16S rRNA sequences and showing the phylogenetic relationships between all 33 γ-proteobacteria is presented in Fig. 4. Bootstraps values lower than 90% are given. *Helicobacter pylori* was used as an outgroup because it is close enough to the γ-proteobacteria so that orthologous 16S rRNA alleles share homology, and distant enough to belong to a different phylogenetic group, the ε-proteobacteria. In addition to the outgroup, eight clusters, each corresponding to a bacterial group, are revealed. These are the *Legionellaceae*, *Xanthomonadaceae*, *Pseudomonadaceae*, *Alteromonadaceae*, *Vibrionaceae*, *Pasteurellaceae* and *Enterobacteriaceae*. Several clusters are further sub-divided into sub-clusters which, in

Legend

Figure 4: Phylogenetic relationships between 33 γ-proteobacteria inferred from 175 16S rDNA allelic sequences. Two *Helicobacter pylori* 16S rDNA allelic sequences were used as outgroups. The multiple alignment of the nucleotide sequences of the 16S rRNA genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 177 16S rRNA alleles X 1647 nucleotides in size. The tree was generated using the Neighbor-Joining method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

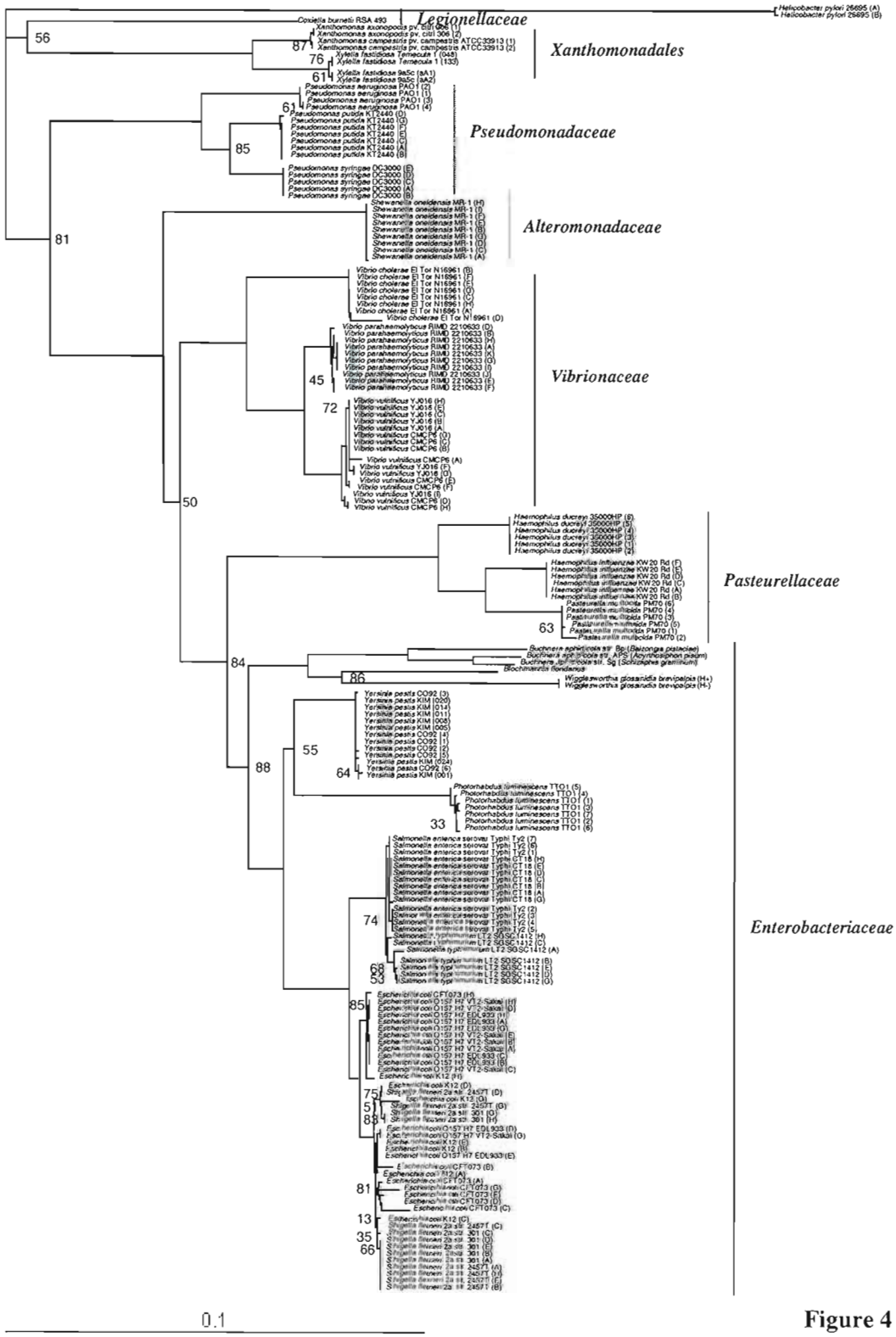

Legionellaceae
Xanthomonadales
Pseudomonadaceae
Alteromonadaceae
Vibrionaceae
Pasteurellaceae
Enterobacteriaceae
0.1

Figure 4

several cases, correspond to γ-proteobacterial species.

This is exemplified by *Xanthomonas axonopodis* and *Xanthomonas campestris*; by *Pseudomonas aeruginosa*, *Pseudomonas putida* and *Pseudomonas syringae*; by *Vibrio cholerae*, *Vibrio parahaemolyticus* and *Vibrio vulnificus*; by *Haemophilus ducreyi* and *Haemophilus influenzae*; and by *Salmonella enterica* and *Salmonella typhimurium*.

The resolving power of this phylogenetic tree at the bacterial strain level, however, is limited and varies with the species and strains under study. 16S rRNA alleles from *Xylella fastidiosa* strains Temecula 1 and 9a5c, can be grouped in different strain-specific sub-clusters. This is also true for *Buchnera aphidicola* strains Bp, APS and Sg. This is not the case, however, for 16S rRNA alleles from *Vibrio vulnificus* strains YJ016 and CMCP6, from *Salmonella enterica* serovars Typhi Ty2 and Typhi CT18; and for *Yersinia pestis* strains CO92 and KIM. In these cases, alleles from different strains or serovars are mixed in a same cluster. The *Escherichia coli* and *Shigella flexneri* strains require different and additional comments. Whereas several 16S rRNA alleles from various *Escherichia coli* strains share identical or nearly identical nucleotide sequences at the intrastrain and intraspecies levels, the same can be said for several 16S rRNA alleles from various *Shigella flexneri* strains. Based on these, *Escherichia coli* and *Shigella flexneri* could almost be separated into different species-specific clusters. However, a few alleles from one species could cluster with an allele from the other as exemplified by *Escherichia coli* K12 allele (D) and *Shigella flexneri* 2a str. 2457T allele (D). Should a single allele be used for phylogenetic analysis,

different relationships between *Escherichia coli* and *Shigella flexneri* could be inferred based on different selected alleles. DNA hybridization studies conducted decades ago have shown, however, that *Shigella* and *Escherichia coli* belong to the same genetic species (Brenner *et al.*, 1972; 1973; 1982). Clearly, here, the intertwining of the *Escherichia coli* and *Shigella flexneri* 16S rRNA alleles was not unexpected.

Next, a second phylogenetic tree was constructed, using a single allele per bacterial strain. When a single allele was present per taxon, or when the alleles were grouped at the species level, the choice was easy. In the few cases where alleles were not discriminatory at the bacterial strain or serovar levels, an allele that grouped with most alleles from same strain was selected. The rooted neighbor-joining tree based on the 16S rRNA sequences and showing the phylogenetic relationships between all 33 γ-proteobacteria inferred from a single 16S rRNA allelic sequence per taxon is presented in Fig. 5. Because of the high homology between intra-strain alleles revealed above, it is likely that a similar tree would have been obtained with other alleles from same bacterial strain. The key point, however, is that in most phylogenies inferred from comparison of 16S rRNA sequences, the usual approach is to use conserved primers to amplify 16S rRNA sequences. The amplified product is cloned and clones are usually selected at random for sequencing and subsequent DNA sequence comparison between taxa. A new field in bacterial ecology was even opened up by the application of this technology for the characterization of unculturable bacteria from various environments (Pace *et al.*, 1986; Ward *et al.*, 1990). Our results on γ-proteobacteria show that alleles are grouped at the species level, with the exception of

Legend

Figure 5: Phylogenetic relationships between 33 γ-proteobacteria inferred from 33 16S rDNA allelic sequences. *Helicobacter pylori* was used as an outgroup. The multiple alignment of the nucleotide sequences of the 16S rRNA genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 34 16S rRNA alleles X 1647 nucleotides in size. The tree was generated using the Neighbor-Joining method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

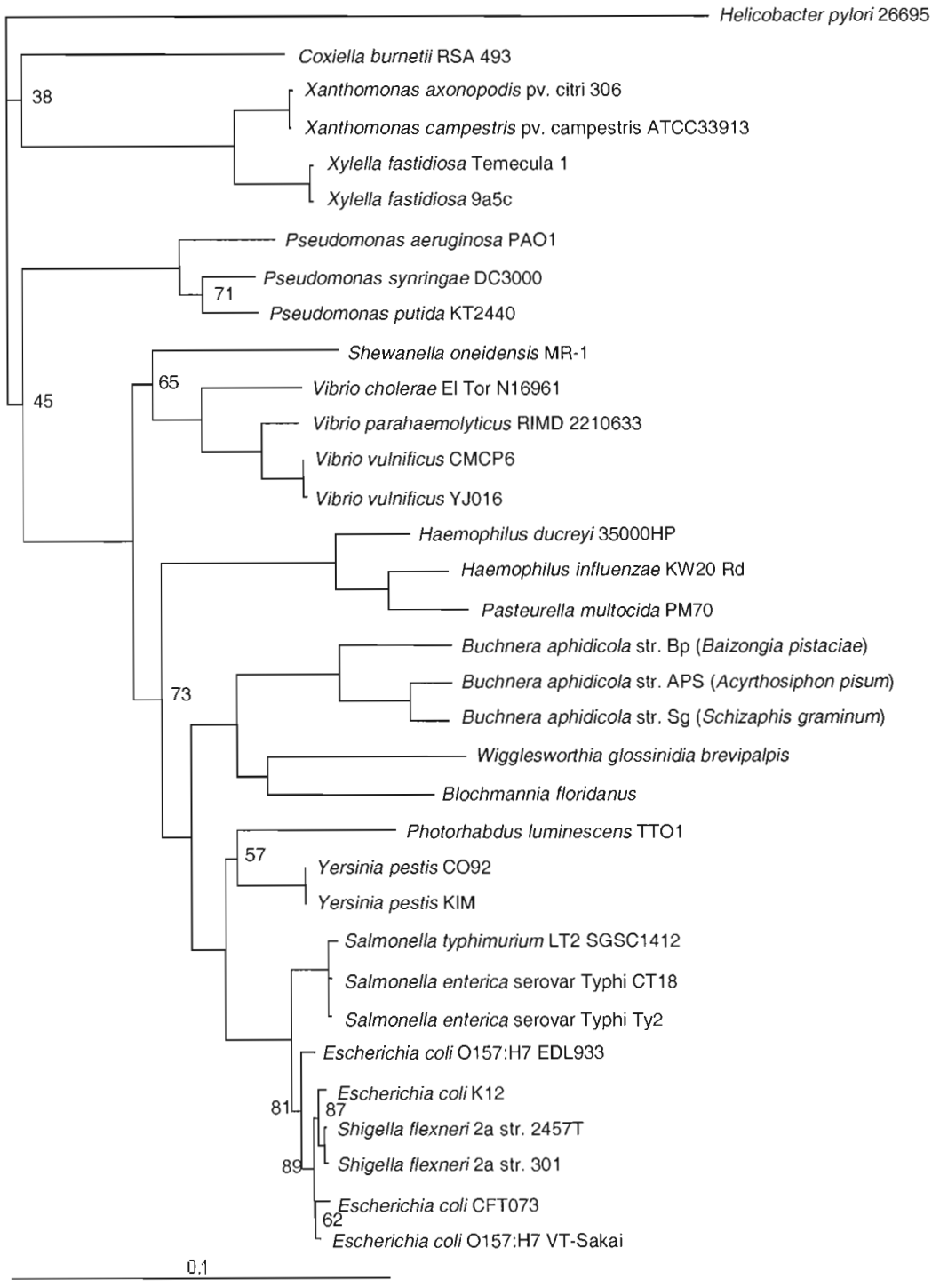


Figure 5 -Phylogenetic relationships between 33 γ-proteobacteria inferred from 33 16S rDNA allelic sequences

Escherichia coli and *Shigella flexneri*. Random selection of 16S rRNA alleles within strain would not generate different phylogenetic trees for the γ-proteobacteria studied here.

