COMPARATIVE AND FUNCTIONAL GENOMICS

OF

ACTINOBACTERIA AND ARCHAEA

By

Beile Gao, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2009) McMaster University

(Biochemistry) Hamilton, Ontario

TITLE: Comparative and Functional Genomics of Actinobacteria and Archaea

AUTHOR: Beile Gao, B.Sc. (McMaster University)

SUPERVISOR: Professor Radhey S. Gupta

NUMBER OF PAGES: xvii, 200

DEDICATION

This thesis is dedicated to my mother, who has devoted all herself to the family, especially her tremendous love to my brother and me; to my father, who set a good example in my pursuit of career; and to my brother for his always encouragement and attentiveness. Finally, to my grandparents, who brought me up and always treat me like a little child. Your proudness is the biggest driving force for me to head forward in science.

ABSTRACT

The higher taxonomic groups within Prokaryotes are presently distinguished mainly on the basis of their branching in phylogenetic trees. In most cases, no molecular, biochemical or physiological characteristics are known that are uniquely shared by species from these groups. Comparative genomic analyses are leading to discovery of molecular characteristics that are specific for different groups of Bacteria and Archaea. These markers include conserved inserts and deletions in universal proteins and lineage-specific proteins, which provide novel means for identifying and circumscribing these groups of prokaryotes in clear molecular terms and for understanding their evolution. Because of their taxa specificities, further studies on these newly discovered molecular characteristics should lead to discovery of novel biochemical and physiological characteristics that are unique to different groups of microbes. The focus of my project was phylogenomic studies for two large prokaryotic group: Actinobacteria and Archaea. My goals were to a) identify molecular markers that are specific to Actinobacteria and Archaea at different taxonomic levels, which will help to understand the phylogenetic relationship within these two major groups; b) understand the functional significance of Actinobacteria-specific proteins. By comparative genomics approach, a number of conserved indels in various proteins (viz. CoxI, GluRS, CTPsyn, Gft, GlyRS, TrmD, Gyrase A, SahH and SHMT) have been identified that are specific for all Actinobacteria and additional indels were found to be unique to its major subgroups, such as Corynebacterineae, Bifidobacteriaceae, etc. In parallel, a large number of proteins were discovered to be restricted to Actinobacteria at different phylogenetic depths. These identified conserved indels and proteins for the first time provide useful markers for defining and circumscribing the Actinobacteria phylum or its subgroups in clear molecular terms. Similar comparative genomic studies have been carried out on Archaea and a vast number of proteins have been identified that are unique to Archaea or its various lineages. Lastly, I have performed functional studies on one of the Actinobacteria-specific proteins (ASP1). The structure of ASP1 was determined and structural comparison indicates that the function of this protein might be novel since it does not match any known protein with or without known function.

ACKNOWLEDGEMENTS

First and foremost, I wish to thank my supervisor, Dr. Radhey S. Gupta, for his remarkable guidance and support. Since I joined his research group, I have greatly benefited from his mentorship and the freedom and support that he has provided me to research independently, convert ideas to actual projects, collaborate with other labs, and also supervise and direct the learning of other students. More importantly, he set me a good example of being a good researcher and also a great educator.

I would also like to acknowledge and thank the members of my supervisory committee, Drs. Justin R. Nodwell and Paul G. Higgs. They have provided me with numerous helpful suggestions and feedbacks, throughout various courses, meetings and seminars.

My gratitude is extended to my collaborator Dr. Murry S. Junop, who spent so much time teaching me and working with me to solve the protein structure. I admire him in the way that he is patient and kind to every person who is eager to learn, in which I will work very hard in the future to be like him.

Special thank to our research associate, Dr. Bhag Singh, who is remarkable in troubleshooting all kinds of problems. To previous graduates Emma Griffiths, Tim Bowes, Jae Park, and my current fellow graduate students Xianying Amy Cui, Emily Lorezini, Madhanki Kumaresan at the Gupta lab, for their inspirations and idea exchange, both on science and life in general. I am forever grateful to have met such wonderful friends, whose presence made my research experience furthermore pleasant.

Lastly, I would like to give my gratitude to my dear lovely friends whom I met in Canada, Kiana Setoodehnia, Yongfang Zhu, Mingfang Li and Ye Xu, whose friendship and warm-heart made my life here enjoyable and fun. And to my boyfriend Li Zhang, who gave me the strongest support and encouragement since I met him. His always "So?" question after listening to my findings and thoughts about research help me to think deeply and shape my ideas. I love you all!

PREFACE

This thesis is a sandwich thesis. Chapters 1 and 7 (introduction and discussion) contain literature and figures that have been published in a book chapter: Radhey S. Gupta and Beile Gao. (2010) Recent advances in understanding microbial systematics. In *Microbial Population Genetics*. Xu JP, edit. (Caister Academic Press). Chapters 2, 3, 5 and 6 were each published as primary articles prior to the completion of this thesis work. The preface section in each chapter describes the details of the published article, as well as my contribution to the multiple-authored work.

All chapters have been reproduced with permission of all co-authors. Irrevocable, non-exclusive license has been granted to McMaster University and to the National Library of Canada from all publishers. Copies of permission and licenses have been submitted to the School of Graduate Studies.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

.

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF ABBREVIATIONS

aa: amino acid
AlaRS: alanyl-tRNA synthetase;
ASP: actinobacterial-specific protein;
BLAST: Basic Local Alignment Search Tool;
CarA: carbamoyl phosphate synthase small subunit;
CDP: Cytidylyltransferase;
CFBG: Cytophaga-Flavobacterium-Bacteriodes group
COG: clusters of orthologous groups;
CorA: Mg^{2+} transporter protein;
CoxI: cytochrome-c oxidase subunit 1;
CSI: conserved signature indel;
CSP: conserved signature protein;
CTPsyn: CTP synthetase;
DSMZ: German Collection of Microorganisms and Cell Cultures;
DnaK: chaperone DnaK;
EF-G: elongation factor G;
E value: Expect value;
FabG: ketoacyl reductase;
Ga: billion years;
Gft: glucosamine--fructose-6-phosphate aminotransferase;
GluRS: glutamyl-tRNA synthetase;
GroEL: chaperonin GroEL;
HGT: horizontal gene transfer;
HolB: DNA polymerase III subunit delta;
IF-2: initiation factor IF-2;
Indel: insert or deletion;
IPTG: Isopropyl β-D-1-thiogalactopyranoside;
LGT: lateral gene transfer;
ML: maximum likelihood;
MP: maximum parsimony;
NCBI: National Center for Biotechnology Information;
NJ: neighbor joining;
ORF: open reading frame;
PCR: polymerase chain reaction;
RecR: recombination protein RecR;
RGC: rare genomic change;
S3: 30S ribosomal protein S3;
S9: 30S ribosomal protein S9;
SahH: S-adenosyl-L-homocysteine hydrolase;
SeMet: Selenomethionine;
SHMT: serine hydroxymethyltransferase;
SRP: signal recognition particle;
TrmD: tRNA (Guanine-1)-methyltransferase.

GLOSSARY:

Bootstrapping: a statistical technique used to estimate the reliability of a result (usually a phylogenetic tree) that involves sampling data with replacement from the original data set.

Clade: a group of species including all the species descending from an internal node of a tree and no others.

Homologs: sequences that are evolutionarily related by descent from a common ancestor (cf. orthologs and paralogs)

Last universal common ancestor: the most recent organism from which all organisms now living on earth descend. Thus it is the most recent common ancestor of all current life on Earth.

Long branch attraction: a phenomenon in phylogenetic analyses (most commonly those employing maximum parsimony) when rapidly evolving lineages are inferred to be closely related, regardless of their true evolutionary relationships.

Monophyletic: adjective describing a group of species on a phylogenetic tree that share a common ancestor that is not shared by species outside the group. A clade is a monophyletic group.

Orthologs: sequences from different species that are evolutionarily related by descent from a common ancestral sequence and that diverged from one another as a result of speciation.

Outgroup: a species (or group of species) that is known to be the earliest-diverging species in a phylogenetic analysis. The outgroup is added in order to determine the position of the root.

Paralogs: sequences from the same organism that have arisen by duplication of one original sequence.

Phylogeny: an evolutionary tree showing the relationship between sequences or species.

Phylum: a taxonomic rank below Kingdom and above Class. The minimal requirement is that all organisms in a phylum should be related closely enough for them to be clearly more closely related to one another than to any other group.

Polyphyletic: adjective describing a group of species on a phylogenetic tree for which there is no common ancestor that is not also shared by species outside the group. A polyphyletic group is evolutionarily ill-defined.

CHAPTER 1.

Introduction

1.1 Preface

"The affinities of all the beings of the same class have sometimes been represented by a great tree."

~Charles Darwin (the Origin of Species, Chapter IV, 1859)

1.2 Current "rRNA-based"concept of prokaryotic phylogeny and unresolved critical isssues

An understanding of the evolutionary history of life, which spans a period of more than 3.5 billion years (Ga), constitutes one of the most fascinating problems in life sciences (Gupta, 1998b; Cavalier-Smith, 2002; Woese, 1987; Woese et al., 1990). As the pioneers of life forms, prokaryotic organisms were the sole inhabitants of this planet for the first 2-2.5 Ga (Kasting and Siefert, 2002; Schopf, 1978; Knoll, 1999). Thus, a sound understanding of prokaryotic evolution is essential, which can help us understand many fundamental issuess such as: the nature and origin of the first cell, the origin of different types of metabolism and information transfer processes, photosynthesis, the origin of eukaryotic cells, and the evolution of pathogenic and beneficial microbes (Gupta and Golding, 1996; Gupta, 2000a; Raskin et al., 2006; Xiong, 2006). In view of their small and simple morphology, there is no effective way to study the ecology or diversity of prokaryotes without an understanding of their phylogenetic relationships (Fox et al., 1980; Woese, 2006; Stackebrandt, 2006). In the past 25-30 years, with the introduction of 16S rRNA for reconstructing phylogenies, much has been learned about the diversity of prokaryotic organisms, which opened the door to the elucidation of the evolutionary history and systematics of the prokaryotes (Woese et al., 1985; Woese, 1987; Olsen et al., 1994).

Due to its functional constancy, ubiquitous distribution, information content and easy sequencing, 16S rRNA has become a powerful standand method for the identification of microorganisms and for defining and restructuring prokaryotic taxa (Ludwig and Klenk, 2001; Woese, 1987). Based on the branching pattern of the 16S rRNA tree, the prokaryotic organisms are presently divided into two main domains: Archaea and Bacteria (Woese et al., 1990). Archaea are further divided into two phyla Crenarchaeota and Euryarchaeota. However, Archaea are comprised by many more phyla based on analysis of rRNA sequences from environmental samples, although it is difficult to study them in the laboratory (Brochier et al., 2005b; Brochier-Armanet et al., 2008). On the other hand, Bacteria comprise the vast majority (>95%) of known prokaryotic organisms and cultured bacteria are currently divided into about 25 phyla as depicted in Figure 1 (Ludwig and Klenk, 2001). Some of these phyla (viz. Thermodesulfobacteria, Dictyoglomi, Fusobacteria, Acidobacteria, Fibrobacteres and Nitrospira) consist of only a few species, whereas other phyla (viz. Proteobacteria, Actinobacteria, Firmicutes, Cyanobacteria and Bacteroidetes) contain thousands of species accounting for more than 90-95% of all known bacteria.

Although the 16S rRNA approach allowed for a tremendous expansion in our knowledge of prokaryotic relationships during recent years, its resolving power is somewhat limited. Some limitations and drawbacks of the 16S rRNA trees have been pointed out in earlier studies, such as: the GC contents of the rRNA genes are strongly correlated with the optimal growth temperatures of prokaryotes; depending on functional importance, the individual structural elements of rRNAs cannot be freely changed, therefore, sequences changes in the rRNAs occurs in jumps rather than as a continuous process; they are highly conserved, so lack variation to discriminate closely related species; etc (Ludwig and Klenk, 2001; Garrity et al., 2005; Stackebrandt, 2006). In the post-genomic era, inferences based on the rRNA trees are widely questioned. For example, some species contain heterologous 16S rRNA in their genome (Rainey et al., 1996; Nubel et al., 1996). Further, some species that share almost identical rRNA sequence are more more divergent at the whole genome level (Welch et al., 2002; Tettelin et al., 2005). Besides, the tree topology of 16S rRNA is substantially dependent on treeing methods which all have limitations. The three most commonly used treeing methods include distance method, maximum parsimony (MP), and maximum likelihood (ML) (Delsuc et al., 2005). The distance method first converts the character matrix into a distance matrix that represents the evolutionary distances between all pairs of species; the phylogenetic tree is then inferred from this distance matrix using algorithms such as neighbour joining (NJ) (Saitou and Nei, 1987). The drawback of this method is that it only relies on matrices of distance values, while the character of change is not taken into account. MP method is based on a model of evolution that assumes preservation to be more likely than change, which means the parsimony tree requires the minimal number of base changes (Steel and Penny, 2000). Thus, it only infers branching patterns but does not calculate branch lengths *per se*. The most sophisticated treeing method to date is ML, which utilizes more of the information content of the underlying sequences, such as transition/transversion ratio, positional variability, character state probability per position and many others (Felsenstein, 1981). But an accompanying disadvantage of this method is the need for expensive computing time and performance. Even if powerful computing facilities are accessible, only a limited number of sequences can be handled within a reasonable time (Delsuc et al., 2005).

The division of Bacteria into these 25 phyla and the hierarchical classification system within each phylum are solely based on the branching pattern in the 16S rRNA tree. In addition, the distinction of taxa higher than the rank of genus was only based on taxon-specific 16S rRNA signature nucleotides (Ludwig and Klenk, 2001; Stackebrandt, 2006). Except these, currenly there are no objective criteria as to what constitutes a phylum or other higher taxonomic groups such as Class, Order or Family (Gupta and Griffiths, 2002; Stackebrandt, 2006). Most taxonomic ranks were defined in the early days of comparative rRNA sequencing when the data set was small and long "naked" branches enabled clear-cut delimitation (Woese, 1987). With the rapidly increasing rRNA database, most of these "naked" branches have been filled and in some cases it is no longer possible to demonstrate a monophyletic structure or to clearly delimit traditional groups (Ludwig and Klenk, 2001; Maidak et al., 2001). In addition, the 16S rRNA signature nucleotides were based on published 16S rRNA sequences of type strains, so they change when new sequences were added to the database (Stackebrandt, 2006; Zhi et al., 2009). Moreover, the relative branching order cannot be unambiguously determined for the majority of the phyla in the Bacteria, or for many of the lineages within each phylum, as indicated by multifurcations in the trees (Ludwig and Klenk, 2001; Woese, 2006).

More importantly, except for their branching pattern in phylogenetic trees, for most bacterial groups, no molecular, biochemical or physiological characteristics are known that can define each group. Hence, a central aspect of fundamental importance to microbiology that remains to be understood is: "*In what aspects do different main groups of Prokaryotes differ from each other; do species from these groups share any unique molecular, biochemical, structural or physiological characteristics that are distinctive of them* (Gupta and Gao, 2010)?" Another central issue in prokaryotic phylogeny is how different main groups within prokaryotes are related to each other and evolved from a common ancestor (Gupta, 2001; Gupta and Griffiths, 2002). Phylogenetic trees based on rRNA and other gene/protein sequences have not been able to resolve these relationships, leading to the notion that this important problem is insolvable (Doolittle, 1999; Ludwig and Klenk, 2001).

Based on this brief overview, it is evident that in order to develop a reliable understanding of microbial systematics and phylogeny, it is necessary to first develop new well-defined (molecular or biochemical) criteria for identifying all of the main groups or divisions within Prokaryotes in a precise and definitive manner (Gupta and Griffiths, 2002). These new criteria or properties should enable identification and circumscription of all of the major taxa (at various taxonomic levels) in clear molecular and/or biochemical terms. Further, these critieria should also provide insights to improve our understanding of how different groups of Bacteria are related to each other and how they have branched off from a common ancestor.

1.3 Prokaryotic evolution in the light of genomics

1.3.1 The prokaryotic genome is plastic and dynamic

The availability of genome sequences from large numbers of microbes in recent years has opened up new dimensions for studying the evolution of prokaryotes. To date, 924 prokaryotic genomes (64 archaea and 860 bacteria) have been completely sequenced, representing hundreds of species from different lineages. Microbial genomes are believed to be plastic and dynamic as seen by the large variation in their genome size, from 0.49 Mb (*Nanoarchaeum equitans*) to 13.0 Mb (*Sorangium cellulosum*) (Lawrence and Hendrickson, 2005; Ochman, 2005; Snel et al., 2002). The main driving force for genome expansion or reduction is niche adaptation (Lerat et al., 2005; Raskin et al., 2006). In the case of Actinobacteria, most isolated species are free-living and from complex and densely populated soil environments. Thus, their genomes are generally large (5~9 Mb) in order to combat with the environmental changes and species competition (Chater and Chandra, 2006; Ventura et al., 2007). Some species, particularly parasites and symbionts, have undergone extensive genome reduction to settle down in the much more stable conditions within the host as compared to inferred ancestral conditions (Cole et al., 2001; Raskin et al., 2006). While host association favored genome contraction, host diversification favored genome expansion, such as the closely related *Frankia* strains with narrow host range or broad host range showing divergent genome evolution (Normand et al., 2007; Bentley et al., 2008). Moreover, although it is debated whether genome reduction is a strategy to reduce the energy cost of maintaining genome integrity at extreme environments, some species isolated from harsh conditions have relatively smaller genomes compared to their evolutionarily close relatives inhabiting in mild environments (Ciaramella et al., 2005; Freilich et al., 2009; Ranea, 2006).

A number of comparative analyses suggest that selection does not act on genome size; rather, it acts on individual genes and determines the gene repertoire, which in turn influences the genome size (Kuo and Ochman, 2009; Froula and Francino, 2007; Koonin, 2003). Gene duplication, gene loss and gene transfer are the three major events that impact genomic contents (Figure 2) (Abby and Daubin, 2007; Lawrence and Hendrickson, 2005; Lerat et al., 2005). For the gene repertoire of a particular bacterial cell, some genes have been transmitted vertically for very long periods of time, perhaps from the time of the common ancestor of all cellular life forms. These genes are so called "core-set" of genes, most of which are involved in translation, transcription, replication, and central metabolic pathways, and are indispensable for maintaining cellular integrity (Abby and Daubin, 2007). Thus, they are subject to strong purifying selection and highly conserved across lengthy periods of life history. In this way, they act as molecular clock as the 16S rRNA, which could be used for the reconstruction of prokaryotic phylogeny. However, these universal genes make up only a tiny fraction of the entire gene repertoire; altogether, this central core of genes consists of, at most, ~70 genes, which is, no more than 10% of the genes in even the smallest genomes of cellular life forms (Koonin, 2009a; Koonin, 2003). The rest of the genome is mostly composed of the genes belonging to a genetic "shell" where genes were acquired or generated at various points in the history of the lineage, including some very recently acquired (Daubin and Ochman, 2004; Koonin and Wolf, 2008). Some of these "shell" genes are shared by closely related species, or found in distant lineages most likely introduced by horizontal/lateral gene transfer (HGT or LGT) (Abby and Daubin, 2007). Some genes are real ORFans (i.e. ORFs that have no known homologs) that are only found in certain species, strain or even genome and do not have homologues found in other organisms so far (Siew and Fischer, 2003; Siew et al., 2004). Compared to the "core-set" of genes, the genes of the "shell" are more fluid and are subject to the influx and outflux of the genome (Lawrence and Hendrickson, 2005). Comparative genomic studies have shown that gene acquisitions are prevalent at the tips of the phylogeny, which evolve fast and are prone to loss if not conferring advantage to the host (Kuo and Ochman, 2009; van Passel et al., 2008).

1.3.2 Genomic data to explore prokaryotic phylogeny

The most common strategy employed in genomic studies is to make alignments of concatenated DNA or protein sequences and produce trees that are interpreted as reflecting the evolutionary relationships among the organisms (Snel et al., 2005; Forterre and Philippe, 1999). If trees are produced for individual genes or proteins rather than for concatenated sequences, the topologies of all of the trees are further examined to identify evolutionary lineages, and so-called "supertrees" are constructed (Wolf et al., 2002). It has been shown that the resolution and accuracy of the combined protein trees are superior to trees based on a single gene such as16S rRNA, because the combined dataset have more informative sites both in length and variation, and can also dilute the effect of lateral gene transfer (Brochier et al., 2005a; Snel et al., 2005). However, there are several limitations to this approach. First, the identification of orthologous sequences for analysis is a major problem due to the presence of paralogous genes and in some cases mysterious loss of one of the copies (Lin and Gerstein, 2000; Korbel et al., 2002). Second, concatenation forces a single sequence change model on all proteins (including branch lengths and intra-protein variability of evolutionary rates), which in general, is not necessarily true (Wolf et al., 2002). Third, genome sequences are only available for one or a few strains of each species and it is not known to what extent the genome of the sequenced strain is representative of the genetic variation in the higher taxa to which that species belongs (Wolf et al., 2002; Snel et al., 2005). Finally, it has been discussed in section 1.2 that the treeing methods are not flawless and computational power also needs to be improved for large datasets.

Despite the above difficulties in reconstructing phylogeny based on combined datasets, numerous attempts to improve the tree have been published (Korbel et al., 2002; Daubin et al., 2002). Ciccarelli et al. have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains (Ciccarelli et al., 2006). The tree is based on a concatenation of 31 proteins from 191 species, for which orthologs could be unambiguously identified and products of HGT have been excluded after systematical tests. The final tree supports a Gram-positive origin of Bacteria with Firmicutes (low GC gram-positive bacteria) as the earliest division, and it suggests a thermophilic last universal common ancestor. Some relationships revealed in the tree, which are novel, debated, or difficult to reproduce by other methods, were also pointed out in the paper.

Another obvious way of comparing genomes is the analysis of gene content (Huynen and Bork, 1998). Closely related species share a large proportion of genes; by contrast, distantly related species should have lost a significant fraction of the genes inherited from their last common ancestor, rendering the proportion of shared genes low (Wolf et al., 2002). Idealy, this process should continue in a regular fashion. However, this approach is only useful for assessing genetic variation within species, while the plasticity of prokaryotic genomes diminish the trend at higher divergence level (Snel et al., 2005). The most difficult aspect of this method is the variable genome sizes. The extreme case is intracelluar pathogenic bacteria, which have undergone extensive gene loss, resulting from strong selective pressure and not following the uniform rates of change (Snel et al., 1999; Yang et al., 2005). For example, *E. coli* (gamma-proteobacteria) and *Bacillus subtilis* (low GC gram-positives) from different phyla, share more genes than with their more closely related but smaller size cousins, such as *E. coli* with *Buchnera aphidicola* or *B. subtilis* with *Mycoplasma genitalium* (Snel et al., 2005). In order to overcome this, some studies have simply left out the small genomes or normalized the intragenomic distances by dividing the number of shared genes by the number of genes in the smaller genome (House and Fitz-Gibbon, 2002; Korbel et al., 2002). Second, similar to the combined protein tree method, it needs a large-scale definition of orthology (Korbel et al., 2002). Another factor that alters the gene content is gene acquistion by horizontal transfer. Especially for species inhabiting in same extreme environment, gene transfer will be more frequent during their adaptation (Nelson et al., 1999; Doolittle, 1999). Moreover, the number of shared genes is phenetic rather than phylogenetic character, which is the same as the distance matrix method for phylogenetic analysis (Doolittle, 1999).

When only a small number of genomes were available, gene order was analyzed to approach the phylogenetic relationship of prokaryotes (Lathe et al., 2000; Wolf et al., 2001). Similar to the gene content, rearrangements continuously shuffle the genomes, gradually breaking ancestral gene strings. However, gene order evolves faster than gene content: for example, *E. coli* and *Haemophilus influenzae* share 78% of their genes, while their gene order is conserved for only 36% (Huynen and Bork, 1998). Gene order is extensively conserved between some closely related species, but rapidly becomes less conserved among more distantly related organisms (Snel et al., 2005). The prokaryotic genome is featured by the presence of operons, which contain a set of genes that are co-transcribed and co-regulated. Studies have shown that the organization of small groups of functionally linked genes in operons are conserved in all or most of the bacterial and archaeal genomes, in part, owing to the extensive HGT (Koonin, 2009b; Lawrence, 2003; Novichkov et al., 2009).

1. 4 Do Horizontal Gene Transfer (HGT) undermine the "Tree of Life"?

Nowadays, it is believed that horizontal gene transfer (HGT), also referred to lateral gene transfer (LGT), is an important force in bacterial evolution (Syvanen, 1985; Lawrence and Hendrickson, 2005; Gogarten and Townsend, 2005). HGT refers to any process in which an organism incorporates genetic material from another organism without being the offspring of that organism. There are three common mechanisms for horizontal gene transfer: (1) transformation, the process by which bacterial cells take up naked DNA molecules; (2) transduction, the process in which bacterial DNA is moved from one bacterium to another by a bacterial virus (bacteriophage); (3) bacterial conjugation, a process in which a living bacterial cell transfers genetic material through cell-to-cell contact (Thomas and Nielsen, 2005). HGT was first described in 1959 and this study demonstrated the transfer of antibiotic resistance genes between different species of bacteria (Ochiai et al., 1959). For nearly 50 years, HGT events have been discovered in numerous bacteria or archaea species (Ochman et al., 2000; Gogarten and Townsend, 2005). With the completion of prokaryotic genome sequence, the search for horizontally acquired genes was carried out by scanning the genome sequence for regions of atypical base composition, such as GC content and codon usage pattern (Marri et al., 2006; Teichmann and Mitchison, 1999). The genomes are also surveyed for genes whose best matches (as detected by Basic Local Alignment Search Tool, BLAST) lie outside their closest sequenced relatives, and in the case of certain sequenced bacteria (e.g., *Thermotoga maritime* and *Aquifex aeolicus*), substantial fractions of their genes were found to be most similar to genes present in Archaea (Nelson et al., 1999; Deckert et al., 1998; Koski and Golding, 2001). Another common way to identify cases of transfer and exchange is by searching for evidence of discordance among gene trees (Koonin et al., 2001; Boucher et al., 2003).

With the large numbers of cases of HGT identified from genome sequences, the Darwinian tree-like representation of relationships between species has been questioned by some scientists, asserting that HGT events are so "rampant" that genes cannot be used as reliable phylogenetic markers (Walsh and Doolittle, 2005; Doolittle, 1999). They proposed a network of species, arguing that a signal of vertical inheritance cannot be unraveled from horizontal signals due to HGT (Doolittle, 1999; Kunin et al., 2005). To test whether HGT truly diminish the tree of life, two questions need be answered: To what extent is the HGT affecting the prokaryotic genomes and, whether the core of the genome still exists (Snel et al., 2005). Additionally, the concept or definition of HGT/LGT needs to be revised in accordance with the evolutionary process of species. It is known that not all genes were inherited from the last universal common ancestor of life forms. Although the mechanism by which new genes evolve in the cell is not known, it is believed that gene transfer as a source of genome expansion occur throughout the evolutionary process (see section 1.3.1) (Daubin and Ochman, 2004; Lerat et al., 2005). Thus, the ancient gene acquisition, which happend in the progenitor cell of a specific lineage and subsequently passed on by vertical inheritance, should not be regarded as "horizontal" or "lateral" gene transfer. A follow-up question is: for a long-term evolutionary process, does HGT randomly obscure the prokaryotic phylogeny or actually promote and record the divergence of the species via the introduction of new genes at different stage?

The first two questions can be answered together. The extent of HGT is hotly debated, and due to different species sampling and detection methods/standards, a bacterial genome is suggested to have 0 to >20% genes obtained from alien sources (Ochman et al., 2000; Nakamura et al., 2004; Ragan, 2001). However, among these alien genes, some of them have detected homologs in other species so the transfer is evidenced, and the others are only found in that genome that could originate in that specific genome

(Abby and Daubin, 2007; Lerat et al., 2005). These later genes perhaps constitute a large fraction of the currently identified HGT products because of their lack of homologs. A number of comparative genomic studies have been carried out to carefully examine the HGT event. For example, Novichkov et al. describe a framework for identifying orthologous sets of genes that deviate from a clock-like model of evolution (Novichkov et al., 2004). For several hundred analyzed orthologous sets, they found that 70% of the genes did not show aberrant levels of sequence similarity potentially caused by HGT. Explicit phylogenetic analysis of the remaining 30% indicated that only half of these could be due to HGT, while the other half was due to lineage-specific acceleration of evolution (Novichkov et al., 2004). Recently Kunin et al. analysed a 165-genome dataset and found 4.7–5.2% of events to be LGT, 11.1–11.6% gene losses and 83.4–83.6% vertical transfers (Kunin et al., 2005; Kunin and Ouzounis, 2003). Additionally, Beiko et al. have performed a rigorous phylogenetic analysis of >220,000 proteins from 144 prokaryotic genomes to determine the contribution of gene sharing to current prokaryotic diversity, and the inferred relationships suggest a pattern of inheritance that is largely vertical except among some closely related taxa and among distantly related organisms that live in similar environments (Beiko et al., 2005).

It is known that genes involved in translation and transcription, show fewer indications of transfers (Koonin, 2003). For example, the 31 ortholog genes employed by Ciccarelli et al. are all involved in translation (Ciccarelli et al., 2006). These proteins are highly connected in the cellular network, require tighter stoichiometric and expression control, are less exposed to immediate selective pressure or are evolutionarily conservative, thus less susceptible to homologous replacement via HGT (Ragan and Beiko, 2009; Aris-Brosou, 2005). Although some studies have detected HGT events for some core genes even in the rRNA, the detected species are very few and the number of publications is countable (Gogarten and Townsend, 2005; Gogarten et al., 2002). Besides, single-gene based trees are already questioned and the current trend is to use a core set of nontransferred or rarely transferred genes to track the history of prokaryotes.

The last question points to the history of gene transfer. Gene transfer should occur all the time during the evolutionary process of species (Daubin and Ochman, 2004; Koonin, 2009a). Studies have shown that the acquisition of new genes by gene transfer may be one of the dominant ways of adaptation in bacterial genome evolution (Hao and Golding, 2006). Gene transfer provides the bacterial genome with a new set of genes that help it to explore and adapt to new ecological niches (Kuo and Ochman, 2009). Besides, genes acquired via lateral gene transfer will over time acquire the molecular characteristics of the host genome (Marri and Golding, 2008; Hao and Golding, 2006). If the genes transfer occured at deeper clade and the new genes are retained by all the descendents from the progenitor, then the gene transfer event has likely contributed to the divergence of the clade and also incorporated into the cellular protein interaction network (Lerat et al., 2005; Narra et al., 2008; Ragan and Beiko, 2009). Besides, as their incorporation time increase, these acquired genes will be ameliorated and grow to resemble the native genes, such as their GC content or codon usage (Lawrence and Ochman, 1997; Marri and Golding, 2008; Koski et al., 2001). Thus, the gene transfers which occured at early stages actually record the divergence of the clade or lineage.

Furthermore, after introduction to the progenitor cell, these transfered genes follow vertical inheritage, which is different from the current concept of HGT or LGT.

In summary, HGT contributed to genome evolution of prokaryotes. It brought an extra layer of complexity to the study of the phylogenetic relationship among prokaryotes but it did not preclude the reconstruction of the history of life (Abby and Daubin, 2007; Snel et al., 2005). In addition, the tree is being refined towards high resolution and congruence by incorporating more evolutionary factors.

1.5 Rare Genomic Changes as novel phylogenetic markers for evolutionary studies

The current unresolved issues regarding prokaryotic phylogeny make it necessary to search for novel genomic characteristics that are unique to phylogeneticlly related prokaryotic lineages and also record the divergence of their common ancestor. The ideal characteristics for such studies should meet the following requirments: "*These markers should be homologous apomorphic characters that evolved only once (synapomorphy) but not by convergence*" (Stackebrandt, 2006). Such markers should also not be affected by factors such as multiple changes at a given site, long-branch attraction effect, differences in evolutionary rates between and among species, HGT, etc., which confound the inferences from phylogenetic trees (Delsuc et al., 2005; Philippe et al., 2005a).

1.5.1 Conserved Signature Indels (CSIs) in protein sequences

Conserved inserts and deletions (Indels) in gene/proteins sequences provide an important category of rare genetic changes (RGCs) for understanding bacterial phylogeny (Gupta, 1998b; Rokas and Holland, 2000; Delsuc et al., 2005). The indels, which provide useful phylogenetic markers, are generally of defined size and flanked on both sides by conserved regions to ensure their reliability (Gupta and Griffiths, 2002; Gupta, 1998b). Because of the rarity and highly specific nature of such changes, it is less likely that they could arise independently by either convergent or parallel evolution (i.e. homoplasy) (Gupta, 2000a; Rokas and Holland, 2000). Other confounding factors such as differences in evolutionary rates at different sites or among different species should not also affect the interpretation of a conserved indel. Hence, when a conserved signature indel (CSI) of defined size is uniquely found in a phylogenetically defined group(s) of species, the simplest explanation for this observation is that the genetic change responsible for this CSI occurred once in a common ancestor of this group of species and then passed on to various descendents. Because the presence or absence of a given CSI in different species is not affected by factors such as differences in evolutionary rates, CSIs, which are restricted to particular clade(s), have generally provided good phylogenetic markers of common evolutionary descent (Gupta, 1998b; Gupta, 2003). Also, since genetic changes leading to CSIs could be introduced at various stages during evolution, it allows the identification of CSIs in gene/protein sequences at different phylogenetic depths corresponding to various higher taxonomic groupings (e.g., phylum, order, family or genus) (Gupta, 2001; Gupta and Griffiths, 2002; Gupta, 1998b; Gupta and Gao, 2010; Gao and Gupta, 2005). Such CSIs, in turn, can provide well-defined markers for identifying different taxonomic groups of bacteria in molecular terms.

Recent work from our lab has identified a large number of CSIs that are restricted to many higher taxonomic groups within the Bacteria, such as: Alphaproteobacteria, Gammaproteobacteria, Epsilonproteobacteria, Aquifiales, Chlamydia, Cyanobacteria, Deinococcus–Thermus, Bacteroidetes, etc (Gao et al., 2009a; Griffiths and Gupta, 2004b; Griffiths and Gupta, 2004a; Griffiths et al., 2005; Griffiths and Gupta, 2001; Griffiths and Gupta, 2006; Gupta, 1998b; Gupta, 2004). These newly discovered CSIs provide useful markers for defining or circumscribing the bacterial groups in clear molecular terms. Additionally, identified CSIs that are commonly shared by species from a number of different phyla provide valuable information regarding the branching order and interrelationships among different main groups of bacteria (Gupta, 2001; Gupta, 2003; Gupta, 2009; Gupta and Mok, 2007; Gupta, 2000a). With the greatly expanded microbial genome database, the statistical study of large numbers of such RGCs certainly represents a promising avenue for unraveling the prokaryotic phylogeny.