We conclude that, for these γ-proteobacteria, the selection of a single 16S rRNA allele is sufficient for inferring phylogenies, and that orthologous alleles from same strain would have generated similar phylogenetic trees. It would be hazardous, however, to extend this conclusion to other bacterial taxa. As mentioned earlier, the heterogeneities between 16S rRNA alleles at the intra-strain level varies among taxa. Acinas *et al.*, (2004) have shown that some intra-strain 16S rRNA alleles from *Desulfotomaculum kuznestovii* (*Clostridales*) or *Thermoanaerobacter tengcongensis* (*Thermoanaerobacteriales*) could exhibit up to 8.3% and 11.6% nucleotide sequence divergence, respectively. Whether the positioning of these species on a phylogenetic tree may vary with the selected 16S rRNA allele, although likely, was not studied. Our conclusions on γ-proteobacteria raise a few more questions. How accurate are the DNA sequences deposited in GenBank? It is generally assumed that sequences deposited in GenBank have a 0.1% error rate on the average (Clark and Whittam, 1992). Did all 16S rRNA sequences presented here fall within the error rate? And perhaps as importantly, were they all free of post-sequencing artefacts, corrections, modifications? To answer these questions, at least in part, all *Escherichia* spp. and *Shigella* spp.16S rRNA allelic sequences deposited in GenBank will be retrieved, irrespective whether or not the host genome was fully sequenced, and compared with the sequences presented here. Different laboratories might have used different DNA sequencing methods which in turn might have generated different error rates or different post-sequencing artefacts. In addition,

in-house generated 16S rRNA allelic sequences will be included.

CHAPTER II

PHYLOGENETIC ANALYSIS OF γ-PROTEOBACTERIA INFERRED FROM NUCLEOTIDE SEQUENCE COMPARISONS OF THE HOUSE-KEEPING GENES *ADK*, *AROE* AND *GDH*: COMPARISONS WITH PHYLOGENY INFERRED FROM 16S rRNA GENE SEQUENCES.

Hoon-Yong Lee and Jean-Charles Côté

JOURNAL OF GENERAL AND APPLIED MICROBIOLOGY, vol. 52, p. 147-158, 2006

Summary

Nucleotide sequence comparisons of three house-keeping genes, adenylate kinase (*adk*), shikimate dehydrogenase (*aroE*), and glucose-6-phosphate dehydrogenase (*gdh*), were used to infer the phylogeny of 33 γ-proteobacteria. Phylogenetic trees inferred from each gene, and of the concatenated sequences of all three genes, are, in general, similar to a 16S rRNA gene-inferred tree. Similar grouping of bacteria are revealed at the family, genus, species and strain levels in all five trees. The house-keeping genes, however, show a higher rate of nucleotide sequence substitutions. Consequently, they can possibly probe deeper branches of a phylogenetic tree than the 16S rRNA gene. However, because their nucleotides sequences are not as highly conserved among γ-proteobacteria, family- or genus-specific primers would need to be designed for the amplification of any of these three house-keeping genes. Since these genes are used in multilocus sequence typing, it is expected that the number of sequences publicly available for many taxa will increase over time proving them very useful either at complementing 16S rRNA-inferred phylogenies or for specific, targeted, phylogenetic analysis.

2. 1. Introduction

The proteobacteria, often referred to as "purple bacteria and their relatives", are a major division of the eubacterial tree and encompass bacteria with a great diversity of phenotypes, physiological attributes, and habitats (Stackebrandt *et al.*, 1988; Gupta, 2000). This division of predominantly Gram-negative bacteria contains more than 200 genera. Proteobacteria have been classified based on homology of 16S ribosomal RNA or by hybridization of ribosomal DNA with 16S and 23S ribosomal RNA (Fox *et al.*, 1980; Woese *et al.*, 1985a; Woese, 1987; De Ley, 1992). The division has been sub-divided into five major groups, α- (Woese *et al.*, 1984a), β- (Woese *et al.*, 1984b), γ- (Woese *et al.*, 1985b), δ- and ε-. The γ-proteobacteria include several families of utmost biological importance, most notably the *Enterobacteriaceae*, *Coxiellaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae*, *Shewanellaceae* and *Xanthomonadaceae*, to name a few. Some of these are human-, animal-, or plant pathogens, others are obligate endosymbionts, etc. Because of their biological importance, the genomes of more than 33 γ-proteobacteria have been fully sequenced and are freely available for analyses.

The small subunit ribosomal RNA (16S rRNA) has been established as the macromolecule of choice for single-gene phylogenetic analyses (Fox *et al.*, 1980; Woese, 1987; Woese *et al.*, 1990). The 16S rRNA are essential components of protein synthesis and are present in all bacteria. Because it is under highly constrained function, its gene is highly conserved throughout bacteria and even very distant bacterial species can be compared. It is

assumed that the homology between 16S ribosomal RNA sequences from different bacteria reflects the phylogenetic relationship between same organisms.

The use of 16S rRNA sequences for single-gene phylogenetic analyses is not without limits. First, because 16S rRNA genes are highly conserved, the classification of closely related bacterial species may be problematic (Gürtler and Stanisich, 1996; Kolbert and Persing, 1999). Second, 16S rRNA genes may be subjected to lateral gene transfer (Reischl *et al.*, 1998; Ueda *et al.*, 1999; Yap *et al.*, 1999; Schouls *et al.*, 2003). Conversely, other lateral gene(s) transfer can create regions or islands within a bacterial genome with origins or phylogenies different from the 16S rRNA genes (Lawrence, 1999; Ochman *et al.*, 2000). Fourth, copy number of 16S rRNA genes per bacterial genome ranges between 1 and 15 (Klappenbach *et al.*, 2001). Heterogeneities between copies are not a rare occurrence and paralogous copies may infer different phylogenies. In addition, in recent years, sequence comparisons of orthologous house-keeping genes have revealed incongruence between 16S rRNA sequence-inferred phylogenies and house-keeping gene sequence-inferred phylogenies (Turner and Young, 2000; Parker *et al.*, 2002).

Several genes have been proposed to complement 16S rRNA genes in bacterial phylogenetic analysis. These genes are not transmitted horizontally, their evolutionary rate is higher than the one of 16S rRNA, and they are present in most bacteria.

They include *gyrB* (subunit B protein of DNA gyrase, topoisomerase type II; Yamamoto and Harayama, 1995), *atpD* (beta subunit of the membrane ATP synthase; Ludwig *et al.*, 1993; Gaunt *et al.*, 2001), *infB* (translation initiation factor 2; Hedegaard *et al.*, 1999), *recA* (RecA protein; Gaunt *et al.*, 2001), and *rpoB* (RNA polymerase beta subunit; Mollet *et al.*, 1997) to name a few.

Multilocus sequence typing (MLST) has recently been developed as a molecular typing method (Maiden *et al.*, 1998). It is based on the principles of multilocus enzyme electrophoresis, but characterizes the alleles present at multiple house-keeping genes. Maiden and colleagues have selected six house-keeping genes for MLST, *abcZ* (putative ABC transporter), *adk* (adenylate kinase), *aroE* (shikimate dehydrogenase), *gdh* (glucose-6-phosphate dehydrogenase), *pdhC* (pyruvate dehydrogenase subunit), and *pgm* (phosphoglucomutase), on the basis that they are unlinked, variations accumulate slowly and are likely to be selectively neutral, and they are present in many bacteria. With the increasing number of research groups using MLST as a typing method, the number of sequences publicly available for these six house-keeping genes is rapidly increasing for a wide variety of bacteria. Whether some of these genes could be used to complement 16S rRNA genes to infer phylogenies is worth addressing.

We present here five phylogenies of 33 γ-proteobacteria, a first one inferred from nucleotide sequence comparisons of the adenylate kinase gene (*adk*); a second one inferred from shikimate dehydrogenase (*aroE*); and a third one inferred from glucose-6-phosphate

dehydrogenase (*gdh*). The fourth phylogeny is inferred from the concatenated *adk*, *aroE* and *gdh* gene sequences. The fifth phylogeny is inferred from 16S rRNA gene nucleotide sequences. All five phylogenies are compared, congruencies and differences are discussed.

2.2. Materials and Methods

Bacterial species and strains. The 33 γ-proteobacterial species and strains analysed in this study are listed in Table 2. They were selected on the basis that their complete genome sequences were freely available in GenBank, at the National Center for Biotechnology Information (NCBI) completed microbial genomes database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html; May 2004). These include one *Shewanellaceae* species, 17 *Enterobacteriaceae* strains (encompassing 9 species), one *Coxiellaceae* species, three *Pasteurellaceae* species, three *Pseudomonadaceae* species, four *Vibrionaceae* strains (three species), and four *Xanthomonadaceae* strains (three species). An ε-proteobacterium, *Helicobacter pylori*, was included as an outgroup.

Genes and sequences selections. The house-keeping genes chosen for the phylogenetic analyses were *adk* (encoding adenylate kinase - Adk), *aroE* (shikimate dehydrogenase - AroE) and *gdh* (glucose-6-phosphate dehydrogenase - Gdh). The GenBank accession numbers of the fully sequenced bacterial genomes are given along with the locations of the three house-keeping genes as annotated in GenBank (Table 2). The amino acid sequences were retrieved for each house-keeping protein in each γ-proteobacterium.

The amino acid sequences were aligned using ClustalW (Thompson *et al.*, 1994; http://www.ddbj.nig.ac.jp/search/clustalw-e.html; version 1.83). This first multiple sequence alignment revealed that in few cases, two cases in Adk - *Escherichia coli* CFT073 and *Xylella fastidiosa* 9a5c -, and three in Gdh - *Vibrio vulnificus* YJ016, *Xylella fastidiosa* Temecula 1 and *Yersinia pestis* KIM -, locations had been improperly annotated. Supernumerary amino acids were present at the amino-terminal end of the protein, upstream of the initial methionine. They were deleted to optimize the amino acid sequences alignment. The corrected alignment was used to retrieve the *adk*, *aroE* and *gdh* gene nucleotide sequences. Corrected annotations of the gene locations are given in Table 2.

Percentage sequence similarity between the core Escherichia-Shigella-Salmonella enterics. The multiple alignments of the nucleotide sequences of the orthologous 16S rRNA, *adk*, *aroE*, *gdh* and concatenated genes for the nine core *Escherichia-Shigella-Salmonella* enterics under study were done using ClustalW. Pairwise comparisons revealed respective percentage sequence similarities. The discriminatory power (DP) was defined here as the median value of percentage sequence similarities among the nine core *Escherichia-Shigella-Salmonella* enterics. Lower DP values reflected lower percentage sequence similarities and, consequently, higher discriminating power among the core enterics.

Phylogenetic Analysis. The multiple alignments of the nucleotide sequences of the 34 orthologous *adk*, 33 *aroE* and 32 *gdh* genes were done using ClustalW. Phylogenies were estimated by maximum likelihood using PHYML (Guindon and Gascuel, 2003;

http://atgc.lirmm.fr/phyml/; version 2.4.4, February 2005). The maximum likelihood tree was bootstrapped using 1000 random samples of sites from the alignment. TreeView (Page, 1996; http://taxonomy.zoology.gla.ac.uk/rod/treeview.html; version 1.6.6, September 2001) was used to display and print the phylogenetic tree. In addition to these three single-gene based inferred phylogenies, all three house-keeping genes were concatenated and a phylogeny was inferred as described above. Another phylogeny was inferred from a fourth single gene, the 16S rRNA gene. Construction of the phylogenetic tree of the 33 γ-proteobacteria inferred from the 16S rRNA gene sequences was described elsewhere (Olivier *et al.*, 2005). Because the number of 16S rRNA alleles per bacterial stain varied from 1 to 10, a total of 175 16S rRNA genes nucleotide sequences were retrieved from GenBank for the 33 γ-proteobacteria species and strains. In the present study, a single 16S rRNA allele per bacterial strain, deemed representative of that strain (Olivier *et al.*, 2005), was used to infer the phylogeny. Identities and locations of the selected 16S rRNA alleles are given in Table 2. The phylogeny was inferred as described above. This 16S rRNA gene-based phylogeny was compared with the *adk*-, *aroE*- and *gdh*-gene nucleotide sequences inferred phylogenies.

In addition to these nucleotide-based inferred phylogenies, the multiple alignments of the amino acid sequences of the 34 orthologous adenylate kinase, 33 shikimate dehydrogenase and 32 glucose-6-phosphate dehydrogenase proteins were done using ClustalW. Phylogenies were inferred for each protein and the concatenated sequences as described above.

Table 2 - List of γ-proteobacteria species and strains used in this study. Complete bacterial genome GenBank accession number, locations of the start and end points of the *adk*, *aroE*, *gdh* and 16S rRNA alleles as indicated in GenBank, and locations after correction are presented.