1.5.2 Conserved Signature Proteins (CSPs) that are lineage-specific

Another type of RGCs that are useful for taxonomic classification as well as for understanding evolutionary relationships among different organisms are whole proteins that are uniquely present in particular groups or subgroups of bacteria but not found anywhere else (Kainth and Gupta, 2005; Dutilh et al., 2008). Recent analyses of genomic sequences have indicated that such conserved signature proteins (CSPs), which are also referred to as lineage-specific proteins, arise throughout the evolutionary process of a bacterial lineage (Gao and Gupta, 2007; Lerat et al., 2005). A vast number of lineage-specific proteins unique to certain species, strain or even genome, which are also called "ORFans", are introduced recently during speciation or strain divergence (Daubin and Ochman, 2004). Studies have shown that these proteins present at the tips of the phylogeny evolve fast and are subject to loss if not conferring advantages to the host

(Narra et al., 2008; Kuo and Ochman, 2009). However, if the lineage-specific proteins originate at deeper clade and are retained by all the descendents from the progenitor, they are confined to the monophyletic group (Gao et al., 2006; Dutilh et al., 2008; Gupta and Gao, 2010). Thus, such ORFans are no more solitary "orphans", they are conserved signature proteins (CSPs) uniquely shared by every daughter lineage of that group, which can serve as molecular markers for defining or distinguishing that group from other bacteria (Gupta and Gao, 2009; Gao et al., 2009a). Furthermore, based on a number of CSPs that are specific to different lineages, it is possible to infer their branching order or interrelationship (Gupta and Mok, 2007; Kainth and Gupta, 2005; Gupta and Griffiths, 2006).

Comparative genomic studies have been carried out on several major bacterial phyla to identify CSPs that are unique to them, such as Alphaproteobacteria, Gammaproteobacteria, Epsilonproteobacteria, Chlamydia, Cyanobacteria, Deinococcus–Thermus, Bacteroidetes, etc (Kainth and Gupta, 2005; Gao et al., 2009a; Gupta, 2006; Griffiths et al., 2006; Griffiths and Gupta, 2007; Gupta and Lorenzini, 2007; Gupta, 2009; Gupta and Mok, 2007). The identified CSPs unique to different bacterial groups have proved of great value in defining these major groups and have also provided useful information regarding the branching order of different lineages within them. More interestingly, a majority of identified CSPs are of hypothetical functions, which point out our lack of knowledge on many building blocks in the prokaryotic cell (Gupta and Gao, 2010). Studies on these lineage-specific CSPs that originate at deeper clade are very meaningful, and the reasons are as follows. First, because of their retention in all daughter lineages and compared to the easily lost ORFans, they must perform important function in the species of the clade. Second, due to their uniqueness, their function might confer some distinctive characteristics that make the clade different from other bacteria. Third, a thorough understanding of their evolution as individual and components in the protein network should provide insight into the mechanisms of genesis or speciation, since their incorporation process into the existing cellular network might be the trend that a large number of newly introduced ORFans are evolving towards (Daubin and Ochman, 2004; Kuo and Ochman, 2009).

1.6 The diversity and phylogeny of Actinobacteria and Archeae

Gram-positive bacteria with high GC DNA content are currently recognized as a distinct phylum, Actinobacteria, on the basis of their branching pattern in 16S rRNA trees (Embley and Stackebrandt, 1994; Garrity et al., 2005). This phylum constitutes one of the largest groups among Bacteria. The most updated taxonomy of Actinobacteria by Zhi X *et al.* suggest that this phylum encompass 219 genera in 48 families, much more enriched than the previously defined 130 genera in *Bergey's Manual* 2001 (Zhi et al., 2009). Actinobacterial species exhibit high level of diversity in terms of their morphology and physiology and play important roles in medicine, industry and environment; some species can produce bioactive secondary metabolites while many others can cause serious human, animal and plant diseases (Embley and Stackebrandt, 1994; Stackebrandt and Schumann, 2006). The most extensively studied representatives from this group include soil-dwelling *Streptomyces* which are the major antibiotics producers and important pathogen *Mycobacterium* that are responsible for the largest number of human deaths by infection

(Chater and Chandra, 2006; Smith, 2003). However, except for their branching pattern in the 16S rRNA tree, until recently no other biochemical or molecular characteristics were known that could distinguish species of this group from all other bacteria (Zhi et al., 2009; Garrity et al., 2005; Ludwig and Klenk, 2001).

Within the Actinobacteria, the hierarchical classification system was deduced from the clustering of genera in 16S rRNA trees and the distinction of taxa higher than the rank of genus was solely based on taxon-specific 16S rRNA signature nucleotides (Zhi et al., 2009). However, the 16S rRNA sequence diversity in case of Actinobacteria is somewhat lower than that found with some other phyla, leading to a tight clustering of Actinobacteria, which make it impossible to determine either a stable or a significant branching order within this phylum (see Fig. 3) (Ludwig and Klenk, 2001; Garrity et al., 2005). In addition, the 16S rRNA signature nucleotides were based on published 16S rRNA sequences of type strains, so the specificity of these nucleotides change when new sequences were added to the database (Zhi et al., 2009). Moreover, although phenotypic characteristics such as morphological, physiological and chemotaxonomic features are valuable in preliminary classification and identification of many spore-forming actinobacteria, the level of congruence of these characteristics with the phylogeny is low (Embley and Stackebrandt, 1994; Stackebrandt and Schumann, 2006). Hence, novel characteristics are desired to define and distinguish *Actinobacteria* and its different lineages.

Archaea are widely regarded as one of the three main domains of life, although their origin is a subject of debate (Woese et al., 1990; Graham et al., 2000; Gupta, 1998b;

Gupta, 1998a). Archaeal species were earlier believed to inhabit only extreme environments such as extremely hot, or hot and acidic, extremely saline, or very acidic or alkaline conditions (Woese, 1987; Barns et al., 1994; Kennedy et al., 2001; Gonzalez et al., 1999; Futterer et al., 2004). However, recent studies provide evidence that they are widespread in different environments (Schleper et al., 2005; Pace, 1997). The archaea also include methanogens, which grow under strictly anaerobic and often thermophilic conditions, and are the only organisms that derive all of their metabolic energy by reduction of CO_2 by hydrogen to produce methane (Jones et al., 1987; Lange and Ahring, 2001). The archaeal species branch distinctly from all other organisms in phylogenetic trees based on 16S rRNA and many other gene/protein sequences (Olsen et al., 1994; Olsen and Woese, 1997; Brown and Doolittle, 1997; Brendel et al., 1997). In addition, many morphological or physiological characteristics such as the presence of branched-chain ether-linked lipids in their cell membrane, lack of peptidoglycan in their cell wall, characteristic subunit pattern of RNA polymerase, presence of modified bases in tRNA, and the presence of a unique form of DNA polymerase, have previously been indicated to be defining characteristics of archaea (Woese et al., 1990; Woese, 1987). However, as noted by Walsh and Doolittle, many of these features are either not shared by all archaea or they are also present in various eukaryotes or some thermophilic bacteria, indicating that they do not constitute distinctive characteristics of all Archaea (Walsh and Doolittle, 2005).

The phylogenetic analyses of Archaea have led to their division into two major groups or phyla designated as Crenarchaeota and Euryarchaeota (Woese et al., 1990;

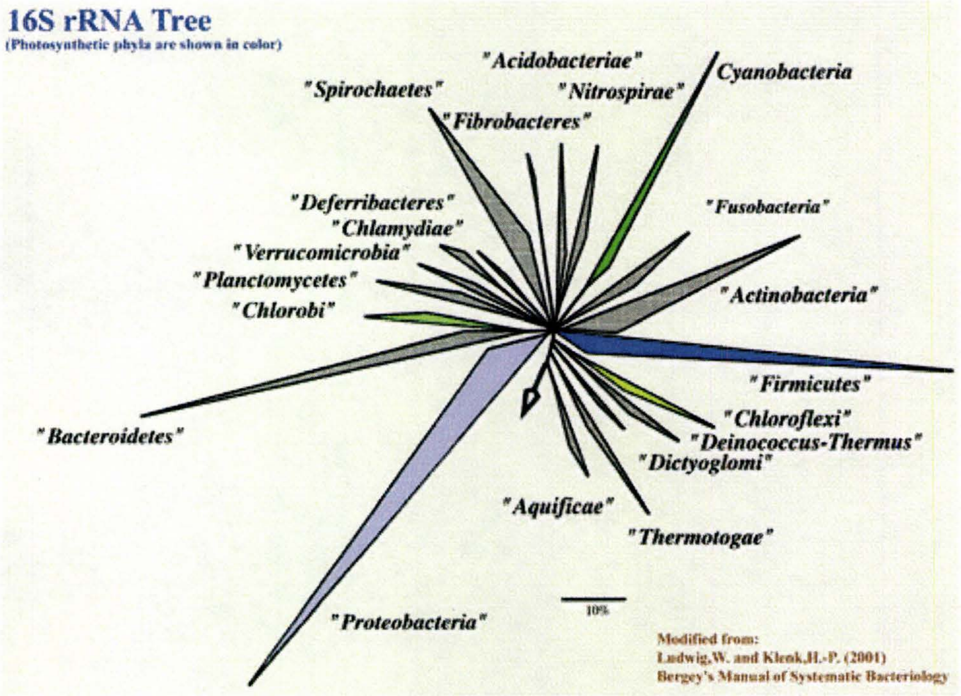
Ludwig and Klenk, 2001; Olsen et al., 1994; Bapteste et al., 2005). The species from both groups, particularly Euryarchaeota, are highly diverse in terms of their metabolism and physiology. Based on their metabolic and physiological characteristics and other unique features, five functionally distinct groups within Euryarchaeota are currently recognized: methanogens, sulfate reducers, extreme halophiles, cell wall-less archaea, and extremely thermophilic sulfur metabolizing archaea (Ludwig and Klenk, 2001; Gribaldo and Brochier-Armanet, 2006; Brochier et al., 2004). Some of these groups, such as methanogens, are polyphyletic in different phylogenetic trees (see Fig. 4) (Brochier et al., 2005a; Matte-Tailliez et al., 2002; Brochier et al., 2004). If methanogenic archaea are polyphyletic, then does methanogenesis evolve only once or multiple times? The origin and evolution of methanogenesis is a very important issue, since this energy production process is suggested to be ancestral (Gribaldo and Brochier-Armanet, 2006; Bapteste et al., 2005; Reeve et al., 1997). Therefore, it is necessary to identify molecular characteristics that can correlate with the special physiological features of different archaeal lineages.

1.7 Research objectives

The objectives of my research project are two folds: phylogenomic studies of Actinobacteria and Archaea to identify CSIs and CSPs; functional studies of CSPs that are specific to actinobacteria. Both Actinobacteria and Archaea are very large and highly diverse in terms of their morphology, physiology, and ecology. Currently, except their clustering pattern in the 16S rRNA tree, there are no distinctive characteristics that are known to be unique to these two prokaryotic groups. Besides, the phylogeny within these two groups is unclear due to insufficient resolution of 16S rRNA and the limitations of the current tree construction methods. Thus, comparative genomic studies were carried out to identify two kinds of molecular markers: CSIs and CSPs, which are specific to Actinobacteria and Archaea at different taxonomic levels. These markers will not only help to define these two prokaryotic groups but also allow the delineation of hierarchical relationships among these groups. Also, phylogenetic tree based on large datasets of combined protein sequences will be constructed to compare with the results from different CSIs and CSPs.

For the second part of the project, functional studies were carried out on CSPs that are specific to all actinobacteria. Because of their specificity, these molecular markers likely play important roles in the cell that distinguish the Actinobacteria from other bacteria. Functional studies on these CSPs might reveal unique physiological characteristics that are shared by all Actinobacateria.

1.8 Figures 1-4


16S rRNA Tree
(Photosynthetic phyla are shown in color)

Modified from:
Ludwig,W. and Klenk,H.-P. (2001)
Bergey's Manual of Systematic Bacteriology

Figure 1.1. The current 16S rRNA tree of the Bacteria domain, which appears to be more bush-like than tree-like. The branching order of the main groups is unresolved according to this model. This figure was taken from Bacterial (Prokaryotic) Phylogeny Webpage (March 2006): http://www.bacterialphylogeny.com/index.html, which is a modified version from *Bergey's Manual of Systematics Bacteriology* (Ludwig and Klenk, 2001). The colored branches refer to bacterial groups that include species with photosynthetic ability.

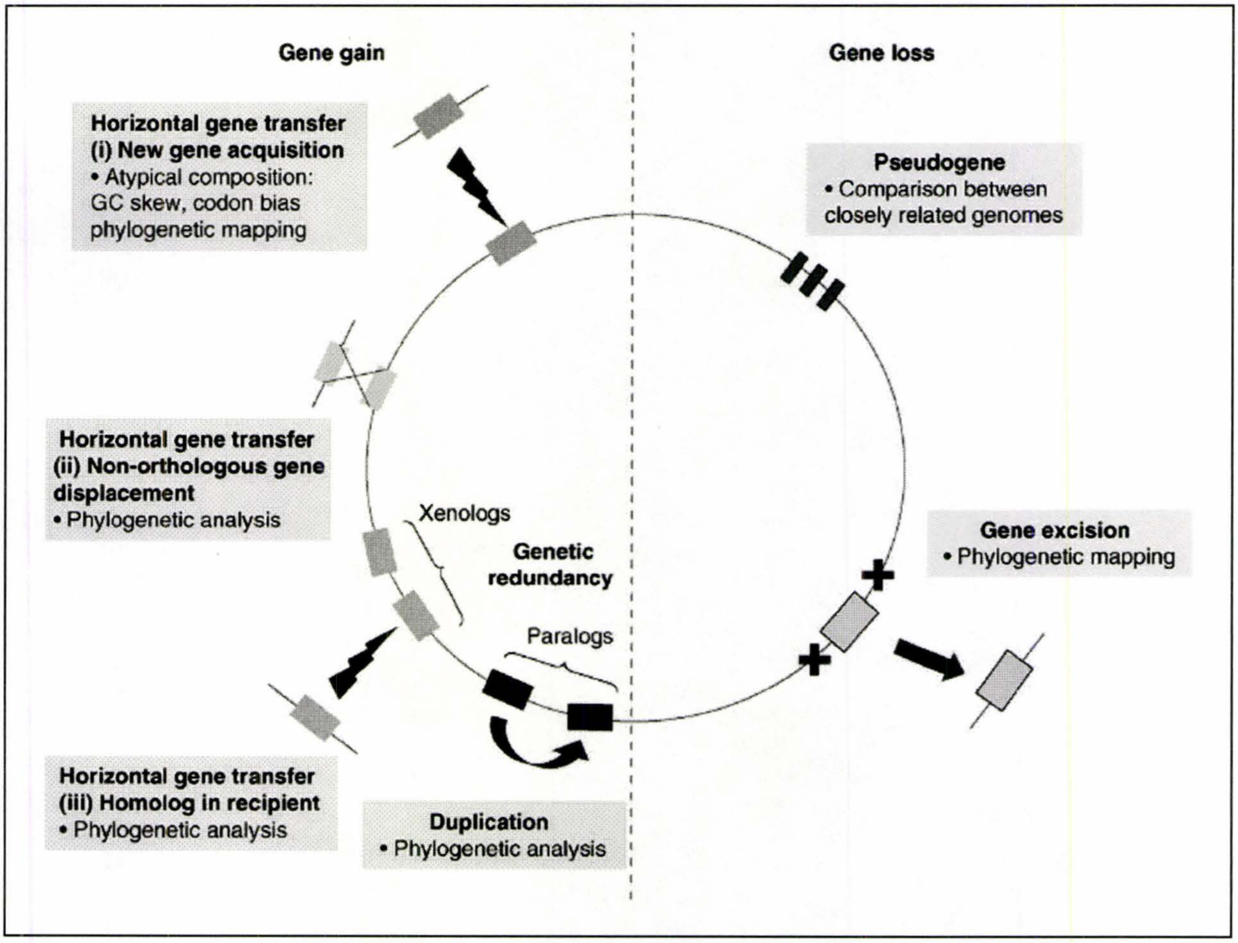


Figure 1.2. The dynamics of genome repertoire. Bacterial genomes are dynamic entities that constantly gain (left; blue boxes) and lose genes (right; beige boxes). This figure is taken from Figure 2, (Abby and Daubin, 2007).

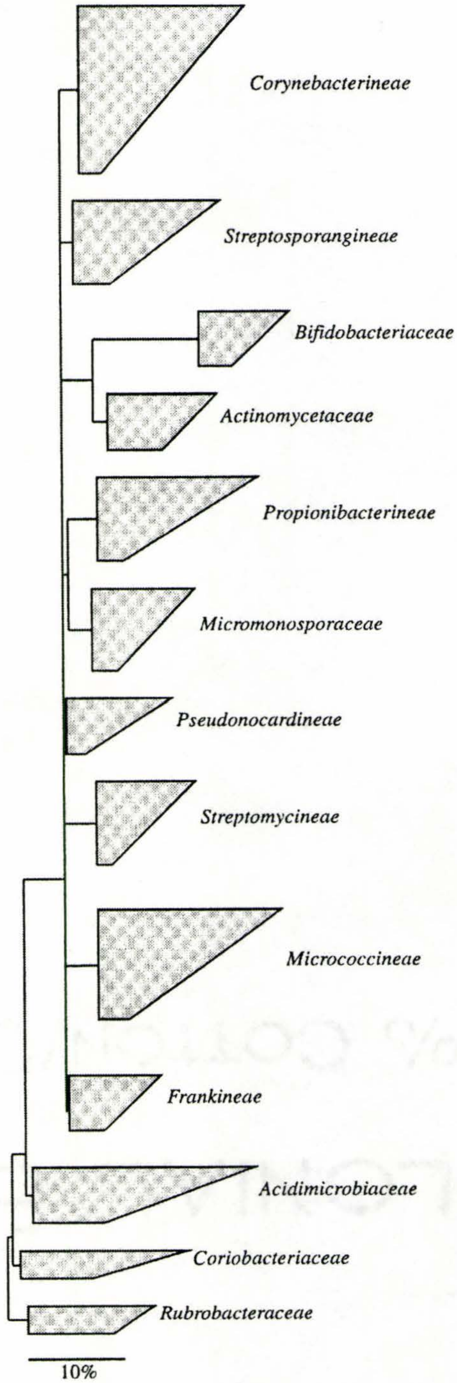


Figure 1.3. 16S rRNA-based tree depicting the major phylogenetic groups within the Actinobacteria. This figure was taken from *Bergey's Manual of Systematics Bacteriology* (Ludwig and Klenk, 2001).

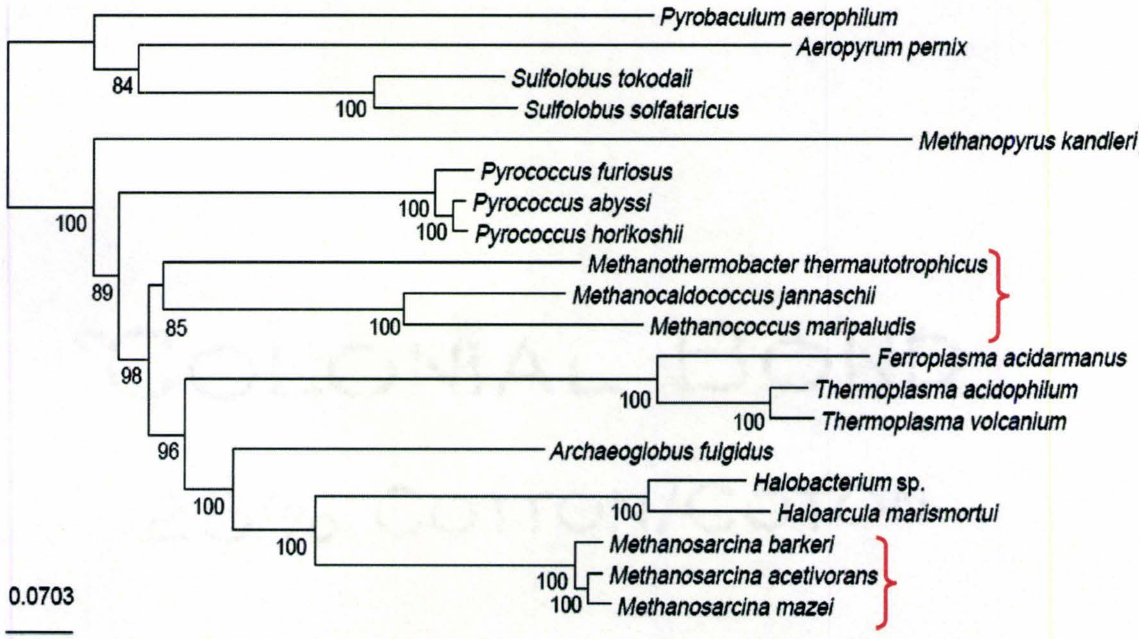


Figure 1.4. Unrooted phylogenetic trees for Archaea obtained from the transcription dataset. Methanogens were split into 3 clusters, which were labeled by red brackets. (Taken from Figure 2(a), (Brochier et al., 2004))

CHAPTER 2.

CSIs that are Characteristic of the Phylum Actinobacteria and its different lineages

2.1 Preface

This chapter describes conserved signature indels (CSIs) that are unique to Actinobacteria or its various subgroups. Some of the results were reproduced from the published manuscript (Gao and Gupta, 2005), which reported 3 CSIs in 3 widely distributed proteins that are distinctive characteristics of the Actinobacteria and are not found in any other groups of bacteria. The complete reference is: Gao B & Gupta RS. Conserved indels in protein sequences that are characteristic of the phylum Actinobacteria. Int J Syst Evol Microbiol. 2005 Nov;55(Pt 6):2401-12. In addition, a number of CSIs that are uniquely shared by certain actinobacterial subgroups are also discussed, which provide useful information regarding the branching order or interrelatioship within the Actinobacteria phylum.

2.2 Introduction

As briefly mentioned in section 1.6, Actinobacteria constitute one of the main phyla within the Bacteria. This phylum encompasses genera covering a wide range of morphology: Some species are coccoid (e.g. *Micrococcus*) or rod-coccoid (e.g. *Arthrobacter*) in shape, while others display fragmenting hyphal forms (e.g. *Nocardia*) or permanent and highly differentiated branched mycelium (e.g. *Streptomyces*) (Ventura et al., 2007; Stackebrandt and Schumann, 2006). Previously, organisms containing the mycelia which resemble the unrelated fungi, were classified as the Actinomycetes (Embley and Stackebrandt, 1994; Goodfellow and Williams, 1983). Spore formation is common among the Actinobacteria, although not ubiquitous, and spores range from motile zoospores to specialized propagules. They are also physiologically diverse bacteria, as evidenced by their production of numerous extracellular enzymes and by the thousands of metabolic products they synthesize and excrete (Schrempf, 2001), many of which are antibiotics (Lechevalier and Lechevalier, 1967). Actinobacteria especially members of *Streptomycetaceae* are the major antibiotic producers in the pharmaceutical industry (Bentley et al., 2002; Chater and Chandra, 2006). However, a few Actinobacteria are important human, animal and plant pathogens. For example, *Mycobacterium tuberculosis* infection result in tuberculosis; *Corynebacterium diphtheriae* causes diphtheria; *Propionibacterium acnes* is the causative agent of acne (Cole, 2002; Cerdeno-Tarraga et al., 2003; Bruggemann et al., 2004). Actinobacterial species are widely distributed in both terrestrial and aquatic ecosystems, especially in soil (Goodfellow and Williams, 1983; Ward and Bora, 2006). In nature, Actinobacteria play an important role in decomposition and humus formation, which is an integral part of the recycling of biomaterials (Lechevalier and Lechevalier, 1967; Goodfellow and Williams, 1983).

Due to their pharmaceutical, industrial and environmental importance, the taxonomy and phylogeny of the Actinobacteria are of great interest (Embley and Stackebrandt, 1994; Stackebrandt et al., 1997; Ahmad et al., 2000; Bull and Stach, 2007). Earlier attempts to determine Actinobacteria phylogeny based on morphological and chemotaxonomic traits were found to be unreliable indicators of phylogenetic relationships above the family level (Embley and Stackebrandt, 1994). Hence, our current understanding of the taxonomy and evolutionary relationships of Actinobacterial divisions is mainly based on the branching patterns of these species in the 16S rRNA trees (Stackebrandt et al., 1997; Ludwig and Klenk, 2001). Different species have been placed in this group based on 16S rRNA oligonucleotide catalogs and phylogenetic analysis of full and partial 16S rDNA sequences, although some species do not possess a high GC content (Woese, 1987; Stackebrandt et al., 1997; Ludwig and Klenk, 2001). In 1997, a new taxonomic hierarchical classification for Actinobacteria was proposed by Stackebrandt, which recognized this group as a distinct class, Actinobacteria classis nov., within Gram-positive bacteria (Stackebrandt et al., 1997). In the latest *Bergey's Manual* (2001), the Actinobacteria have been assigned the rank of a phylum, recognizing that the phylogenetic depth represented in this lineage is equivalent to that of existing phyla and that the group shows clear separation from the Firmicutes (Garrity and Holt, 2001). However, except for their distinct clustering in the 16S rRNA trees, presently no other reliable biochemical or molecular characteristics are known which can clearly distinguish species belonging to the Actinobacteria phylum from other bacteria.

The availability of genome seqeuence opened a new window for discovering novel molecular characteristics that are useful for biochemical, taxonomic and phylogenetic purposes (Karlin et al., 2003; Ventura et al., 2007). Our recent work has focused on identifying CSIs in widely distributed proteins that are characteristics of the different groups of bacteria and are also helpful in understanding the interrelationships among them (Gupta and Griffiths, 2002; Gupta, 2000b; Gupta, 2009). In the present work, we identified a number of CSIs in widely distributed conserved proteins that are distinctive characteristics of the Actinobacteria phylum, and not found in any other bacteria. The sequence information for these proteins was previously available from only a limited number of actinobacteria, whose genomes have been sequenced. One possible signature for actinobacteria, consisting of a large insert in the 23S rRNA, has previously been described (Roller et al., 1992). However, the validity and specificity of this signature also needs to be further tested using sequence data from additional species. Thus, we have examined the presence of the newly identified CSIs in several proteins, and also the 23S rRNA indel, from a broad range of actinobacterial species by PCR amplifying and sequencing the corresponding gene fragments.

2.3 Materials and Methods

2.3.1 Bacterial strains and chromosomal DNA isolation

The various actinobacterial strains that were used for PCR amplification in this study and their taxonomic positions, i.e. Orders, Suborders or Families, within the Actinobacteria phylum are given in Table 2.1. Together with the actinobacterial strains that have been completely sequenced in the GenBank, these strains cover a broad range of the diversity represented by this phylum. The typed versions of various new strains were purchased from the DSMZ (German Collection of Microorganisms and Cell Cultures) and cultured under the recommended conditions. The chromosomal DNA was purified by the following method. A 100 μl pellet was transferred to a microcentrifuge tube and washed 3 times with 0.5 ml 10.3% sucrose. The cell pellet was resuspended in 0.5 ml buffer containing 0.3 M sucrose, 25 mM Tris-HCl, 25 mM EDTA, pH8.0, 2 mg/ml lysozyme, and incubated for 2 hours at 37 ^{0}C. Then 50 μl 20% SDS was added and the cell suspension was incubated in 65 ^{0}C for 30 min. The cell lysate was extracted with chloroform twice, and the aqueous layer was separated during centrifugation at

14,000 g for 15 min. The DNA was precipitated with 2x ethanol and dissolved in sterile H_2O for PCR amplification. The chromosomal DNA of *Rhodococcus rhodochrous* and *P. acnes* was generously provided to us by Dr. L D Eltis (UBC, British Columbia) and Dr. Mark Farrar (Leeds University, UK) (Warren et al., 2004; Farrar et al., 2000).

2.3.2 Identification of CSIs in protein sequences

Multiple sequence alignments for a large number of proteins have been created in our earlier work (Gupta, 2000b; Gupta et al., 2003; Griffiths and Gupta, 2004b) (Griffiths and Gupta, 2004a). To search for Actinobacteria-specific CSIs, these alignments were visually inspected to identify any indel that was uniquely present in only actinobacterial homologs, and which was flanked by conserved sequences. The indels which were not flanked by conserved regions or which were also present in other bacterial groups were omitted from further consideration in this study. The specificity of potentially useful indels for Actinobacteria was further evaluated by carrying out detailed BLAST searches on short sequence segments (usually between 60-100 aa) containing the indel and the flanking conserved regions. The purpose of these BLAST searches was to obtain sequence information from all available bacterial homologs to ensure that the identified signatures are only present in the actinobacterial homologs. The sequence information for various useful CSIs was compiled into alignment files.

2.3.3 PCR amplification and sequencing

Degenerate oligonucleotide primers, in opposite orientations, were designed for highly conserved regions that flanked the identified CSIs in sequence alignments. The sequences of various PCR primers used in these studies are given in Table 3.2. The PCR reaction was performed in 30 μl of solution and all primer sets were optimized for Mg^{2+} concentration (in the range of 1.5 to 4 mM) for each DNA strain tested. PCR amplification was carried out in a Techne Progene thermocycler. After an initial denaturation step at 97 ^{0}C, the DNA was amplified for 30 cycles (30 s at 97 ^{0}C, 30 s at 55 ^{0}C, 1 min at 72 ^{0}C). The last cycle was followed by a 15 min extension at 72 ^{0}C. Because all the genomic actinobacterial DNA we tested contain a high GC content, the denature temperature is set as 97 ^{0}C so that the DNA was mostly denatured in a short time with less damage. The DNA fragments of the expected size were purified from 0.8% (w/v) agarose gels and subcloned into the plasmid pDRIVE using a UA cloning kit (Qiagen). After transforming *E. coli* JM109 cells with the plasmids, inserts from a number of positive clones were sequenced. The sequence information for various actinobacterial species have been deposited in the GenBank and their accession numbers are given in Figure 3.1-3.4.

2.4 Results

2.4.1 Description of novel Actinobacteria-specific CSIs and examination of their specificity

Our work has identified a number of CSIs in different protein sequences that are limited to only the actinobacterial species. The sequence information for most of these genes/proteins was mainly available from only those actinobacterial species whose genomes have been sequenced when this work was carried out. Many of the sequenced species are closely related and some belong to the same genus. Hence, to determine the Actinobacteria specificities of the identified CSIs, we have cultured and extracted chromosomal DNA of 23 actinobacterial type strains, covering a large number of orders and suborders (e.g. *Rubrobacterineae, Micrococcineae, Corynebacterineae, Micromonosporineae, Propionibacterineae, Pseudonocardineae, Streptomycineae and Streptosporangineae*) within this phylum (Table 2.1). The sequence information for the identified CSIs was obtained from many of these species to confirm and validate their specificity. A brief description of the newly identified CSIs that are specific to all atinobacteria is given below.

One of the Actinobacteria-specific CSIs that we have identified is present in the subunit 1 of the Cytochrome c oxidase (Cox1) protein. The cytochrome c oxidase is an intrinsic membrane protein, composed of 3 subunits, that functions as the terminal enzyme of the respiratory electron transport chain (Michel et al., 1998). In the Cox1 subunit, a 2-aa gap is present in a conserved region (boxed) that is unique to various actinobacterial species, but not seen in any other bacteria (Figure 2.1)(Gao and Gupta, 2005). Because this sequence gap is absent in Cox1 homologs from all other groups of bacteria, it likely constitutes a deletion in the actinobacterial homologs. By means of PCR amplification, we were successful in obtaining sequence information for the Cox1 gene from 22 additional actinobacterial species belonging to different orders and families. All of these new fragments were also found to contain this 2-aa indel at the same position. The sequence information of this region for various actinobacterial species (including those sequenced in the present work) and some representatives from other groups of bacteria is presented in Figure 2.1. A complete alignment of all available Cox1 homologues in the GenBank were also carried out, which includes 44 sequences from actinobacterial species and 224 sequences from other bacterial groups such as Firmicutes, Deinococcus-Thermus, Cyanobacteria, Chlamydia-CFBG, Spirochetes, Aquifex, Proteobacteria, etc. The observed 2-aa indel is unique to various actinobacteria, and no exceptions were observed (Gao and Gupta, 2005). The shared presence of this 2-aa indel in various actinobacteria strongly indicates that it was introduced only once in a common ancestor of Actinobacteria and passed on to all their descendents. Because of its unique presence in various Actinobacteria, this signature provides a good molecular marker for distinguishing the actinobacterial species from other bacterial phyla. In the structure of the Cox1 protein from *Paracoccus denitrificans*, the region where this deletion is present (residues 332-333), is located on the periplasmic surface, but its functional significance remains to be determined (Michel et al., 1998). It should be mentioned that the 2-aa indel in Cox1 is absent in *Symbiobacterium thermophilum*, which is grouped with Actinobacteria when this work was carried out in 2005 (Gao and Gupta, 2005). However, recent genomic analyses indicate that this species is much more closely related to Bacilli and Clostridia than to Actinobacteria (Ueda et al., 2004).

Another signature for Actinobacteria is present in the enzyme CTP synthetase which catalyzes the conversion of UTP into CTP by transferring an amino group to the 4-oxo group of the uracil ring (Endrizzi et al., 2004). Except for the mycoplasma species, the gene encoding CTP synthetase is present in all other microbial genomes and has only one copy in the genome. A 10-aa indel which is distinctive of various proteobacterial species has previously been identified in this protein (Gupta, 2000b). Interestingly, in the same position where this proteobacterial insert is present, a smaller 4-aa indel is found in all of the actinobacterial species except *Tropheryma whipplei*, which contain 3-aa indel at the same position (Figure 2.2). This indel is again highly specific for Actinobacteria and not present in the CTP synthetase homologs from any other bacteria whose sequences are available in the databases (Gao and Gupta, 2005). In the present work, we have successfully amplified the CTP synthetase gene fragments from 15 additional actinobacterial stains and all of these were found to contain this 4-aa indel in the same position (Figure 2.2). Based on its shared presence in various actinobacterial species, the identified CSI in CTP synthetase was likely introduced in a common ancestor of the Actinobacteria phylum, independently of the CSI in the Proteobacteria, and it provides a specific and useful molecular marker for the Actinobacterial phylum. The CTP synthetase homolog from *Symbiobacterium thermophilum* again did not contain the 4-aa indel, supporting the inference from Cox1 CSI and other studies that this species is distinct from other Actinobacteria. The enzyme CTP synthetase consists of a single polypeptide containing two domains: the C-terminal glutamine amide transfer (GAT) domain catalyses the hydrolysis of glutamine, whereas the N-terminal synthase domain is responsible for the amination of UTP (Endrizzi et al., 2004). The observed indel is located in the GAT domain of the enzyme and it is indicated to be present on the outside surface of the protein. However, its functional significance for these groups of bacteria remains to be determined.

Roller et al. have previously described a large insert of about 100 nt in the 23S rRNA that was present in a large number of actinobacterial species, but not found in other groups of bacteria (Roller et al., 1992). The original paper describing this signature included 64 high GC content Gram-positive strains from the currently recognized phylum Actinobacteria. However, many of these strains were from the same genus, even the same species, and they represented only 22 genera. To examine the specificity of this signature further, we have successfully amplified and sequenced the 23S rRNA insert region from 13 additional actinobacterial species representing 13 additional genera, covering all of the major groups within this phylum. Sequence information for these sequences as well as various other actinobacterial species and a few other bacterial groups is presented in the partial sequence alignment shown in Figure 2.3. As seen, an insert of between 90 and 100 bp is present in all of the actinobacterial species that we sequenced but it is not found in any other bacterial groups (Figure 2.3). Of these species, smaller inserts were present in this position in *Tropheryma whipplei* (79 bp) and *Microbispora bispora* (30 bp) and *Rubrobacter xylanophilus* was found to be lacking the insert entirely. We have also confirmed the absence of this insert in another Rubrobacter species, *Rubrobacter radiotolerans*, by amplifying and sequencing of the insert from the type strain of this species. Because Rubrobacter represents a deep branch within the phylum Actinobacteria (Stackebrandt et al., 1997; Ludwig and Klenk, 2001), the large insert in 23S rRNA gene was very likely introduced in an actinobacterial ancestor, after the divergence of Rubrobacter.

Besides the 3 CSIs described above, we have identified 7 additional CSIs in various proteins that are uniquely shared by almost all actinobacteria, including: 4-aa indel in glutamyl-tRNA synthetase (GluRS), 4-aa indel in glucosamine--fructose-6-phosphate aminotransferase (Gft), 3-aa indel in glycyl-tRNA synthetase (GlyRS), 4-aa indel in tRNA (Guanine-1)-methyltransferase (TrmD), 4-aa indel in Gyrase A, 9-aa indel in S-adenosyl-L-homocysteine hydrolase (SahH), and 5-aa indel in serine hydroxymethyltransferase (SHMT) (Table 2.3). These identified CSIs provide novel molecular means for defining and circumscribing the phylum Actinobacteria. Similar to the large insert in 23S rRNA, all of these CSIs are not found in the homologues from *Symbiobacterium thermophilum* and *Rubrobacter xylanophilus*.

2.4.2 CSIs that are specific to different actinobacterial subgroups

In addition to CSIs that are specific to all actinobacteria, we have also identified a number of CSIs that are unique to different subgroups within this phylum. A summary of all the identified CSIs that are specific to different actinobacterial subgroups and their specificities can be found in Table 2.3. The CSIs that are exclusive to two major suborders *Corynebacterineae* and *Bifidobacterinenae* are described below.

Corynebacterineae is one of the biggest suborder within Actinobacteria and include many species that are important human and animal pathogens such as species from genera *Mycobacterium, Nocardia, Corynebacterium, Gordonia*, etc (Embley and Stackebrandt, 1994; Stackebrandt et al., 1997). Members from this suborder share similar ultrastructure and chemical composition of their cell envelopes, which is composed of a tripartite structure consisting of the ubiquitous cytoplasmic membrane, the cell wall and an outer layer ((Daffe and Draper, 1998; Brennan, 2003; Hutchings et al., 2009). Mycolic acids is a unique component in their cell wall and is a defining feature of the *Corynebacterineae* (Brennan, 2003; Sutcliffe and Harrington, 2004). We have discovered several CSIs in various proteins that are likely unique to this subgroup. One of these signatures is found in Carbamoyl-phosphate synthase small subunit (CarA). By aligning the homologous sequences, we found a 2-aa indel in this protein that was only shared by *Mycobacterium* and *Corynebacterium* (Figure 2.4). To verify its specificity, I amplified 709-bp fragments that contained the indel region from 12 additional actinobacterial strains. Among these 12 strains, *Tsukamurella paurometabola* and *Rhodococcus rhodochrous* belong to the suborder *Corynebacterineae*, and both of them contain an indel at the same position as *Mycobacterium* and *Corynebacterium*, while the other 9 strains that are different from *Corynebacterineae* all lack this indel. The newly sequenced *Corynebacterineae* species, such as *Gordonia bronchialis, Rhodococcus jostii,* etc., also contain the indel. More interestingly, all 4 *Rhodococcus* species contain a larger indel of 9 aa instead of 2 aa at the specific position. It is likely that the 2-aa indel was introduced in a common ancestor of the *Corynebacterineae* and passed on to all descendants of this lineage. Subsequently, another 7-aa indel was introduced at the same position in the common ancestor of *Rhodococcus* species. Overall, this CSI provide a useful molecular marker to distinguish the *Corynebacterineae* from other actinobacterial subgroups. Moreover, we have identified two more CSIs that are unique to *Corynebacterineae*, including 4~20 aa indel in recombination protein RecR and 1-aa indel in initiation factor IF-2 (see Table 2.3). All these 3 CSIs will be very helpful in the identification of new species related to this subgroup.

Bifidobacterial species are generally found in the human gastrointestinal tract (GIT) and are important for establishing and maintaining homeostasis of the intestinal ecosystem to allow for normal digestion (Lee et al., 2008; Ventura et al., 2006b).

Currently, they form a distinct order *Bifidobacteriales* within the Actinobacteria phylum (Stackebrandt et al., 1997). This order comprised a single family *Bifidobacteriaceae*, which in turn consists of four genera, *Bifidobacterium, Gardnerella, Scardovia* and *Parascardovia* (Ventura et al., 2004; Ludwig and Klenk, 2001). Except for the *Bifidobacterium* genus, which contains 29 species, the other genera each contain just a single species. Due to their importance for human health, currently, 7 strains from this family have been completely sequenced and another 28 strains are in assembly. By aligning homologous sequences, we have identified two CSIs that are exclusive to members from this family. One of the CSI is 5-aa indel in Cytidylyltransferase (CDP) homologues (Figure 2.5), and the other is 1-aa indel in signal recognition particle (SRP) proteins (Figure 2.6). Both of these two CSIs are uniquely shared by *Bifidobacterium* and *Gardnerella* species, but not found in any other actinobacteria or other bacterial groups. Thus, they provide useful molecular markers for all *Bifidobacterium* and *Gardnerella*, and it will be worthwhile to test whether they are also shared by other genera from this family.

2.4.3 CSIs that are uniquely shared by certain actinobacterial lineages provided molecular evidence for their phylogenetic relationship

By identifying CSIs that are shared by only certain subgroups within a main phylum, it should also be possible to reliably determine the interrelationships among different subgroups (viz. families, orders, etc.) within a given phylum. In the case of Actinobacteria, we have identified several CSIs that are uniquely shared by two or more subgroups within this phylum, which provide additional molecular evidence for their clustering in phylogentic trees and suggest their common origin. Two examples are given below.

In the 16S rRNA tree, species from the suborder *Corynebacterineae* and *Pseudonocardineae* (represented by genera *Saccharopolyspora* and *Actinosynnema*) branch together although their clustering is weakly supported (Zhi et al., 2009). Presently, except the branching pattern revealed in the 16S rRNA tree, there are no other studies that examine the relationship for these two suborders. We have identified two CSIs that are found uniquely in species of *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae*, but not found in any other species outside these 3 actinobacterial suborders. These two CSIs are 2-aa indel in DNA polymerase III subunit delta (holB) (Figure 2.7) and 1-aa indel in 30S ribosomal protein S3 (Figure 2.8). The position of *Micromonosporineae* (represented by genus *Salinispora*) is not resolved in the 16S rRNA tree as lack of significant bootstrap support and also incongruence in trees constructed by different researchers (Zhi et al., 2009; Ludwig and Klenk, 2001). However, in the most comprehensive tree for Actinobacteria based on concatenated alignement of 35 conserved proteins by this work, *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae* form a well-defined cluster with a high bootstrap score 84% (see Figure 4.1 and section 4.4). The identified CSIs in protein holB and S3 provide additional evidence that these 3 suborders are more closely related than other actinobacteria and likely evolved from a common ancestor.

Micrococcineae is the most diverse subgroup within the phylum Actinobacteria containing ecologically, morphologically and chemotaxonomically divergent species.

The most updated taxonomic outline of this suborder encompass 15 families and 86 genera (Schumann et al., 2009; Zhi et al., 2009). By doing alignment for universal proteins from different bacterial group, we found two CSIs that are uniquely shared by species of *Micrococcineae, Bifidobacteriaceae* and *Actinomycycineae*. These two CSIs include a 5-aa indel in chaperone DnaK proteins (Figure 2.9) and another 5-aa indel in ribosomal protein S3 sequences (Figure 2.10). Although *Bifidobacteriale*s is regarded as an independent order in the taxonomic structure of Actinobacteria, its placement in the phylogenetic tree is questioned (Stackebrandt et al., 1997). In the NJ tree based on 16S rRNA, *Bifidobacteriales* forms a deep-branching lineage within Actinobacteria but it was positioned within the suborder *Micrococcineae* in both ML and MP tree using the same dataset (Zhi et al., 2009). While in the 16S rRNA tree by Ludwig and Klenk, *Bifidobacteriaceae* and *Actinomycycineae* branch together and indicate a common origin for these two subgroups (Ludwig and Klenk, 2001). Besides, in the TrmD protein, we found a 4-aa indel that is specific to all actinobacteria, and adjacent to this CSI, an extra aa is uniquely present in *Bifidobacteriaceae* and *Actinomycycineae*, which is not found in any other species (Table 2.3). The identified CSIs in S3 and DnaK suggest that these 3 suborders are closely related and they likely evolved from a common ancestor exclusive of other actinobacteria. More discussions based on this result and also the combined protein tree are presented in section 4.4.

2.5 Discussion

In this work, we identified 9 novel molecular signatures consisting of conserved inserts and deletion in widely distributed proteins that are unique to all Actinobacteria.

These CSIs are only found in actinobacterial homologs and not found in homologs from any other groups of bacteria for which extensive sequence information is now available. In addition, we have also tested the Actinobacteria-specificity of a large insert in 23S rRNA that was previously described (Roller et al., 1992). We have tested the hypothesis that these CSIs are distinctive characteristics of Actinobacteria by obtaining sequence information from a large number of actinobacterial species, covering a wide range of families and suborders within this phylum, for which no sequence information was available. All of these species were found to contain the indicated CSIs confirming that they are distinctive characteristics of this group.

The placement of a new bacterial species into the Actinobacteria phylum is at present based solely on their branching pattern in the 16S rRNA trees. Based on phylogenetic trees, it has proven difficult to reliably circumscribe a given phylum (Ludwig and Klenk, 2001; Gupta and Griffiths, 2002; Oren, 2004). The problems that one face in these regards are that in the 16S rRNA tree, most of the actinobacterial species formed a well-defined cluster, branching together in 100% of the bootstrap replicates (Ludwig and Klenk, 2001; Stackebrandt et al., 1997). However, both *Rubrobacter radiotolerans* and *Symbiobacterium thermophilum* are separated by a large genetic distance from the other actinobacterial species, forms the outgroup of this cluster (Gao and Gupta, 2005). In the current taxonomy based on 16S rRNA trees, *Rubrobacter radiotolerans* is recognized as a deep branch within the Actinobacteria phylum and *Symbiobacterium thermophilum* is also classified as part of Actinobacteria when this species was isolated (Ohno et al., 2000).

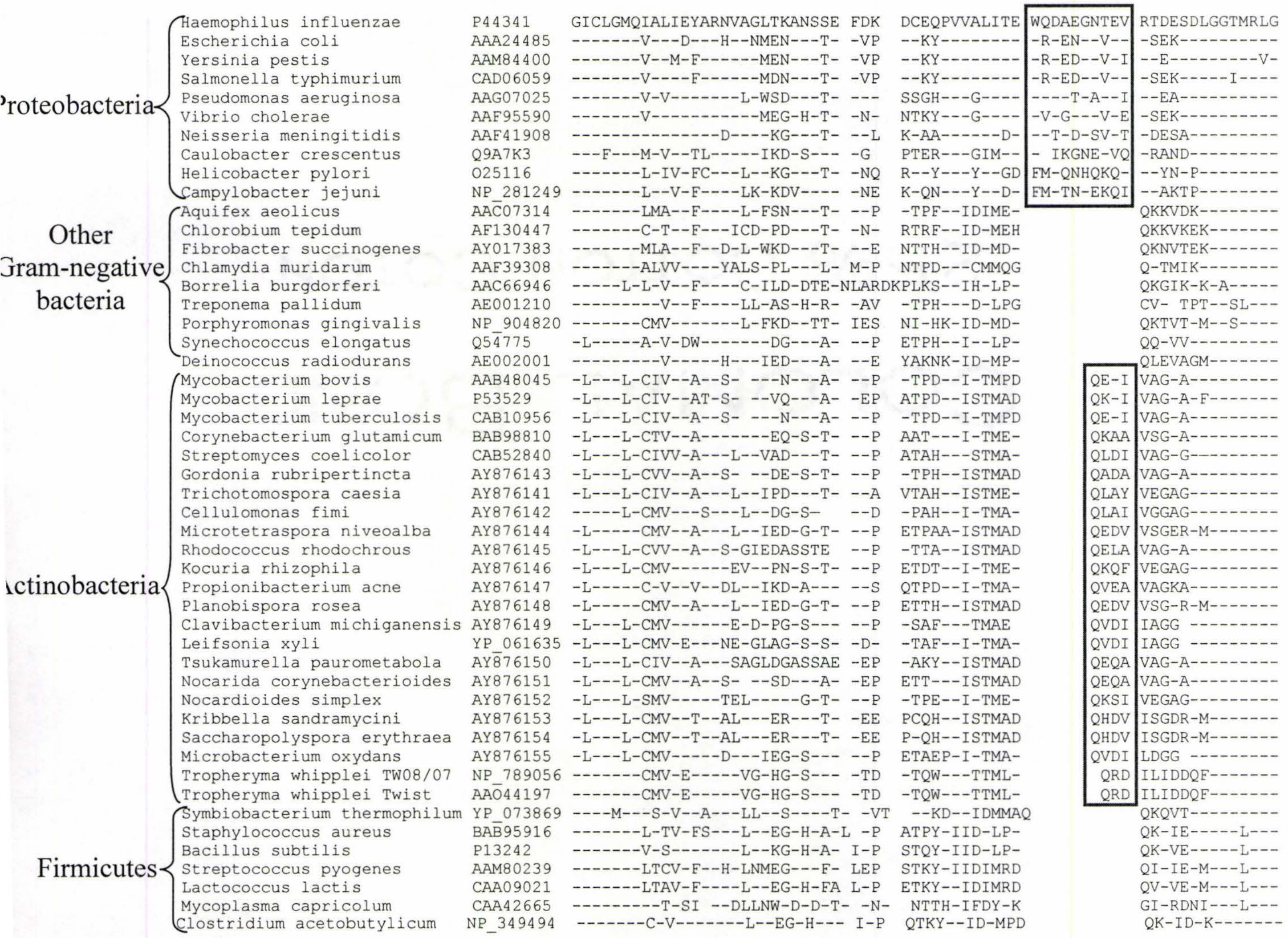
In our analyses, all 9 CSIs in various proteins and the large insert in 23S rRNA that are present in most other Actinobacteria were found to be lacking in both *Symbiobacterium thermophilum* and *Rubrobacter radiotoleran.* The analysis of various genes in *Symbiobacterium thermophilum* genome indicate that they are most closely related to Firmicutes rather than Actinobacteria (Ueda et al., 2004). The absence of these 10 CSIs indicates strongly that this species should not be placed in the Actinobacteria phylum despite its high GC content. For *Rubrobacter*, our comparative genomic analyses reveal other reliable characteristics CSPs that are uniquely shared by *Rubrobacter* species and other Actinobacteria, which support its placement in the Actinobacteria phylum (see details in section 3.4.1). However, based on the absence of the actinobacteria-specific CSIs, and the fact that this species is distantly related to other Actinobacteria in the 16S rRNA tree (Stackebrandt et al., 1997; Gao and Gupta, 2005), it is suggested that the species from this genus do not also comprise typical actinobacterial species.

In addition to CSIs that are specific to all Actinobacteria, we have also another 15 CSIs that are either unique to a specific actinobacterial subgroup or uniquely shared by two or more subgroups (Table 2.3). These subgroup-specific CSIs provide useful molecular markers for inferring the relationship among different actinobacterial lineages. Moreover, it is of much interest to understand the biological and functional significance of these rare genomic changes in important proteins that are limited to only the actinobacterial species. Because most of these CSIs have not been lost from any of the species belonging to this phylum, it is reasonable to assume that they play important biological roles in these organisms. Hence, the studies aimed at understanding their functional significance should be of much interest.

2.6 Figures 1-10 and Tables 1-3

```
                                                         326                                         368
                Escherichia coli             NP_752476   LSFIVWLHHFFTMG AG ANVNAFFGITTMIIAIPTGVKIFNWLF
                Bordetella parapertussis     NP_886335   ---L--A--M--T- IP VVGQL--MYA--L-S---------VA
                Vibrio vulnificus            NP_762523   ---V--A--M--T- MP VFAEL--MYC--M--V---------VA
                Caulobacter crescentus       NP_420580   --Y----------- S- -S-----------S----A-------
                Agrobacterium tumefaciens    NP_353177   --YL---------- S- -------------S----A-------
Gram-negative   Pseudomonas aeruginosa       ZP_00138941 -G-T---------- S- GD--G---VA--L-S------L-----
bacteria        Prochlorococcus marinus      NP_895169   -GLV--A--MF-S- TP PWMRL--T-A-SF--V---I-F----A
                Themus thermophilus          YP_005640   -GTM--A--M--V- ES TLFQIA-AFF-AL--V-----L--IIG
                Deinococcus radiodurans      NP_296339   V-C------M-AV- IP EAWQIA-M-S-L-V-V--------LIG
                Chloroflexus aurantiacus     ZP_00355855 -G-L--G--M-VSS QS VYAGLI-SFI--LV---SAI-V---TA
                Cytophaga hutchinsonii       ZP_00309629 ------A--M-VT- MN PFLGSI-MFL-L---V-SA--A--YIA
                Aquifex aeol                 NP_214504   V--FL-I--M-VS- VP NWTRVL-SY--LL--V---I-----ML
                Staphylococcus aureus        YP_040448   ---L--V------- N- -LI-S--S-S--L-G------L----L
                Listeria innocua             NP_469361   ---L--V------- S- -L--S--S----M------I-------
Firmicutes      Geobacillus kaustophilus     YP_149311   ------V------- -- PA--SA-S----A-------------
                Oceanobacillus iheyensis     NP_693175   ---V--V------- Q- -LT-SI-S----A--V---I------L
                Exiguobacterium sp           ZP_00182647 -GFM--V--M--V- L- PVA--I-AVA--A--V----------
                Bacillus cereus              NP_830510   ---V---------- -- PA--S--S-S--A-S-----------
                Symbiobacterium thermophilum YP_075926   MG-T--S--M--V- M- PV--SI-SL---A--V----------S
                Corynebacterium efficiens    NP_739028   --MAV-A--MFVT-    -VLLP--SFM-FL-SV-----F---VG
                Corynebacterium glutamicum   NP_601724   --MA--A--M-VT-    -VLLP--SFM-FL-SV-----F---VG
                Gordonia westfalica          NP_954783   --VA--A--MYVT-    -VLLP--SFMTFL--V-----F---IG
                Leifsonia xyli               YP_062417   --IT--A--MYVT-    SVLLPW-SLM--L--V---------IG
                Mycobacterium leprae         NP_302190   --VA--A--M-AT-    -VLLP--SFM-YL--V-----F---VG
                Thermobifida fusca           ZP_00057854 --MT--A--M-PT-    -VLLP--SFM-FL--V-----F---IG
                Mycobacterium tuberculosis   NP_217559   --VA--A--M-AT-    -VLLP--SFM-YL--V---I-F---IG
                Streptomyces avermitilis     NP_827224   --VT--A--MYVT-    -VLLP--SFM-FL--V-----F---IG
                Rhodococcus erythropolis     NP_898717   --IA--A--MYAT-    -VLLPY-SFM-FL--V---------IG
                Tropheryma whipplei          NP_787372   --VT--A--MYVT-    -VLLP--SFM--L--V---------VG
                Trichotomospora caesia       AY876119    --VV--A--M-AT-    -VLLP--SVLSFL--V-----F---AG
                Cellulomonas fimi            AY876120    --VT--A-RMYVT-    SVLLP--AFM--L--V-----F---IG
                Kocuria rhizophila           AY876121    --VT--A--MYVT-    -VALG--SFM--M--V-----F---IG
                Gordonia rubripertincta      AY876122    --VA--A--MYVT-    -VLLP--SFM-FL--V-----F---IG
                Microtetraspora niveoalba    AY876123    --MT--A--M-AT-    -ALLP--SMLSFL------I-F---TG
Actinobacteria  Arthrobacter nicotinovorans  AY876124    --VT--A--MYVT-    SVLLP--AFM--L--V-----F---IG
                Micromonospora chersina      AY876125    --MS--A--M-AT-    QVLLP--SFLSYL--V---M-F-S-IG
                Rhodococcus rhodochrous      AY876126    --IA--A--MYAT-    -VLLPY-SFM-FL--V-----F---IG
                Propionibacterium acnes      AY876127    --VS--A--M-VT-    -VSLP--SFMTFT--V-----F---IG
                Streptosporangium roseum     AY876128    --IT--A--M-PT-    QVLLP--SFM-FL--V-----F---IG
                Pseudonocardia halophobica   AY876129    --AA--A--MYAT-    -VLL---SF--LL------I-FV--IG
                Tsukamurella paurometabola   AY876131    --VA--A--MYAT-    -VLLP--SFM-FL--V-----F---I-
                Nocardioides simplex         AY876138    --VA--A--M-VT-    -VNLP--SGM-FL--V-----F---IG
                Planobispora rosea           AY876132    --VA--A--M-PT-    QVLLP--SFM-FL--V-----F---IG
                Clavibacterium michiganensis AY876133    --VT--A--MYVT-    SVLLP--SLM--L--V---------IG
                Nocardia corynebacterioides  AY876134    --IA--A--MYAT-    -VLLP--SFM-FL--V-----F---IG
                Saccharopolyspora erythraea  AY876135    --VV--A--MYAT-    -VLLP--AF--FL--V---M-F---IG
                Microbacterium oxydans       AY876136    --VA--A--MHVT-    SVLLP--ALM--L--V---------IG
                Kribbella sandramycini       AY876137    --VA--A--I-VT-    -MNLP--SFM-FL--V-----F---IG
                Williamsia murale            AY876139    --VA--A--MYAT-    -VLLP--SFM-F---V-----F---IG
                Oerskovia turbata            AY876140    --V---A--MYVT-    -VLLP--AFM--L--V-----F---IG
                Streptomycoides glaucoflavus AY876130    --IT--A--M-VT-    QVLLP--SFM-FL--V-----F---VG
```

Figure 2.1 Partial alignment of Cytochrome c oxidase subunit 1 (CoxI) sequences showing a 2-aa-indel (outer box), which is specific for all Actinobacterial species. Dashes in all sequence alignments indicate identity with the amino acid on the top line. The accession numbers of various sequences are shown in the second column. The sequences whose accession number start with "AY" were amplified in this work. Only representative sequences from different Bacteria are shown.



```
                   Haemophilus influenzae        P44341     GICLGMQIALIEYARNVAGLTKANSSE FDK  DCEQPVVALITE WQDAEGNTEV RTDESDLGGTMRLG
                   Escherichia coli              AAA24485   -------V---D---H--NMEN---T- -VP  --KY-------- -R-EN--V-- -SEK----------
                   Yersinia pestis               AAM84400   -------V--M-F------MEN---T- -VP  --KY-------- -R-ED--V-I --E---------V-
                   Salmonella typhimurium        CAD06059   -------V----F------MDN---T- -VP  --KY-------- -R-ED--V-- -SEK-----I----
                   Pseudomonas aeruginosa        AAG07025   -------V-V-------L-WSD---T- ---  SSGH---G---- ----T-A--I --EA----------
'roteobacteria     Vibrio cholerae               AAF95590   -------V-----------MEG-H-T- -N-  NTKY---G---- -V-G---V-E -SEK----------
                   Neisseria meningitidis        AAF41908   ---------------D----KG---T- --L  K-AA------D- --T-D-SV-T -DESA---------
                   Caulobacter crescentus        Q9A7K3     ---F---M-V--TL-----IKD-S--- -G   PTER---GIM-- - IKGNE-VQ -RAND---------
                   Helicobacter pylori           O25116     -------L-IV-FC---L--KG---T- -NQ  R--Y---Y--GD FM-QNHQKQ- --YN-P--------
                   Campylobacter jejuni          NP_281249  -------L--V-F----LK-KDV---- -NE  K-QN---Y--D- FM-TN-EKQI --AKTP--------
                   Aquifex aeolicus              AAC07314   -------LMA--F----L-FSN---T- --P  -TPF--IDIME-            QKKVDK--------
        Other      Chlorobium tepidum            AF130447   -------C-T--F---ICD-PD---T- -N-  RTRF--ID-MEH            QKKVKEK-------
                   Fibrobacter succinogenes      AY017383   -------MLA--F--D-L-WKD---T- --E  NTTH--ID-MD-            QKNVTEK-------
Gram-negative      Chlamydia muridarum           AAF39308   -------ALVV----YALS-PL---L- M-P  NTPD---CMMQG            Q-TMIK--------
                   Borrelia burgdorferi          AAC66946   -----L-L-V--F----C-ILD-DTE-NLARDKPLKS--IH-LP-            QKGIK-K-A-----
      bacteria     Treponema pallidum            AE001210   ---------V--F----LL-AS-H-R- -AV  -TPH---D-LPG            CV- TPT--SL---
                   Porphyromonas gingivalis      NP_904820  -------CMV-------L-FKD--TT- IES  NI-HK-ID-MD-            QKTVT-M--S----
                   Synechococcus elongatus       Q54775     -L-----A-V-DW-------DG---A- --P  ETPH--I--LP-            QQ-VV---------
                   Deinococcus radiodurans       AE002001   ---------V-----H---IED---A- --E  YAKNK-ID-MP-            QLEVAGM-------
                   Mycobacterium bovis           AAB48045   -L---L-CIV--A--S- ---N---A- --P  -TPD--I-TMPD      QE-I VAG-A---------
                   Mycobacterium leprae          P53529     -L---L-CIV--AT-S- --VQ---A- -EP  ATPD--ISTMAD      QK-I VAG-A-F-------
                   Mycobacterium tuberculosis    CAB10956   -L---L-CIV--A--S- ---N---A- --P  -TPD--I-TMPD      QE-I VAG-A---------
                   Corynebacterium glutamicum    BAB98810   -L---L-CTV--A-------EQ-S-T- --P  AAT---I-TME-      QKAA VSG-A---------
                   Streptomyces coelicolor       CAB52840   -L---L-CIVV-A---L--VAD---T- --P  ATAH---STMA-      QLDI VAG-G---------
                   Gordonia rubripertincta       AY876143   -L---L-CVV--A--S- --DE-S-T- --P  -TPH--ISTMAD      QADA VAG-A---------
                   Trichotomospora caesia        AY876141   -L---L-CIV--A---L--IPD---T- --A  VTAH--ISTME-      QLAY VEGAG---------
                   Cellulomonas fimi             AY876142   -----L-CMV---S---L--DG-S-   --D  -PAH--I-TMA-      QLAI VGGAG---------
                   Microtetraspora niveoalba     AY876144   -L-----CMV--A---L--IED-G-T- --P  ETPAA-ISTMAD      QEDV VSGER-M-------
                   Rhodococcus rhodochrous       AY876145   -L---L-CVV--A--S-GIEDASSTE  --P  -TTA--ISTMAD      QELA VAG-A---------
                   Kocuria rhizophila            AY876146   -L---L-CMV------EV--PN-S-T- --P  ETDT--I-TME-      QKQF VEGAG---------
\ctinobacteria     Propionibacterium acne        AY876147   -L-----C-V--V--DL--IKD-A--- --S  QTPD--I-TMA-      QVEA VAGKA---------
                   Planobispora rosea            AY876148   -L-----CMV--A---L--IED-G-T- --P  ETTH--ISTMAD      QEDV VSG-R-M-------
                   Clavibacterium michiganensis  AY876149   -L---L-CMV------E-D-PG-S--- --P  -SAF---TMAE       QVDI IAGG ---------
                   Leifsonia xyli                YP_061635  -L---L-CMV-E---NE-GLAG-S-S- -D-  -TAF--I-TMA-      QVDI IAGG ---------
                   Tsukamurella paurometabola    AY876150   -L---L-CIV--A---SAGLDGASSAE -EP  -AKY--ISTMAD      QEQA VAG-A---------
                   Nocarida corynebacterioides   AY876151   -L---L-CMV--A--S- --SD---A- -EP  ETT---ISTMAD      QEQA VAG-A---------
                   Nocardioides simplex          AY876152   -L---L-SMV-----TEL-----G-T- --P  -TPE--I-TME-      QKSI VEGAG---------
                   Kribbella sandramycini        AY876153   -L---L-CMV--T--AL---ER---T- -EE  PCQH--ISTMAD      QHDV ISGDR-M-------
                   Saccharopolyspora erythraea   AY876154   -L---L-CMV--T--AL---ER---T- -EE  P-QH--ISTMAD      QHDV ISGDR-M-------
                   Microbacterium oxydans        AY876155   -L---L-CMV-----D---IEG-S--- --P  ETAEP-I-TMA-      QVDI LDGG ---------
                   Tropheryma whipplei TW08/07   NP_789056  -------CMV-E-----VG-HG-S--- -TD  -TQW---TTML-       QRD ILIDDQF-------
                   Tropheryma whipplei Twist     AAO44197   -------CMV-E-----VG-HG-S--- -TD  -TQW---TTML-       QRD ILIDDQF-------
                   Symbiobacterium thermophilum  YP_073869  ----M---S-V--A---LL--S----T- -VT  --KD--IDMMAQ            QKQVT---------
                   Staphylococcus aureus         BAB95916   -------L-TV-FS---L--EG-H-A-L -P  ATPY-IID-LP-            QK-IE-----L---
                   Bacillus subtilis             P13242     -------V-S-------L--KG-H-A- I-P  STQY-IID-LP-            QK-VE-----L---
   Firmicutes      Streptococcus pyogenes        AAM80239   -------LTCV-F--H-LNMEG---F- LEP  STKY-IIDIMRD            QI-IE-M---L---
                   Lactococcus lactis            CAA09021   -------LTAV-F----L--EG-H-FA L-P  ETKY--IDIMRD            QV-VE-M---L---
                   Mycoplasma capricolum         CAA42665   ---------T-SI --DLLNW-D-D-T- -N-  NTTH-IFDY-K            GI-RDNI---L---
                   Clostridium acetobutylicum    NP_349494  -------C-V-------L--EG-H--- I-P  QTKY--ID-MPD            QK-ID-K-------
```