Family Species	Accession Number	*adk* Location Annotation	*adk* Location Correction	*aroE* Location Annotation	*gdh* Location Annotation	*gdh* Location Correction	16S rRNA Location Annotation	16S rRNA Location Correction
Shewanellaceae								
Shewanella oneidensis MR-1	NC_004347	2117461..2118105		43717..44580	2611113..2612585(-)		4380831..4382359(-)	4380826..4382368
Enterobacteriaceae								
Buchnera aphidicola str. APS	NC_002528	535058..535705		542838..543659(-)	355650..357125		274065..275524	274037..275584
Buchnera aphidicola str. Bp	NC_004545	503441..504088		511465..512313(-)	349167..350639		266932..268450	266925..268485
Buchnera aphidicola str. Sg	NC_000061	535946..536590		543669..544490(-)	357259..358731		275522..277031	275515..277066
'*Candidatus* Blochmannia floridanus'	NC_005061	334622..335302		237684..238568	492629..494125(-)		616478..618033(-)	616527..618111
Escherichia coli CFT073	NC_004431	571765..572469	571826..572469	3861387..3862205	2086516..2087991(-)		2992309..2993859(-)	
Escherichia coli K12	NC_000913	496399..497043		3427657..3428475	1932863..1934338		2727638..2729179(-)	
Escherichia coli O157:H7 VT2-Sakai	NC_002695	563070..563714		4165032..4165850(-)	2536670..2538145(-)		3449735..3451276(-)	
Escherichia coli O157:H7 EDL933	NC_002655	563073..563717		4232254..4233072(-)	2611878..2613353(-)		3519578..3521124(-)	
Photorhabdus luminescens TTO1	NC_005126	4507364..4508008(-)		4507364..4508008(-)	2507490..2508965		1472782..1474326	
Salmonella enterica subsp. *enterica* ser. Typhi CT18	NC_003198	539066..539710		4255418..4256236	1945364..1946746(-)		2715999..2717540(-)	
Salmonella enterica subsp. *enterica* ser. Typhi Ty2	NC_004631	2441391..2442035(-)		4240066..4240884	1079699..1081174		3541972..3543505(-)	3541971..3543512
Salmonella typhimurium LT2 SGSC1412	NC_003197	546041..546685		3573266..3574084(-)	1979667..1981142(-)		2800118..2801660(-)	2800119..2801660
Shigella flexneri 2a str. 2457T	NC_004741	433501..434145		3402985..3403803(-)	1867906..1869381(-)		2720026..2721566(-)	
Shigella flexneri 2a str. 301	NC_004337	433640..434344		3412810..3413628(-)	1898712..1900188(-)		2726674..2728214(-)	
Wigglesworthia glossinidia brevipalpis	NC_004344	611679..612311					684192..685742(-)	
Yersinia pestis CO92	NC_003143	3475023..3475667(-)		244319..245140	2345985..2347469		4178944..4180432(-)	4178917..4180459
Yersinia pestis KIM	NC_004088	1200105..1200749		4465423..4466244	2472015..2473604	2472120..2473604	3415745..3417329(-)	3415788..3417330
Coxiellaceae								
Coxiella burnetii RSA 493	NC_002971	399357..400052		10986..11804			165579..167035	165577..167116
Pasteurellaceae								
Pasteurella multocida PM70	NC_002663	323720..324364(-)		1463132..1463941	1751723..1753213		1941231..1942772(-)	
Haemophilus ducreyi 35000HP	NC_002940	656284..656931(-)		656284..656931(-)	667336..668823(-)		616741..618277	616740..618280
Haemophilus influenzae KW20 Rd	NC_000907	375895..376539(-)		698718..699536(-)	576891..578375(-)		623825..625364	
Pseudomonadaceae								
Pseudomonas aeruginosa PAO1	NC_002516	4126947..4127594(-)		26711..27535	3572323..3573792(-)		4792196..4793731(-)	
Pseudomonas putida KT2440	NC_002947	1712263..1712913		84198..85022	1165614..1167083		5311162..5312679(-)	5311158..5312694
Pseudomonas syringae DC3000	NC_004578	1668011..1668658		189618..190442(-)	1429651..1431120		3873151..3874668(-)	3873144..3874682
Vibrionaceae								
Vibrio cholerae O1 biovar eltor str. N16961	NC_002505	1050494..1051138		51668..52504	850221..851726		2931745..2933279(-)	2931744..2933286
Vibrio parahaemolyticus RIMD 2210633	NC_004603	851266..851910		3241072..3241905(-)	1820846..1822351(-)		3071548..3073018(-)	3071468..3073020
Vibrio vulnificus CMCP6	NC_004459	179953..180597(-)		1058641..1059474	2731240..2732745(-)		1389748..1391290	
Vibrio vulnificus YJ016	NC_005139	997760..998404		997760..998404	1650475..1652055	1650400..1652055	3041406..3042940(-)	3041405-3042947
Xanthomonadaceae								
Xylella fastidiosa Temecula 1	NC_004556	284089..284652		1775750..1776598(-)	436386..437888(-)	436386..437819	171124..172668	
Xylella fastidiosa 9a5c	NC_002488	285623..286219	285656..286219	602559..603407	1023042..1024475(-)		172274..173818	
Xanthomonas campestris pv.campestris ATCC33913	NC_003902	3907725..3908288(-)		4681319..4682170(-)	2520183..2521613		4561295..4562841(-)	
Xanthomonas axonopodis pv. Citri 306	NC_003919	4047498..4048061(-)		4724103..4724954(-)	2419342..2420772(-)		4580055..4581601(-)	
OUTGROUP								
Helicobacter pylori 26695	NC_000915	663843..664418		1324695..1325486(-)	1162198..1163475		1207583..1209081	

2.3. Results and Discussion

Genes and sequences selections. A total of 33 adenylate kinase (*adk*), 32 shikimate dehydrogenase (*aroE*) and 31 glucose-6-phosphate dehydrogenase (*gdh*) genes nucleotide sequences were retrieved from GenBank for the 33 γ-proteobacteria species and strains under study. The *adk* gene varied in length between 564 nucleotides for the two *Xylella fastidiosa* strains and the two *Xanthomonas* species and 696 nucleotides for *Coxiella burnetii* RSA 493. The *aroE* gene varied in length between 810 nucleotides for *Pasteurella multocida* PM70 and 885 nucleotides for '*Candidatus* Blochmannia floridanus'. The *gdh* gene varied in length between 1431 nucleotides for the two *Xanthomonas* species and 1506 nucleotides for the four *Vibrio* strains. No *aroE* gene could be retrieved from *Wigglesworthia glossinidia brevipalpis*, and no *gdh* gene could be retrieved from *W. glossinidia brevipalpis* and *Coxiella burnetii* RSA 493 using different retrieval softwares. This is not surprising for *W. glossinidia brevipalpis* considering it is an endosymbiont of the tsetse fly and its genome is only 0.7 Mbp in size, about 1/7 that of *Escherichia coli*. Similarly, *C. burnetii* is an obligate intracellular pathogen that can infect reptiles, birds, and mammals. Its genome is only 2.0 Mbp in size. Sequences of each three house-keeping genes were retrieved from the ε-proteobacterium *Helicobacter pylori* which served as the outgroup. *H. pylori* was chosen as an outgroup because it is close enough to the γ-proteobacteria so that orthologous *adk*, *aroE* and *gdh* alleles share homology, and distant enough to belong to a different phylogenetic group, the ε-proteobacteria.

Percentage sequence similarity between the core Escherichia-Shigella-Salmonella enterics. The discriminatory power (DP) of the 16S rRNA gene was compared with the DP of the three house-keeping genes, *adk*, *aroE* and *gdh*, and concatenated genes, among the nine core *Escherichia-Shigella-Salmonella* enterics available (Table 3). The 16S rRNA gene showed the highest DP value (highest median value of percentage sequence similarity) of 98.2%. All three house-keeping genes, *adk*, *aroE*, and *gdh*, and the concanenated sequences, showed lower DP values of 91.9%, 96.2%, 90.7%, and 89%, respectively. In addition, the two least similar 16S rRNA sequences - *E. coli* O157:H7 VT2-Sakai and *Salmonella typhimurium* LT2 SGSC1412 - still shared a relatively high 96.9% identical nucleotides. This is to be compared with the least similar *adk* - the three *Salmonella* strains on the one hand and the two *Shigella* strains on the other hand -, *aroE* - *E. coli* O157:H7 EDL933 and *Salmonella typhimurium* LT2 SGSC1412 -, and *gdh* - also the three *Salmonella* strains on the one hand and the two *Shigella* strains on the other hand - sequences, which shared lower 85.3%, 68.7% and 83% identical nucleotides, respectively. Clearly, any of the three house-keeping genes has a lower percentage of gene sequence similarity than the 16S rRNA gene and can better distinguish species among the core *Escherichia-Shigella-Salmonella* enterics.

Phylogenetic Analysis. The matrices generated after multiple nucleotide sequence alignments were 34 *adk* alleles X 722 nucleotides in size, 33 *aroE* alleles X 948 nucleotides in size, and 32 *gdh* alleles X 1560 nucleotides in size, each including the outgroup allelic sequence. The *H. pylori*-rooted maximum likelihood trees inferred from the *adk*, *aroE* and *gdh* nucleotide sequences are presented in Figs. 6, 7 and 8, respectively. Bootstrap values

Table 3- Percentage sequence similarity between 16S rRNA, selected house-keeping, and concatenated house-keeping genes, among "core" *Escherichia-Shigella-Salmonella* enterics.

	Percentage similarity				
Gene	**16S rRNA**	***adk***	***aroE***	***gdh***	**Concatenated**
Median value	98.2	91.9	96.2	90.7	89
Average	98.3	92.2	84.8	91.1	89.5
Lowest value	96.9	85.3	68.7	83	80.2
Highest value	99.9	100	100	100	100

lower than 90% are given. A fourth matrix was generated using the concatenated *adk*, *aroE* and *gdh* nucleotide sequences. This matrix was 32 allelic sequences X 3170 nucleotides in size. *W. glossinidia brevipalpis* and *C. burnetii* RSA 493 were not included. A rooted maximum likelihood tree inferred from the concatenated *adk*, *aroE* and *gdh* sequences is presented in Fig. 9. A fifth matrix was generated using a 16S rRNA allelic sequence from each of the 33 γ-proteobacteria and the ε-proteobacterium outgroup. This matrix was 34 allelic sequences X 1647 nucleotides in size. The rooted maximum likelihood tree inferred from the 16S rRNA sequences is presented in Fig. 10.

As shown in Fig. 6, based on the *adk* gene nucleotide sequence, all γ-proteobacteria strains are grouped at the species level. This is true for both *Xylella fastidiosa* strains, the three *Buchnera aphidicola*, both *Vibrio vulnificus*, both *Yersinia pestis*, both *Salmonella enterica*, the four *Escherichia coli* and both *Shigella flexneri* strains, respectively. Here, the two *S. flexneri* strains appear to be very close to *E. coli*. DNA hybridization studies conducted decades ago, however, have shown that *Shigella* spp. and *E. coli* belong to the same genetic species (Brenner *et al.*, 1972; 1973). Clearly, here, the intertwining of the *E. coli* and *S. flexneri adk* alleles in Fig. 6 is not unexpected. All γ-proteobacteria species are also grouped at the genus level. This is true for the two *Xanthomonas* species, the two *Haemophilus* species, the three *Vibrio* species, the three *Pseudomonas* species and the two *Salmonella* species. Certainly, some species appear closer than others. *Salmonella typhimurium* appears closer to *S. enterica* than *Haemophilus influenzae* to *H. ducreyi*. Fig. 6

Legend

Figure 6: Phylogenetic relationships between 33 γ-proteobacteria inferred from 33 *adk* allelic sequences. *Helicobacter pylori adk* allelic sequence was used as outgroup. The multiple alignment of the nucleotide sequences of the *adk* genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 34 *adk* alleles X 719 nucleotides in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