Figure 2.2. Partial alignment of CTP synthetase sequences showing a 10-aa indel (outer box) that is specific for proteobacteria. And at the same position a smaller 4-aa indel for all Actinobacteria. The sequences whose accession number start with "AY" were amplified in this work.

```
Actinobacteria
Trichotomospora caesia       AY956800  GCGTAGTCGATGGA-CAACCGGTTGA-TATTCCGGTACCCGCTTTGAAACGCCCAATATTGAATCCTCTGATGCTAAGTCCGTGAA----
Thermomonospora chromogena   AF116563  GCGTAGTCGATGGA-TAACGGGTTGA-TATTCCCGTACCCGCCGTGGTGCGCCCA-CGTCGAGGCCGTTGATGCTAACCCGTCGAG----
Pseudonocardia halophobica   AY956793  GCGTAGGCGATGGA-TAACGGGTTGA-TATTCCCGTGCTCGTGATAGTGCGTCCA-TGCCGAGGCTGTTGATGCTAACCATCCGAA----
Saccharopolyspora erythraea  AY956803  GCGTAGGCGATGGA-TAACGGGTTGA-TATTCCCGTACCCGTGCGCATGCGTCCA-TGGTGAAACGGTTGAGACTAACCATCCG------
Tsukamurella paurometabola   AY956797  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCCGTGTCAGATCGCCCC-TGATGAATCAGTTG-TACTAACCGTCCTGA----
Gordonia rubripertincta      AY956799  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCCGTGTCAGATCGCCCC-TGATGAATCAGTTG-TACTAACCGTCCTGA----
Streptomycoides glaucoflavus AY956805  GCGTAGGCGATGGA-CAACGGGTTGA-TATTCCCGTACCCGTGTATCCGCGCCCA-TGCTGAATCAGTTG-TACTAACCATCCAGA----
Williamsia murale            AY956801  GCGTAGGCGATGGA-CAACGGGTCGA-TATTCCCGTACCCGTGTAGTCGCGTCCG-TGATGAATCAGCAG-TACTAACCATCCTGA----
Micromonospora cherisina     AY956804  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCCGCGAAAGAGCGACCC-TGACGAACCTCGTTGTGCTAACCACCCAAA----
Kocuria rhizophila           AY956795  GCGTAGGCGATGGA-CAACGGGTTGA-TATTCCCGTACCGATGAAGAACCGACCC-TACTGA-GCCGGGATACTAACCACCCGAGCCAC
Renibacterium salmoninarum   AF143477  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCGGCGAAGAACCGCCCA-TACTGA-GCAGGTGATACTAACCGCCAGA---AG
Microbacterium oxydans       AY956796  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCGGCGAAGAACCGCCCA-AGCTAA-TCCAGTAGTGCTAAGTGTCTGA---AT
Leifsonia xyli               AE016822  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCGGCGAAGAACCGTCCA-AGCTAA-TCCAGTGGTGCTAAGAGTCCTA---AT
Oerskovia turbata            AY956798  GCGTAGGCGATGGA-CAAGGAGTTGA-TATTCTCCTACCGGCGAAGAACCGCCCA-TACCGAACCCGGTGATGCTAAGCGCCCTT---AA
Kribbella sandramycini       AY956794  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCGGCATTAACACGACCC-GACCGAACCTGCTGATGCTAAGTCTTCGA-----
Tropheryma whippelii         AE016850  GCGTAGTCGATGGA-CAACGGGTTGA-TATTCCCGTACCGGCAAAGAACCGCCCA-TATT---CCGCGTAGT------------------
Microbispora bispora         U83912    GCGTAGTCGATGGG-CAACGGGTTGA-TATTCCCGTACCCGCCGTGGCGCGTTCT-G---------------------------------
Rubrobacter radiotolerans    AY956802  GCGTAGGCGATGGA-AAACAGGTTAA-TATTCCTGTACTTCCAGTTATTGGTCGCT----------------------------------
Symbiobacterium thermophilum AP006840  GCGTAGGCGATGGG-AAACAGGTCGA-CATTCCTGTACCACCTACG--------------------------------------------
Other bacteria
Mycoplasma penetrans         NC_000912 GCGTAGCTGATGGA-TAACAGGTTAA-TATTCCTGTACCAATGTAT-G------------------------------------------
Listeria innocua             X92949    GCGTAGGCGATGGA-CAACAGGTAGA-GATTCCTGTACCAGTGCTAATT-GTTTA-----------------------------------
Bacillus subtilis            Z99104    GCGTAGGCGATGGA-CAACAGGTTGA-TATTCCTGTACCACCTCCTCACCATTTG-----------------------------------
Lactococcus lactis           X64887    GCGTAGTCGATGGA-CAACTGGTTGA-TATTCCAGTACTAGATAT-GATCGT--------------------------------------
Staphylococcus aureus        BA000017  GCGTAGGCGATGGA-TAACAGGTTGA-TATTCCTGTACCACCTAT-AATCGTTTT-----------------------------------
Clostridium tetani           NC_004557 GCGTAGGTGATGGA-CAATCGGTTGA-TATTCCGATACCGCCAACTTTC-GTTTG-----------------------------------
Synechocystis sp.            BA000022  GCGTAGTCGATGGA-CAACCGGTCAA-TATTCCGGTACTGATTATAGATTGT--------------------------------------
Thermus thermophilus         X12612    GCGTAGCCGAAGGG-CAGCCGGTTAA-TATTCCGGCCCTTCCCGCAGGT-----------------------------------------
Chlamydia muridarum          U68437    GCGTAGACGATGGAGCAGCAGGTTAAATATTCCTGCACCACCTAAAACTAT---------------------------------------
Aquifex aeolicus             AE000751  GCGTAGCCGATGGG-AAGCGGGTCAA-CATTCCCGCGCCAGCTCGGTGG-----------------------------------------
Escheriachia coli            NC_000913 GCGTAGTCGATGGG-AAACAGGTTAA-TATTCCTGTACTTGGTGTTACT-----------------------------------------
                                       ******  ** **   *    **  *  ****     *

Actinobacteria
Trichotomospora caesia       GCCGCCCCTGATCTCTTCGGAG--TGAGGGGAGTGGTGGAGCCGACGACCCGA-GGTGGTA----GTAGGTAAGCGA--TGGGGTGACGCAGGAAGGTA
Thermomonospora chromogena   TCCGGCCACTCTCTTCTTTGAG--GGGGGTGGTGTGGAGGAGCCGGGGACCCGA-GGCGGTA----GTAGGCGAGCGA--TGGGGTGACGCAGGAAGGTA
Pseudonocardia halophobica   -CCCGCCTCTGAGTCCTTCGGG-ACGAGGGGAGTGGGGAGCGTGGGGTCCGA-TTCGGTA----GTAGGCAAGCGA--TGGGGTGACGCAGGAGGGTA
Saccharopolyspora erythraea  --CTGGCTGTGTGAGTCTTCGG-ACGAGCGTAGTTGGT---GCATGGGACCTGA-TTCCGCG----GTAGTCAAGTGA--TGGGGTGACGCAGGAGGGTA
Tsukamurella paurometabola   -AGCACCTTGATCACCTTCGGG-TGACGGTGGTGTGGAT--GCACGGGACCTCG-GCTGGTA----GTAGTCAAGCGA--TGGGGTGACGCAGGAAGGTA
Gordonia rubripertincta      -AGCACCTTGATCACCTTCGGG-TGACGGTGGTGTGGAT--GCACGGGACCTCG-GCTGGTA----GTAGTCAAGCGA--TGGGGTGACGCAGGAAGGTA
Streptomycoides glaucoflavus -TCCCCCAGGAGCACCTTCGGG-TGCCGGGTGGGGTGAT--GCATGGGACCTTG-GCTGGTA----GTAGGCAAGCGA--TGGGGTGACGCAGGAAGGTA
Williamsia murale            -AGATCAAGGGTTACCTTCGGG-TTTCGTTTGGTTGGAT--GCATGGGACCTTT-GCTGGTA----GTAGTCAAGCGA--TGGGGTGACGCAGGAAGGTA
Micromonospora cherisina     -CCAGCCAAGGT---CTTCGGA-CTGAGGTTGGGGA-----GCGTGGGAACCTG-GCGGGTA----GTAGTCAAGCGA--TGGGGTGACGCAGGAAGGTA
Kocuria rhizophila           CATGACCGTGACCCCTTGTGGGTCGCGGGGTGTGTGGTGAGGC-TGGGACCTGATCCGGGGA---GGTAAACGTGTTAACAGGTGTGACGCAGGAAGGTA
Renibacterium salmoninarum   CATGATCGATCACCCTTGTGG--TGTGAGGTTTTTGTGGATCGCGGGACCTTATCCTGGGA---GGTAAGCGTATTAACAGGTGTGACGCAGGAAGGTA
Microbacterium oxydans       CCCAGTGACTGATCCCTTCGGG--GTGACGCTCTGGGCCTAGCGCACGACCCCATTCTGGTGC--GGTTAGCGTATTAACAGGTGTGACGCAGGAAGGTA
Leifsonia xyli               CCTGGACACCGATCCCTTCGGG--GTGACGGTCCAGGTCTAACGCTCGACCCCATGCTGGTGC--GGCTAGCGTATTAACAGGTGTGACGCAGGAAGGTA
Oerskovia turbata            CCCG--CACCGTCTCCTTCGGG--A-GACATCGCGGGAGCGGCGCGCGACCCGA-ACCGGTACTAGGTAAGCGTATTAACAGGGGTGACGCAGGAAGGTA
Kribbella sandramycini       ----------AACCATGAGGCCTTCGGGTTGAGTGGCGGAGCAGACGGCCCGA-GGTGGTA----GTAGGTGCATTG--AGGAGTGACGCAGGAGGGTA
Tropheryma whippelii         ------------TTCGTTCAGC--ATTTTG--CTGTGCTTACGGTAC--CCCTTTTACGGTGT--GCGGGATAAGTGTTCAGGTGTGACGCAGGAAGGTA
Microbispora bispora         --------------------------------------------------CCGCTTCGGTG----GCAGGTAAGCGA--TGGGGGGACGCAGGAAGGTA
Rubrobacter radiotolerans    -------------------------------------------------------------------------GCGA--TGGAGGGACGGAGAAGGCTA
Symbiobacterium thermophilum ----------------------------------------------------------------------TGACCGA--TGCGGGGACGCAGGAGGCTA
Other bacteria
Mycoplasma penetrans         ------------------------------------------------------------------------AATGA--TGGAGTGACGGAGAAGGTTA
Listeria innocua             ------------------------------------------------------------------------ACCGA--TGGGGTGACACAGAAGGATA
Bacillus subtilis            ------------------------------------------------------------------------AGCAA--TGGGGGGACGCAGGAGGATA
Lactococcus lactis           ---------------------------------------------------------------------------GA--TGGAGGGACGCAGTAGGCTA
Staphylococcus aureus        -----------------------------------------------------------------------AATCGA--TGGGGGGACGCAGTAGGATA
Clostridium tetani           ------------------------------------------------------------------------ACAAA--TGGGGTGACACAGAAGGATA
Synechocystis sp.            ---------------------------------------------------------------------------GG--CGG-GGGACGGAGAAGGCTA
Thermus thermophilus         -------------------------------------------------------------------------GCGA--TGGGGGGACGCTCTAGGCTA
Chlamydia muridarum          ------------------------------------------------------------------------AGCAA--AGGAATGACGGAGTAAGTTA
Aquifex aeolicus             -----------------------------------------------------------------------AGCCGG--TGTCGTGACGCAGGAGGCTA
Escheriachia coli            -------------------------------------------------------------------------GCGA--AGGGGGGACGGAGAAGGCTA
                                                                                                             *    ***     * * **
```

Figure 2.3 Partial alignment of 23S rRNA sequences showing a large insert (99~110nt) specific for Actinobacteria. Dashes in the alignment represent lack of nucleotide at that position. The sequences in the braces are from different actinobacterial strains and the left sequences are from representatives of other bacterial groups. The sequences whose accession number start with "AY" were amplified in this work.

```
                  Rhodococcus jostii          YP_707073    TDPSYHRQIVVATAPQIGNTGWNDEDDES VGPTGSSAE AKIWVAGYVVRDPARRTSSWR
                  Rhodococcus opacus          YP_002784124 ----------------------------- --------- ---------------------
                  Rhodococcus erythropolis    ZP_04387376  -----------S----------------- -----E--G -------------S-V-----
                  Rhodococcus rhodochrous         *        -----------S----------------- -----E--G -------------S-V-----
                  Mycobacterium tuberculosis  ZP_02550712  -----------------------G--S--        RG ER------A----SP-A-N--
                  Mycobacterium avium         NP_960052    -----------------------G--A--        RG D-------A----SP-V-N--
                  Mycobacterium ulcerans      YP_905731    --------------------------G--        RG D-------A----SP-A-N--
Corynebacterineae Tsukamurella paurometabola  ZP_04028273  -----D-----------------------        RG D-------A--N-T--V-N-
                  Tsukamurella paurometabola      *        -----D-----------------------        RG D-------A--N-T--V-N--
                  Gordonia bronchialis        ZP_03884043  -----------------------T-----        HG G-------A--N-T--V-N--
                  Corynebacterium glutamicum  YP_001138553 --------------------------N--        RD G------L-I--L-A-V-N--
                  Corynebacterium jeikeium    YP_250805    --------------------------S--        RG D------L-I--LS-SV-N--
                  Corynebacterium efficiens   NP_738340    -----------------------E--N--        HD GS-----L-I--L-V-V-N--
                  Nocardia farcinica          YP_119832    -----------------------------           QR------A-------V-N--
                  Janibacter sp.              ZP_00995511  --------V--M---HV------------           SR---S--------L-P-N--
                  Kribbella flavida           ZP_03864675  ----------TQ---H-----V-------           QR-------------VP-N--
                  Salinispora tropica         YP_001158695 --------V--Q---H-----V-G-----           SR-------------IG-N--
                  Streptomyces albus          ZP_04700923  --------V--M---HV----V----P--           RR---S----------A-N--
                  Brevibacteriumlinens        ZP_00381474  ----------IQ---H-----IT-D----           ------------AS-VS-N--
                  Thermobifida fusca          YP_289117    -----------M---H-----V--D-A--           HR---S-----E-S-IV-N--
     Other        Frankia sp. CcI3            YP_482285    S---F-G-V-IM---H-----V----Y--           ER-Q---F--------A-N--
 Actinobacteria   Arthrobacter aurescens      YP_948006    -----A--L--Q---H-----V-S--A--           RR----------A---P-N--
                  Kocuria rhizophila          YP_001855186 -----AG----Q---H-----V-T--M--           RR---N-F--------P-N--
                  Bifidobacterium animalis    YP_002470017 -----D-----Q-F-H--D--V--D-L--           SR-----------SMVV-N--
                  Clavibacter michiganensis   YP_001710721 -----AG---LQ---H-----M--D-M--           RR-----------S-VV-NF-
                  Kytococcus sedentarius      ZP_04042894  ----FAG---AL---HV------------           RR-FTE-I------L-P-N--
                  Acidothermus cellulolyticus YP_873057    ------K-V--M---H-----V----P--           -R---S---------IP-N--
                  Leifsonia xyli              YP_062065    -----AG---LM---H-----T----M--           RQ-----F---E-S-VV-NF-
                  Lactobacillus casei         YP_806673    --Q--NG--ITF-Q-L---V-I-RD-Y--           IDPTAK-I----V--V-GN--
                  Clostridium cellulolyticum  YP_002504976 -----CG---CM-Y-L---Y-V-I--I--           L-PQ-K-FI--ELCKTP-N--
                  Listeria monocytogenes      YP_014457    -----YG--ITF-Y-LV--Y-V-RD-F--           INPA-K-V---EA-EFP-N--
   Firmicutes     Bacillus sp.                ZP_01171604  S----CS----L-Y-L---Y-I-RD-F--           ILPAIH-F--KEA-EYP-N--
                  Thermotoga maritima         NP_228368    -----TG----M-Y-E--IY-V----V--           DG-K---F--YRSVDTP-N--
                  Moorella thermoacetica      YP_429741    -----CG---AL-Y-L---Y-I----L--           DGPR---F--HEACP-P-N--
                  Alkaliphilus metalliredigen YP_001321950 ------G---TM-Y-LM--Y-I-L--V--           S-VR-KALI--EQCVKP-N--
                  Aquifex aeolicus            NP_213287    -----KG----M-YT----Y-V----I--           KSVQ-N-F--KEAFF-Y-N--
                  Xanthomonas campestris      YP_001903533 -----A--M-TL-Y-H-----F--Q---A           KQV-A--LI---VP--P----
                  Roseiflexus sp.             YP_001278349 -----YG----M---H-----V-R--E--           RHP----F---AASPYV-N--
                  Xylella fastidiosa          ZP_00679926  -----AY-L-TL-Y-H-----CT-Q---A           N-V-A--LI---VP--P-N--
  Gram-negative   Methylococcus capsulatus    YP_114290    -----A----AL-Y-H-----I-E--T--           TG-FAS-L-I--LPL-A-N--
    Bacteria      Marinomonas sp.             YP_001342866 -----A--M-TL-Y-H-----V-S--E--           D-V-CE-LII--LPLVA----
                  Alcanivorax borkumensis     YP_692037    -----AK---TL-Y-H-----VTA--E--           SR--AS-L-I--L-MTV-N--
                  Chloroflexus aggregans      YP_002463948 -----CG-L-TM---H-----V--F-P--           IQPQ---FI--NYSEHY----
                  Synechococcus elongatus     YP_401139    -----CG---SF-Y-EL----V-P--E--           -RPQ-S-LIA-NV---H-N--
                  Pseudomonas putida          NP_746832    -----AQ---TL-Y-H-----TTP--A--           SRV-S--L-I--LPLLA-N--
```

Figure 2.4. Partial sequence alignment of carbamoyl phosphate synthase small subunit (CarA) homologues showing a 2-aa indel (boxed) in a conserved region that is found uniquely in species of *Corynebacterineae*. Another 7-aa is only found in *Rhodococcus* species. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequences denoted by * were amplified by the author. Sequence information for only representative species is presented.

```
                      Bifidobacterium dentium      ZP_02918696   SPKKSWEGLAGSIVFAMVGAFVVMFCTYDAS VWMSR WWVPVIAGIMIGIAGTFGDLCASMLKR
                      Bifidobacterium angulatum    ZP_04447950   ------------VV-A-V-A--------TPQ --AT- ----IIM-ILV-VV----------I--
                      Bifidobacterium bifidum      ZP_03646485   -----V---F-----A---A---F----DGA --A-- ----I-M-VLT-V--------------
                      Bifidobacterium longum       ZP_00120405   -----V---V--M--A-A-A---FA---DAS K-AT- ----I---IL--AV------------
                      Bifidobacterium longum       NP_696666     -----V---V--M--A-A-A---FA---DAS K-AT- ----I---IL--AV------------
Bifidobacteriaceae    Bifidobacterium longum       ZP_03977351   -----V---I--M--A-A-A---FA---DAS K-AT- ----I---IL--AV------------
                      Bifidobacterium longum       YP_002322530  -----V---I--M--A-A-A---FA---DAS K-AT- --M-I---IL--AV------------
                      Bifidobacterium breve        ZP_03619258   -----V---I--M--A-------FIF--DPS LQT-- ----I---IL--AV------------
                      Bifidobacterium animalis     ZP_02963058   -----V---C--M--AVA-AY--FA---EPA L-AT- --A-LLT-VI---V--Y---------
                      Bifidobacterium gallicum     ZP_03447661   ---------------A-A--Y--FA---DAA I-AT- ----II---I----------------
                      Gardnerella vaginalis        ZP_03937379   -----I---L--M--AL--AY-IISGF---S  --KF --M-AIM-VAS-MV------------
                      Mycobacterium gilvum         YP_001135380  ------------L--GVTA-VLAVVFLLDKP       --AG-AL-L-LV-T--L---VE-QF--
                      Corynebacterium efficiens    NP_738519     --------F-----LGSLTGAITVHFLLHHH       --LG-IL-ICLVVCA-L---VE-QF--
                      Nocardioides sp.             YP_924428     --------F---VV-CLAAGW-LVVYLLEG-       ---GLLL-LIAVVMA-L----E-VI--
                      Stackebrandtia nassauensis   ZP_04485170   --------MG--LVACA--GAVTLQLMFDVP       --QGA-F-VALA-TA-V---TE-LI--
                      Propionibacterium acnes      YP_056222     --------F---VITAAFVGW-CLGGLLSAP       --AGI-L-VVLALT--A---VE--I--
                      Renibacterium salmoninarum   YP_001623696  --------F---LAGAILIGVLAAIFLLHEQ       --IGI-LAIGLVA-A-T---AE--V--
                      Mobiluncus curtisiiATCC      ZP_03923891   --R-T---F---V--AVTV-V-GT-WMGIPW        FYGIIL-VLMAVV--I---SE-LM--
                      Streptomyces coelicolor      NP_629762     --G-TR---L-A-A-A-VAGALC-QFLIDDG      A--QGLLL-LVVAVSA-L---GE--I--
                      Brevibacterium linens        ZP_00380132   --------YF--VV-AAAVATILALTLFDAP       F-TGL-F-VV-PAFA-L--FSE--I--
                      Kytococcus sedentarius       ZP_04042358   --------F---VALAAL-GV-S-AWLVDGP       L--G-PL-IAVAL-S-V--FAE-A---
       Other          Micromonospora sp.           ZP_04608036   --------F---VTAAAV--ALLIWLLFDVA       P-WGALF-VAVSC-AVL---AE--I--
                      Actinomyces odontolyticus    ZP_02044796   --------F---A-TAIAVGVVGLWLLGA-W        -WG----IS-AFV--M---TE-LI--
Actinobacteria        Actinosynnema mirum          ZP_03819236   --------F---LVAG-A-GVLTVGALL-GQ       --HG-LF-AA-VVTA-T---IE-LI--
                      Gordonia bronchialis         ZP_03883620   ---------V--LVVGTT-AVCCVIFLL--H       --IGA-L-PILVVCA-L---VE-QV--
                      Nakamurella multipartita     ZP_04349609   --------FG--LAAC-LAGALCVWLLL-GH       ---GLLL-VA-AVTA-V---TE-LI--
                      Kineococcus radiotolerans    YP_001361171  --------T---LAVGAVTGAIALPLALDGS       --GGALL-LVTVVVAVL---SE-----
                      Tsukamurella paurometabola   ZP_04026609   --------F---MVAGA--AVL-LKYLLDVN       PLWGLIL-PVVV-TA-L---LE-QV--
                      Arthrobacter sp.             YP_830865     --------F----GGAIAVGVLACLFLLDKP       ---G--LAVGMVA-S-I---SE--V--
                      Janibacter sp.               ZP_00997184   --------F---AFTCAVVGAVGIATLLDGP       --KGA-I-VLVAA-A-I---IE-SI--
                      Salinispora arenicola        YP_001536131  ------------LGAAAF--ALLIWLLLDVA       P-WGALF-VAVSA-AVL---AE--I--
                      Jonesia denitrificans        ZP_03867243   ------------FIL-LAVGIVGSAVVLDYN       PM-G-AL-LLTP-TA-V---AE-LI--
                      Frankia sp.                  YP_482659     --G-----F---AVTCV-VAGI-LAWPLEAE       A-QG-LL-LAVACTA-I---GE-L---
                      Geodermatophilus obscurus    ZP_03891094   --------M---V-GCVLVATPIVTLAL-GP       --GG-LF-VALAVSA-A---GE-LI--
                      Eubacterium biforme          ZP_03489530   --------FV-G-V-GF-L-L--S-SYVSNL       NPVLNTLLCLLCP-TAEL----F-AI--
                      Bacillus pumilus             ZP_03053273   --N-TV--FI-G-VTAVVL--VFQAI-GFLP       SYLLVMFITLLLS-F-QL---VE-A---
     Firmicutes       Geobacillus sp.              YP_002949272  --N-TV--SI-G-VCAVVVAIIYQL--NLF-       SFALLI-MTIVLS-F-QL---VE-AF--
                      Bacillus subtilis            NP_389536     --N-TV--F--G-VIALVLATIFQLVAQLPI       PYIYLLLITLFLSVF-QL---VE-A---
                      Lactobacillus reuteri        ZP_04022480   --N-T---SI-GT-AAV-ILAIYCYFIPVGA       G-VTMIFVTLILS-F-Q----IE-S---
                      Bacteroides coprocola        ZP_03011970   --------SI-GAV--I-AAIVLAHFFTFLS        TG-WIGL-LTVVVF--W---TE-LM--
                      Roseiflexus sp.              YP_001276350  --R-T---A--GMVGALA-A-VALALFGLPL       SL-ATTLI-IAA--V-PI---SE-FI--
                      Legionella drancourtii       ZP_05108845   --G-----VL-GVILA--IAGVGCIYFAPVA        KIYWF-LALYTV-ISI----FI-I---
   Gram-negative      Chloroflexus aurantiacus     YP_001637143  --------F--GMVAAIATALFCVPLLGLPI       TLLEAAIL-VIA--F-PL---AE-LI--
                      Myxococcus xanthusDK         YP_630775     --N-T---FF-GM-G-VG-M-IARQFFFPVF       TV-DC-LL-IAG--L-PV----E-----
      Bacteria        Caulobacter crescentus       NP_420725     --N-T-A-FV-GLAAAT--AVV-ASLAKLDL       I-QAAALI-LLG-L-TMA---WE-----
                      Cytophaga hutchinsonii       YP_676988     ----T---SI-GLATA-LF-YLLYYFYGIFS        IP-WMGLCVIVV-S-SL---VE--I--
                      Escherichia coli             ZP_03048092   --G-TL---V-GVITT--A-LIIGPLLTPLN       TSQALL--LL---S-FC--VVM-AI--
                      Desulfovibrio piger          ZP_03312394   --N--S--AV--LVGCV-FCTIYGA-LGSAS        --AFALL-IA-NAFAQL---FE-A---
                      Nitrosococcus oceani         YP_343728     --G-TV---I-G-ATTIVL-WSLASWLTPLN        VPQS-A--AL-----FV--VTI-A---
```

Figure 2.5. Partial sequence alignment of Cytidylyltransferase (CDP) homologues showing a 5-aa indel (boxed) in a conserved region that is found uniquely in species of *Bifidobacteriaceae*. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

```
                    Bifidobacterium dentium      ZP_02917203  QQVVKIVNEELTDVLG|A|GVDRPLNFAKNPPTIIMLAGLQGAGKT
                    Bifidobacterium breve        ZP_03619965  ----------------|Q|---------------------------
                    Bifidobacterium longum       ZP_00206561  ----------------|Q|---------------------------
                    Bifidobacterium bifidum      ZP_03645746  ----------------|Q|---------------------------
                    Bifidobacterium angulatum    ZP_04448583  ------------S---|-|---------------------------
Bifidobacteriaceae  Bifidobacterium adolescentis YP_909074    ------------A---|-|---------------------------
                    Bifidobacterium catenulatum  ZP_03324872  ------------A---|-|---------------------------
                    Bifidobacterium animalis     ZP_02963208  ----R---D----I--|Q|---------------------------
                    Bifidobacterium gallicum     ZP_03446040  --------D----I--|Q|-----I----Q----------------
                    Gardnerella vaginalis        ZP_03937056  ----R---D---N---|S|----------Y---V------------
                    Streptomyces avermitilis     NP_823824    ---L-------VTI--   -ET-R-R-------V------------
                    Thermomonospora curvata      ZP_04031383  -----------IEI--   -ET-R-R---T---V------------
                    Kineococcus radiotolerans    YP_001361148 -----------VAI--   -ET-R-R-S-----V------------
                    Rhodococcus erythropolis     YP_002765889 -----------VE---   -ET-R-Q---T---V--------S---
                    Corynebacterium glutamicum   YP_001138843 ---I-------VQI--   -ET-R-SL------V------------
                    Nocardia farcinica           YP_120373    -----------VGI--   -ET-R--L--T---V------------
Other               Leifsonia xyli               YP_062408    ----Q------IGI--   -EQ-R-Q---K---V------------
Actinobacteria      Salinispora arenicola        YP_001536089 --II-------IN---   -EG-R-Q---Q---V--------S---
                    Thermobifida fusca           YP_288722    ---I---H---IE---   -ET-TIR---T---V------------
                    Clavibacter michiganensis    YP_001222104 ----Q------VGI--   -QQ-RIQ---K---V------------
                    Mycobacterium tuberculosis   ZP_03538017  -----------ISI--   -ET-E-A---T---VV-------S---
                    Propionibacterium acnes      YP_056151    --I--------VEI--   -QT-TVR---T----V-----------
                    Brevibacterium linens        ZP_00378147  -----V--D--VGI--   -ET-R--Y--R---V------------
                    Frankia alni                 YP_715946    ---I-------VAI--   -GTTT-R---T---V-L------T---
                    Rubrobacter xylanophilus     YP_644162    -----------ANLM-   -SAHK-SY-SR---VV-----N-H---
                    Moorella thermoacetica       YP_429824    -------H----ALM-   -GESKI-W-SQ---V------------
                    Alkaliphilus oremlandii      YP_001513003 ---I--------ELM-   TSQSKI--SSK-------V--------
                    Lactobacillus fermentum      ZP_03945747  --I-----D---EMM-   ETAT---KSAHI--V--MV--------
Firmicutes          Exiguobacterium sp.          YP_002887251 -------H---SNLM-   SDVV-IT-SQK---VV-MV--------
                    Bacillus subtilis            NP_389480    ---I-V-Q----ELM-   -EESKIAV--R---V--MV--------
                    Clostridium butyricum        ZP_02948600  ---I--------NLM-   -SESK-SYNSSG--V---V--------
                    Symbiobacterium thermophilum YP_075294    -M-I---Y---VALM-   -ESVG--M-DR-------C--------
                    Streptococcus suis           YP_001198242 --II---D----A---   SETSEIIKSPKI-----M---------
                    Magnetococcus sp.            YP_864434    ---I---HD--VA-M-   AANES--L-NQ--VVV-M-----S---
                    Fusobacterium sp.            ZP_04571566  --FI-L--D--VEL--   -TSSK-TKGLRN---------------
Gram-negative       Roseiflexus castenholzii     YP_001431777 ---I---HQ--I-L--   QANV--AE-RPG------I----S---
                    Prochlorococcus marinus      YP_001484713 -KFIEV--K--INIM-   NENS---EN--S--V-LM---------
Bacteria            Leptotrichia buccalis        ZP_04338628  --F-----D--VE---   -SNVSIAK-DKN---V--S--------
                    Synechococcus sp.            YP_001227964 --FI-L-H---VE-M-   -ANA--AK-EQS--VVLM---------
                    Bacteroides capillosus       ZP_02036484  -MI---------ALM-   -ESAK-TISPK---VV--V--N-----
                    Marinobacter algicola        ZP_01894823  -VF--V-QQ--ER-M-   DGNES--L-VQ--AV--M---------
```

Figure 2.6. Partial sequence alignment of signal recognition particle (SRP) homologues showing a 1-aa indel (boxed) in a conserved region that is only found in *Bifidobacteriaceae* species. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

Corynebacterineae/ Pseudonocardineae/ Micromonosporineae	Mycobacterium smegmatis	YP_885825	LSNRVAFRRAMRKAIQSAMRQP	N	VKGIRVQCSGRLGGAE
	Mycobacterium leprae	NP_302259	----------------------	-	----------------
	Mycobacterium avium	YP_883600	----------------------	-	----------------
	Mycobacterium tuberculosis	ZP_03431583	M---------------------	-	----------------
	Gordonia bronchialis	ZP_03883087	----------------------	-	----------------
	Rhodococcus erythropolis	YP_002765305	----------------------	-	----------------
	Nocardia farcinica	YP_116948	--------------------S-	-	----------------
	Saccharomonospora viridis	ZP_04507339	----------------TS--S-	Q	----------------
	Stackebrandtia nassauensis	ZP_04484461	-AS--N-------S----L-N-	T	----K-A---------
	Corynebacterium efficiens	NP_737138	----------------------	Q	----K-V---------
	Corynebacterium glutamicum	NP_599754	-T--------------------	Q	----K-V---------
	Corynebacterium diphtheriae	NP_938858	-T--------------G-----	Q	----K-V---------
	Micromonospora sp.	ZP_04605824	--S--S--------M----KN-	V	C------V--------
	Tsukamurella paurometabola	ZP_04025728	----------------------	-	----------------
	Saccharopolyspora erythraea	YP_001108919	-----S--------------S-	Q	--------G-------
	Nakamurella multipartita	ZP_04349023	-----S-------SM---Q-S-	Q	-----I----------
	Actinosynnema mirum	ZP_03818686	--------------------SS	Q	--------G-------
	Geodermatophilus obscurus	ZP_03890919	--S--S--------M---Q-S-	Q	--------------T-
	Salinispora arenicola	YP_001539081	--S--S--------M----KN-	V	C------V--------
	Salinispora tropica	YP_001160725	--S--S--------M----KN-	V	C------V--------
Other Actinobacteria	Frankia alni	YP_711346	--S--S--------M-T--KGG		A---------------
	Acidothermus cellulolyticus	YP_872072	--Q--S--------L---LKAG		A--V---VA-------
	Janibacter sp.	ZP_00993899	--A--S-------SM---T-AG		A------V--------
	Thermobifida fusca	YP_290696	--S----------------KSG		A---------------
	Nocardiopsis dassonvillei	ZP_04331909	--S-----------M-T--KSG		A----I--G-------
	Cellulomonas flavigena	ZP_04366346	-AS--S-------G----Q-AG		A------V--------
	Streptomyces avermitilis	NP_826109	--S--S-------SM----KAG		A---KI--G-------
	Actinomyces odontolyticus	ZP_02043500	-AA--S-------G----Q-AG		A------V--------
	Renibacterium salmoninarum	YP_001625305	--S--------K--M---Q-AG		A---------------
	Kineococcus radiotolerans	YP_001360447	-AS--S-------GM-TT--SG		A-------A-------
	Micrococcus luteus NCTC	YP_002957751	-AS--------K--------AG		AQ---I----------
	Arthrobacter aurescens	YP_948651	-TS--------K--M---Q-AG		A-----A---------
	Bifidobacterium longum	YP_002323668	-T---T--------Q-D---AG		A----IKL--------
	Kocuria rhizophila	YP_001854475	-AS--------K------Q-AG		A----I--A-------
Firmicutes	Clostridium bartlettii	ZP_02210561	IER--------KQ-V-R-LKSG		A---K-AA--------
	Eubacterium hallii	ZP_03716676	-E--IS-----KSCM-RT--NG		AL--KTS--------D
	Pediococcus pentosaceus	YP_804892	-EG--------KQ-M-RS--SG		A---KT-VA---N--D
	Natranaerobius thermophilus	YP_001916387	-ER--------KQ-VGR----G		A---K-M---------
	Ruminococcus sp.	ZP_04857889	-E--IS-----KSTM-RT-KAG		A---KTSV-------D
	Oceanobacillus iheyensis	NP_6910461	-E--IS----QKQ---R---GG		A---KT-V-------D
	Bacillus clausii	YP_1736601	-E--IS-----KQ---RT--AG		A---KT-V-------D
	Moorella thermoacetica	YP_4312791	-EK-I------KQ-VGR---LG		AQ--KIA-G---A---
Gram-negative Bacteria	Campylobacter coli	ZP_00370770	-EK-I------K-V--G-QKAG		A---K-SV--------
	Vibrio cholerae	ZP_01976812	-ER--M-----KR-V-N---LG		A---K-EV--------
	Helicobacter bilis	ZP_04581710	-ER--------K-VM-Q--KSG		A---K-KV----A---
	Myxococcus xanthus	YP_631504	-ER-I------K--L-T--KFG		A-----A---------
	Shigella flexneri	NP_709102	-ER--M-----KR-V-N---LG		A---K-EV--------
	Photobacterium sp.	ZP_01162659	-ER--M-----KR-V-N---LG		A---K-EV--------
	Haemophilus influenzae	NP_438942	-ER--M-----KR-V-----LG		A---K-EV--------
	Neisseria lactamica	ZP_03721539	-EK--Q-----KR-M-N---SG		A---KIMT----N--D
	Chlamydia trachomatis	YP_002888145	IER--S-----K--L--V-DAG		AL-VK--V----A---

Figure 2.7. Partial sequence alignment of 30S ribosomal protein S3 homologues showing a 1-aa indel (boxed) in a conserved region that is found uniquely in species of *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae*. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

Corynebacterineae/ Pseudonocardineae/ Micromonosporineae	Mycobacterium tuberculosis	NP_338293	LKVVEEPPPSTVFLLCAPS	VD	PEDIAVTLRSRC
	Mycobacterium avium	YP_879801	-------------------	--	---V-I------
	Mycobacterium leprae	NP_301270	--------S----------	-A	------------
	Corynebacterium glutamicum	NP_599564	--T----TE---MI----T	T-	-R---I------
	Corynebacterium efficiens	NP_736921	--T-----ER-III-----	T-	----M-------
	Corynebacterium diphtheriae	NP_938721	--T-----AH--II-----	T-	-T--IP-----S
	Rhodococcus opacus	YP_002781456	--------DR---------	--	-Q--S-------
	Rhodococcus jostii	YP_704302	--------DR---------	--	-Q--S-------
	Nocardia farcinica	YP_116574	--------ER---------	--	----S-------
	Gordonia bronchialis	ZP_03883864	--------SR---------	--	-D--S-------
	Saccharopolyspora erythraea	YP_001102653	--A-----DR---------	DH	---VS--I----
	Salinispora arenicola	YP_001539171	--A------R---------	TH	-D--S--I----
	Actinosynnema mirum	ZP_03816576	--A-----DR---------	EH	-Y-VS--I----
	Tsukamurella paurometabola	ZP_04025248	--M-----AQ-IV-----T	--	----S---K---
	Stackebrandtia nassauensis	ZP_04482602	--SI----ER---------	SN	-AE-S--I----
	Nakamurella multipartita	ZP_04346586	--A----AEH---------	TH	-D-VS--I----
	Geodermatophilus obscurus	ZP_03891755	--ML----AR---------	LH	-D-VP--I----
Other Actinobacteria	Jonesia denitrificans	ZP_03867625	--AI-----R-IW------		-Q-VM--I----
	Actinomyces odontolyticus	ZP_02044122	--AI----EH--W------		---MIA-I----
	Frankia sp.	YP_001504679	--AL---AERA--------		VD-VLP-I----
	Kocuria rhizophila	YP_001855735	--SI-----H-IW------		-M-VL--I----
	Propionibacterium acnes	ZP_03388497	--AI---A-K--W-----T		---VI--I----
	Arthrobacter aurescens	YP_946681	--AI---T-R-IWM-----		-A-VL--I----
	Nocardioides sp.	YP_921617	--AI---ADR--WM----T		V--VLP-V----
	Brevibacterium linens	ZP_00379802	--AI-----A--W------		-I-VLT-I----
	Kytococcus sedentarius	ZP_04041913	--AL---S-R-IWM-----		-Q-VI--I---S
	Streptomyces coelicolor	NP_627739	--A----A-R--W------		V--VLP-I----
	Micrococcus luteus	YP_002956470	--AI----AR-IWM--T--		-A-VLP-V----
	Nocardiopsis dassonvillei	ZP_04332090	--A----S-R--W---T-T		-D-LL--I----
	Thermobifida fusca	YP_290838	--AI---A-R--W---T-T		AH-LL--I----
	Kineococcus radiotolerans	YP_001360244	--TI-----Q--W-----A		VD-VLP-I---A
	Janibacter sp.	ZP_00994154	--AL---T-R--WM-----		V--VII-I---S
	Leifsonia xyli	YP_061516	--AL----ER--WI-----		EA-LLP-I---V
Firmicutes	Clostridium botulinum	ZP_04863530	--TL----EKV--I-ATTD		-QKLPI-IL---
	Bacillus thuringiensis	ZP_04068199	--L-----ENVL-MF-TTD		-DKMLD--LN--
	Alkaliphilus metalliredigens	YP_001322506	--TL-----HVI-I-ATTE		-QKLPA-IL---
	Thermosinus carboxydivorans	ZP_01666179	--LL-----RLI-I-LTAL		-HAVLS-I----
	Thermoanaerobacter sp.	YP_001661692	--TL-----HVI-I-ATTE		-DKLPD-IL---
	Bacillus cereus	ZP_04245393	--TL----AHVI-I-ATTD		-QKVPK-II---
	Eubacterium ventriosum	ZP_02026736	--TL----AYVI-I-ATTE		-HK-PI-IL---
	Staphylococcus carnosus	YP_002633223	--FL----EN-IAI-LSTK		--Q-LS-IH---
	Oceanobacillus iheyensis	NP_6909501	--TL----KHV--I-ATTE		-HK-PL-II---
Gram-negative Bacteria	Escherichia coli	ZP_03069580	--TL----EHVK---ATTD		-QKLP--IL---
	Salmonella enterica	YP_150888	--TL----EQ-W-F-AS-E		-ARLLA------
	Flavobacteria bacterium	ZP_01201256	--LI----AK-----I-ED		EDQ-IN-IK---
	Buchnera aphidicola	NP_777941	--II----QN-Y---INYL		-HK-IT------
	Synechococcus elongatus	YP_171340	--TL----DRV--V-ATTD		-QRVLP-II---
	Neisseria cinerea	ZP_03745288	---L-----QV----VSHA		VDKVLP-VK---
	Bordetella avium	YP_786284	---L----AH-----V-DA		-DRLLP--L---
	Aquifex aeolicus	NP_214275	--TL-----R---V--TTE		YDK-LP-IL---

Figure 2.8. Partial sequence alignment of DNA polymerase III subunit delta (holB) homologues showing a 2-aa indel (boxed) in a conserved region that is found uniquely in *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae* species. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

```
                    Arthrobacter chlorophenolicus YP_002489666 VFDLGGGTFDVSLLEVGKD EDNFS TIQVRATAGDNRLGGDDWD
                    Jonesia denitrificans         ZP_03868290  ------------------- ----- -------S-----------
                    Brevibacterium linens         ZP_00380512  ------------------- ----- -------S-----------
                    Beutenbergia cavernae         YP_002883713 ------------------- ----- -------N-----------
                    Micrococcus luteus            YP_002957918 ----------------A-- ----A -------------------
                    Actinomyces odontolyticus     ZP_02043593  ------------------- D-G-- -------S---H-------
                    Actinomyces urogenitalis      ZP_03926684  ------------------- --G-- -------N-----------
                    Renibacterium salmoninarum    YP_001625056 ------------------- D-G-- -------S-----------
                    Cellulomonas flavigena        ZP_04364840  ------------------- D-G-- --E----------------
                    Clavibacter michiganensis     YP_001220890 -------------------  ---- -----S-------------
Micrococcineae/     Catenulispora acidiphila      ZP_04368872  ------------------- D-G-- --E-K--N---H-------
Actinomycineae/     Mobiluncus mulieris           ZP_03994510  ------------------- D-G-- -------N---K-------
Bifidobacteriaceae  Leifsonia xyli                YP_063121    -------------------  ---- -------------------
                    Tropheryma whipplei           NP_787878    -------------------  ---- -------S-----------
                    Brachybacterium faecium       ZP_04532122  ----------------S--  -EG- ----Q--------------
                    Streptosporangium roseum      ZP_04471590  --------------D--QE  -GHG FVE-K--S---H-------
                    Gardnerella vaginalis         ZP_03936511  ---------------I--- D-G-- ----Q--N---H-------
                    Bifidobacterium animalis      AAT90388     ---------------I--- D-G-- ----Q--S---H-------
                    Bifidobacterium adolescentis  ZP_02029998  ---------------I--- D-G-- ----Q--N---H-------
                    Bifidobacterium dentium       ZP_02917609  ---------------I--- D-G-- ----Q--N---H-------
                    Bifidobacterium longum        ZP_00121343  ---------------I--- D-G-- ----Q--N---H-------
                    Bifidobacterium breve         ZP_03619714  ---------------I--- D-G-- ----Q--N---H-------
                    Bifidobacterium bifidum       AAV84826     ---------------I--- D-G-- ----K--N---H-------
                    Micrococcus luteus            YP_002956178 --------------D--DG       VVE--S----TH-----F-
                    Thermomonospora curvata       AAL87093     -----------------DG       VVE-K--S---H-------
                    Geodermatophilus obscurus     ZP_03889552  ---------------I-EG       V-E-K------H-------
                    Mycobacterium leprae          NP_302613    ---------------I-EG       VVE----S---H-------
                    Saccharopolyspora erythraea   YP_001109293 ---------------I-EG       VVE----S---H-------
                    Rhodococcus jostii            YP_706860    ---------------I-EG       VVE----S---H-------
                    Saccharomonospora viridis     ZP_04508196  ---------------I-EG       VVE----S---H-------
Other               Nocardia farcinica            YP_121625    ---------------I-EG       VVE----S---H-------
Actinobacteria      Gordonia bronchialis          ZP_03887600  ---------------I-DG       VVE----S---H-------
                    Streptomyces sviceus          ZP_05019219  ---------------I-DG       VVE-K--N---H-------
                    Acidothermus cellulolyticus   YP_873873    ---------------I-EG       IVE-K--S--TH-------
                    Nocardioides sp.              YP_925547    ---------------I-EG       VVE-K--S---H-------
                    Frankia alni                  YP_716766    ---------------I-DG       VVE-KS-S--TH-------
                    Tsukamurella paurometabola    ZP_04026140  ---------------I-DG       VVE----S---N-------
                    Nocardiopsis dassonvillei     ZP_04334206  -Y---------------DG       VVE-K--N---H-------
                    Kribbella flavida             ZP_03860673  ---------------I-DG       VFE-K--S---H-------
                    Rubrobacter xylanophilus      YP_643564    ------------I--L-DG       VFE-K--S-N-H-----F-
                    Ureaplasma urealyticum        ZP_03004057  ------------V-DMADG       -FE-LS-S---H-------
                    Clostridium spiroforme        ZP_02867788  ------------II-I-NG       V-E-LS-S---H-----FN
Firmicutes          Eubacterium rectale           YP_002935959 -Y----------VI-I-DN       L-E-L------H-----F-
                    Thermotoga lettingae          YP_001469651 -Y----------I--I-EG       V---V--S-N-H-----F-
                    Geobacillus sp.               YP_002950399 -Y----------I--L-DG       VFE-K------H-----F-
                    Bacillus coagulans            ZP_04432863  ------------I--L-DG       VFE-H------H-----F-
                    Dehalococcoides ethenogenes   YP_182108    -Y--------I-I--L-EG       -F--KS----TH-----F-
                    Methylacidiphilum infernorum  YP_001940829 -Y--------I-V--I-EG       VFE-K--N--TH-------
Gram-negative       Chloroflexus aggregans        YP_002464252 ------------I----DG       V-E-K--S--TH-----Y-
                    Nostoc punctiforme            YP_001869143 ------------I----EG       VFE-K--S--TH-----F-
Bacteria            Synechococcus sp.             YP_474256    ------------V-QL-SG       VFE-Q----N-H-----F-
                    Thermus aquaticus             ZP_03496130  -----------TV--I-EG       VFE-KS-S--TH---S-M-
                    Myxococcus xanthus            YP_630964    -Y----------I--I---       VFE-L-----TY-----F-
                    Gloeobacter violaceus         NP_927210    ------------I----DG       VFE-KS-S--TH-----F-
```

Figure 2.9. Partial sequence alignment of chaperone DnaK omologues showing a 5-aa indel (boxed) in a conserved region that is only found in species of *Micrococcineae, Actinomycineae* and *Bifidobacteriaceae*. Dashes (-) in the sequence alignment denote identity with the amino acid on the top line. Sequence information for only representative species is presented.

```
                    Kocuria rhizophila          YP_001854475 MGQKINPNGFRLGITTDHVSHWYAD SNQPG QRYKDYIREDVKIR
                    Clavibacter michiganensis   YP_001223356 ----V--Y------------R-FS- -TKK- ---S--VA---R--
                    Micrococcus luteus          YP_002957751 ----------------------F-- -HKE- ---A-FLK------
                    Beutenbergia cavernae       YP_002883143 ----V--L----------R-R-F-- -TK-- ---R--V----Q--
                    Xylanimonas cellulosilytica ZP_03911421  ----VH-H-Y--------R-R-F-- -TK-- ---R--V----E--
                    Janibacter sp.              ZP_00993899  ----V--H-------SE-R-R-F-- -TKE- ---R--VK---A--
                    Cellulomonas flavigena      ZP_04366346  ----V--L-Y--------R-R-F-- -TK-- ---R--V----Q--
                    Leifsonia xyli              YP_062848    ----V--Y------------R-FS- -TKK- ---S--LA------
                    Tropheryma whipplei         NP_787676    -------Y-L-------------S- -TR-- ---A--VS--I---
                    Jonesia denitrificans       ZP_03868814  ----VH-H-Y--------R-R-F-- -TK-- ---R--V----A--
                    Rothia mucilaginosa         ZP_05367924  -----H--------------K-F-- --K-- E--A-FV-------
Micrococcineae/     Mobiluncus mulieris         ZP_03994325  ----V--T--------E-R-R-F-- -TK-- ---R-FVK---E--
Actinomycycineae/   Brevibacterium linens       ZP_00379557  ------------------K-K-F-- -TK-- ---S--VL------
Bifidobacteriaceae  Kineococcus radiotolerans   YP_001360447 ----V--F----------R-R-F-- -TKT- ---A--VK---A--
                    Brachybacterium faecium     ZP_04530668  ----V-----------E-S-R-F-- -SKE- ---R--VK---A--
                    Kytococcus sedentarius      ZP_04042438  -------H----------K-R-F-- -SAE- ---A-FVG---A--
                    Mobiluncus curtisii         ZP_03922755  ----VH-T-----V-AE-R-R-F-- -TKS- ---R-FVK---E--
                    Actinomyces odontolyticus   ZP_02043500  ----V--R----------R-R-FS- -TTK- ---A--VA---A--
                    Actinomyces coleocanis      ZP_03924867  ----V--T--------E-R-R-F-- -TTK- ---S--VA---A--
                    Gardnerella vaginalis       ZP_03936909  ------F-Y-----EN-R-R-FS-- TKA-E ---R-FVL--D---
                    Bifidobacterium longum      ZP_00121721  -------F-Y-----EN-R-K-FS- --KA- E--R-FVL--DQ--
                    Bifidobacterium breve       ZP_03618113  -------F-Y-----EN-R-K-FS- --KA- E--R-FVL--DQ--
                    Bifidobacterium dentium     ZP_02917074  -------F-Y-----EN-R-K-FS- --KA- E--R-FVL--DA--
                    Bifidobacterium bifidum     ZP_03647037  -------F-Y-----ES-R-K-FS- --KV- E--S-FVL--D---
                    Bifidobacterium animalis    ZP_02963319  -------F-Y-----ES-R-K-FS- --KV- E--R-FVL--DA--
                    Catenulispora acidiphila    ZP_04373371  ----V--H---------FK-R----       KL----VK---A--
                    Nocardiopsis dassonvillei   ZP_04331909  ----V--H-----V---FK-R----       KS----VK---A--
                    Thermobifida fusca          YP_290696    ----V--H-----V---FK-R-F--       KL----VK---A--
                    Streptomyces coelicolor     NP_628867    ----V--H---------FK-R----       KL----VK---A--
                    Nocardia farcinica IFM      YP_116948    -------H---------WK-R----       KQ-A--VK---A--
                    Acidothermus cellulolyticus YP_872072    ----V--H--------EFS-R----       RM-R--VK---A--
Other               Mycobacterium avium         YP_883600    -------H---------WK-R----       KQ-A--VK---A--
Actinobacteria      Gordonia bronchialis        ZP_03883087  -------H---------WK-R----       KQ-A--VK---A--
                    Saccharopolyspora erythraea YP_001108919 -------H---------WK-R----       KQ-SE-VA------
                    Frankia sp.                 YP_001510293 ----V--H-------SEFT-R----       KQ--A-VG------
                    Rhodococcus jostii          YP_706074    -------H---------WK-R----       KQ-AE-VK---A--
                    Salinispora arenicola       YP_001539081 ----VH-H------S--WK-R-F--       KL-----G------
                    Corynebacterium diphtheriae NP_938858    ----H-H-L-----S-WK------        KN-AE-LA--IRV-
                    Bacillus selenitireducens   ZP_02171790  ---------L-V-VIKGWE-K---G       KD-A-LLH--IR--
                    Lactobacillus johnsonii     NP_964365    --------------VNR-WEAK----       KN-A-TLN--LR--
                    Thermoan. mathranii         ZP_05380129  ----VH-Y-L-V-V-Q-WLAK----      DKNFSKFLI--I---
Firmicutes          Oceanobacillus iheyensis    NP_691046    -------T-L-V--IK-WE-K---G       KD-A-LLH--I---
                    Pediococcus pentosaceus     YP_804892    ----V------V-VIR-WQAK----       KDFSKFLA--I---
                    Geo. stearothermophilus     P23309       ----V--I-L-I--IR-WE-R---E       KD-A-LVH--L---
                    Helio. modesticaldum        YP_001679972 ----V--K-L-I--IK-WDAR-F-G       KN-AELLH--L---
                    Synechococcus elongatus     YP_172581    -------V-----V-QE-R-R-F--       PN--PQLLQ--K---
                    Anabaena variabilis         YP_321216    -----H-V-------QE-Q-R-F-E       PS--PELLQ--H-L-
                    Legionella drancourtii      ZP_05108633  ----V--I-I----IK-WN-K-F-G        K--AEFLNQ-I-L-
                    Pseudomonas mendocina       YP_001189382 ----VH-V-I----VK--T-V----       -RN-A--LNA-L-V-
Gram-negative       Hahella chejuensis          YP_437281    ----V--V-I----VK--N-V----       KKN-S-HLLT-I-V-
Bacteria            Sinorhizobium medicae       YP_001326680 -------I------NRTWD-R-F--       NAE-GQLLH--L---
                    Rhizobium sp.               YP_002825732 -------I------NRTWD-R-F--       NAE-GQLLH--L---
                    Neisseria lactamica         ZP_03721539  -------T----AV-K-WA-K-F-K       STDFSAVLKQ-IDV-
                    Agrobacterium radiobacter   YP_002544211 -------I------NRTWD-R-F--       NAE-GQLLH--L-M-
                    Magnetococcus sp.           YP_864778    ----VH-T-----T-KTWDTR-F--        RN-A-LLL--I---
                    Bacteroides capillosus      ZP_02037267  ----V--H-L-V-VIK-WD-R---R       NEKVG-LLV--K---
```

Figure 2.10. Partial sequence alignment of 30S ribosomal protein S3 homologues showing a 5-aa indel (boxed) in a conserved region that is found uniquely in species of *Micrococcineae, Actinomycycineae* and *Bifidobacteriaceae*. Sequence information for only representative species is presented.

Table 2.1. Actinobacterial strains used in this study
(Taxonomy based on Bergey's Maual 2nd Edition, 2001)

Subclass II. Rubrobacteridae			Rubrobacter radiotolerans^{T} (DSM 5868)
Subclass V. Actinobacteridae			
	Suborder VI. Micrococcineae		
		Family I. Micrococcaceae	Arthrobacter nicotinovorans^{T} (DSM 420)
		Family III. Cellulomonadaaceae	Cellulomonas fimi^{T} (DSM 20113) Oerskovia turbata^{T} (DSM 20577)
		Family VIII. Microbacteriaceae	Microbacterium oxydans^{T} (DSM 20578) Clavibacter michiganensis (DSM 340)
	Suborder VII. Corynebacterineae		
		Family III. Gordoniaceae	Gordonia rubripertincta^{T} (DSM 43197)
		Family V. Nocardiaceae	Nocardia corynebacterioides^{T} (DSM 20151) Rhodococcus rhodochrous (116)
		Family VI. Tsukamurellaceae	Tsukamurella paurometabola^{T} (DSM 20162)
		Family VII. Williamsiaceae	Williamsia murale^{T} (DSM 44343)
	Suborder VIII. Micromonosporineae		Micromonospora chersina^{T} (DSM 44151)
	Suborder IX. Propionibacterineae		
		Family I. Propionibactiaceae	Propionibacterium acnes (AT1)
		Family II. Nocardioidaceae	Nocardioides simplex^{T} (DSM 20130) Kribbella sandramycini^{T} (DSM 15626)
	Suborder X. Pseudonocardineae		Pseudonocardia halophobica^{T} (DSM 43089) Saccharopolyspora erythraea^{T} (DSM 40517)
	Suborder XI. Streptomycineae		Streptomycoides glaucoflavus^{T} (DSM 43891) Trichotomospora caesia (DSM 43890)
	Suborder XII. Streptosporangineae		Streptosporangium roseum^{T} (DSM 43021) Microtetraspora niveoalba^{T} (DSM 43174) Planobispora rosea^{T} (DSM 43051)

Table 2.2 PCR primers for amplifying different sequences that contain CSIs

Gene	Primer	Primer sequence* 5'-3'	Fragment size
Cytochrome c Oxidase subunit 1	Forward	TGGTTYTTYGGSCACCCYGARGT	581bp
(CoxI)	Reverse	CCVAVCCARTGCTGBAYSADRAA	
Glutamyl tRNA synthetase	Forward	ACBGCSCTKTTYAACTGG	773bp
(GluRS)	Reverse	AGRTARTTSARMAKRCCYTC	
CTP synthetase	Forward	AARACVAARCCVACHCAGCA	986bp
(CTPsyn)	Reverse	TCVGGRTGNGCCTGBGT	
23S rRNA insert	Forward	CCGANAGGCGTAGBCGATGG	361bp
	Reverse	CCWGWGTYGGTTTVSGGTA	
Carbamoyl-phosphate synthase	Forward	SSATGWCCGGBTAYCARGA	709bp
(CarA)	Reverse	TGRTTSCCRAARCARATRCC	

*Where N=A,T,C or G; Y=C or T; S=G or C; R=A or G; V=A,C or G; B=C,G or T; D=A,T or G; K=G or T; M=A or C; H=A,C or T; W=A or T.

Table 2.3

Summary of CSIs that are specific to all Actinobacteria or its various subgroups

Group specificity	Protein	CSI
all actionobacteria	CoxI: cytochrome-c oxidase subunit 1	2-aa indel
	GluRS: glutamyl-tRNA synthetase	4-aa indel
	CTPsyn: CTP synthetase	4-aa indel
	23S RNA	~100 nt indel
	Gft: glucosamine--fructose-6-phosphate aminotransferase	4-aa indel
	GlyRS: glycyl-tRNA synthetase	3-aa indel
	TrmD: tRNA (Guanine-1)-methyltransferase	4-aa indel, and an extra aa is unique to Bifidobacterium and Actinomyces
	Gyrase A	4-aa indel
	SahH: S-adenosyl-L-homocysteine hydrolase	9-aa indel
	SHMT: serine hydroxymethyltransferase	5-aa indel
Corynebacterineae	CarA: carbamoyl phosphate synthase small subunit	2-aa indel and longer indel found in Rhodococcus and Nakamurella
	RecR: recombination protein RecR	4~20 aa indel
	IF-2: initiation factor IF-2	1-aa indel
Corynebacterium	GroEL: chaperonin GroEL	5-aa indel
Some Corynebacterium	EF-G: elongation factor G	5-aa indel
	AlaRS: alanyl-tRNA synthetase	2-aa indel
Frankia	Gyrase B	7-aa indel
Micrococcineae/ Actinomycineae/ Bifidobacteriaceae	DnaK: chaperone DnaK	5-aa indel
	S3: 30S ribosomal protein S3	5-aa indel
Streptomycineae/ Frankineae/ Streptosporangineae	FabG: ketoacyl reductase	2-aa indel
Corynebacterineae/ Pseudonocardineae/ Micromonosporineae	HolB: DNA polymerase III subunit delta	2-aa indel
	S3: 30S ribosomal protein S3	1-aa indel
Streptomycineae/ Streptosporangineae/ Propionibacterineae/ Micrococcineae/ Actinomycineae/ Bifidobacteriaceae	S9: 30S ribosomal protein S9	1-aa indel
Bifidobacteriaceae	SRP: signal recognition particle	1-aa indel
	CDP: Cytidylyltransferase	5-aa indel

CHAPTER 3.

CSPs that are Distinctive Characteristics of Actinobacteria and its different lineages

3.1 Preface

This chapter describes Actinobacteria-specific CSPs that are identified in our comparative genomic studies. Most description was reproduced from the published manuscript (Gao et al., 2006): Gao B, Paramanathan R, Gupta RS. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. Antonie Van Leeuwenhoek. 2006 Jul;90(1):69-91. Since this paper was published in 2006 when only limited number of actinobacterial genomes was available, the specificity of the identified CSPs was examined in May 2009 and additional 8 new actinobacterial genomes were investigated by BLAST search of each ORF to identify new CSPs that are specific to different subgroups within this phylum.

3.2 Introduction

Comparative genomic studies have previously been carried out only on some closely related actinobacterial species. Extensive work has been done on *Mycobacterium* genomes to identify possible virulence factors or new drug targets (Domenech et al., 2001; Cole, 2002; Stinear et al., 2008). Sutcliffe and Harrington have analyzed the *M. tuberculosis* genome to identify various genes/ proteins that are involved in the synthesis and regulation of cell envelope lipoproteins (Sutcliffe and Harrington, 2004). Studies have also been done on the *Streptomyces* genomes to identify proteins/enzymes that are possibly involved in production of useful secondary metabolites (Zazopoulos et al., 2003; Ikeda et al., 2003; McAlpine et al., 2005). However, none of these studies aimed at identifying different gene/proteins that are uniquely present either in all Actinobacteria or in various subgroups that make up this large phylum. In addition to CSIs which have been shown to be reliable molecular markers for the Actinobacteria phylum and its various subgroups, we have also discovered a number of whole proteins (CSPs) that are unique characteristics of the Actinobacteria (Gao et al., 2006). Such studies are of much interest in order to understand what unifying molecular characteristics are shared by various actinobacterial species beneath their highly diverse phenotypes.

3.3 Methods

3.3.1 Identification of CSPs that are specific to Actinobacteria

To identify proteins which are specific for Actinobacteria or its various subgroups, 12 genomes representing species from different subgroups have been selected as probes to do the BLAST search, including: *Mycobacterium leprae* TN (Cole et al., 2001), *Leifsonia xyli* CTCB07 (Raoult et al., 2003), *Bifidobacterium longum* NCC2705 (Schell et al., 2002), *Thermobifida fusca* YX (Lykidis et al., 2007), *Saccharopolyspora erythraea* NRRL2338 (Oliynyk et al., 2007), *M. avium* 104 (Horan et al., 2006), *Rhodococcus jostii* RHA1 (Mcleod et al., 2006), *Corynebacterium glutamicum* ATCC13032 (Kalinowski et al., 2003), *Streptomyces coelicolor* A3(2) (Bentley et al., 2002), *Frankia* sp. CcI3 (Normand et al., 2007), *B. dentium* Bd1 and *Clavibacterium michiganensis* subsp. sepedonicus (Bentley et al., 2008). BLAST searches were carried out on each individual protein in these genomes to identify all other organisms containing proteins with similar sequences (Altschul et al., 1997; Schaffer et al., 2001). Protein – protein BLAST was performed with default parameters as set by the BLAST program against sequences from all organisms in the GenBank and the results were visually inspected for homologues showing specificity to Actinobacteria. Expected values (E-values) were analyzed for putative Actinobacteria-specific proteins (Kainth and Gupta, 2005; Gao et al., 2006). The E-values, which are calculated by the BLAST software, indicate the probability that the observed similarity between the query protein and any other protein detected by the BLAST search arose by chance (Altschul et al., 1997; Schaffer et al., 2001). In BLAST searches, the E values are lowest (closer to 0) for BLAST hits with a high degree of homology to the query sequence and they increase as BLAST hits are detected with lower similarity. The results of BLAST searches were inspected for sudden increase in E-values from the last actinobacterial species in the search to the first non-actinobacterial organism. This increase in E-values was important when the first non-actinobacterial BLAST hit was in a higher range, such as more than 10^{-4}. Scores above this value suggest that the BLAST matches represent a weak level of similarity that could occur by chance. However, higher E-values are sometimes acceptable for smaller proteins as the magnitude of the E-value depends upon the length of the query sequence (Altschul et al., 1997). A protein was considered to be Actinobacteria-specific if all BLAST hits with acceptable E-values corresponded to actinobacterial species. We have retained a few proteins where, besides Actinobacteria, 1 or 2 isolated species from other groups of bacteria also had acceptable E-values. We consider these proteins to be also Actinobacteria-specific and the presence of a related homologue in isolated other species is very likely due to lateral gene transfer (LGT). All proteins indicated in the Tables 3.1 –3.3 are specific for the Actinobacteria based on these criteria unless otherwise mentioned.

3.4 Results

3.4.1 CSPs specific for all Actinobacteria

We have previously identified 29 CSPs that are present in nearly all actinobacterial species and are not found in any other Bacteria with a few exceptions (Gao et al., 2006). As the number of sequenced actinobacterial species tripled compared to 2006, the specificity of these proteins was examined by BLAST search of these 29 proteins against the updated GenBank. As expected, most of these proteins are also found in the newly sequenced actinobacterial species and also retained their actinobacterial specificity. Only 3 out of these 29 proteins (ML0257, ML2073 and ML1666, "ML" refers to gene ID from *M. leprae* genome) were detected in other bacterial groups, thus, should not be regarded as actinobacteria-specific CSPs. Among the confirmed 26 CSPs, four proteins ML0642, ML1009, ML1029, and ML1306 are present in all sequenced actinobacterial genomes including *Rubrobacter xylanophilus* DSM 9941 (see Table 3.1). The observed E-values for these proteins from actinobacterial species are very low, close to 0 (i.e. $<e^{-200}$), indicating that the proteins in various actinobacteria are homologous to the query sequence. In the 16S rRNA tree, Rubrobacter species are distantly related to other actinobacterial species and form an outgroup of the other actinobacteria (Stackebrandt et al., 1997; Gao and Gupta, 2005; Ludwig and Klenk, 2001). Presently, there are no biochemical or molecular characteristics (other than the 16S rRNA gene sequence analyses) known that support a specific relationship of *Rubrobacter* species to the Actinobacteria. In Chapter 2, 9 CSIs in various universal proteins and a large insert in 23S rRNA were described that were uniquely shared by various other actinobacteria. However, these indels were not present in *Rubrobacter* species, thus failing to reveal a

specific relationship of this group to Actinobacteria (Gao and Gupta, 2005). In this context, the shared presence of these four CSPs in *R. xylanophilus* and various other actinobacteria is of much interest. The simplest and most logical explanation for the shared presence of these four CSPs is that the genes for these proteins evolved only once in a common ancestor of *R. xylanophilus* and various other actinobacteria and then passed on to all members of the Actinobacteria phylum through vertical descent. This observation, in conjunction with the phylogenetic relationship of *R. xylanophilus* to other Actinobacteria in 16S rRNA gene sequence analyses, provides evidence that this species is a part of the phylum Actinobacteria.

The remaining 9 proteins in Table 3.1 are found in almost all sequenced actinobacterial species except *R. xylanophilus*. Based upon its deep branching in the rRNA trees, the most likely explanation for the absence of these 9 proteins in *R. xylanophilus* will be that the genes for these proteins have evolved in a common ancestor of Actinobacteria after the divergence of *Rubrobacter* (Stackebrandt et al., 1997; Gao and Gupta, 2005; Zhi et al., 2009). In addition, we also identified 13 proteins that show similar distribution as the proteins listed in Table 3.1, but which are missing in the two *T. whipplei* strains. *T. whipplei* is an intracellular pathogen and the genomes of these strains have undergone massive gene decay (to only 0.93 Mb), as many proteins are not required in the intracellular environment (Moran and Wernegreen, 2000; Raoult et al., 2003; Bentley et al., 2003). Thus, the absence of these genes in the two *T. whipplei* strains represents a special situation, which is not characteristic of other Actinobacteria. Therefore, despite their absence in *T. whipplei*, we still regard these proteins as distinctive characteristics of various other Actinobacteria. For all of the 26 Actinobacteria-specific CSPs, no homologues were detected in the *S. thermophilum* genome, which support our results from CSIs that *S. thermophilum* is distinct from all other actinobacteria and it should not be placed in the phylum Actinobacteria (Gao and Gupta, 2005).

Among the CSPs listed in Table 3.1, ML0760 and ML0804 are very similar to each other and they are homologous to the developmental regulator gene *whiB* in *S. coelicolor*. In addition to these two proteins, there are five copies of *whiB* in *M. leprae*, which also include ML0639, ML2307 and ML0382. The WhiB protein family was previously suggested to be essential for sporulation of aerial hyphae in *Streptomyces* but its role in many other non-sporulating actinobacterial was unclear (Soliveri et al., 2000; Chater and Chandra, 2006). Our observation that *whiB*-like genes are present in all sequenced actinobacterial genomes including the non-spore-forming intracellular pathogens *T. whipplei* and *L. xyli*, suggests that this protein, in addition to its role in sporulation, also performs a more generalized function common to all Actinobacteria. Recent studies on *Mycobacterium* indicate that WhiB proteins are differentially expressed which are important in regulating virulence, cell division, antibiotic resistance and other stress response (Brosch et al., 2007; Morris et al., 2005).

3.4.2 CSPs specific for actinobacterial subgroups

In addition to the CSPs that are specific to all Actinobacteria, we also identified some CSPs that are uniquely shared by two or more particular lineages. Based on their species distribution, these CSPs can be sorted into different groups and a summary of these actinobacterial subgroup-specific CSPs are summarized in Table 3.3. Compared to the phylogenetic tree, they provide molecular evidence for the interrelationship or branching orders of different lineages within Actinobacteria. Several examples are given below.

Our comparative genomic analyses uncovered 14 CSPs that are uniquely shared by the suborder *Corynebacterineae* and *Pseudonocardineae*, providing strong molecular evidence that support their phylogenetic relationship seen in the tree (Table 3.2A). In the 16S rRNA tree, species from the suborder *Corynebacterineae* and *Pseudonocardineae* form a compact cluster, indicating that these two suborders are more closely related than other actinobacteria and likely evolved from a common ancestor (Zhi et al., 2009) (Figure 3.1). Additionally, 6 CSPs were found to be unique to *Corynebacterineae* and *Pseudonocardineae* but not found in *Corynebacterium* species (Table 3.2B). It is likely that these 6 CSPs evolved in the common ancestor of *Corynebacterineae* and *Pseudonocardineae*, but subsequently lost in the progenitor of *Corynebacterium* species when it diverged from others. In Chapter 2, 3 CSIs that are specific to *Corynebacterineae* were described, which provide useful molecular characteristics for defining this subgroup. Besides, we have identified 4 CSPs (MAV_1296, MAV_0225, MAV_0229 and MAV_4967, "MAV" refer to gene ID from *M. avium* 104 genome) that are uniqurely shared by all members from this suborder but not found in species outside this group. Among these four proteins, MAV_0225 and MAV_0229 are functionally characterized and they are involved in the biosynthesis of their unique cell envelope (Belanger et al., 1996; Berg et al., 2005). In *Mycobacterium*, these proteins are the sites of resistance to the anti-tuberculosis drug ethambutol (EMB).

Within the suborder *Corynebacterineae*, we have identified 30 CSPs that are uniquely shared by species from *Mycobacterium, Rhodococcus, Nocardia, Gordonia* and *Tsukamurella*, but not found in *Corynebacterium* (Table 3.3 & Figure 3.1). In the phylogenetic tree based on 16S rRNA, *Corynebacterium* genus form the outgroup within the *Corynebacterineae* suborder, while the other genera mentioned above cluster together suggesting a closer relationship (Zhi et al., 2009). There are two possibilities for the presence of these 30 CSPs. First, it is likely that these 30 CSPs origniate in the common ancestor of *Mycobacterium, Rhodococcus, Nocardia, Gordonia* and *Tsukamurella*, after the divergence of *Corynebacterium*. The other possibility is that these CSPs evolved in the progenitor cell of all *Corynebacterineae*, and subsequently lost in *Corynebacterium*. Although currently we do not have further evidence to favor either possibility, the unique presence of these 30 CSPs provide useful molecular markers for the above 5 genera, which support their monophyletic clustering in the 16S rRNA tree. Besides, we also discovered 14 CSPs that are uniquely shared by *Nocardia* and *Rhodococcus*, which suggest that these two genera are closely related. What's more, our analyses have also identified 32 CSPs that are unique to the genus *Mycobacterium*, and 21 CSPs that are exclusive to the genus *Corynebacterium* (Table 3.3 & Figure 3.1). These CSPs provide molecular markers to define these two genera. Moreover, many CSPs restricted to *Mycobacterium* genus (Gao et al., 2006) were found to be virulence factors, such as the prevalent PE/PPE family protein, *mce* family, etc., of which the exact functions are examined recently (Gao et al., 2006; Joshi et al., 2006; Strong et al., 2006).