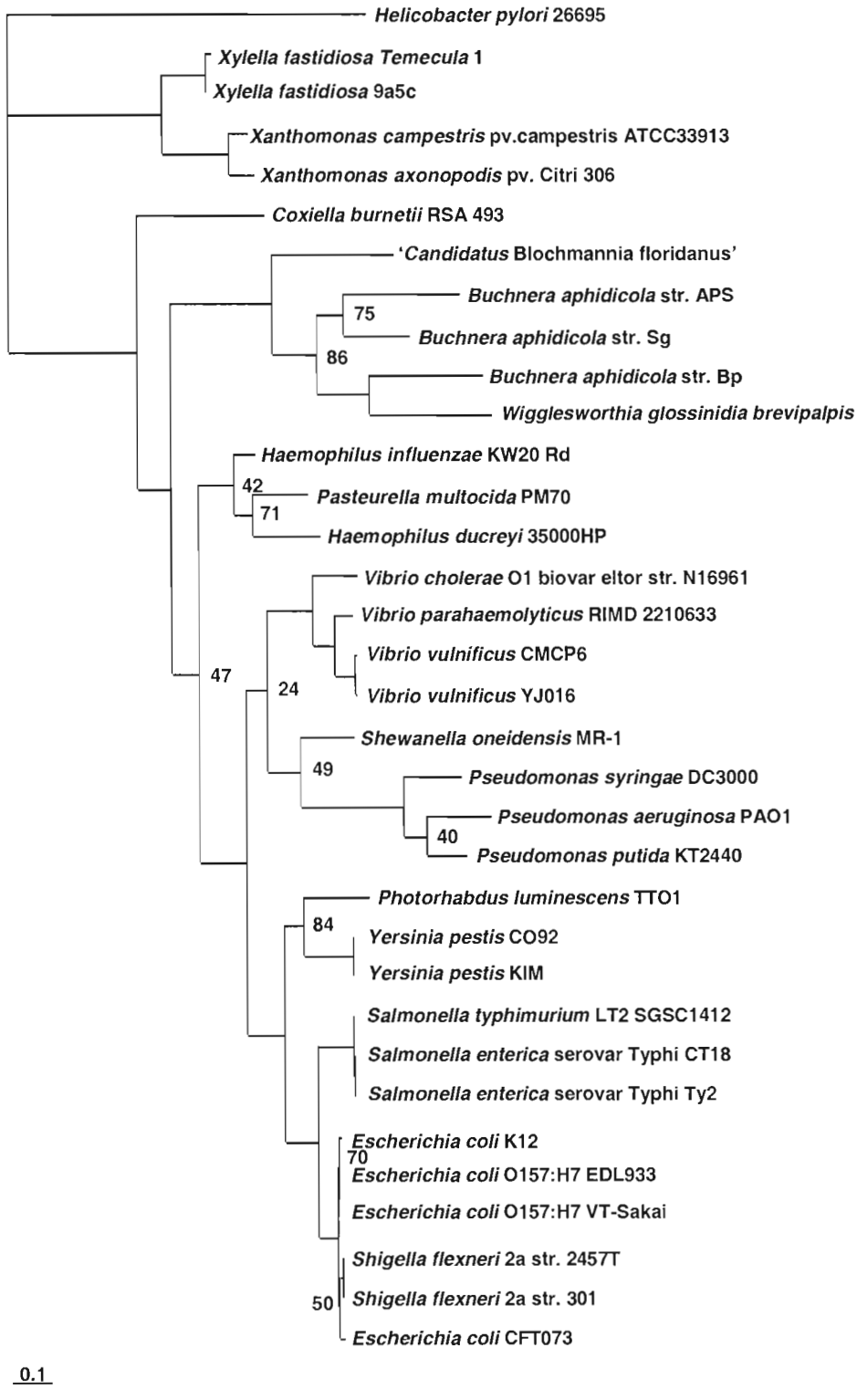


Figure 6 -Phylogenetic tree based on *adk* gene sequences using ML method

also groups related genera at the family level, with exceptions. As expected, the two *Xanthomonadaceae* genera - *Xanthomonas* and *Xylella*; the plant pathogens - are grouped together. Likewise, the two *Pasteurellaceae* genera - *Pasteurella* and *Haemophilus*; the etiological agents of some diseases in humans and other vertebrates - are grouped together. Interestingly, however, the *adk* gene of *H. ducreyi* 35000HP appears more closely related to its orthologous sequence in *Pasteurella multocida* PM70 than to the orthologous sequence in *H. influenzae* KW20 Rd. Grouping of γ-proteobacteria related genera at the family level, however, is not fully resolved for the *Enterobacteriaceae*. Whereas a single cluster of *Enterobacteriaceae* genera might have been expected, two sub-groups are revealed in Fig. 6. A first one in which *W. glossinidia brevipalpis* is grouped with the three *B. aphidicola* strains, APS, Sg and Bp. '*Candidatus* Blochmannia floridanus' also groups closely to these species. A second one which encompasses *Photorhabdus*, *Yersinia*, *Salmonella*, *Escherichia* and *Shigella*. Composition of both sub-groups is not surprising in itself. All three species in sub-group 1, *W. glossinidia brevipalpis*, *B. aphidicola*, and '*Candidatus* Blochmannia floridanus' are classified as *Enterobacteriaceae* and share common lifestyles by being obligate endosymbionts of tsetse flies, aphids and carpenter ants, respectively. Indeed, it has been suggested that orthologous sequences in the other two *B. aphidicola* strains, APS and Sg. Sub-group 2 contains the core *Escherichia-Shigella-Salmonella* enterics and *Yersinia* and *Photorhabdus*. The surprise is that both *Enterobacteriaceae* sub-groups do not form a single cluster.

The topology of Fig. 7, based on the *aroE* gene nucleotide sequence, is similar to Fig. 6, with exceptions. Here again, strains are grouped at the species level, most species are grouped at the genus level, and most genera are grouped at the family level. Here, however, the *aroE* gene in *H. influenzae* appears closer to its ortholog in *P. multocida* than to the ortholog in *H. ducreyi*. The two *Enterobacteriaceae* sub-groups revealed in Fig. 6, the obligate endosymbionts and the core enterics, are also present here. *Yersinia* and *Photorhabdus* appear somehow equally distant from these two *Enterobacteriaceae* sub-groups.

The topology of Fig. 8, based on the *gdh* gene nucleotide sequence, is similar to Fig. 6 and Fig 7, with exceptions. Here however, the *Enterobacteriaceae* form a single cluster although sub-groups can be detected: *Photorhabdus*, the obligate endosymbionts, *Yersinia* and the core enterics.

Fig. 9 is inferred from the concatenated *adk*, *aroE* and *gdh* nucleotide sequences. Here again, its topology is similar to the topology to the other three previous figures with regards to strains grouping at the species level, species grouping at the genus level, genera grouping at the family level, and positioning of families respective to other families. Again, two *Enterobacteriaceae* sub-groups are revealed, the obligate endosymbionts and the core enterics, with *Photorhabdus* and *Yersinia* clustering with the core enterics.

Legend

Figure 7: Phylogenetic relationships between 32 γ-proteobacteria inferred from 32 *aroE* allelic sequences. *Helicobacter pylori aroE* allelic sequence was used as outgroup. The multiple alignment of the nucleotide sequences of the *aroE* genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 33 *aroE* alleles X 942 nucleotides in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

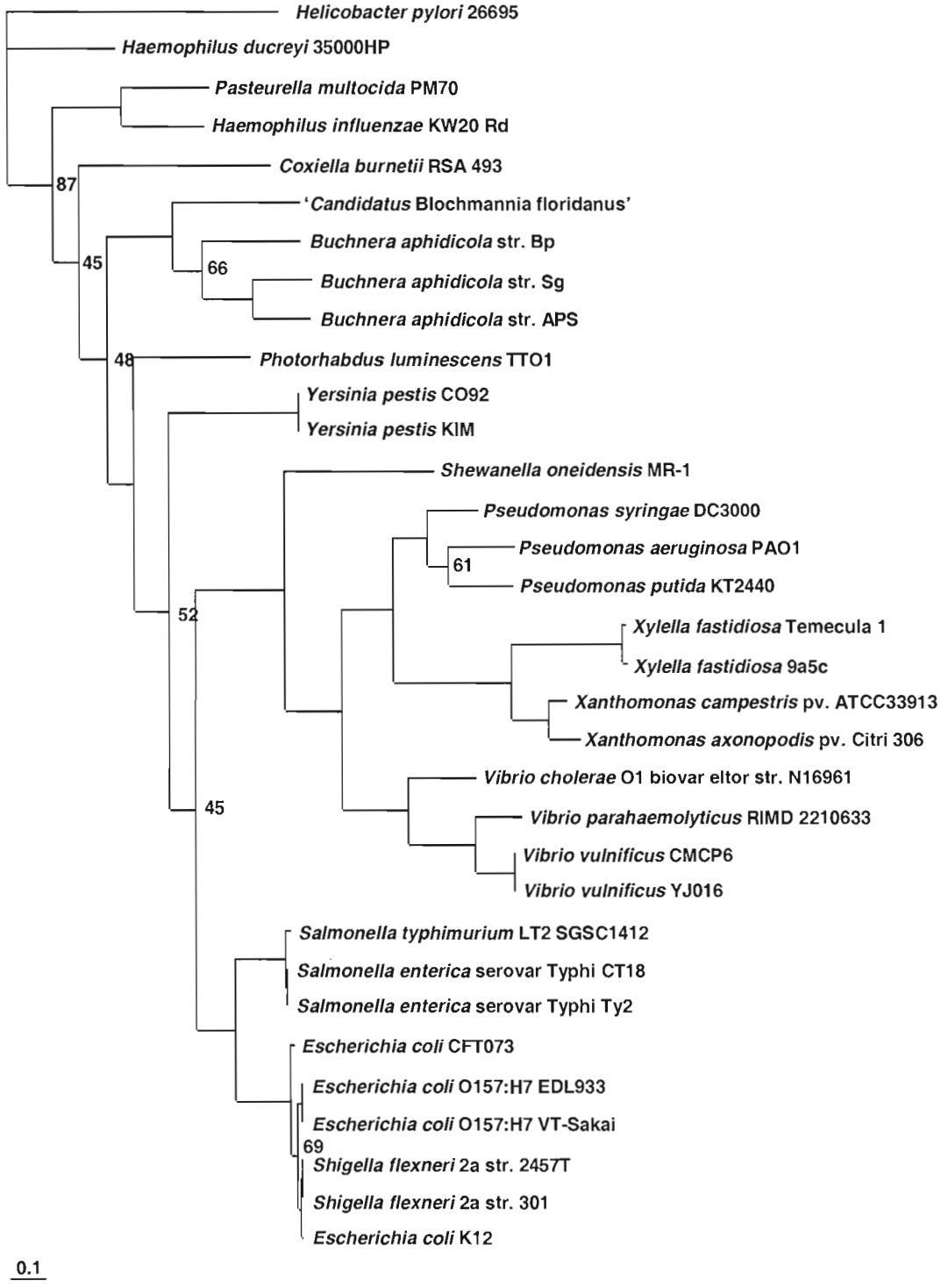


Figure 7 -Phylogenetic tree based on *aroE* gene sequences using ML method

Legend

Figure 8: Phylogenetic relationships between 31 γ-proteobacteria inferred from 31 *gdh* allelic sequences. *Helicobacter pylori gdh* allelic sequence was used as outgroup. The multiple alignment of the nucleotide sequences of the *gdh* genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 32 *gdh* alleles X 1560 nucleotides in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

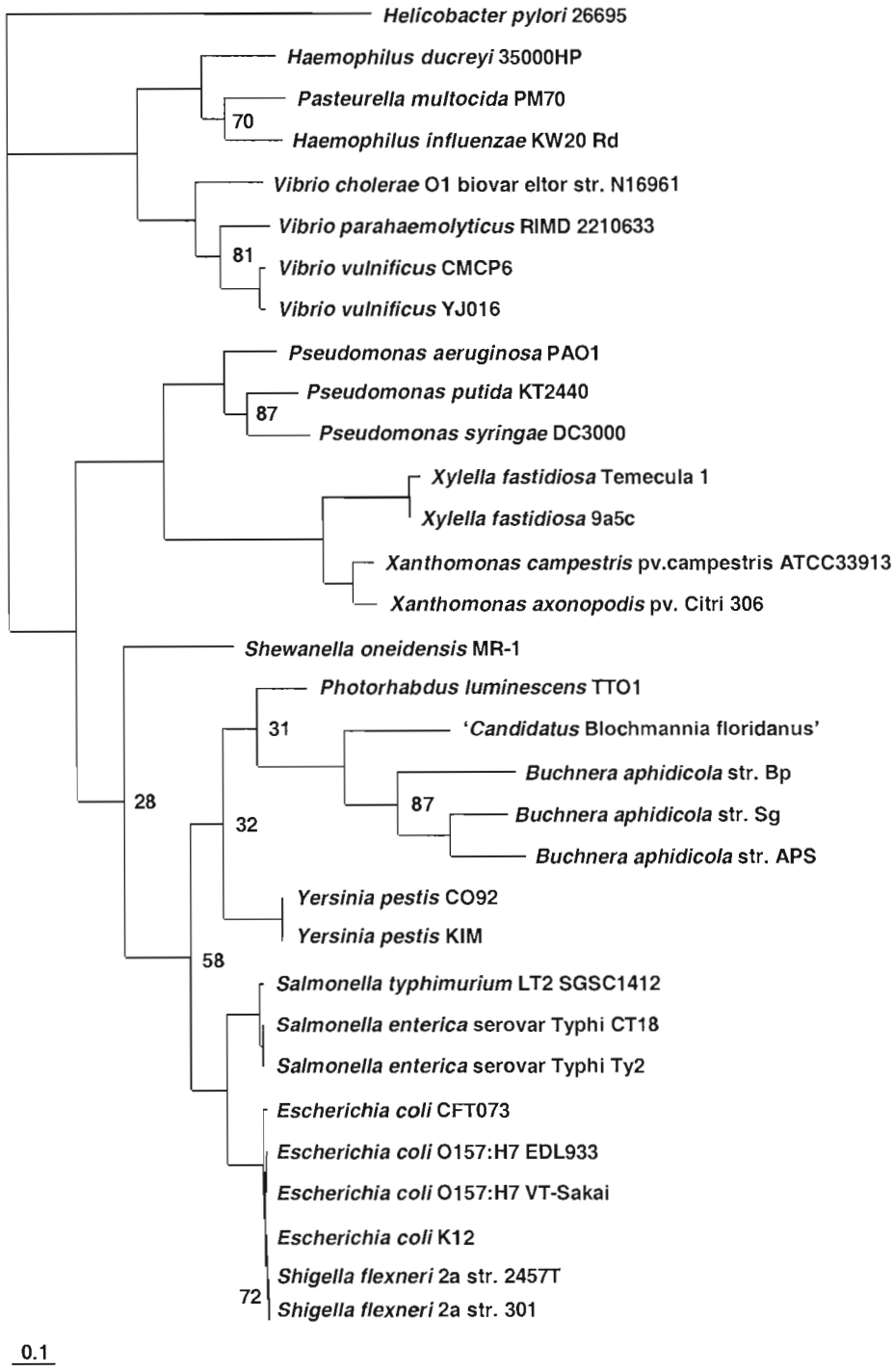


Figure 8 -Phylogenetic tree based on *gdh* gene sequences using ML method

Legend

Figure 9: Phylogenetic relationships between 31 γ-proteobacteria inferred from 31 *adk*, *aroE* and *gdh* concatenated sequences. *Helicobacter pylori* was used as an outgroup. The multiple alignment of the concatenated nucleotide sequences of the *adk*, *aroE* and *gdh* genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 32 *adk*, *aroE* and *gdh* alleles X 3170 nucleotides in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