Analysis of BLAST results on each ORF from a new genome *B. dentium* Bd1 uncovered a number of CSPs that are either specific to *Bifidobacterium* genus, *B. dentium* species, or Bd1 strain (Table 3.3 & Figure 3.2). In addition, 24 CSPs are found to be uniquely shared by *B. longum* and *B. adolescentis*, which suggest that these two species are very closely related than other *Bifidobacterium* species. What's more, for the genus *Streptomyces*, 3 genomes are completely sequenced, and another 26 genomes are still in assembly but a lot of protein sequences are available in the GenBank. With this rich resource and the large genome size of *Streptomyces*, a vast number of CSPs were identified that are restricted to *Streptomyces* species but not found in any other bacteria. However, sorting them into groups based on their distribution pattern is a big problem since many CSPs are presently missing or not available in the unfinished genomes, while others always have homologues in the several new genomes. To make it simple, 27 CSPs were identified that are uniquely shared by 8 *Streptomyces* species, 54 CSPs found in 7 Streptomyces species, and 26 CSPs only missing in two *Streptomyces* species (Table 3.3 & Figure 3.3A). Furthermore, many CSPs were identified to be specific to the genus *Frankia* as indicated in Figure 3.3B (Table 3.3).

3.4.3 Gene transfer from Actinobacteria to *Magnetospirillum magnetotacticum*

One interesting and surprising observation from the present work is that for a number of proteins that are Actinobacteria-specific, homologous proteins (as indicated by their low E-values and similar protein lengths) are also present in the genome of *M.*

magnetotacticum MS-1. *M. magnetotacticum* is a magenetotactic bacteria belonging to the alpha-proteobacteria subdivision (Bazylinski and Frankel, 2004; Kainth and Gupta, 2005). It forms internal crystals of magnetite in membrane enclosed bodies which it uses to swim along geomagnetic field lines (Bazylinski and Frankel, 2004). In the present work, we have identified a total of 14 proteins (viz. ML1029, ML1666, ML0761, ML0762, ML1781, Lxx08190, Tfu_1340, Tfu_2483, BL0895, Lxx08745, ML1526, Tfu_2164, BL1224 and Lxx11715) for which a related homologue is found in *M. magnetotacticum* (Gao et al., 2006). Most of these genes/proteins from *M. magnetotacticum* exhibit highest similarity to the corresponding genes/proteins from Streptomyces species. When BLAST searches were carried out on these proteins from *M. magnetotacticum*, all of the hits with highest similarity were from actinobacterial species and no proteobacterial hits with low E-values were observed (results not shown). In view of the fact that besides *M. magnetotacticum*, no other a-proteobacterial species was found to contain any of these proteins, it is very likely that these genes in *M. magnetotacticum* have been acquired from actinobacterial species by means of LGT. It is known that *M. magnetotacticum* has a very large genome (ca. 9.2 Mb) with very high GC content (66.4%), similar to those of Actinobacteria (Matsunaga et al., 2005). The lateral transfer of these genes to *M. magnetotacticum* seems to have occurred in a highly specific manner as, other than *M. magnetotacticum*, very few and only isolated examples of the presence of these gene/proteins in other groups of bacteria were observed. The possible functional significance of the genes, which have been apparently laterally transferred from Actinobacteria to *M. magnetotacticum* remains to be determined.

3.5 Discussion

Our comparative analyses of actinobacterial genomes have identified a large number of proteins that are uniquely found in Actinobacteria. Some of these CSPs are present in all sequenced actinobacterial genomes, whereas others are limited to various subgroups of Actinobacteria at different phylogenetic depths. They provide novel molecular markers that are distinctive characteristics of the entire phylum. For the suborder *Corynebacterineae* that encompass many important pathogens, a number of CSPs were identified that are unique to either all members or certain genera within this suborder. The absence of all of these proteins in *S. thermophilum* indicates that this species should not be grouped with Actinobacteria, an inference which is also supported by other lines of evidences (Ueda et al., 2004; Gao and Gupta, 2005).

Most of the actinobacteria-specific CSPs identified in the present work are of unknown function. The GC contents of these proteins are very similar to the rest of their genomes and their Ka/Ks ratios (i.e., substitution rates at non-synonymous versus synonymous sites) are less than 0.1 (results not shown) (Yang and Nielsen, 2000). These results strongly indicate that the identified ORFs very likely correspond to functional proteins and they are not due to errors in gene annotation (Daubin and Ochman, 2004) (Yang et al., 2005). Because of the specificity of these CSPs for either all Actinobacteria or certain subgroups within this phylum, it is highly likely that these proteins carry out certain unique functions that are limited to these groups of bacteria. Therefore, studies aimed at understanding the functions of these Actinobacteria-specific proteins should be of great interest, as they will likely provide important insights into unique biochemical and physiological characteristics that distinguish these bacteria (or specific subgroups among them) from all other bacteria. Because of their specificity for Actinobacteria or certain groups within this phylum, many of which are important human pathogens (e.g. *M. leprae, M. tuberculosis* and *N. farcinica*), these proteins potentially also provide novel targets for development of drugs that are specifically directed against these bacteria.

3.6 Tables 1-3 and Figures 1-3

Table 3.1. CSPs specific for all Actinobacteria

Protein	ML0642	ML1009	ML1029	ML1306	ML0760	ML0804	ML0857	ML0869	ML1016	ML1026	ML2137	ML2204	ML0013
Length	**479 aa**	**326 aa**	**273 aa**	**274 aa**	**89 aa**	**84 aa**	**250 aa**	**124 aa**	**107 aa**	**100 aa**	**251 aa**	**62 aa**	**93 aa**
Possible function	Unknown	Unknown	Unknown	Unknown	whiB	whiB	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Mycobacterium	1	1	1	1	1	1	1	1	1	1	1	1	1
Rhodococcus	1	1	1	1	1	1	1	1	1	1	1	1	1
Nocardia	1	1	1	1	1	1	1	1	1	1	1	N	1
Gordonia	1	1	1	1	1	1	1	1	1	1	1	N	1
Corynebacterium	1	1	1	1	1	1	1	1	1	1	1	1	1
Actinosynnema	1	1	1	1	1	1	1	1	1	1	1	1	1
Saccharopolyspora	1	1	1	1	1	1	1	1	1	1	1	N	1
Tsukamurella	1	1	1	1	1	1	1	1	1	1	1	1	1
Streptomyces	1	1	1	1	1	1	1	1	1	1	1	1	1
Geodermatophilus	1	1	1	1	1	1	1	1	1	1	1	1	1
Catenulispora	1	1	1	1	1	1	1	1	1	1	1	1	1
Frankia	1	1	1	1	1	1	1	1	1	1	1	1	1
Beutenbergia	1	1	1	1	1	1	1	1	1	1	1	1	1
Jonesia	1	1	1	1	1	1	1	1	1	1	1	1	1
Janibacter	1	1	1	1	1	1	1	1	1	1	1	1	1
Salinispora	1	1	1	1	1	1	1	1	1	1	1	1	1
Cellulomonas	1	1	1	1	1	1	1	1	1	1	N	1	1
Xylanimonas	1	1	1	1	1	1	1	1	1	1	1	1	1
Kocuria	1	1	1	1	1	1	1	1	1	1	1	1	1
Micrococcus	1	1	1	1	1	1	1	1	1	1	1	1	1
Kytococcus	1	1	1	1	1	1	1	1	1	1	1	1	1
Kineococcus	1	1	1	1	1	1	1	1	1	1	1	1	1
Thermobifida	1	1	1	1	1	1	1	1	1	1	1	1	N
Arthrobacter	1	1	1	1	1	1	1	1	1	1	1	1	1
Nocardioides	1	1	1	1	1	1	1	1	1	1	1	1	1
Acidothermus	1	1	1	1	1	1	1	1	1	1	1	1	1
Kribbella	1	1	1	1	1	1	1	1	1	1	1	1	1
Actinomyces	1	1	1	1	1	1	1	1	1	1	1	1	1
Propionibacterium	1	1	1	1	1	1	1	1	1	1	1	1	1
Thermomonospora	1	1	1	1	1	1	1	1	1	1	1	N	1
Brevibacterium	1	1	1	1	1	1	1	1	1	N	1	1	1
Nocardiopsis	1	1	1	1	1	1	1	1	1	1	1	1	1
Renibacterium	1	1	1	1	1	1	1	1	1	1	1	1	1
Leifsonia	1	1	1	1	1	1	1	1	1	1	1	1	1
Nakamurella	1	1	N	1	1	1	1	1	1	1		1	1
Mobiluncus	1	1	1	1	1	1	1	1	1	1	1	1	1
Clavibacter	1	1	1	1	1	1	1	1	1	1	1	1	1
marine action*	1	1	1	1	1	1	1	1	1	1	1	1	1
Bifidobacterium	1	1	1	1	1	1	1	1	1	1	1	1	1
Tropheryma	1	1	1	1	1	1	1	1	1		1	1	1
Gardnerella	1	N	1	N	1	1	1	1	1	1	1	1	1
Rubrobacter	1	1	1	1	N	N	N	N	N	N	N	N	N

Note: These proteins were identified by BLASTP searches as detailed in section 3.3. The top line is the protein ID number in genome of *M. leprae* TN (ML), which was used as probe to perform the blast search. The second line and the third line describe the sequence length and possible function of each query protein. The left column lists the actinobacterial genera from which one or more species have been completely sequenced or in assembly. "1" means present; "N" means absent. *: marine actinobacterium PHSC20C1.

Table 3.2. Proteins uniquely shared by Corynebacterineae and Pseudonocardineae

A. Proteins uniquely shared by Pseudonocardineae (Saccharopolyspora/ Actinosynnema) and Corynebacterineae (Mycobacterium/ Nocardia/ Rhodococcus/ Gordonia/ Tsukamurella/ Corynebacterium)	
SACE_0086 [YP_001102365][1]	SACE_6570 [YP_001108663]= ML0703[1]
SACE_0562 [YP_001102834][1]	SACE_7174 [YP_001109257]= ML0281[1]
SACE_1059 [YP_001103316]= ML0810 [1]	SACE_7191 [YP_001109274]
SACE_1071 [YP_001103328][1]	SACE_1352 [YP_001103600] (except Nocardia) [1]
SACE_2293 [YP_001104522]	SACE_0938 [YP_001103197] (Thermobifida) [1]
SACE_5423 [YP_001107583]= SACE_5424[2]	SACE_1071 [YP_001103328] (Janibacter) [1]
SACE_5810 [YP_001107919]= ML0107 AftA	SACE_6709 [YP_001108800] (Frankia/Salinispora)
B. Proteins uniquely shared by Pseudonocardineae and Corynebacterineae but not found in Corynebacterium	
SACE_0595 [YP_001102866][1]	SACE_1495 [YP_001103742]
SACE_0609 [YP_001102880][1]	SACE_1695 [YP_001103938]
SACE_1166 [YP_001103419]	SACE_1719 [YP_001103962] LppM

Notes: The protein ID number starting with SACE represents query protein from the genome of *S. erythraea* NRRL2338, which was used as probe to perform the BLAST search. Accession numbers for these proteins are shown in square brackets. "=ML" means the same protein which was identified as *Corynebacterineae*-specific proteins in 2006 (Gao et al., 2006).

1. Protein is also present in the unfinished genome of *Nakamurella multipartita* DSM 44233;
2. A similar protein to SACE_5423 is also found in *Deinococcus geothermalis* DSM 11300. Also, SACE_5424 might be paralogous to SACE_5423 based on sequence similarity.

Table 3.3. Summary of CSPs specific to its various subgroups

Suborder	Subgroups which share CSPs	No. of CSPs
Corynebacterineae/ Pseudonocardineae	Pseudonocardineae (Saccharopolyspora/ Actinosynnema) and Corynebacterineae (Mycobacterium/ Nocardia/ Rhodococcus/ Gordonia/ Tsukamurella/ Corynebacterium)	14
	Only absent in Corynebacterium	6
Corynebacterineae	Mycobacterium/ Nocardia/ Rhodococcus/ Corynebacterium	4
	Mycobacterium/ Nocardia/ Rhodococcus	30
	Mycobacterium/ Nocardia	6
	Mycobacterium/ Rhodococcus	14
	Nocardia/ Rhodococcus	14
	all Mycobacterium	32
	all Mycobacterium except M. leprae	29
	M. avium/ M. tuberculosis/ M. marinum/ M. ulcerans/ M. bovis	19
	M. avium/ M. Marinum/ M. ulcerans	9
	M. avium/ M. tuberculosis/ M. bovis	15
	all Corynebacterium (C. glutamicum/ C. efficiens/ C. diphtheriae/ C. jeikeium/ C. urealyticum)	21
	C. glutamicum/ C. efficiens/ C. diphtheriae	21
	C. glutamicum/ C. efficiens	48
Bifidobacterineae	B. dentium, B. adolescentis, B. longum and B. animalis	43
	B. dentium, B. adolescentis and B. longum	3
	B. dentium and B. adolescentis	24
	B. dentium species (Bd1 and ATCC27678)	166
	B. dentium Bd1	126
Streptomycineae	all Streptomyces (S. coelicolor/ S. avermitilis/ S. sviceus/ S. griseus/ S. pristinaespiralis/ S. clavuligerus/ S. sp. SPB74/ S. sp. Mg1)	27
	all Streptomyces missing in Streptomyces sp. SPB74*	14
	all Streptomyces missing in S. pristinaespiralis*	10
	all Streptomyces missing in S. clavuligerus*	10
	all Streptomyces missing in Streptomyces sp. Mg1*	6
	all Streptomyces missing in S. avermitilis	2
	all Streptomyces missing in S. sviceus	2
	all Streptomyces missing in two species	26
	S. coelicolor/ S. avermitilis/ S. sviceus/ S. griseus and 1 or 2 other Streptomyces	24
	S. coelicolor/ S. avermitilis/ S. sviceus and 1 or 2 other Streptomyces	15
	S. coelicolor/ S. avermitilis/ S. sviceus	8
Frankineae	Frankia sp. CcI3 / Frankia sp. EAN1pec / F. alni	121
	Frankia sp. CcI3 and Frankia alni	39
	Frankia sp. CcI3 and Frankia sp. EAN1pec	42
Micrococcineae	Cluster III (Clavibacter/ marine actinobacterium PHSC20C1/ Leifsonia)	7
	Clavibacter/ marine actinobacterium PHSC20C1	9
	Clavibacter/ Leifsonia	3

Note: * denotes genomes that are still in assembly. It is likely that some of the CSPs that are not found in these genomes at this moment are present in the complete genome sequence.

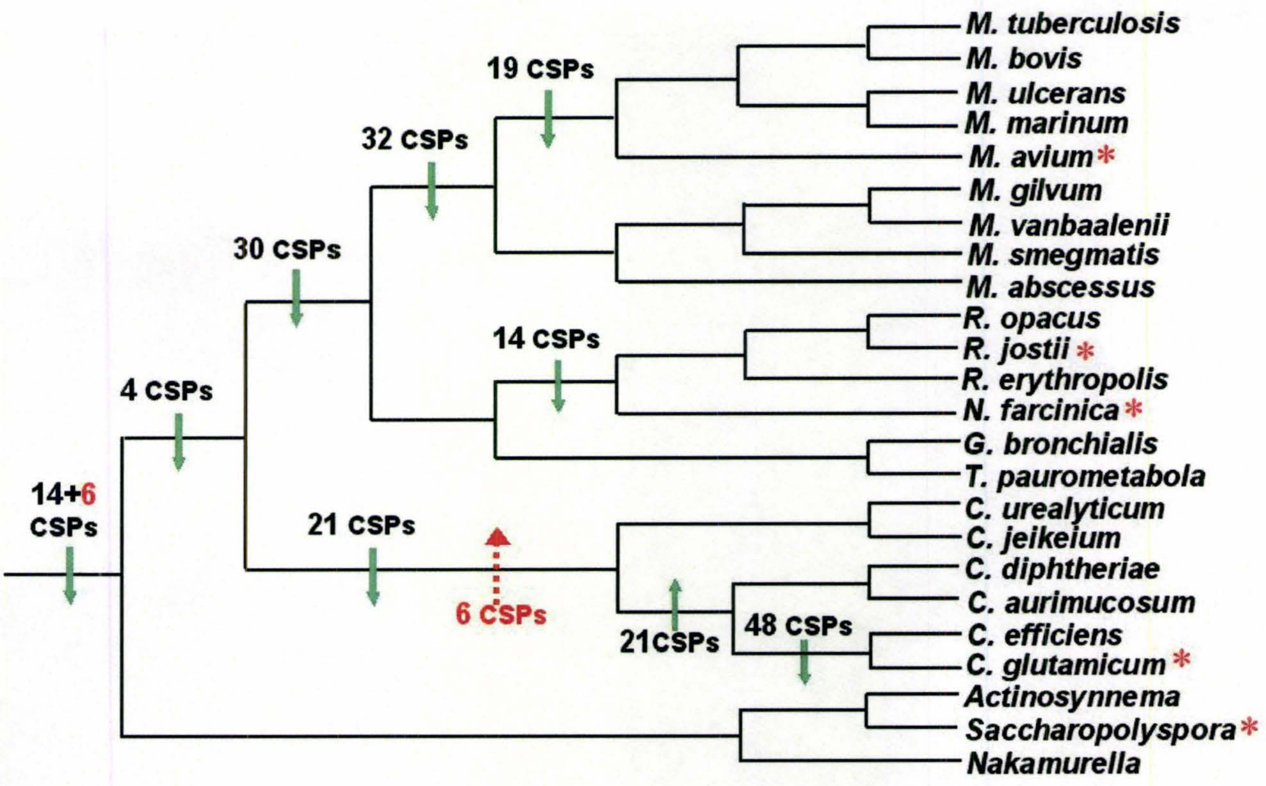


Figure 3.1. A subtree from Figure 4.1 in Chapter 4 showing the distribution pattern of various CSPs that are specific to *Corynebacterineae*. The green arrows indicate the evolutionary stages where many of these CSPs likely evolved while the red dashed arrow indicate the loss of 6CSPs in the last common ancestor of *Corynebacterium*. * denotes the species or genera from which a genome was chosen to perform the BLAST searches.

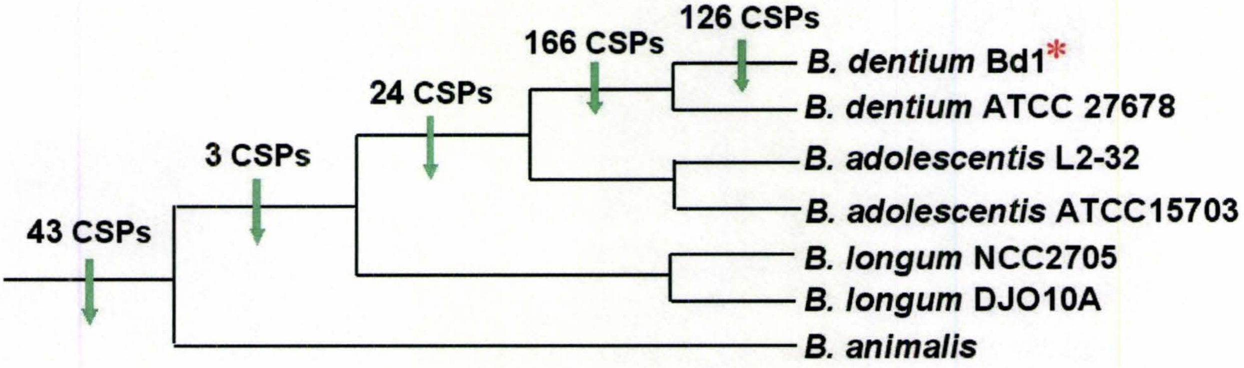


Figure 3.2 A subtree from Figure 4.1 in Chapter 4 showing the distribution pattern of various CSPs that are specific to *Bifidobacterium*.

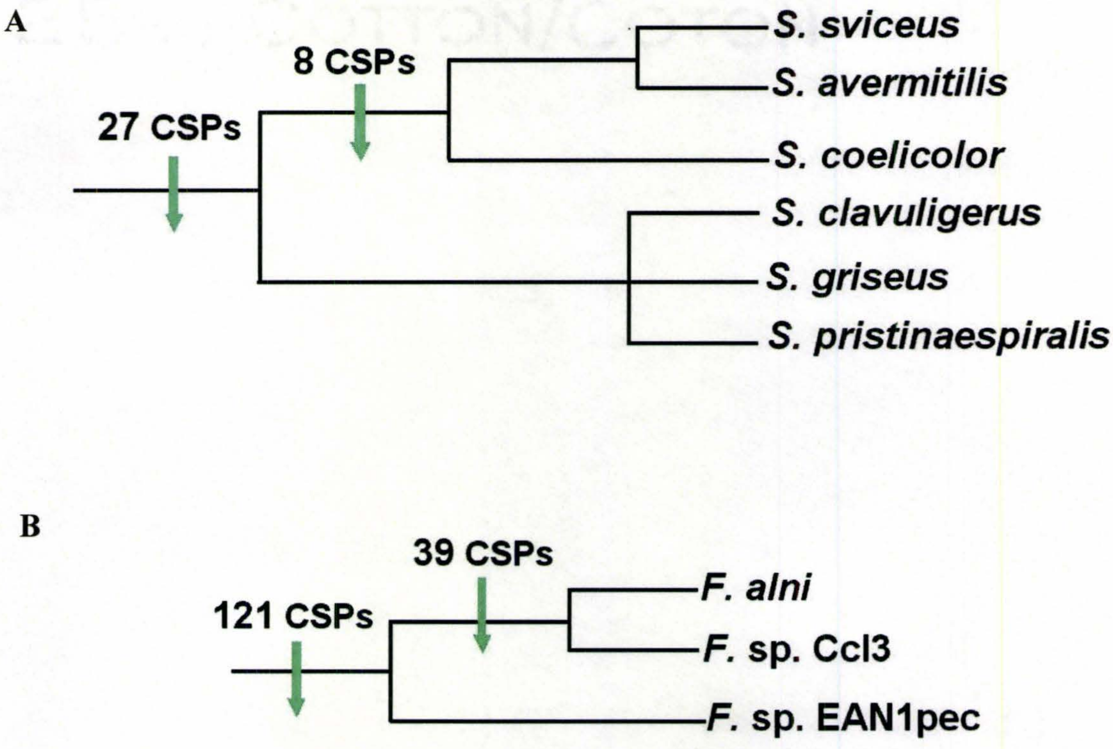


Figure 3.3. Subtrees from Figure 4.1 in Chapter 4 showing the distribution pattern of various CSPs that are specific to *Streptomyces* (A) and *Frankia* (B).

CHAPTER 4.

A Holistic View of Actinobacteria Phylogeny based on CSIs, CSPs and Combined Protein Tree

4.1 Preface

This chapter summarizes the results of the two molecular markers CSIs and CSPs, and also updates the actinobacterial phylogenetic tree by adding more actinobacterial species whose genomes are recently sequenced. Most of the results were reproduced from a manuscript that is in preparation: Gao B & Gupta RS, Comparative genomics of Actinobacteria revealed adaptive evolution and divergence hallmarks for different lineages, in preparation.

4.2 Introduction

The rapidly increasing genome sequences provide us the platform to study actinobacteria from different aspects, which also make it possible to reconstruct their phylogeny on the basis of a much larger data set per species, allowing a more reliable and representative inference of the tree of life. In the past few years, the phylogeny of actinobacteria or its subgroups have been studied using different datasets (Gao and Gupta, 2005; Ventura et al., 2006a; Kunisawa, 2007). The latest comprehensive study in 2007 constructed a phylogenomic tree for 20 available actinobacterial genomes based on 123 protein sequences representing the minimal core gene sequences of Actinobacteria (Ventura et al., 2007). However, in only two years, the number of sequenced actinobacterial genomes tripled (69 complete and 189 in progress) and the newly investigated genomes are from >13 new genera, which greatly enriched the resource and diversity for actinobacterial studies. Therefore, it is necessary to update the holistic view of actinobacterial phylogeny, based upon comparisions of the phylogenetic relationship revealed by phylogenetic trees as well as other novel actinobacteria-specific molecular markers discovered from comparative genomic analysis that are described in Chapter 2 & 3.

4.3 Materials and Methods

4.3.1 Phylogenetic Ananlysis

Phylogenetic analyses were performed on a concatenated sequence alignment for 35 conserved and widely distributed proteins. The information regarding the lengths and clusters of orthologous groups (COG) for these proteins is provided in Table 4.1. For each of these proteins, sequences from 76 actinobacterial genomes (from 42 genera), along with two species from Firmicutes as the root were retrieved. Multiple sequence alignments were created using the CLUSTAL_X 1.83 program (Thompson et al., 1997). The sequence alignments for these proteins were then concatenated and imported into the Gblocks 0.91b program to remove poorly aligned regions (Castresana, 2000). The Gblocks program was used mainly with the default setting (namely, minimum number of sequences for a conserved position, 24; minimum number of sequences for a flank position, 39; maximum number of contiguous non-conserved positions, 8; minimum length of a block, 10; allowed gap positions, half). The final alignment contained a total of 10,369 aa positions after removal of ambiguous regions. A neighbour-joining (NJ) tree based on 1000 bootstrap replicates was constructed by the Kimura model using the TREECON 1.3b program (Kimura, 1980; Van de Peer and De Wachter R, 1994). The maximum-likelihood (ML) analysis was carried out using the WAG+F model with gamma distribution of evolutionary rates with four categories using the TREE-PUZZLE program with 10 000 puzzling steps (Schmidt et al., 2002). A maximum-parsimony (MP) tree based on 1000 bootstrap replicates was computed using the MEGA 4.1 program (Tamura et al., 2007).

4.4 Results

4.4.1 Phylogenetic tree of Actinobacteria based on combined protein dataset

In order to get a comprehensive view of actinobacterial phylogeny covering different lineages, especially for the newly sequenced species, 76 actinobacterial genomes (from 42 genera) were chosen for the analysis including both complete and assembling genomes. A total of 35 universally distributed proteins were extracted from these genomes for phylogenetic analyses, all of which are highly conserved proteins and involved in a broad range of functions in the bacterial cell (Ciccarelli et al., 2006) (see Table 4.1). Phylogenetic trees based on a concatenated sequence alignment of these proteins were constructed using the neighbor-joining (NJ), maximum-parsimony (MP) and maximum-likelihood (ML) methods. The resulting trees from these analyses are shown in Figures 4.1-4.3. As seen in the trees, these three methods generally yielded consistant results regarding the clustering of members belonging to the same major subgroups. However, as for the branching order or interrelationship of the major subgroups, the NJ tree has better resolution while both MP and ML trees show multifurcation pattern at the weakly supported branching points. Compared to other phylogenetic tress (Zhi et al., 2009; Stackebrandt et al., 1997), the bootstrap scores at most of the branch points in these 3 trees are fairly high because of the large dataset of highly conserved proteins. More importantly, since many more new species from different genera were added into the analysis, the interrelationship between different lineages is much clearer, especially for some species whose phylogenetic positions were previously undetermined because of lack of close relatives, such as *T. fusca, Frankia sp., Kineococcus radiotolerans, Propionibacterium acne*, and so on. The following discussion will focus on the NJ tree in Figure 4.1.

Similar to the 16S rRNA tree, *Rubrobacter* represent the earliest branching lineage within the Actinobacteria and are distantly related to other species (Stackebrandt et al., 1997; Gao et al., 2006). Except for bifidobacteria, the other actinobacterial species shown in this tree are currently placed in the order *Actinomycetales*, which is the largest order in the Actinobacteria phylum (Ludwig and Klenk, 2001; Embley and Stackebrandt, 1994). Within this phylogenetically compact order, 7 previously defined major lineages (ranked as suborder in the current taxonomy) were clearly distinguished from each other, including: *Corynebacterineae, Pseudonocardineae, Micromonosporineae, Propionibacterineae, Streptosporangineae, Streptomycineae* and *Actinomycineae*. Among these suborders, two major clusters were revealed as highlighted in the NJ tree (Figure 4.1). *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae* form a well-defined cluster with a high bootstrap score 84, while *Frankineae, Streptosporangineae* and *Streptomycineae* branch together, distinct from other clusters.

In contrast, one of the largest suborder *Micrococcineae* do not form phylogenetically coherent cluster with bifidobacteria interspersed. *Micrococcineae* is the most diverse group within the Actinobacteria and the relationship within this suborder cannot be resolved by 16S rRNA gene with high confidence levels (Embley and Stackebrandt, 1994; Takarada et al., 2008). Based on the branching pattern, this suborder is divided into three clusters as marked in Figure 4.1: Cluster I includes *Arthrobacter, Renibacterium, Micrococcus, Kocuria*, and *Brevibacterium*; Cluster II consists of *Beutenbergia, Jonesia, Cellulomonas*, and *Xylanimonas*; while Cluster III collects *Clavibacter*, marine actinobacterium PHSC20C1, *Leifsonia* and the fast evolving intracellular parasite *Tropheryma*. Since *Tropheryma* show a very long branch in the tree that might result in false clustering pattern due to long-branch attraction (Bergsten, 2005; Philippe et al., 2005b), a phylogenetic analysis was performed with 74 actinobacterial genomes after removing the two *Tropheryma* samples. The resulting tree showed that Cluster III and Cluster II branched together but still distinct form Cluster I (data not shown). What's more, although the NJ tree based on 16S rRNA genes suggests bifidobacteria form one of the deepest branches and currently recognized as a separate order *Bifidobaceriales*, our results indicate that bifidobacteria is clustered within the *Micrococcineae*, showing a closer relationship with *Actinomycineae* (Stackebrandt et al., 1997). The MP and ML analysis based on 16S rRNA also suggest the clustering of bifidobacteria and Actinomycineae (Zhi et al., 2009). Additionally, our analyses of CSIs and CSPs in Chapter 2&3 did not discover any markers that are uniquely shared by all three clusters of *Micrococcineae*. But we have identified two CSIs that are only found in bifidobacteria, *Actinomycineae* and *Micrococcineae*, supporting their common origin (see discussion in section 4.4.2).

The phylogeny within the suborder *Frankineae* remains unclear and under emendation as new species were investigated by 16S rRNA sequence analysis (Zhi et al., 2009; Normand et al., 2007). Although the 3 *Frankia* strains group with *Acidothermus* forming "the core of *Frankineae*", other previously defined members (*Kineococcus, Geodermatophilus, Nakamurella*) branch independently as suggested in Figure 4.1 and by other evidence. The earlier studies regarded *Kineococcus* as a member of *Frankineae* (Bagwell et al., 2008; Stackebrandt et al., 1997; Cole et al., 2009). However, it is proposed by Zhi et al. that the genera *Kineococcus, Kineosporia* and *Quadrisphaera* form a distinct clade in the 16S rRNA tree and should be recognized as a new suborder *Kineosporiineae* (Zhi et al., 2009). Although the genome sequences for *Kineosporia* and *Quadrisphaera* are not available at the moment, the branching pattern in Figure 4.1 support that *Kineococcus* is independent from *Frankineae.* Nevertheless, it forms a separate cluster with another two species from *Kytococcus* and *Janibacter*, which are classified as belonging to *Micrococcineae.* The grouping of *Kineococcus* and *Janibacter* was also observed in the combined protein tree in supplementary Figure S4 by Zhi et al (Zhi et al., 2009). Other questionable species *Geodermatophilus* branch with *Salinispora* (*Micromonosporineae*), whereas *Nakamurella* group with *Saccharopolyspora* and *Actinosynnema* (*Pseudonocardineae*). It is mentioned by Zhi et al. that the families belonging to the suborder *Frankineae* were separated into distinct clades in the MP and ML trees, and it is notable that the branching points in the NJ tree were also not supported by high bootstrap scores (<50%), which suggest that the grouping may not be reliable or confident (Zhi et al., 2009). Besides, it is common that phylogenetic affiliation will be affected by the introduction of novel species. Most importantly, the incongruence of the positions of the above species in the 16S rRNA and combined protein trees may provide a starting point for reevaluation of the phylogeny and taxonomy of the suborder

Frankineae. Therefore, other group-specific characteristics are required to discriminate the different subgroups within the Actinobacteria phylum.

4.4.2 Evolutionary relationship of Actinobacteria resolved by CSIs and CSPs

As described in Chapter 2 & 3, we have identified a number of CSIs in different universal proteins and CSPs that are either specific to all actinobacteria or exclusive to its various subgroups. The shared presence of these rare genomic changes in a particular bacterial group supports its monophyletic relationshiop, which means that all members of that bacterial group came from a common ancestor. To compare the evolutionary relationship revealed by CSIs and CSPs with the phylogenetic tree, we labeled these two markers at different branch points in the tree, which indicate where they likely evolved (Figure 4.4). As seen in the figure, with both CSIs and CSPs, the Actinobacteria phylum and several major lineages can be delineated. For example, we have identified 3 CSIs in 3 different proteins (viz. CarA, RecR, IF-2) and 4 CSPs that are unique to one of the largest suborder Corynebacterineae. Besides, we also discovered 2 CSIs in 2 protein (viz. SRP and CDP) and 43 CSPs that are specific to *Bifidobacteriaceae*. More importantly, for the unresolved placement of species within the Actinobacteria phylum, the two markers provide useful information regarding their affiliation. For example, our studies have identified 4 CSPs that are unique to all actinobacteria including the deepest branch *Rubrobacte*r. These 4 CSPs are the only known characteristics that are shared by *Rubrobacter* and other actinobacteria. Thus, in addition to their branching pattern in the 16S rRNA tree, the newly identified markers provide valuable molecular evidence to define or circumscribe the Actinobacteria phylum or its subgroups in clear terms. Using these markers, it should be possible to identify new species belonging to Actinobacteria or its defined lineages.

In addition to serving as group-specific markers, CSIs and CSPs can also be used to infer the interrelationship among different lineages within a phylum. For example, we have identified 2 CSIs in protein HolB and S3 that are uniquely shared by *Corynebacterineae, Pseudonocardineae* and *Micromonosporineae*. The shared presence of the CSIs in these 3 suborders suggest that they are closely related and likely evolved from a progenitor cell, which is consistent with their clustering in the phylogenetic trees. Besides, 14 CSPs are found to be specific for *Corynebacterineae* and *Pseudonocardineae*, indicating that these two are more closely related than the *Micromonsporineae*. It should be mentioned that both these 2 CSIs are also found in species *Nakamurella multipartita* and *Geodermatophilus obscurus* (Figure 2.7 & 2.8), which provide additional molecular evidence that these two genera do not belong to Frankineae, and because of the absence of the *Corynebacterineae*-spcific CSIs and CSPs in these two genera, they probably belong to *Pseudonocardineae* and *Micromonosporineae*. Additionally, 9 out of 14 CSPs that are unique to *Corynebacterineae* and *Pseudonocardineae*, are also present in *N. multipartita* DSM 44233 although its genome is still in assembly (Table 3.2). The presence of these 9 CSPs in *N. multipartita* further suggests that this species should be placed in *Pseudonocardineae*.

As for the debating placement of bifidobacteia, our analyses have identified two CSIs in protein DnaK and S3 that are uniquely shared by bifidobacteria, *Actinomycineae* and *Micrococcineae* (Figure 2.9 & 2.10). In addition, a 1-aa indel in TrmD is specific to bifidobacteria and *Actinomycineae* (Table 2.3). These CSIs indicate that the above 3 subgroups are closely related, which support their clustering pattern in the combined protein tree (Figures 4.1-4.3). The deep branching pattern of bifidobacteria within the Actinobacteria phylum is only seen in the NJ tree based on 16S rRNA, but not supported by any other evidence. Our results from CSIs and combined protein tree suggest that bifidobacteria should not be regarded as a separate order, but rather it is a suborder that is closely related to *Actinomycineae* (Figure 4.4).

For some suborders, such as *Pseudonocardineae, Micromonosporineae, Streptosporangineae, Propionibacterineae* and *Actinomycineae*, only 1 or 2 species have been completely sequenced, while the genomes of other species shown in the tree are still in assembly. Therefore, currently we have not identified any CSIs or CSPs that can delineate these subgroups. As for the suborder *Micrococcineae*, our analyses did not uncover any CSIs or CSPs that are uniquely shared by all members of this suborder. But we identified two CSIs in DnaK and S3 that are uniquely shared by this suborder, bifidobacteria, and *Actinomycineae.* Thus, the monophyly of the current suborder *Micrococcineae* is questionable, which should be reevaluated by different methods.

4.5 Conclusion

To date, our phylogenetic analyses based on combined protein datasets provide the most comprehensive information regarding the Actinobacteria phylogeny. Compared to the 16S rRNA tree in which Actinobacteira display as a compact cluster, the combined protein tree have more resolving power regarding the relationship among different lineages as denoted by high bootstrap scores at the branch points. Besides, our tree points out some questionable placement of some lineages or new species, which suggest that the current taxonomy structure based on 16S rRNA should be reevaluated. More importantly, the identified CSIs and CSPs in this work provide additional useful molecular markers to define the Actinobacteria phylum or its various subgroups. Furthermore, for some weakly supported relationship in the phylogenetic tree, the two molecular markers provide robust molecular evidence regarding the common origin of some lineages.

4.6 Figures 1-4 and Table

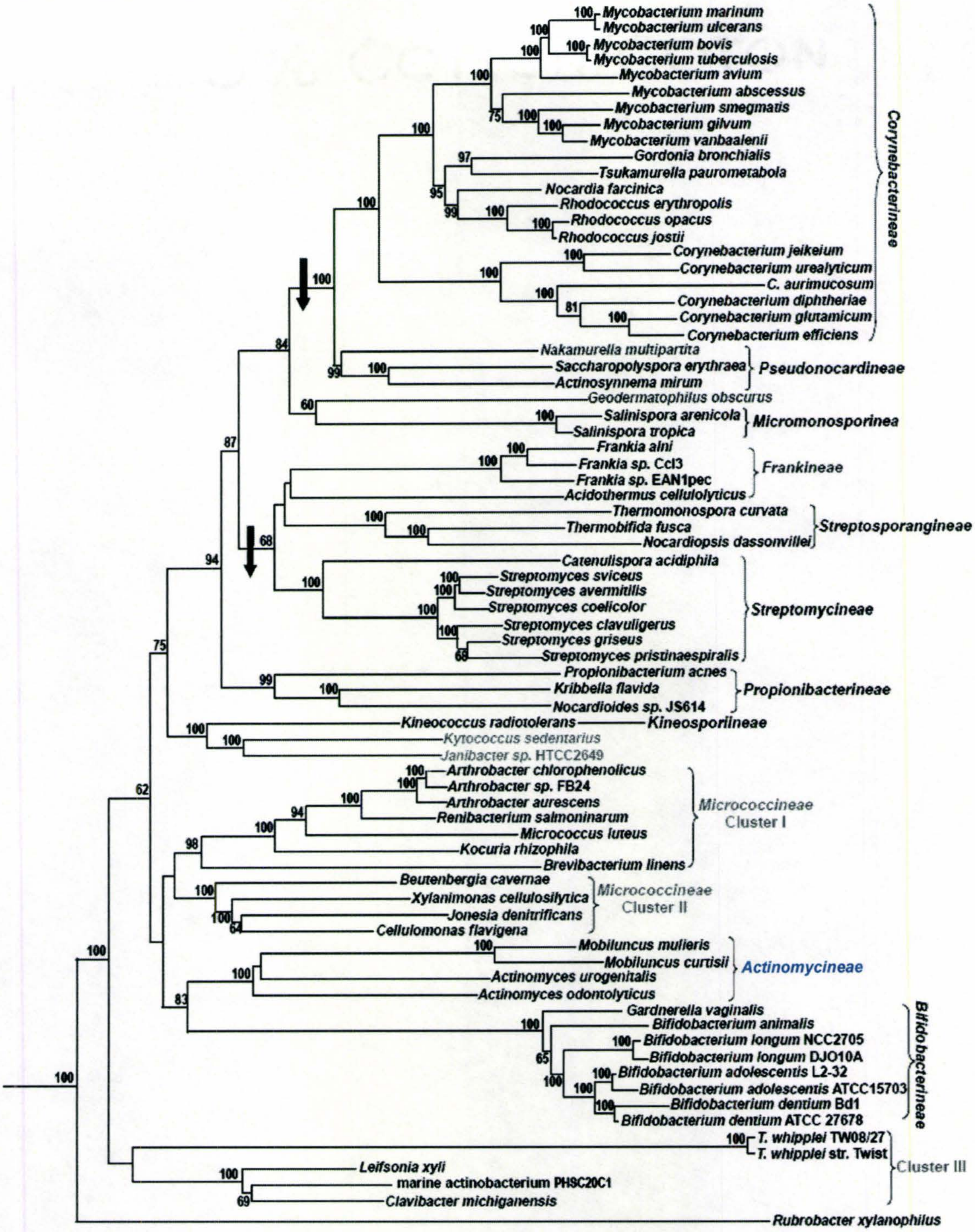


Figure 4.1. NJ tree of 76 actinobacterial species based on combined protein dataset. The numbers on the nodes indicate bootstrap scores that are >50%. Two species *B. subtilis* and *C. tetani* from Firmicutes were used as outgroup to root the tree. The two species colored in green are currently placed in *Frankineae* while the two species in orange belong to *Micrococcineae*.

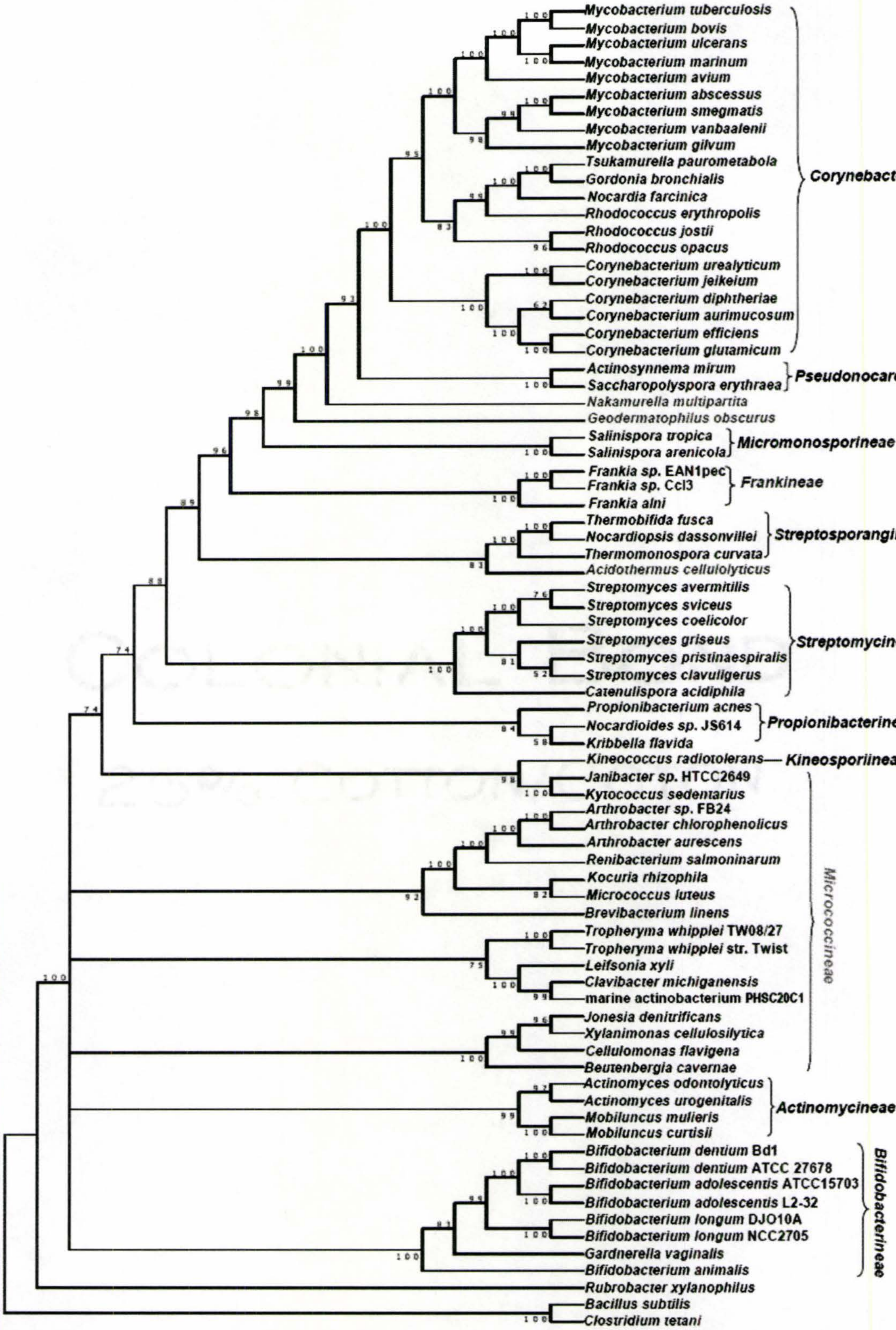


Figure 4.2. MP tree of 76 actinobacterial species based on the same dataset as Figure 4.1.

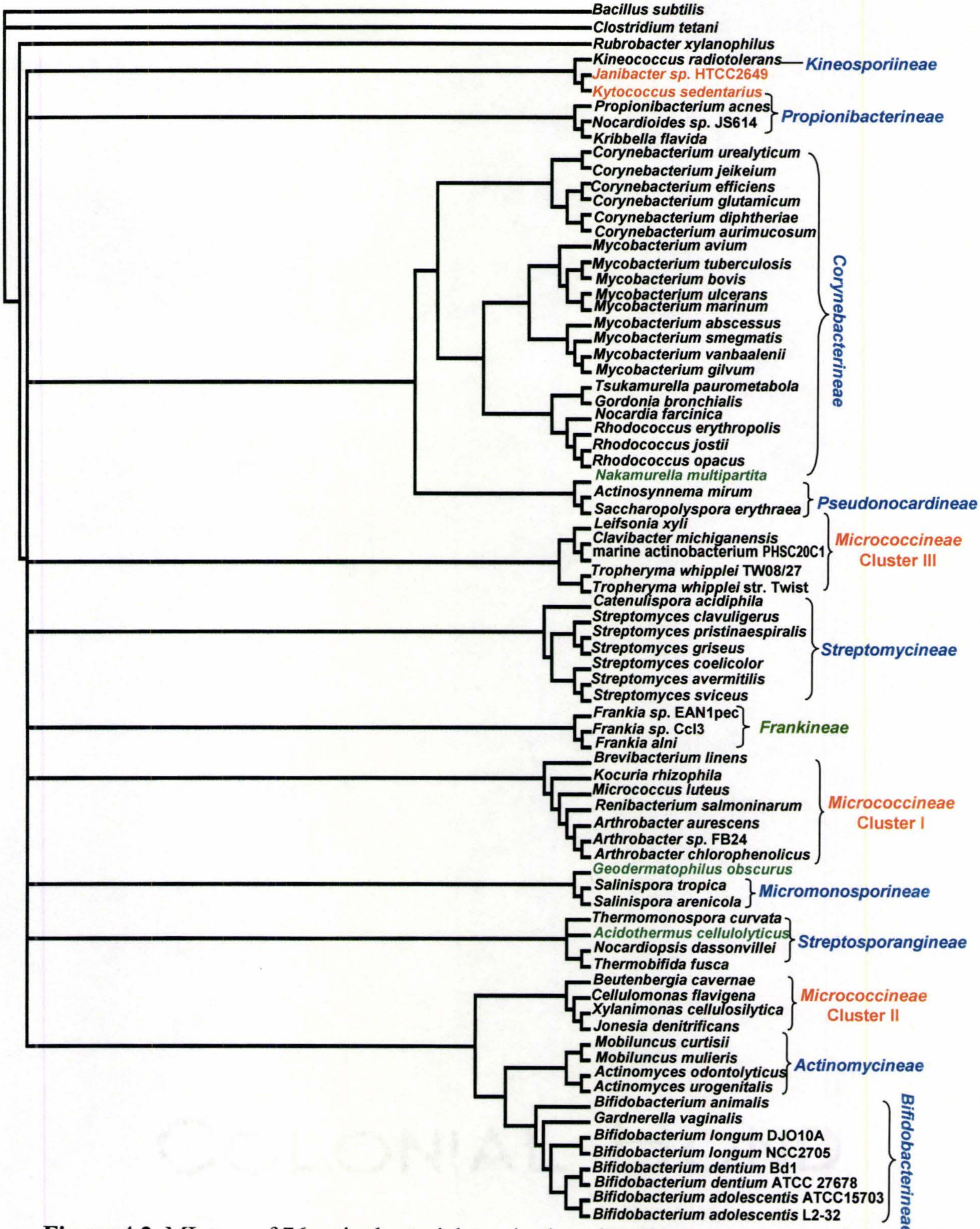


Figure 4.3. ML tree of 76 actinobacterial species based on the same dataset as Figure 4.1.

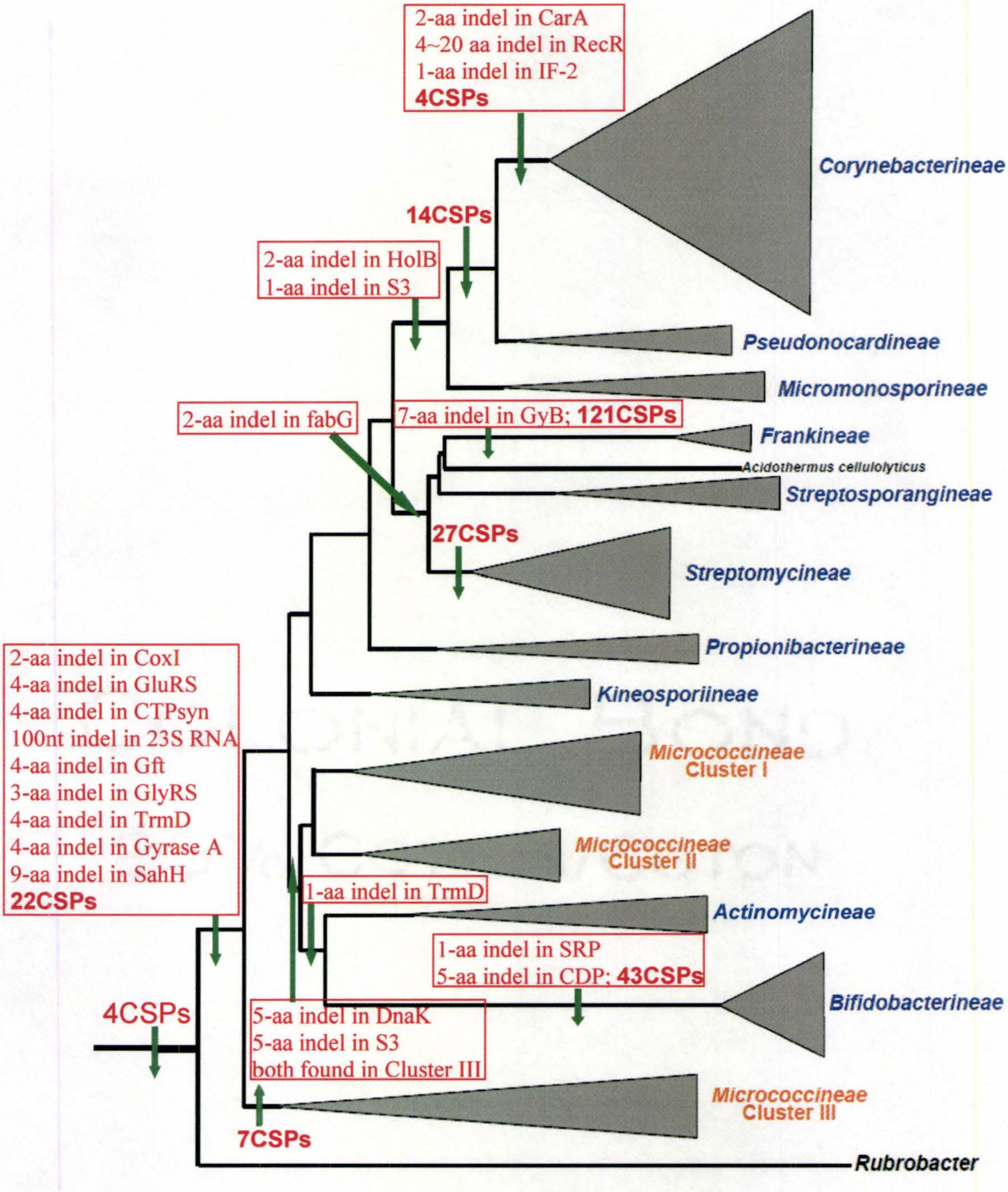


Figure 4.4. Summary of CSIs and CSPs that define the relationship within Actionobacteria phylum. The branching of the tree is a simplified version of Figure 4.1.

Table 4.1. Description of the Proteins used in Phylogenetic Analysis

COG Group	Length (aa)#	Annotation
COG0012*	362	GTP-binding protein, probable translation factor
COG0030	286	Dimethyladenosine transferase (KsgA)
COG0049*	156	30S ribosomal protein S7
COG0050	397	Elongation factor Tu
COG0080*	144	50S ribosomal protein L11
COG0081*	241	50S ribosomal protein L1
COG0085*	1161	RNA polymerase subunit beta (RpoB)
COG0086	1299	RpoC
COG0087*	214	50S ribosomal protein L3
COG0088	219	Ribosomal protein L4
COG0090	278	Ribosomal protein L2
COG0091*	125	50S ribosomal protein L22
COG0092*	277	30S ribosomal protein S3
COG0093*	122	50S ribosomal protein L14
COG0094*	185	Ribosomal protein L5
COG0096*	132	30S ribosomal protein S8
COG0097*	179	50S ribosomal protein L6
COG0098*	201	30S ribosomal protein S5
COG0099*	126	Ribosomal protein S13
COG0102*	147	50S ribosomal protein L13P
COG0103*	170	30S ribosomal protein S9
COG0184*	95	30S ribosomal protein S15
COG0185	93	Ribosomal protein S19
COG0187	686	Gyrase B
COG0188	857	Gyrase A
COG0197*	176	50S ribosomal protein L10/L16
COG0201*	437	Preprotein translocase subunit SecY
COG0202*	340	RNA polymerase subunit alpha
COG0441	658	Threonyl-tRNA-synthetase
COG0443	618	Molecular chaperone DnaK (Hsp70)
COG0459	541	Chaperonin GroEL (Hsp60)
COG0533*	374	O-sialoglycoprotein endopeptidase
COG0575	391	Phosphatidate cytidylyltransferase
COG0575	391	CDP-diglyceride synthase (CdsA)
COG1466	401	DNA polymerase III subunit delta (holB)

Note: The proteins were selected based on the paper (Ciccarelli et al., 2006), which describes 31 proteins that are universally distributed in all organisms and involved in a broad range of functions. However, only 21 proteins (denoted by * in the table) from these 31 were available for all 76 actinobacterial genomes since some are still in assembly. In order to maximize both the number of new species and the size of combined protein dataset for analysis, in addition to the 21, another 14 proteins were selected that are highly conserved and also involved in different cellular functions. #: Protein length is from the representative strain *S. coelicolor* A3(2).

CHAPTER 5.

Structural and Phylogenetic Analysis of a Conserved Actinobacteria-Specific Protein (ASP1; SCO1997) from *Streptomyces coelicolor*

5.1 Preface

This chapter was reproduced from the published manuscript (Gao et al., 2009b): Gao, B., Sugiman-Marangos, S., Junop, M.S., Gupta, R.S. Structural and phylogenetic analysis of a conserved Actinobacteria-Specific Protein (ASP1; SCO1997) from Streptomyces coelicolor. 2009. BMC Struct Biol. 9:40.

This chapter describes the first characterization of one of the 5 actinobacteria-specific proteins, ASP1 (Gene ID: SCO1997) from *Streptomyces coelicolor*. The X-ray crystal structure of ASP1 was determined and compared with its most similar structure of nucleoside phosphorylase enzymes. Sequence analyses were carried out which revealed that ASP1 is paralogous to another actinobacteria-specific protein ASP2 (SCO1662 from *S. coelicolor*). Dr. Murray Junop collected the X-ray diffraction data for this protein and solved the structure with the help of Seiji Sugiman-Marangos. All other work presented in this chapter such as protein crystallization and phylogenetic analysis was conducted by this author.

5.2 Introduction

Our recent comparative genomic studies on available actinobacterial genomes have identified a large number of proteins that are either specific for all actinobacterial species or certain subgroups within this phylum (Gao et al., 2006). BLAST searches with these proteins show no significant hits or similarity to any other protein in the database. These proteins thus provide novel and useful molecular markers for this diverse group of bacteria (Gao et al., 2006; Gupta and Gao, 2010). Among these actinobacteria-specific proteins, four proteins (corresponding to ML1009, ML1306, ML1029 and ML0642 from the genome of *Mycobacterium leprae* TN) were found in every sequenced actinobacterial species including those from the deepest branch *Rubrobacter xylanophilus* and also from intracellular pathogens such as *Tropheryma whipplei* which have highly reduced genomes (Gao et al., 2006; Stackebrandt et al., 1997; Raoult et al., 2003). All four of these proteins are conserved within actinobacteria but have no known function. These four actinobacteria-specific proteins are referred to in this work as ASP-1, 2, 3 and 4. The simplest and most logical explanation for the persistence of these proteins in only actinobacteria is that their genes evolved only once in a common ancestor of all actinobacteria and were subsequently passed on to all their decedents. So these genes/proteins provide among the very few molecular characteristics known that are distinctive of the *Actinobacteria* phylum (Gao et al., 2006; Gao and Gupta, 2005; Roller et al., 1992). In view of their actinobacteria-specificity, it is of great interest to determine the cellular functions of these proteins and the cellular processes in which they participate. These studies are expected to provide novel insights into biochemical processes and physiological characteristics that are unique to actinobacteria.

In an attempt to gain insight into the cellular functions of these proteins, we have initiated structural work on these 4 actinobacteria-specific proteins. We determined the crystal structure of SCO1997 from *S. coelicolor*, which corresponds to the protein ML1009 from *M. leprae* (ASP1) (Gao et al., 2006). Structural and phylogenetic analysis indicates that although ASP1 retains a similar overall fold compared to members of the hydrolase superfamily such as purine nucleoside phosphorylase, the active site region and therefore function of ASP1 are distinct (Pugmire and Ealick, 2002; Mao et al., 1997).

Comparison of the most highly conserved sequences of ASP1 from different actinobacteria with their positions in the crystal structure reveals a potential role for ASP1 in binding and transport of divalent metal ion. Interestingly, additional sequence and structural analyses show that another actinobacteria-specific protein ASP2 (SCO1662; ML1306) is evolutionarily and functionally related to ASP1 (Gao et al., 2006; Maguire, 2006; Payandeh and Pai, 2006).

5.3 Materials and Methods

5.3.1 Protein Expression and Purification

The ASP1 gene (SCO1997) from *S. coelicolor* A3(2) was cloned into the pET-22b vector and expressed in *E. coli* BL21(DE3) as a full length recombinant protein with a C-terminal $(His)_6$-tag. SeMet protein was expressed in the methionine auxotroph *E. coli* B834 using a previously described method (Hendrickson et al., 1990). For expression of both native and SeMet derivatized ASP1, cells were grown at 37 °C to an OD_{600} of ~0.6; induced with 1mM IPTG; harvested after 4 h; resuspended in a binding buffer containing 20 mM Tris, pH 7.4, 500 mM NaCl and 10 mM imidazole; lysed in a French pressure cell; and clarified by centrifugation. Supernatant was loaded on a 1 mL Ni-column, and washed with 200mL binding buffer along with 36mM imidazole, and finally eluted at 300 mM imidazole. The eluted proteins were diluted 5 fold with buffer A (20 mM Tris, pH 7.5) and loaded onto a 5 mL HiTrap Q HP anion exchange column (Amersham) for further purification. Proteins were eluted with a 120 mL linear gradient from 50 to 500 mM NaCl. ASP1 eluted as a single peak at ~260 mM NaCl. Individual fractions from across the peak were pooled and buffer exchanged into a low-salt buffer (25mM KCl, 10mM HEPES, pH 7.5) for crystallization. The buffer used for gel filtration chromatography contained 20 mM Tris (pH 7.4) and 200mM KCl.

5.3.2 Crystallization and Data Collection of ASP1

All crystals were grown at 17 °C using the hanging drop/vapour diffusion method. Hanging drops containing 1 uL of protein solution (5 mg/mL) and 1 uL of mother liquor (0.1M MES, 0.55M magnesium formate, pH6.5~6.8, 0.25~0.5% n-Octyl-beta-D-glucoside, 0~1.5% glycerol) were dehydrated over a reservoir containing 800uL of 1.5M $(NH_4)_2SO_4$. Cubic shaped crystals (100 x 100 x 100 μm^3), suitable for data collection, grew after approximately 3 days incubation. Crystals were flash frozen directly in a nitrogen cold stream (100 K) with no further cryo-protection. Diffraction data sets for native and SeMet crystals were collected at wavelengths of 1.1 and 0.979 Å, respectively. All data was collected at the X25 beamline using an ADSC Q315 CCD x-ray detector (NSLS, Brookhaven, NY).

5.3.3 Structure Determination and Model Refinement

SAD data collected to 2.0 Å was processed using d*TREK (Pflugrath, 1999). All 5 of the expected SeMet sites were located using HYSS (Adams et al., 2002; Grosse-Kunstleve and Adams, 2003). Phasing and density modification were carried out using CNS (Brunger et al., 1998). Iterative rounds of manual model building and refinement were performed with Coot and REFMAC5 until R and R_{free} values converged and could no longer be improved (Emsley and Cowtan, 2004; Vagin et al., 2004). The coordinates of the final APS1 model were deposited in the Protein Data Bank under accession code 3E35. Surface area calculations were performed using the program PISA version 1.15

(Krissinel and Henrick, 2007). Structure similarity searches were performed by DaliLite program v3 (Holm et al., 2008). Structural illustrations presented in figures were generated with PyMOL (Delano, 2002).

5.3.4 Phylogenetic Analysis

Phylogenetic analyses were carried out based on sequence alignments for ASP1 and ASP2 homologous genes from 18 actinobacterial species. Among these selected species, only 8 contain one of the two genes, while the others contain both gene copies. Multiple sequence alignments were created using the ClustalX version 1.83 (Thompson et al., 1997). The alignment was then imported into TREE-PUZZLE version 5.2 for maximum-likelihood (ML) analysis using the WAG+F model with gamma distribution of evolutionary rates with four categories (Schmidt et al., 2002; Whelan and Goldman, 2001).

5.4 Results

5.4.1 Crystal Structure of ASP1 from *S. coelicolor*

The protein ASP-1 is of hypothetical or unknown function. The genes involved in related functions (e.g. those that are part of an operon) are generally clustered in various species or closely related species. Thus, genetic linkage studies can often provide valuable clues regarding possible cellular function of a given gene/protein (Galperin and Koonin, 2000; Doerks et al., 2004). Hence, we have examined the neighboring genes of ASP1 in various sequenced actinobacteria. The genes flanking ASP1 in different actinobacterial genomes are either of unknown function or perform unrelated functions.

Thus, it provides no useful information regarding the possible cellular function of this protein (Gao et al., 2009b).

To gain insight into the cellular function of ASP1, we have cloned, expressed and crystallized the gene for this protein from *S. coelicolor* A3(2) (Gao et al., 2009b). The crystal structure of full length ASP1 was determined using Seleno-methionine (SeMet) derivatized ASP1 and single anomalous diffraction (SAD) techniques. The final model was refined with native data (2.2 Å) to R and R_{free} values of 17.4% and 23.4%, respectively. The structure of ASP1 contained three regions that were unable to be traced into electron density and therefore not included in the final model. These disordered regions included the first 2 residues at the N-terminus, the last 36 C-terminal residues (amino acids 277-312) as well as a short loop region encompassing residues 168-172. A complete list of data collection and model refinement statistics can be found in Table 5.1.

Crystals grew in space group *I*23 and contained a single copy of ASP1 in the asymmetric unit. Upon inspection of crystallographic packing interactions it appeared that ASP1 might exist as a trimer. The amount of surface area buried through the formation of an ASP1 trimer is significant at 7560 Å^2. As well, when analyzed by size exclusion chromatography (Figure 5.1), ASP1 eluted with a Stokes radius consistent with a molecular mass equivalent to ~125 kDa (monomer 36 kDa), further supporting the idea that ASP1 exists as a trimer in solution.

ASP1 contains a single domain comprising a central mixed β-sheet (β1-3-6-7-2-8-10-9) flanked by 4 α-helices on one side and 3 on the other yielding an overall three layered αβα fold (Figure 5.2). Helices F and G form an elbow-shaped extension that is peripheral to the core domain. Based on secondary structure prediction of the missing 36 C-terminal residues, an additional or perhaps extended helix is expected to follow αG. Trimer formation is largely stabilized by interactions between an extended anti-parallel hairpin (β4-5) and the αD region from an adjacent monomer (Figure 5.2). A portion of the extended loop (residues 175-179) preceding αD further stabilizes the trimer through interactions with β4-5, resulting in formation of a 3-stranded anti-parallel sheet (Figure 5.3).

Assembly of the ASP1 trimer results in the formation of a roughly globular complex (~ diameter 70 Å) with three notable features (Figure 5.3). First, one side of the trimer adopts a very flat surface, forming what could perhaps function as a large docking interface. The electrostatic potential on this surface is quite neutral having only a small amount of basic potential. A second unusual feature of the ASP1 trimer is the presence of a large internal cavity (~ 7500 $Å^3$) surrounded by a three-pronged claw-like structure. Given the size of this cavity and overall claw-like structure that surrounds it, it is quite possible that this region acts as a binding surface for another protein(s) and or substrate. The electrostatic surface potential of each claw is negative creating an overall acidic surface on the internal cavity region of ASP1.

The final and most notable feature of the ASP1 trimer is the presence of two well-ordered magnesium ions (see Figure 5.4C for bonding distance and geometry) located at a central pore formed along the central three fold symmetry axis. This pore is ~ 20 Å deep and is lined by six concentric rings of amino acids with alternating charge and polarity (Figure 5.3 and 5.4A). The shape of the pore is conical and is tapered to its narrowest point of 4.14 Å at D71 located within the second layer. The first Mg^{2+} ion is positioned just above a negative ring of amino acids formed by three copies of D71 and D116 (Figure 5.4). Water molecules in the first hydration shell of this metal ion are directly hydrogen bonded to D71 (Figure 5.4C). A second metal ion is located in a hydrophobic pocket lined by V117 at the third layer. Water molecules within the first hydration shell of this metal ion are in direct van der Waals contact with V117. Through its second hydration shell the second Mg^{2+} is further stabilized by hydrogen bonding to D71 and also to the main chain carbonyl of R68. Because there was high concentration of Mg^{2+} in the mother liquor (~0.55M), the specificity and possible role of this metal ion-bound, channel-like pore is unclear, but may be involved in the biological function of the ASP1 trimer.

Although it is tempting to speculate that the presence of two Mg^{2+} ions in the central pore region of ASP1 suggests a role for ASP1 in metal transport, there is no direct evidence to support this idea. Furthermore, a structural comparison of ASP1 with CorA, a well characterized Mg^{2+} transporter whose homologs are present in *S. coelicolor* and various actinobacterial (Lunin et al., 2006; Payandeh and Pai, 2006), shows no obvious similarity between these proteins (results not shown). Therefore, if ASP1 function does involve some aspect of Mg^{2+} binding and/or transporter it does not appear to be similar to that conducted by CorA.

5.4.2 Structural Comparisons of ASP1

To further characterize the structure of ASP1 and gain insight into its possible function, we performed a comparative structural analysis using the program DaliLite version 3 (Holm et al., 2008). This analysis revealed significant structural similarity to a homologue from *Corynebacterium glutamicum* (GeneID: Ncgl1848) [PDB: 2P90], as well as several bacterial purine nucleoside phosphorylases and a number of other glycosidic hydrolases from the larger NP-1 family.

5.4.2.1 Comparison of ASP1 from *S. coelicolor* and *C. glutamicum*

As expected, structural comparison of ASP1 from *S. coelicolor* and *C. glutamicum* showed a high degree of conservation (root mean square deviation (RMSD): 1.6 Å). Importantly, the structure of ASP1 from *C. glutamicum* crystallized as a trimer that is identical to the trimer reported here for ASP1 from *S. coelicolor*. This finding, along with our gel filtration data, provides additional support for the trimeric structure of ASP1 generated through crystallographic symmetry. Another important observation from the comparison of the structure from *C. glutamicum* is the structural conservation of the metal binding pore despite the absence of bound metal ion. The fact that the pore region adopts an identical structure even when a metal ion is not present provides strong evidence to suggest that the binding of metal is not simply required for structure integrity of the ASP1 trimer.