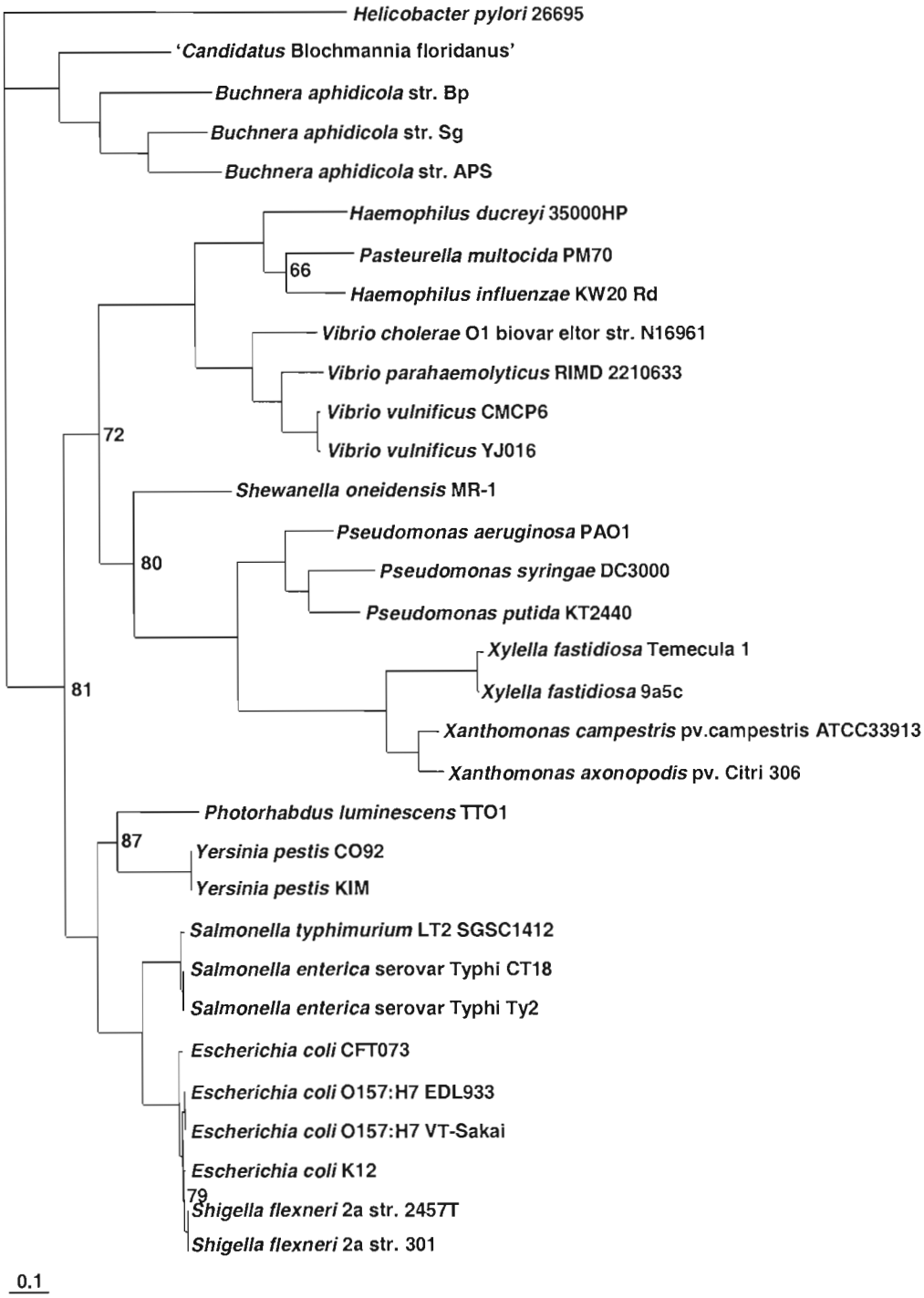


Figure 9 -Phylogenetic tree based on concatenated alignment using ML method

In addition to these nucleotide-based inferred phylogenies, phylogenies were inferred from the amino acid sequences of the 34 orthologous adenylate kinase, 33 shikimate dehydrogenase and 32 glucose-6-phosphate dehydrogenase proteins (see SUPPLEMENTARY DATA). Because of the degeneracy of the genetic code, whereby multiple codons may encode a single amino acid, the amino acid sequences are slightly more conserved among taxa than the corresponding nucleotide sequences. Nevertheless, the overall topologies of the Adk-, AroE-, Gdh- and concatenated amino acid sequence-inferred phylogenetic trees duplicate, in general, the topology of the *adk*-, *aroE*-, *gdh*- and concatenated nucleotide sequence-inferred phylogenetic trees.

Fig. 10 is inferred from 16S rRNA gene sequences. As indicated before, most γ-proteobacteria phylogeny is based on the rRNA sequences. Fig. 10 has to be regarded as the "accepted standard" for γ-proteobacteria phylogeny. Its overall topology was compared with the topology of each is similar to all other four figures based on house-keeping gene nucleotide sequences and found to be, in general, similar. Strains are grouped at the species level, species at the genus level, and genera at the family level. Some differences, however, are worth pointing out. Based on 16S rRNA sequences, *W. glossinidia brevipalpis* is closer to '*Candidatus* Blochmannia floridanus' whereas based on *adk* gene it was closer to *B. aphidicola* strain Bp. The two *Enterobacteriaceae* sub-groups revealed above with the house-keeping genes, sub-group 1 - the obligate endosymbionts -, and sub-group 2 - the core enterics, *Yersinia* and *Photorhabdus* - are now re-grouped into a single, more compact cluster, although the same two sub-groups can be revealed. Based on the scale provided in

Legend

Figure 10: Phylogenetic relationships between 33 γ-proteobacteria inferred from 33 16S rDNA allelic sequences. *Helicobacter pylori* was used as an outgroup. The multiple alignment of the nucleotide sequences of the 16S rRNA genes was done using ClustalW. The matrix generated after the corrected multiple alignment was 34 16S rRNA alleles X 1647 nucleotides in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).

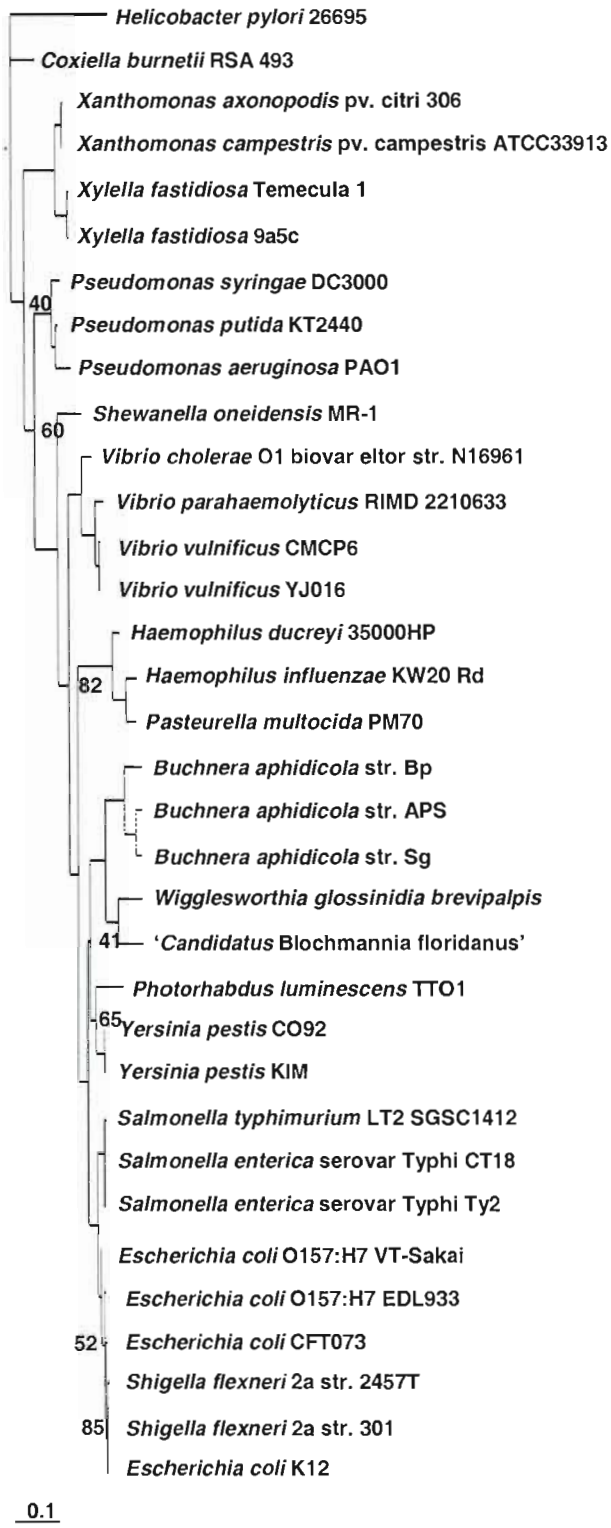


Figure 10 -Phylogenetic tree based on 16S rRNA gene sequences using ML method

each figure, it is very important to note that the 16S rRNA gene shows a lower percentage of nucleotide sequence divergence than the *adk*, *aroE* and *gdh* genes. The *adk*, *aroE* and *gdh* genes, and the house-keeping genes concatenated sequences, show a percentage of nucleotide sequence divergence up to 3-, 4-, 3.7- and 2.8-fold greater than the one of the 16S rRNA gene, respectively. These data should also be regarded in the light of the comparison of the discriminatory power (DP) of the 16S rRNA gene with the DP of the three house-keeping genes, *adk*, *aroE* and *gdh*, among the nine core *Escherichia-Shigella-Salmonella* enterics available. Any of the three house-keeping genes could better discriminate the nine core enterics than the 16S rRNA gene. Certainly, these house-keeping gene sequences could be used to achieve deeper levels of discrimination than the 16S rRNA gene. It would be interesting to study more strains from some species and more species from a given genus to assess the discriminating power of these house-keeping genes, compared with 16S rRNA genes, at the strain and species level. Another consequence of the higher percentage of nucleotide sequence divergence for house-keeping genes, compared with 16S rRNA genes, is that, based on the nucleotides sequences analysed here, three unique pairs of universal primers for the amplification of *adk*, *aroE* and *gdh* genes, respectively, in γ-proteobacteria, cannot be designed. The nucleotide sequences of these genes are not conserved enough at that taxa level. Primers, however, could potentially be designed at the sub-family level. For example, a pair of primers with the following sequences, P1: 5'-ATGCGTATCATTCTGCTTGG-3' and P2: 5'-TTAGCCGAGGATTTTTTCCA-3', would share 19/20 and 19/20 identical nucleotides with the 5' and 3' ends, respectively, of the *adk* gene of the core enterics, and could possibly be used for its amplification in the genera

Salmonella, *Escherichia* and *Shigella*. These primers, however, would share 19/20 but only 14/20 identical nucleotides with the 5' and 3' ends, respectively, of the *adk* gene in the closely related genus *Yersinia*. Likewise, a pair of primers with the following sequences, P1: 5'-ATGCGATTGGTTCTGTTGGGACC-3' and P2: 5'-GAATCGGTATAGACCTGCA-3', would share 22/23 and 18/19 identical nucleotides with the 5' and 3' ends, respectively, of the *adk* gene of the *Xanthomonadaceae*, and thus could possibly be used for the amplification of the *adk* gene in the genera *Xylella* and *Xanthomonas*. Other primers would have to be specifically designed for sub-families for each of the three house-keeping genes under study.

Of the six house-keeping genes used in MLST, three, *adk*, *aroE* and *gdh*, have been studied here for their capability to infer phylogenies compared to 16S rRNA genes. Although the data are not presented here, we also studied, albeit briefly, three more house-keeping genes used in MLST, *abcZ*, *pdhC* and *pgm*. The *abcZ* gene was purposely left out because of complexity and confusion in the nomenclature of putative ABC transporters. For *pdhC*, the gene nucleotide sequences were retrieved from the 33 γ-proteobacteria species and strains under study. However, no *pdhC* gene could be retrieved from the ε-proteobacterium *Helicobacter pylori* 26695 which served as the outgroup, nor from any *Helicobacter* species nor from any ε-proteobacteria for which the full genome had been sequenced. These bacteria are microaerophilic and don't have the aerobic pyruvate dehydrogenase. For *pgm*, nucleotide sequences were retrieved from 25 γ-proteobacteria species and strains. No *pgm* gene could be retrieved from each of the three *Buchnera* strains, *Coxiella burnetii* RSA 493,

Haemophilus ducreyi 35000HP, *Haemophilus influenzae* KW20 Rd, *Pasteurella multocida* PM70 and *Pseudomonas aeruginosa* PAO1. Clearly, a problem associated with some of these house-keeping genes, and possibly with other house-keeping genes, is that they are not universally distributed. When present, however, they can prove very useful. Here, they have reconstructed the γ-proteobacteria phylogeny, very similar to the one inferred from 16S rRNA genes, at the family, genus, species and strain levels. Owing to their higher rate of nucleotide sequence substitutions, they can probe deeper branches of a phylogenetic tree than 16S rRNA genes. In addition, because they are used in MLST, the number of nucleotide sequences publicly available for many taxa is expected to increase rapidly over time, thus increasing the number of potential phylogenetic analysis. These house-keeping genes will prove very useful either at complementing 16S rRNA-inferred phylogenies or for specific, targeted, phylogenetic analyses.

CONCLUSION

We have analysed the heterogeneity in 16S rRNA gene sequences of 175 alleles from 33 strains covering 23 species and 16 genus of the γ-proteobacteria. And then of the six house-keeping genes used in MLST, three, *adk*, *aroE* and *gdh*, have been studied here for their capability to infer phylogenies compared to 16S rRNA genes.

Molecular phylogenies for the 33 γ–proteobacteria have been reconstructed. Sequences from the adenylate kinase gene (*adk*), shikimate dehydrogenase (*aroE*), glucose-6-phosphate dehydrogenase (*gdh*), the concatenated *adk*, *aroE* and *gdh*, and 16S rRNA gene, respectively, were used to infer individual gene trees.

We have obtained the following results:

- A total of 175 16S rRNA genes nucleotide sequences were retrieved from GenBank for the 33 γ-proteobacteria species and strains.
- The number of 16S rRNA alleles varies from 1 to 10.
- The phylogenetic tree of all 33 γ-proteobacteria showed that 16S rRNA allelic sequences were clustered within genera and species, except for *Escherichia coli* and *Shigella flexneri*.
- Random selection of 16S rRNA alleles within strain would not generate different phylogenetic trees for the γ-proteobacteria.
- A total of 33 adenylate kinase (*adk*), 32 shikimate dehydrogenase (*aroE*), 31 glucose-6-phosphate dehydrogenase (*gdh*) and 33 16S rRNA genes nucleotide sequences were

retrieved from GenBank for the 33 γ-proteobacteria species and strains.

- In general, phylogenies inferred from the three house-keeping gene sequences (Figs. 6, 7 and 8) and their concatenated sequence (Fig. 9) were similar to the one inferred from 16S rRNA gene sequence (Fig. 10) with regards to strains grouping at the species level, species grouping at the genus level, genera grouping at the family level, with exceptions. Some of these phylogenies revealed two *Enterobacteriaceae* sub-groups, such as the phylogenies inferred from the *adk* (Fig. 6), *aroE* (Fig. 7) and concatenated *adk*, *aroE* and *gdh* gene sequences (Fig. 9) whereas the other phylogenies (Figs. 8 and 10) form a single *Enterobacteriaceae* cluster.
- Based on the scale provided in each figure (Figs. 6, 7, 8, 9 and 10) the house-keeping genes *adk*, *aroE* and *gdh* show a higher percentage of nucleotide sequence divergence than the 16S rRNA gene.
- Any of the three house-keeping genes has a lower percentage of gene sequence similarity than the 16S rRNA gene and can better distinguish species among the nine core *Escherichia-Shigella-Salmonella* enterics.
- Compared with 16S rRNA genes, because of the higher percentage of nucleotide sequence divergence for house-keeping genes, three unique pairs of universal primers for the amplification of *adk*, *aroE* and *gdh* genes, respectively, in γ-proteobacteria, cannot be designed. However, primers could be designed at the sub-family level.
- Three more house-keeping genes used in MLST, *abcZ*, *pdhC* and *pgm,* are not universally distributed.

In conclusion, for the γ-proteobacteria studied here, the selection of a single 16S rRNA allele is sufficient for inferring phylogenies. Because of the high homology between intra-strain alleles, different alleles from same strains would yield similar phylogenies. And also *adk*, *aroE*, *gdh* gene sequences and their concatenated sequences have reconstructed the γ-proteobacteria phylogeny, very similar to the one inferred from 16S rRNA genes, at the family, genus, species and strain levels, and have proved to be very useful at complementing 16S rRNA-inferred phylogenies. Because the three house-keeping genes *adk, aroE* and *gdh* under study have a higher rate of nucleotide sequence substitutions, they can probe deeper branches of a phylogenetic tree than 16S rRNA genes and will be very useful for specific, targeted, phylogenetic analyses. However, because the nucleotides sequences of these house-keeping genes are not as highly conserved among γ-proteobacteria, family- or genus-specific primers would need to be designed for the amplification of any of these genes.

In addition, several house-keeping genes have been used for phylogenetic analyses. The novel aspect of this study is the use of new house-keeping genes, which have been used in MLST, to infer phylogenies for the γ-proteobacteria.

Since these house-keeping genes are used in MLST, the number of nucleotide sequences publicly available for many taxa is expected to increase rapidly over time, thus certainly increasing the number of potential phylogenetic analysis.

Further study should focus on more strains from some species and more species from a given genus to assess the discriminating power of these house-keeping genes, compared with 16S rRNA genes, at the strain and species level.

APPENDIX A

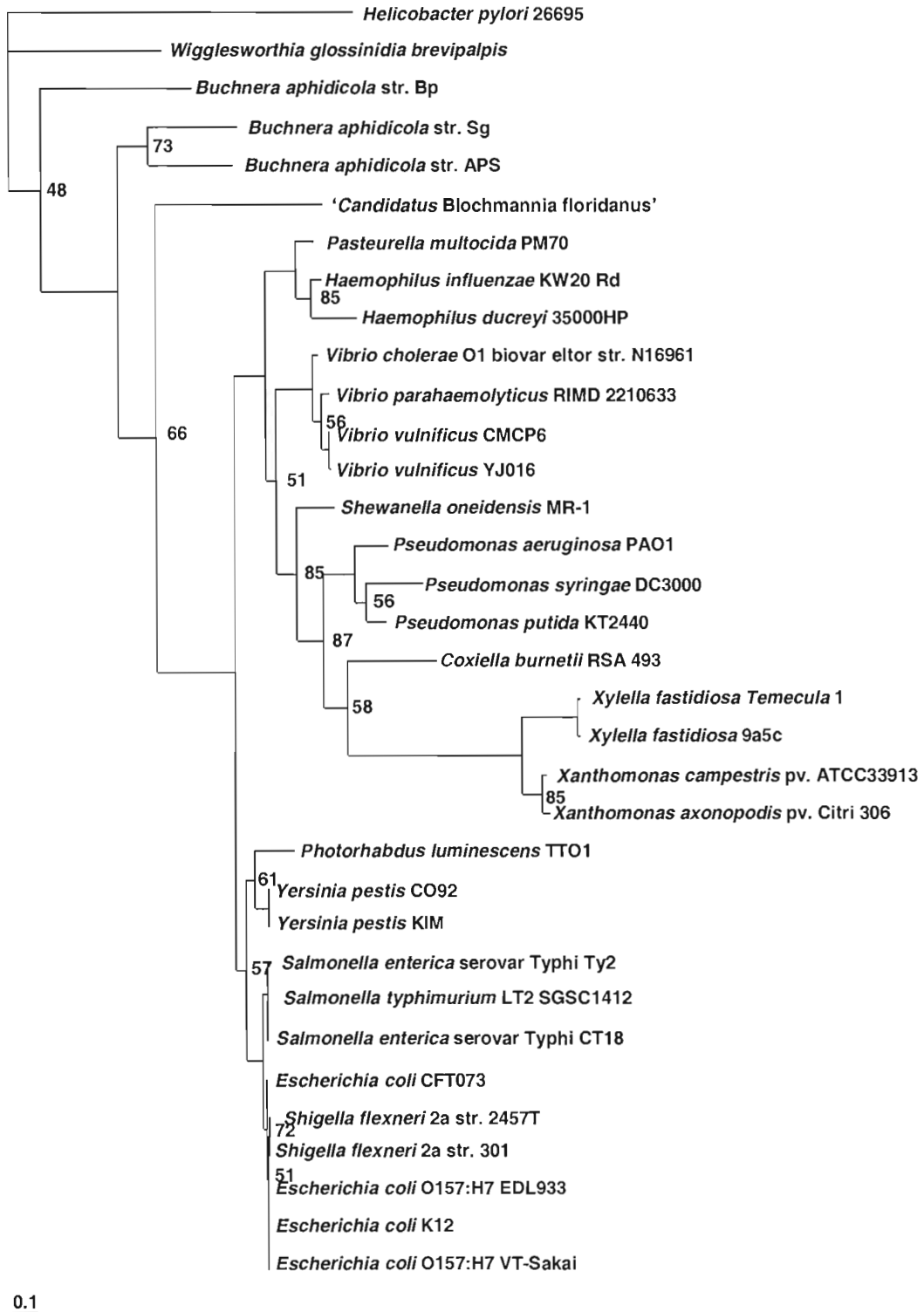
PHYLOGENETIC TREES BASED ON AMINO ACID SEQUENCE COMPARISONS OF THE 33 ADK, 32 AROE, 31 GDH PROTEINS AND THE 31 CONCATENATED SEQUENCES

In addition to these nucleotide-based inferred phylogenies, we have constructed four more phylogenies inferred from the amino acid sequences comparisons of the 33 orthologous adenylate kinase, 32 shikimate dehydrogenase, 31 glucose-6-phosphate dehydrogenase proteins and the 31 concatenated sequences.

The degeneracy of the genetic code, whereby multiple codons may encode a single amino acid, causes the amino acid sequences to be slightly more conserved among taxa than the corresponding nucleotide sequences. Nevertheless, the overall topologies of the Adk-, AroE-, Gdh- and concatenated amino acid sequence-inferred phylogenetic trees duplicate, in general, the topology of the *adk*-, *aroE*-, *gdh*- and concatenated nucleotide sequence-inferred phylogenetic trees (Figs. 6, 7, 8 and 9).

Legend

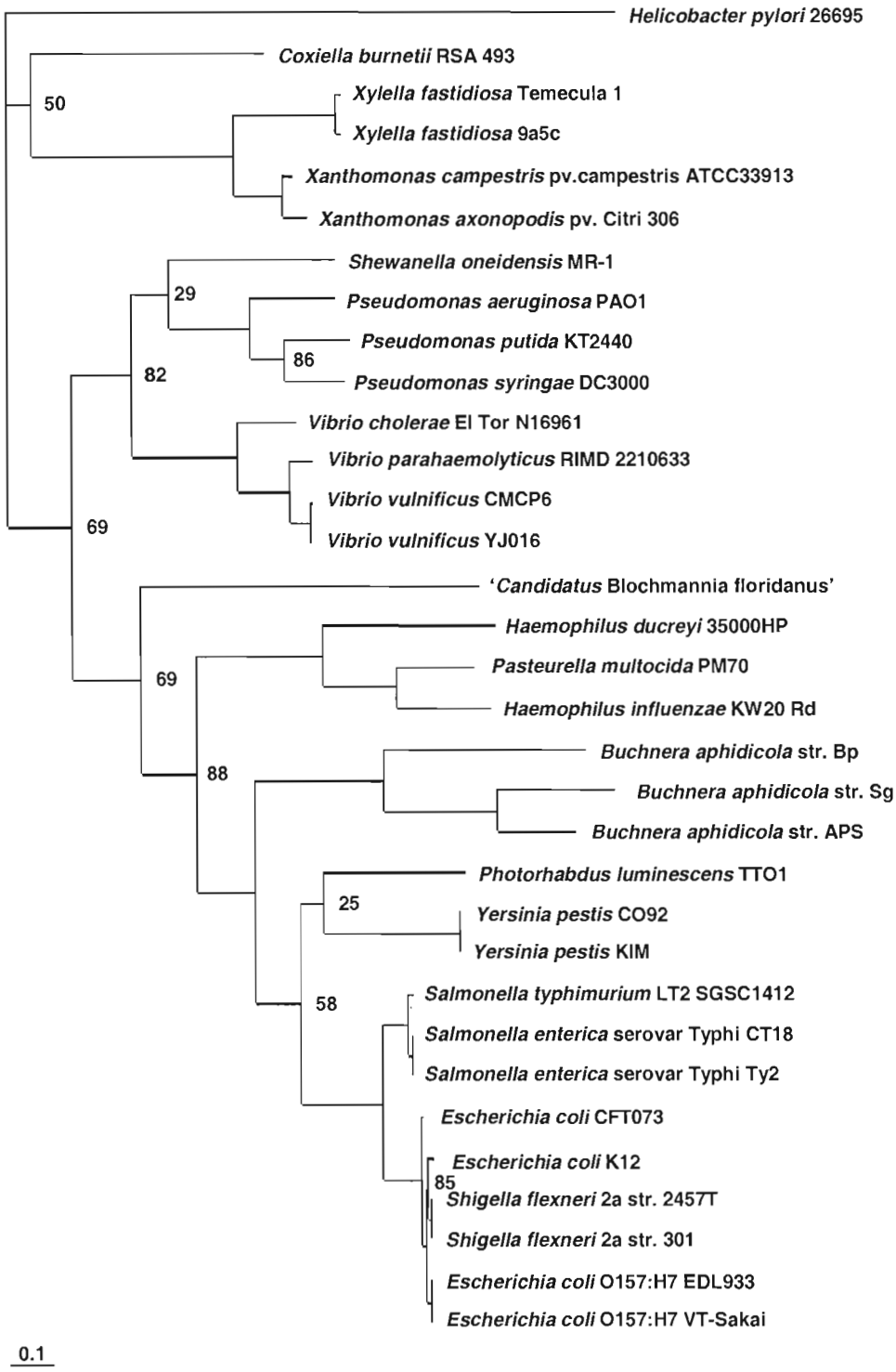
Appendix A.1: Phylogenetic relationships between 33 γ-proteobacteria inferred from 33 Adk allelic sequences. *Helicobacter pylori* Adk allelic sequence was used as outgroup. The multiple alignment of the amino acid sequences of the Adk proteins was done using ClustalW. The matrix generated after the corrected multiple alignment was 34 Adk alleles X 242 amino acids in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).



A.1: Phylogenetic tree based on Adk amino acid sequences using ML method

Legend

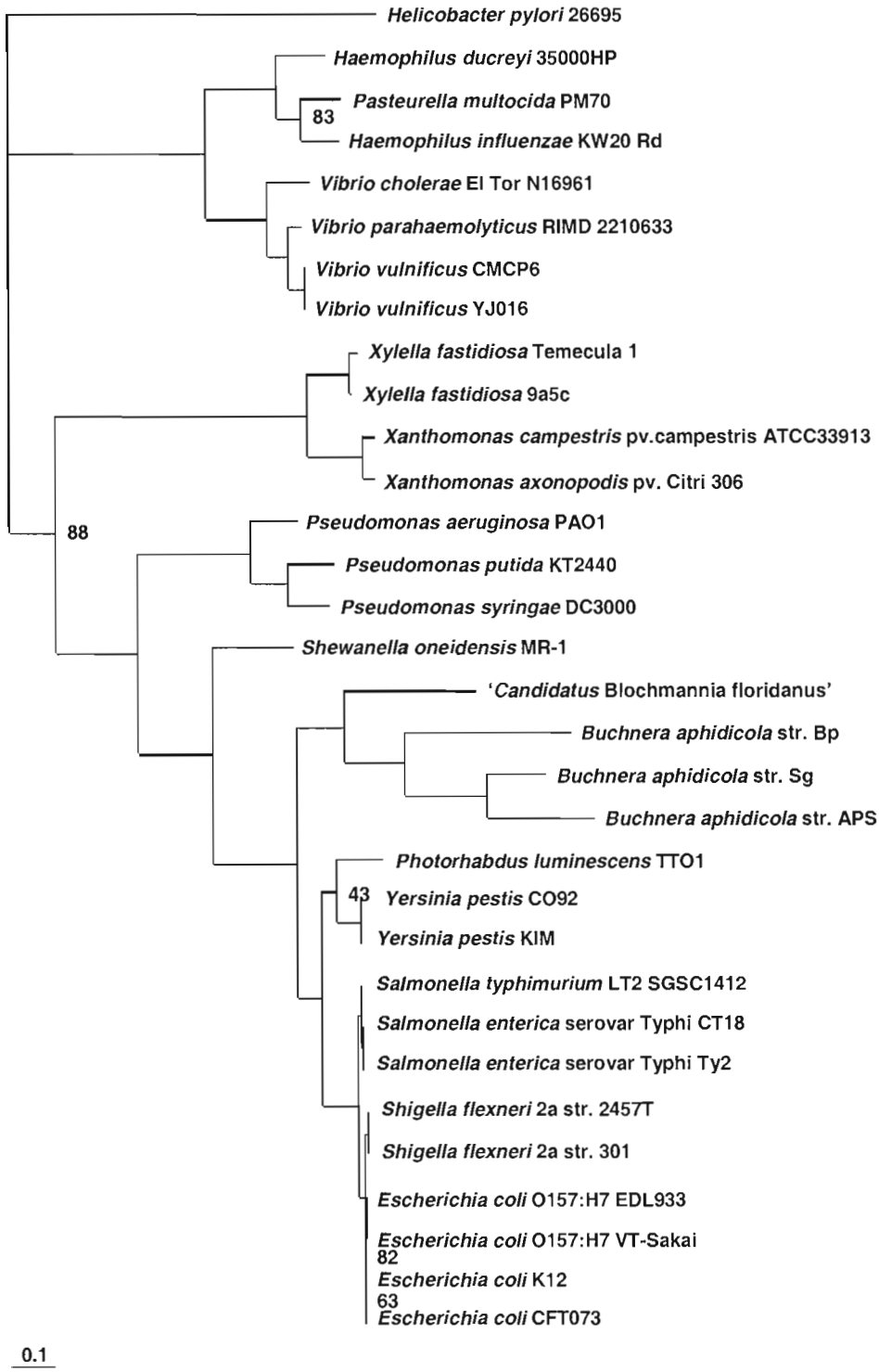
Appendix A.2: Phylogenetic relationships between 32 γ-proteobacteria inferred from 32 AroE allelic sequences. *Helicobacter pylori* AroE allelic sequence was used as outgroup. The multiple alignment of the amino acid of the AroE proteins was done using ClustalW. The matrix generated after the corrected multiple alignment was 33 AroE alleles X 311 amino acids in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).



A.2: Phylogenetic tree based on AroE amino acid sequences using ML method

Legend

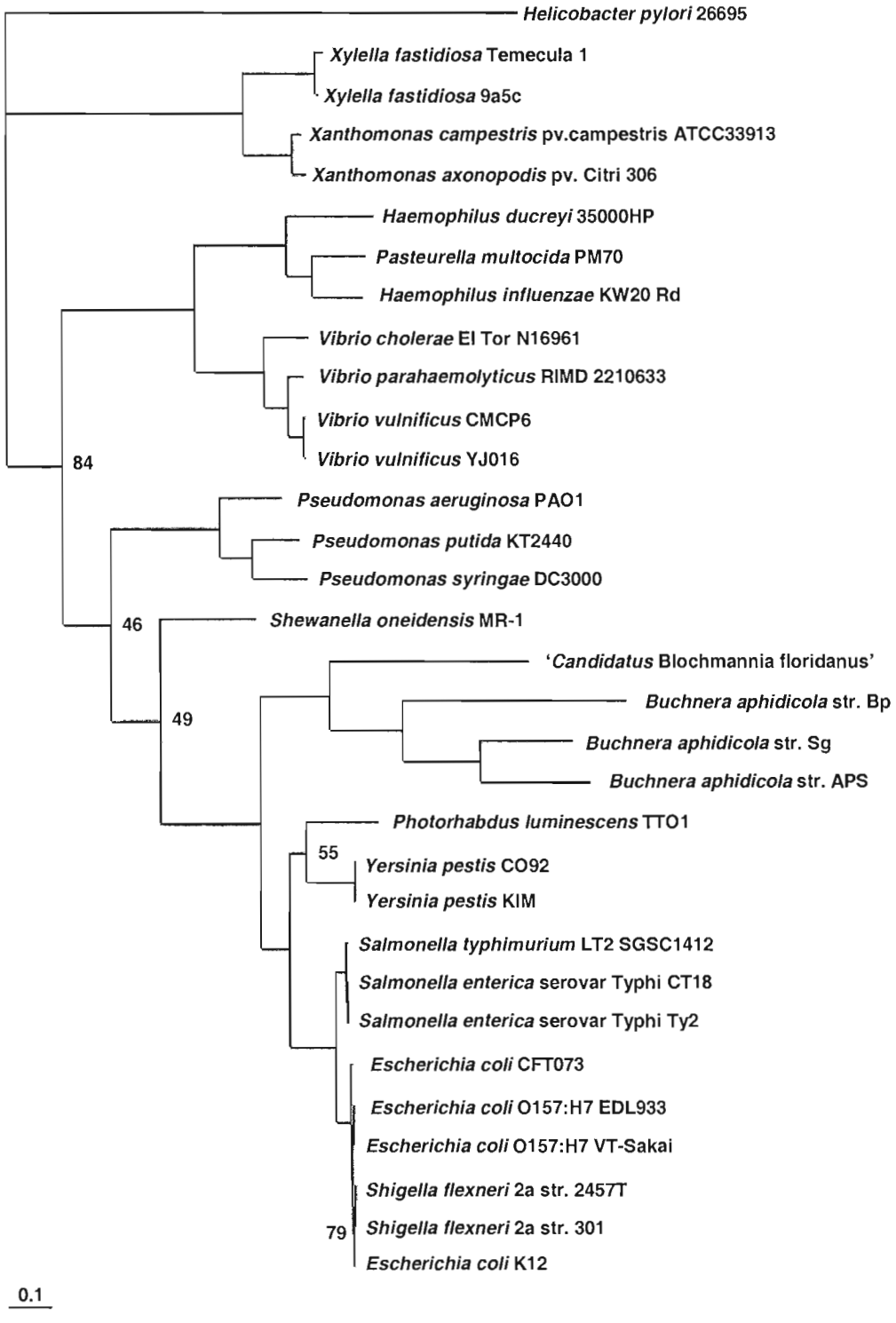
Appendix A.3: Phylogenetic relationships between 31 γ-proteobacteria inferred from 31 Gdh allelic sequences. *Helicobacter pylori* Gdh allelic sequence was used as outgroup. The multiple alignment of the amino acid sequences of the Gdh proteins was done using ClustalW. The matrix generated after the corrected multiple alignment was 32 Gdh alleles X 521 amino acids in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).



A.3: Phylogenetic tree based on Gdh amino acid sequences using ML method

Legend

Appendix A.4: Phylogenetic relationships between 31 γ-proteobacteria inferred from 31 Adk, AroE and Gdh concatenated sequences. *Helicobacter pylori* was used as an outgroup. The multiple alignment of the concatenated amino acid sequences of the Adk, AroE and Gdh proteins was done using ClustalW. The matrix generated after the corrected multiple alignment was 32 Adk, AroE and Gdh alleles X 1095 amino acids in size. The tree was generated using the maximum likelihood method. Numbers indicate bootstrap values lower than 90% (of 1000 cycles).



A.4: Phylogenetic tree based on Adk, AroE and Gdh amino acid sequences using ML method

REFERENCES

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., and Polz, M. F. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.*, **186**: 2629-2635.

Angen O., Ahrens, P., Kuhnert, P., Christensen, H., and Mutters, R. 2003. Proposal of *Histophilus somni* gen. nov., sp. nov. for the three species *incertae sedis* '*Haemophilus somnus*', *Haemophilus agni*' and '*Histophilus ovis*'. *Int. J. Syst. Evol. Microbiol.*, **53:** 1449-1456.

Anzai, Y., Kim, H., Park, J.-Y., Wakabayashi, H., and Oyaizu, H. 2000. Phylogenetic affiliation of the *pseudomonads* based on 16S rRNA sequence. *Int. J. Syst. Evol. Microbiol.*, **50**: 1563-1589.

Baumann, P., Baumann, L., Lai, C.-Y., Roubakhsh, D., Moran, N. A., and Clark, M. A., (1995) Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: Intracellular symbionts of aphids. *Ann. Rev. Microbiol.*, **49**, 55-94.

Baumann, P., Baumann, L., Clark, M. A., and Thao, M. L., (1998) *Buchnera aphidicola*: The endosymbiont of aphids. *ASM News*, **64**, 203-209.

Bosshard, P. P., Zbinden, R., and Altwegg, M. (2002) *Paenibacillus turicensis* sp. nov., a novel bacterium harbouring heterogeneities between 16S rRNA genes. *Int. J. Syst. Evol. Microbiol.*, **52**, 2241-2249.

Brenner, D. J., Fanning, G. R., Miklos, G. V., and Steigerwalt, A. G. (1973) Polynucleotide sequence relatedness among *Shigella* species. *Int. J. Syst. Bacteriol.*, **23**, 1-7.

Brenner, D. J., Fanning, G. R., Skerman, F. J., and Falkow, W. S. (1972) Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J. Bacteriol.*, **109**, 953-965.

Brenner, D. J., Steigerwalt, A. G., Gail Wathen, H., Gross, R. J., and Rowe, B. (1982) Confirmation of aerogenic strains of *Shigella boydii* 13 and further study of *Shigella* serotypes by DNA relatedness. *J. Clin. Microbiol.*, **16**, 432-436.

Bruns, T. D., White, T. J., and Taylor, J. W. (1991) Fungal molecular systematics. *Annu. Rev. Eco. Syst.*, **22**, 525-264.

Cavalier-Smith, T. (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.*, **52**, 7-76.

Cilia, V., Lafay, B., and 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.

Clark, A. G., and Whittam, T. S. (1992) Sequencing errors and molecular evolutionary analysis. *Mol. Biol. Evol.*, **9**, 744-752.

Clayton, R. A., Sutton, G., Hinkle, P. S. Jr., Bult, C., and 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.

Coenye, T., and Vandamme, P. (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett.*, **228**, 45-49.

De Ley, J. (1992) The proteobacteria: ribosomal RNA cistron similarities and bacterial taxonomy. *In* The prokaryotes, 2nd ed., ed. by Balows, A., Trüper, H. G., Dworkin, M., Harder, W., and Schleifer, K. H., Springer Verlag, Germany, pp. 2111-2140.

Drozanski, W. J. (1991) *Sarcobium lyticum* gen. nov., sp. nov., an obligate intracellular bacterial parasite of small free-living amoebae. *Int. J. Syst. Bacteriol.*, **41**, 82-87.

Dunlap, P. V., and Kita-Tsukamoto, K. (2001) Luminous bacteria, chapter 328. *In* M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt (ed.), The Prokaryotes, an evolving electronic resource for the microbiological community. Academic Press, New York, N.Y.

Finkmann, W., Altendorf, K., Stackebrandt, E., and Lipski, A. (2000) Characterization of N_2O-producing *Xanthomonas*-like isolates from biofilters as *Stenotrophomonas nitritireducens* sp. nov., *Luteimonas mephitis* gen. nov., sp. nov. and *Pseudoxanthomonas broegbernensis* gen. nov., sp. nov. *Int. J. Syst. Evol.* Microbiol., **50**, 273-282.

Fitz-Gibbon, S. T., and House, C. H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, **27**, 4218-4222.

Fleischmann, R. D., Adams, M. D., White, O. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496-512.

Fox, G. E., Stackebrandt, E., Hespell R. B., Gibson J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. (1980) The phylogeny of prokaryotes. *Science*, **209**, 457-463.

Gaunt, M. W., Turner, S. L., Rigottier-Gois, L., Lloyd- Macgilp, S. A., and Young, J. P. W. (2001) Phylogenies of *atp*D and *recA* support the small subunit rRNA-based classification of rhizobia. *Int. J. Syst. Evol. Microbiol.*, **51**, 2037-2048.

Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696-704.

Gupta, R. S. (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol. Rev.*, **24**, 367-402.

Gürtler, V., and Stanisich V. A. (1996) New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. *Microbiology*, **142**, 3-16.

Hedegaard, J., Steffensen, S. A., Norskow-Lauritsen, N., Mortensen, K. K., and Sperling-Petersen, H. U. (1999) Identification of *Enterobacteriaceae* by partial sequencing of the gene encoding translation initiation factor 2. *Int. J. Syst. Bacteriol.*, **49**, 1531-1538.

Hill, C., W., and Harnish, B., W. (1981) Inversions between ribosomal RNA genes of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **78,** 7069-7072.

Hillis, D. M., Moritz, C., Porter, C. A., and Baker, R. J. (1991) Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science*, **251**, 308-310.

Holt, J. G., Krieg, N. R., Sneath, P. H. A., Staley, J. T., and Williams, S. T. (1994) Bergey's Manual of Determinative Bacteriology, 9th ed., Williams and Wilkins, Baltimore, MD, pp. 65-69.

Hookey, J. V., Saunders, N. A., Fry, N. K., Birtles, R. J., and Harrison, T. G. (1996) Phylogeny of *Legionellaceae* based on small-subunit ribosomal DNA sequences and proposal of *Legionella lytica* comb. nov. for *Legionella*-like amoebal pathogens. *Int. J. Syst. Bacteriol.*, **46**, 526–531.

Kersters, K., Ludwig, W., Vancanneyt, M., De Vos, P., Gillis, M., and Schleifer, K.-H. (1996) Recent changes in the classification of the *Pseudomonads*: An overview. *Syst. Appl. Microbiol.*, **19**, 465-477.

Kimura, M. (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111-120.

Klappenbach, J. A., Saxman, P. R., Cole, J. R., and Schmidt, T. M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res.*, **29**, 181-184.

Kolbert, C. P., and Persing, D. H. (1999) Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr. Opin. Microbiol.*, **2**, 299-305.

Lane, D. L., Pace, B., Olsen, G. J., Stahl, D., Sogin, M. L., and Pace, N. R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analysis. *Proc Nat Acad. Sci. USA,* **82,** 6955-6959.

Lawrence, J. G., Ochman, H., and Hartl, D. L. (1991) Molecular and evolutionary relationships among enteric bacteria, *J. Gen. Microbiol.*, **137**, 1911-1921.

Lawrence, J. G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.*, **2**, 519-523.

Lee, H.-Y., and Côté, J.-C. (2006) Phylogenetic analysis of γ-proteobacteria inferred from nucleotide sequence comparisons of the house-keeping genes *adk*, *aroE* and *gdh*: Comparisons with phylogeny inferred from 16S rRNA gene sequences. *J. Gen. Appl. Microbiol.*, **52**, 147-158.

Lerat, E., Daubin, V., and Moran, N. A. (2003) From gene trees to organismal phylogeny in Prokaryotes: the case of the γ-proteobacteria. *PLoS Biol.*, **1**, E19.

Li, W. H., and Graur, D. (1991) Fundamentals of Molecular Evolution. Sinauer, Mass. Lipman, D.J., Altschul, S.F., and Kececioglu, J.D. (1989) *Proc. Natl. Acad. Sci. USA*, **86,** 4412- 4415.

Liefting, L. W., Andersen, M. T., Beever, R. E., Gardner, R. C., and Forster, R. L. (1996) Sequence heterogeneity in the two 16S rRNA genes of *Phormium* yellow leaf phytoplasma. *Appl. Environ. Microbiol.*, **62**, 3133-3139.

Loughney, K., Lund, E., and Dahlberg. J. E. (1983) Deletion of an rRNA gene set in *Bacillus subtilis*. *J. Bacteriol.*, **154**, 529-532.

Ludwig, W., Neumaier, J., Klugbauer, N., Brockmann, E., Roller, C., Jilg, S., Reetz, K., Schachtner, I., Ludvigsen, A., Bachleitner, M., Fischer, U., and Schleifer, K. H. (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.

Ludwig, W., and Klenk H. –P. (2001) Overview: A phylogenetic backbone and taxonomic framework for procaryotic systematics *In:* G. Garrity (Ed.) *Bergey's Manual of Systematic Bacteriology, 2nd ed.* Springer-Verlag Baltimore, MDvol. 1, pp. 49-65.

MacDonell, M. T., and Colwell, R. R. (1985) Phylogeny of the *Vibrionaceae*, and recommendation of two new genera, *Listonella* and *Shewanella. Syst. Appl. Microbiol.* **6**, 171-182.

Maiden, M. C. J., Bygraves, J. A., Feil, E. J., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zurth, K., Caugant, D., Feavers, I. M., Achtman, M., and Spratt, B. G. (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.

Marchandin, H., Teyssier, C., Simeon de Buochberg, M., Jean-Pierre, H., Carriere, C., and Jumas-Bilak, E. (2003) Intra-chromosomal heterogeneity between the four 16S rRNA gene copies in the genus *Veillonella*: implications for phylogeny and taxonomy. *Microbiology*, **149**, 1493-1501.

Martínez-Murcia, A. J., Anton, A. I., and Rodriguez-Valera, F. (1999) Patterns of sequence variation in two regions of the 16S rRNA multigene family of *Escherichia coli. Int. J. Syst. Bacteriol.*, **49**, 601-610.

Mollet, C., Drancourt, M., and Raoult, D. (1997) *rpoB* sequence analysis as a novel basis for bacterial identification. *Mol. Microbiol.*, **26**, 1005-1011.

Moran, N. A., Munson, M. A., Baumann, P., and Ishikawa, H. (1993) A molecular clock in endosymbiotic bacteria is calibrated using the insect host. *Proc. Royal Soc. London Ser.*

B Biol. Sci., **235**, 67-171.

Myers, C. R. and Nealson, K. H. (1988) Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor. *Science*, **240**, 1319-1321.

Ninet, B., Monod, M., Emler, S., Pawlowski, J., Metral, C., Rohner, P., Auckenthaler, R., and Hirschel, B. (1996) Two different 16S rRNA genes in a mycobacterial strain. *J. Clin. Microbiol.*, **34**, 2531-2536.

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

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299-304.

Olivier, A., Lee, H. Y., and Côté, J.-C. (2005) Study of the heterogeneity of 16S rRNA genes in γ-proteobacteria: implications for phylogenetic analysis. *J. Gen. Appl. Microbiol.*, **51**, 395-405.

Pace, N. R., Stahl, D. A., Lane, D. L., and Olsen, G. J. (1986) The analysis of natural microbial populations by rRNA sequences. *Adv. Microbiol. Ecol.*, **9**, 1-55.

Page, R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Comp. Appl. Biosci.*, **12**, 357-358.

Palleroni, N. J. (1984) *Pseudomonas* Migula 1894, 237 *In*: N. R. Krieg and J. G. Holt (Eds.) Bergey's Manual of Systematic Bacteriology Williams and Wilkins Baltimore, MDvol. 1, pp. 141-199.

Parker, M. A., Lafay, B., Burdon, J. J., and van Berkum, P. (2002) Conflicting phylogeographic patterns in rRNA and *nifD* indicate regionally restricted gene transfer

in *Bradyrhizobium*. *Microbiology*, **148,** 2557-2565.

Petrovskis, E. A., Vogel, T. M., and Adriaens, P. (1994) Effects of electron acceptors and donors on transformation of tetrachloromethane by *Shewanella putrefaciens* MR-1. *FEMS. Microbiol. Lett.*, **121**, 357-363.

Pettersson, B., Bolske, G., Thiaucourt, F., Uhlen, M., and Johansson, K. E. (1998) Molecular evolution of *Mycoplasma capricolum* subsp. *capripneumoniae* strains, based on polymorphisms in the 16S rRNA genes. *J. Bacteriol.*, **180**, 2350-2358.

Rahn, O. (1937) New principles for the classification of bacteria. *Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg., Abt.*, **2; 96**, 273-286.

Reischl, U., Feldmann, K., Naumann, L., Gaugler, B. J. M., Ninet, B., Hirschel, B., and Emler, S. (1998) 16S rRNA sequence diversity in *Mycobacterium celatum* strains caused by presence of two different copies of 16S rRNA gene. *J. Clin. Microbiol.*, **36**, 1761-1764.

Saitou, N., and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406-425.

Sauer, C., Stackebrandt, E., Gadau, J., Hölldobler, B., and Gross, R. (2000) Systematic relationships and cospeciation of bacterial endosymbionts and their carpenter ant host species: proposal of the new taxon *Candidatus* Blochmannia gen. nov. *Int. J. Syst. Evol. Microbiol.*, **50**, 1877-1886.

Schouls, L. M., Schot, C. S., and Jacobs, J. A. (2003) Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the *Streptococcus anginosus* Group. *J. Bacteriol.*, **185**, 7241-7246.

Semple, K. M., and Westlake, D. W. S. (1987) Characterization of iron reducing *Alteromonas putrefaciens* strains from oil field fluids. *Can. J. Microbiol.*, **35**, 925-931.

Simpson, A. J. G., Reinach, F. C., Arruda, P. *et al.* (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, **406**, 151–157.

Stackebrandt, E. (1995) Bacterial phylogeny *In:* Molecular Basis of Virus Evolution. A. J. Gibbs, C. H. Calisher, and F. Garcia-Arenal (Eds.) Academic Press London, UK, pp. 15-28.

Stackebrandt, E., Murray, R. G. E., and Trüper, H. G. (1988) Proteobacteria classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives". *Int. J. Syst. Bacteriol.*, **38**, 321-325.

Stackebrandt, E., Witt, D., Kemmerling, C., Kroppenstedt, R., and Liesack, W. (1991) Designation of streptomycete 16S and 23S rRNA-based target regions for oligonucleotide probes. *Appl. Environ. Microbiol.*, **57**, 1468-1477.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., Nesme, X., Rossello-Mora, R., Swings, J., Truper, H. G., Vauterin, L., Ward, A. C., and Whitman, W. B. (2002) Report of the ad hoc committee for the reevaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.*, **52,** 1043-1047.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Turner, S. L., and Young, J. P. W. (2000) The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. *Mol. Biol. Evol.*, **17**, 309-319.

Ueda, K., Seki, T., Kudo, T., Yoshida, T., and Kataoka, M. (1999) Two Distinct Mechanisms Cause Heterogeneity of 16S rRNA. *J. Bacteriol.*, **181**, 78-82.

Ward, D., Weller, R., and Bateson, M. M. (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, **345,** 63-65.

Willems, A., De Ley, J., Gillis, M., and Kersters, K. (1991) *Comamonadaceae*, a new family encompassing the acidovorans rRNA complex, including *Variovorax paradoxus* gen. nov., comb. nov., for Alcaligenes paradoxus (Davis 1969). *Int. J. Syst. Bacteriol.*, **41**, 445-450.

Woese, C. R. (1987) Bacterial Evolution. *Microbiol. Rev.*, **51**, 221-271.

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

Woese, C. R., Stackebrandt, E., Macke, T. J., and Fox, G. E. (1985a) A phylogenetic definition of the major eubacterial taxa. *Syst. Appl. Microbiol.*, **6**, 143-151.

Woese, C. R., Stackebrandt, E., Weisburg, W. G., Paster, B. J., Madigan, M. T., Fowler, V. J., Hahn, C. M., Blanz, P., Gupta, R., Nealson, K. H., and Fox; G. E. (1984a) The phylogeny of purple bacteria: the alpha subdivision. *Syst. Appl. Microbiol.*, **5**, 315-326.

Woese, C. R., Weisburg, W. G., Hahn, C. M., Paster, B. J., Zablen, L. B., Lewis, B. J., Macke, T. J., Ludwig, W., and Stackebrandt, E. (1985b) The phylogeny of purple bacteria: the gamma subdivision. *Syst. Appl. Microbiol.*, **6**, 25-33.

Woese, C. R., Weisburg, W. G., Paster, B. J., Hahn, C. M., Tanner, R. S., Krieg, N. R., Koops, H.-P., Harms, H., and Stackebrandt, E. (1984b) The phylogeny of purple bacteria: the beta subdivision. *Syst. Appl. Microbiol.*, **5**, 327-336.

Yamamoto, S., and 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.

Yap, W. H., Zhang, Z., and Wang, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomyce *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.*, **181**, 5201-5209.

Zinder, S. H. (1998) Bacterial diversity. *In* Topley and Wilson's Microbiology and Microbial Infections, Vol. 2, Systematic Bacteriology, 9th ed., ed. by Balows, A., and Duerden, B. I., Arnold, London, pp. 125-147.