A comprehensive sequence alignment of ASP1 homologues from a broad range of actinobacterial species (Figure 5.5) clearly demonstrates that residues contributing to the formation of two distinct regions (the central pore and C-terminal elbow) within the structure of ASP1 represent the most highly conserved sequence of the protein (Figure 5.6) (Gao et al., 2009b). Figure 5.6 illustrates the importance of conserved residues (absolutely conserved in purple, highly conserved in yellow) in forming the pore and elbow regions. While most of these residues are involved in structural stabilization others, such as D71 and L268, are not. As suggested elsewhere, absolutely conserved amino acids that do not directly contribute to structure stability and are solvent exposed, are expected to define key regions for protein function (Schueler-Furman and Baker, 2003; George et al., 2005; Livingstone and Barton, 1993; Lichtarge et al., 1996). At this point it is difficult to infer what function the elbow region might serve. Given its distal location, however, it seems likely to mediate interaction with other proteins or perhaps the missing C-terminal region of ASP1. The C-terminal region of ASP1 contains a number of highly conserved residues (I296, E302, F304, L305). Interestingly, this region is not observed in either of the currently available structures suggesting that it may only become ordered upon binding another molecule.

5.4.2.2 Comparison of ASP1 and PNP

As mentioned above, comparative structural analysis revealed significant similarity (Z score ~ 10) between ASP1 and members of the NP-1 family of nucleoside phosphorylase enzymes. This family of enzymes participates in the salvage pathway of purines and pyrimidines biosynthesis and catalyzes the reversible phosphorolysis of purine and pyrimidine nucleosides (Pugmire and Ealick, 2002; Mao et al., 1997). The NP-1 family member that shares greatest structural similarity to ASP1 is purine nucleoside phosphorylase (PNP) from *E. coli* [PDB: 1ECP]. Despite having very low sequence similarity (8% identity), ASP1 and $\text{PNP}_{E.coli}$ structures could be aligned with an overall RMSD of ~ 3.0 Å. With the exception of a few insertions and deletions, these proteins share identical overall topology (Figure 5.7). Two notable insertions include: the

C-terminal highly conserved elbow (αF-G) and the extended arm region (β4-5) essential for ASP1 trimer stability. In addition, there is an insertion of sequence that significantly increases the loop size between αB-β3, occluding much of the normal PNP substrate-binding surface (Figure 5.7C). While these insertions are expected to contribute to ASP1 function and certainly quaternary structure, members of the NP-1 family are characterized by different oligomeric arrangements ranging from dimer, to trimer and in some instances hexameric structures. Therefore, these observed differences do not necessarily preclude shared function between PNP and ASP1.

In contrast, the following evidence strongly suggests that ASP1 does not function as a nucleoside phosphorylase. First, a large region of PNP responsible for forming an entire side of its active site cleft (residues ~ 100-180 encompassing β7-8-9 and αC-D-E; Figure 5.7B) is completely missing in the ASP1 structure, rendering ASP1 incompatible of binding nucleoside. Second, a sequence alignment of ASP1 homologues fails to identify any of the highly conserved residues involved in substrate binding or catalysis within the NP-1 family (Gao et al., 2009b). Furthermore, from sequence and structural alignments it is equally clear that those regions of ASP1 which are most highly conserved, are not present within NP-1 family members. Finally, a PNP homologue in the *S. coelicolor* genome has already been identified (SCO4917) and shows no significant similarity to ASP1. Taken together, the observations from both sequence and structural comparison indicate that while ASP1 and PNP share similar overall structure and topology, their functions are different.

5.4.3 Phylogenetic Analysis of ASP1 and ASP2

Of the 4 actinobacteria-specific genes previously identified through comparative genomic analysis of 19 actinobacterial species, two genes ASP1 (SCO1997; ML1009) and ASP2 (SCO1662; ML1306) appear to encode structurally related proteins (Gao et al., 2006). These proteins have comparable length and share significant sequence similarity (25% identity and 43% similarity). The question remains, are these two conserved actinobacteria-specific proteins functionally related?

We have conducted a search for ASP1 and ASP2 homologues in all available sequenced actinobacterial genomes (61 strains). Interestingly, while most actinobacterial species contain homologues of both ASP1 and ASP2, some species contain only one homologue (Gao et al., 2009b). The single homologue by definition shares similarity to both ASP1 and ASP2. Species (18 in total), which only contain one homologue, are found in 7 divergent genera (*Corynebacterium, Actinomyces, Saccharopolyspora, Brevibacterium, Bifidobacterium, Tropheryma* and *Rubrobacter*). Further phylogenetic analysis of ASP1 and ASP2 homologues from different actinobacterial species was conducted to determine how these two genes are related. In the phylogenetic tree shown in Figure 5.8, two distinct clusters are observed with a strong bootstrap score (98%) indicating that the observed branch pattern is highly reliable. One cluster collected all genes homologous to ASP2 while the other cluster, grouped only those genes homologous to ASP1. The genes from the 18 species containing only one homologue do not form a third branch, but rather fall into either the ASP1 and ASP2 clusters. The two distinct clusters observed in the phylogenetic tree suggest that ASP1 and ASP2 are paralogues that evolved from a gene duplication event in a common ancestor of actinobacteria. Therefore, most members from this phylum contain both ASP1 and ASP2 except those species, which have lost one copy later in the evolutionary process. The fact that ASP1 and ASP2 are paralogues, yet either can be lost, suggests that these two paralogues perform similar functions. Based on their sequence and functional similarity, these two proteins are also expected to share significant structural similarity. Preliminary X-ray crystallographic analysis indicates that the tertiary and quaternary structure of ASP2 is in fact similar to ASP1 (data not shown).

Sequence alignment of ASP1 and ASP2 homologues demonstrate that important residues which are highly conserved in ASP1 homologues and likely involved in protein function are also conserved in ASP2 homologues (Gao et al., 2009b). 8 of the 15 absolutely conserved residues from ASP1 homologues are also absolutely conserved amongst ASP2 homologues. The remaining 7 are still highly conserved and are only substituted with similar amino acids (Gao et al., 2009b). This finding further underscores the importance of these residues in mediating the function of both paralogs. As stated earlier, amino acids that fall within the category of absolutely conserved and solvent exposed are particularly predictive of regions important for mediating interactions with other functionally important molecules (Schueler-Furman and Baker, 2003; George et al., 2005; Lichtarge et al., 1996). D71 is most interesting in this regard because it not only fits this category, but is also found bound to two magnesium ions in the ASP1 structure. We know that the binding of magnesium is not required for overall structural stability since the structure of ASP1 from *C. glutamicum* does not contain metal ion. The precise function of this region within ASP1 and ASP2 will require further investigation.

5.5 Discussion

The *Actinobacteria* phylum represents one of the largest groups of bacteria. Amazingly this diverse collection of bacteria can be characterized genetically to a first approximation by the presence of only 5 unique genes. All of these 5 genes, are of unknown function but they are expected to encode for function(s) that ultimately control actinobacteria-specific and important biological process(es). Understanding the cellular function of a protein of unknown function is not a straightforward task (Galperin and Koonin, 2004). However, structure determination often provides the most useful information in this regard (Danchin, 1999). In this work, we report the structure of the first actinobacteria-specific protein. Our structural data in combination with sequence analysis further supports the idea that this protein carries out a novel function. This function is novel in the sense that the structure of this protein does not match any known protein, with or without known function. Given the immense number of structures that are now available and the wide coverage of function, it is reasonable to propose that ASP1 may mediate a function highly specific to Actinobacteria. Although it is unclear from the structural data alone, it seems possible that ASP1 function may involve some aspect of divalent metal ion interaction. It will be intriguing to determine what contribution, if any, this highly conserved 'pore' region makes toward ASP1 function. Our phylogenetic analysis also shows that another actinobacteria-specific protein ASP2, which is a paralogue of ASP1, may also have similar structure and function. Future genetic and biochemical studies of these proteins is therefore of great interest in linking the conservation of the biology of actinobacteria and their 4 unique genes.

5.6 Table and Figures 1-10

Table 5.1. Crystallographic data and model refinement statistics

	Native[a]	Se-SAD[a]
Data collection		
Space group	*I*23	*I*23
Cell dimensions		
a, *b*, *c* (Å)	135.1, 135.1, 135.1	135.4, 135.4, 135.4
α, β, γ (°)	90, 90, 90	90, 90, 90
Wavelength	1.1000	0.9794
Resolution (Å)[b]	50.0-2.0 (2.07-2.0)	50.0-2.3 (2.38-2.3)
R_{merge}[b]	8.4 (89.5%)	22.1 (63.0%)
$I / \sigma(I)$[b]	37.9 (4.6)	6.2 (2.2)
Completeness (%)[b]	100.0 (100.0)	99.9 (100.0)
Redundancy[b]	22.1 (22.2)	22.3 (22.5)
Refinement		
Resolution (Å)	50.0-2.0	
No. reflections	25,482	
R_{work} / R_{free}	20.7%/24.1%	
No. atoms		
Protein	2085	
Ligand / ion	2	
Water	272	
B-factors	49.8	
R.m.s deviations		
Bond lengths (Å)	0.01	
Bond angles (°)	1.25	

[a] One crystal was used for data collection.
[b] Values in parentheses are for highest-resolution shell.

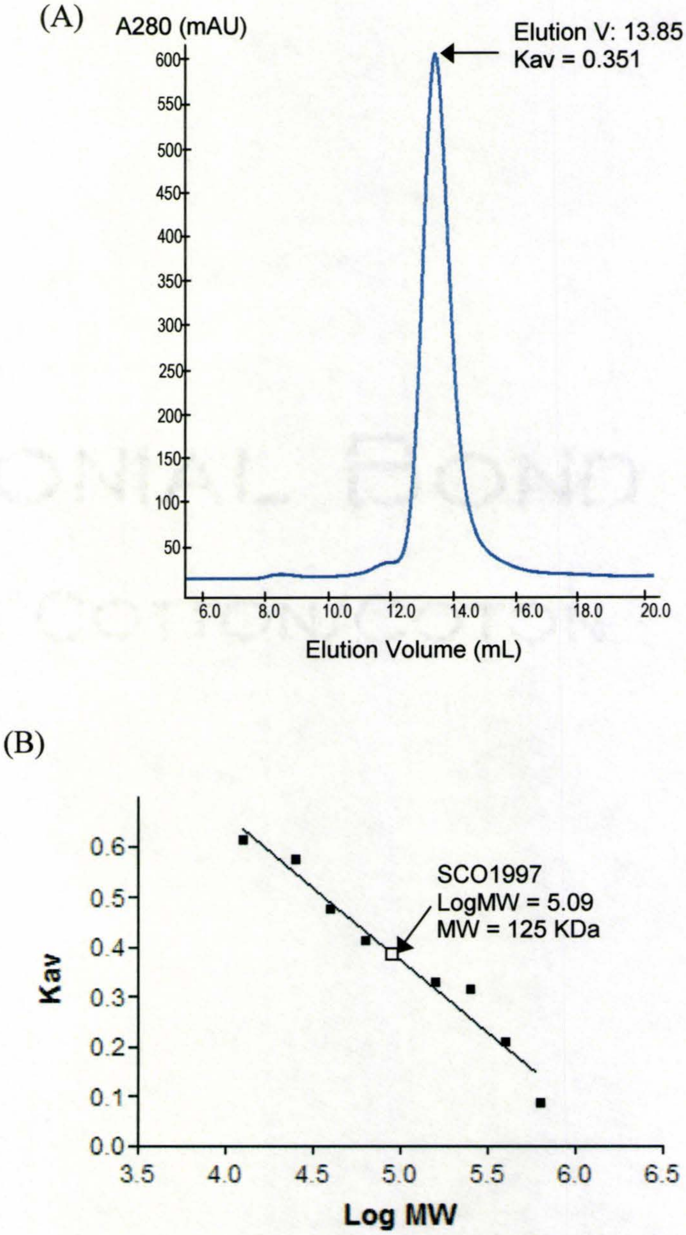


Figure 1. Size exclusion chromatography of ASP1. (A) Size exclusion chromatographic analysis of full length ASP1 at 5 mg/mL. A single peak was eluted at 13.8 mL, consistent with the expected elution volume of a roughly globular ~125 kDa protein. (B) Standard curve for calibration of S200 size exclusion column.

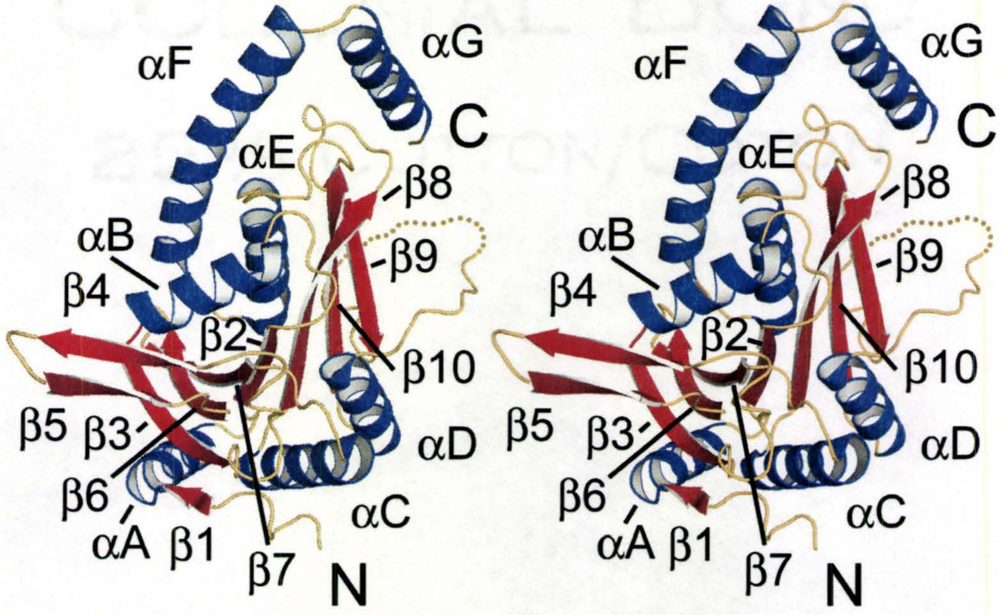


Figure 5.2. Stereo image of ASP1 monomer structure. β strands and α helices are in red and blue, respectively. A single disordered loop between β9-αD is shown as a dotted line.

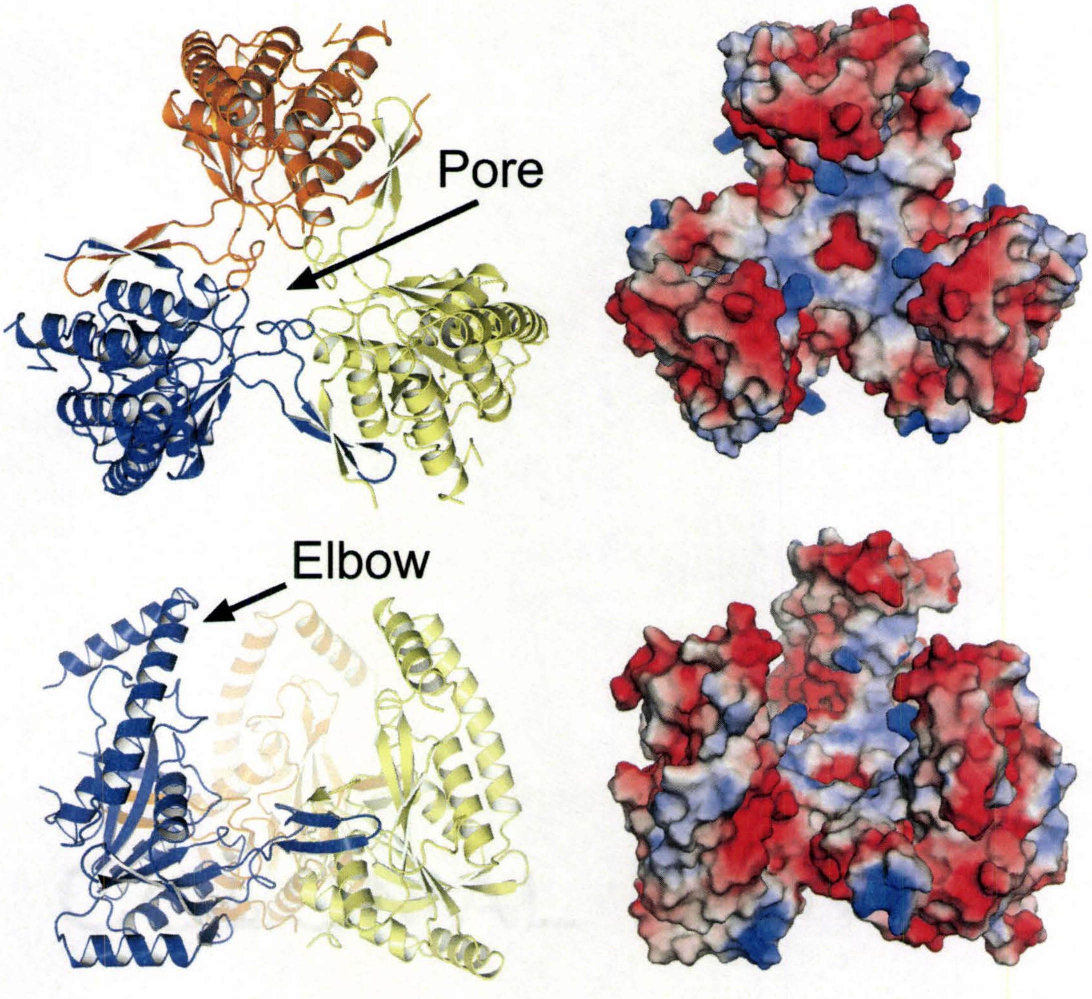


Figure 5.3. Structure of ASP1 trimer. (A) and (C) Orthoganol views of ASP1 trimer shown in ribbon. Individual subunits are colored, yellow, blue and orange. (B) and (D) Surface representations corresponding to views of ASP1 in (A) and (C), respectively. Positive and negative electrostatic potential are indicated in blue and red surface, respectively.

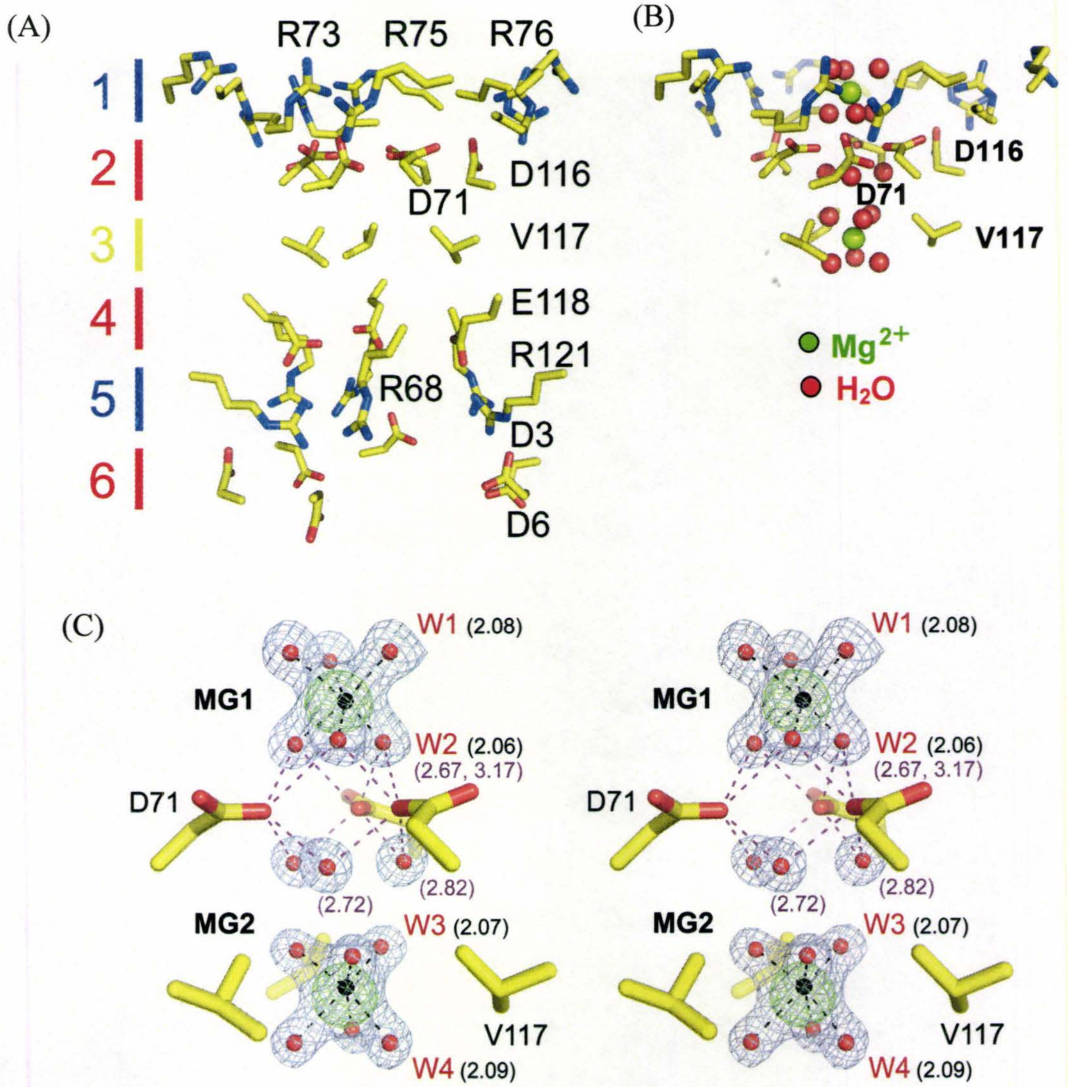


Figure 5.4. Central metal-binding pore. (A) Amino acids lining the metal-binding pore are shown in stick representation. Concentric layers of amino acids are numbered and corresponding polarity indicated by color. Red, blue and yellow indicate negative, positive and neutral charge, respectively. (B) Interaction between upper 3 layers of central pore and hydrated Mg^{2+} ions. (C) Magnesium coordination binding analysis. An Fo-Fc Mg^{2+}omit map contoured at 5 σ is shown in green mesh. For reference, a 2Fo-Fc map contoured at 1.5 σ (blue mesh) is also shown for Mg^{2+} ions (black sphere) and water molecules (red sphere) in the central pore region. Water molecules bond to Mg^{2+} ions are labeled in red W1 to W4. Distances in Å are indicated in parenthesis with black and purple corresponding to water-metal and water-side chain distances, respectively. Interactions with both metals are indicated as black dashed bonds, while those involving D71 are shown in light purple.

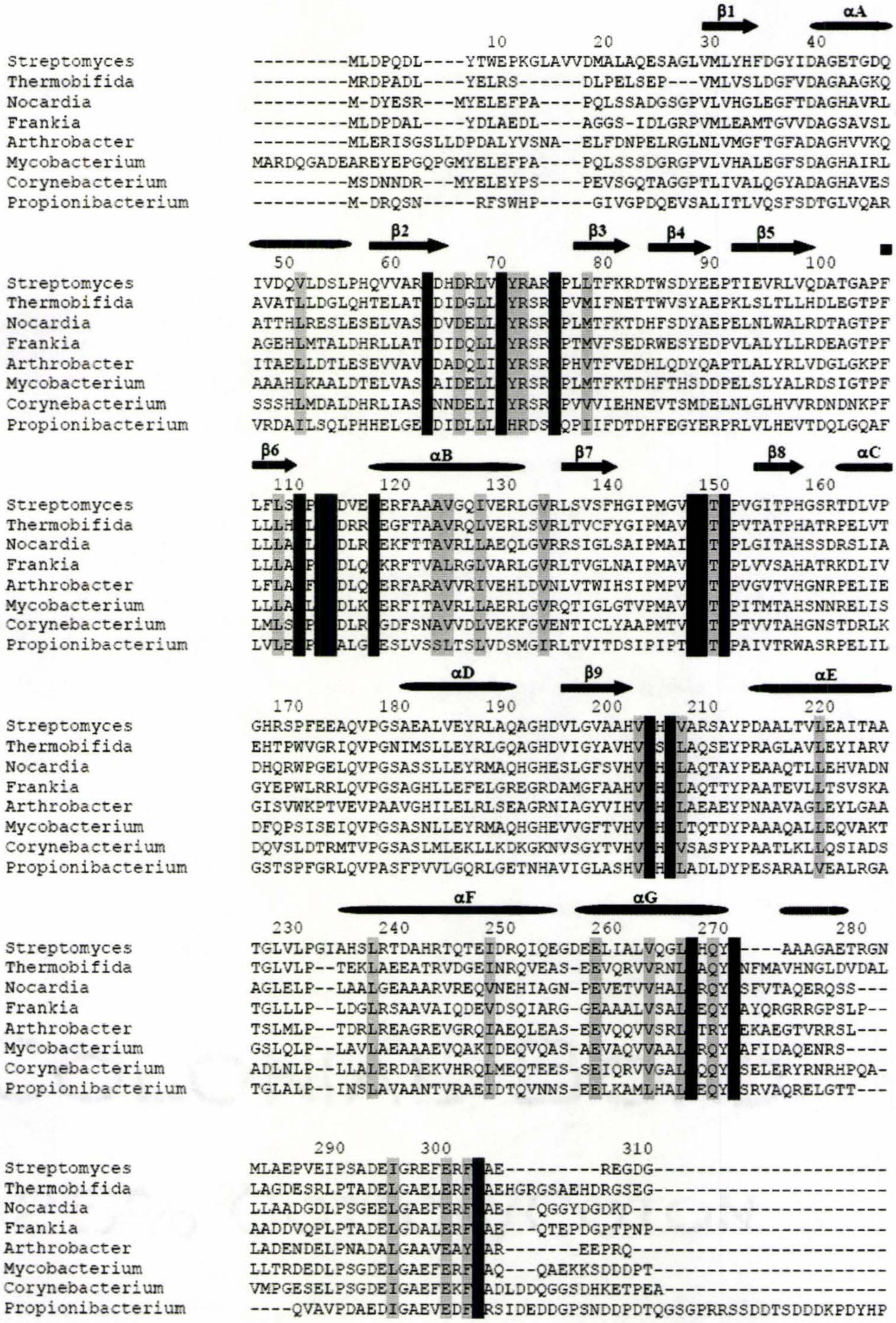


Figure 5.5. Multiple sequence alignments of ASP1 homologues from 8 representative actinobacterial species. Conserved residues are colored based on the complete alignment of all 43 available actinobacterial homologues: purple, absolutely conserved residues; yellow, highly conserved residues. β strands and α helices are labeled in red and blue, corresponding to the ribbon diagram in Figure 5.2.

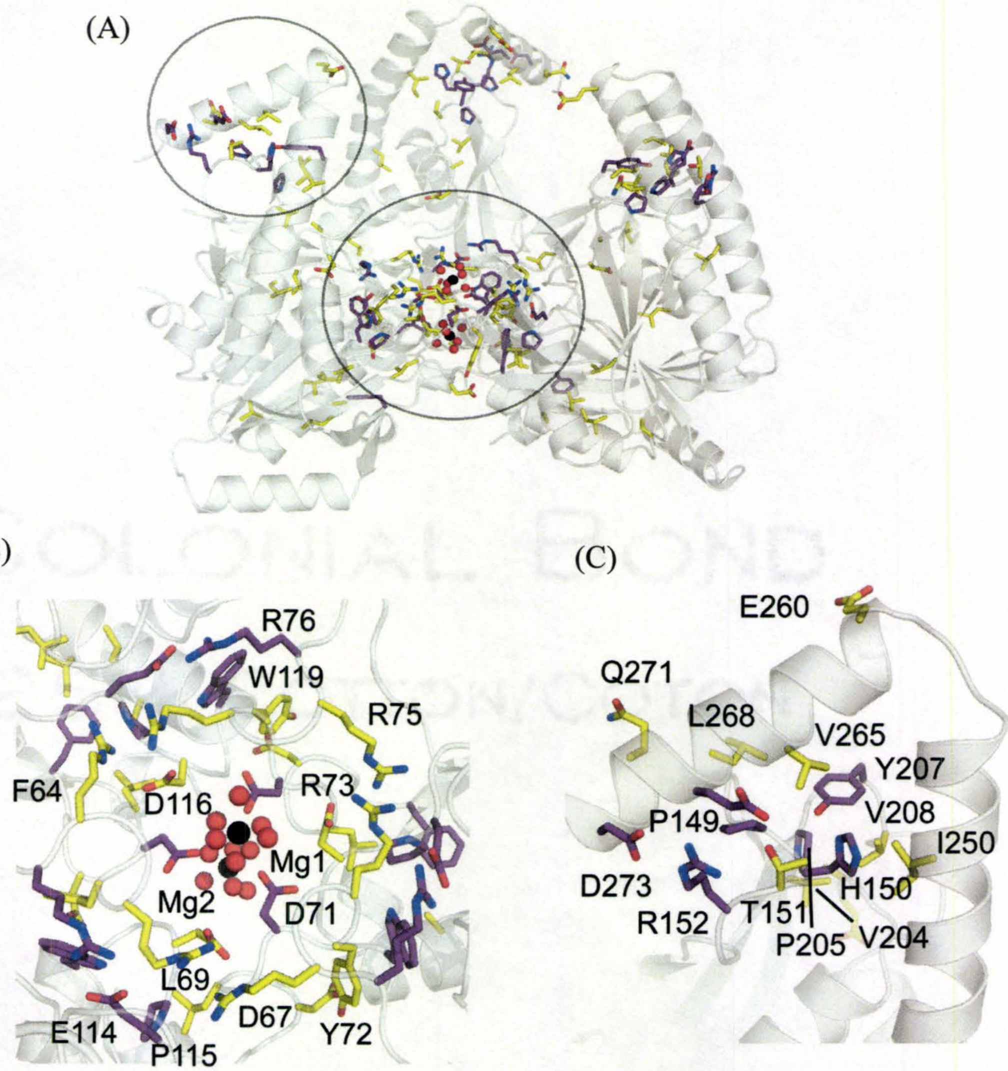


Figure 5.6. Highly conserved regions within the ASP1 trimer. (A) Conserved residues important for forming the central pore and elbow regions. A circle delineates each region. Absolutely conserved residues are colored in purple and highly conserved ones are colored in yellow. (B) and (C) Details about the central pore region and C-terminal elbow regions, respectively. Conserved residues are indicated as in (A). Magnesium ions are shown as black spheres with their first hydration shell of water molecules shown as red spheres.

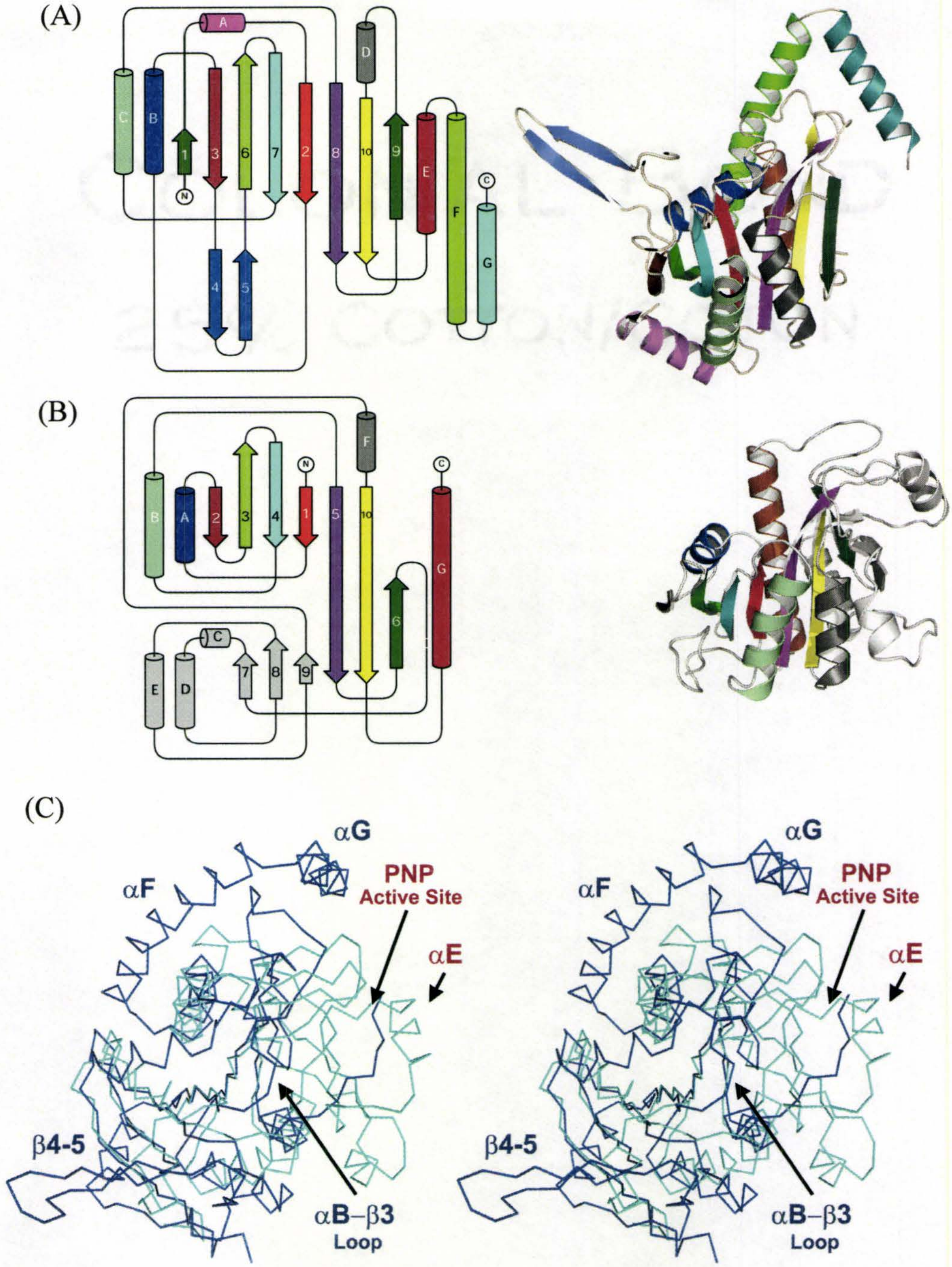


Figure 5.7. Structural and topological comparison of (A) ASP1 and (B) PNP$_{E.\ coli}$ [PDB: 1ECP]. (C) Structural comparison of ASP1 monomer (dark blue) with PNP$_{E.\ coli}$ (light blue) [PDB: 1ECP]. Regions defining ASP1 structure are labeled blue, while those referring to PNP$_{E.\ coli}$ are labeled red.

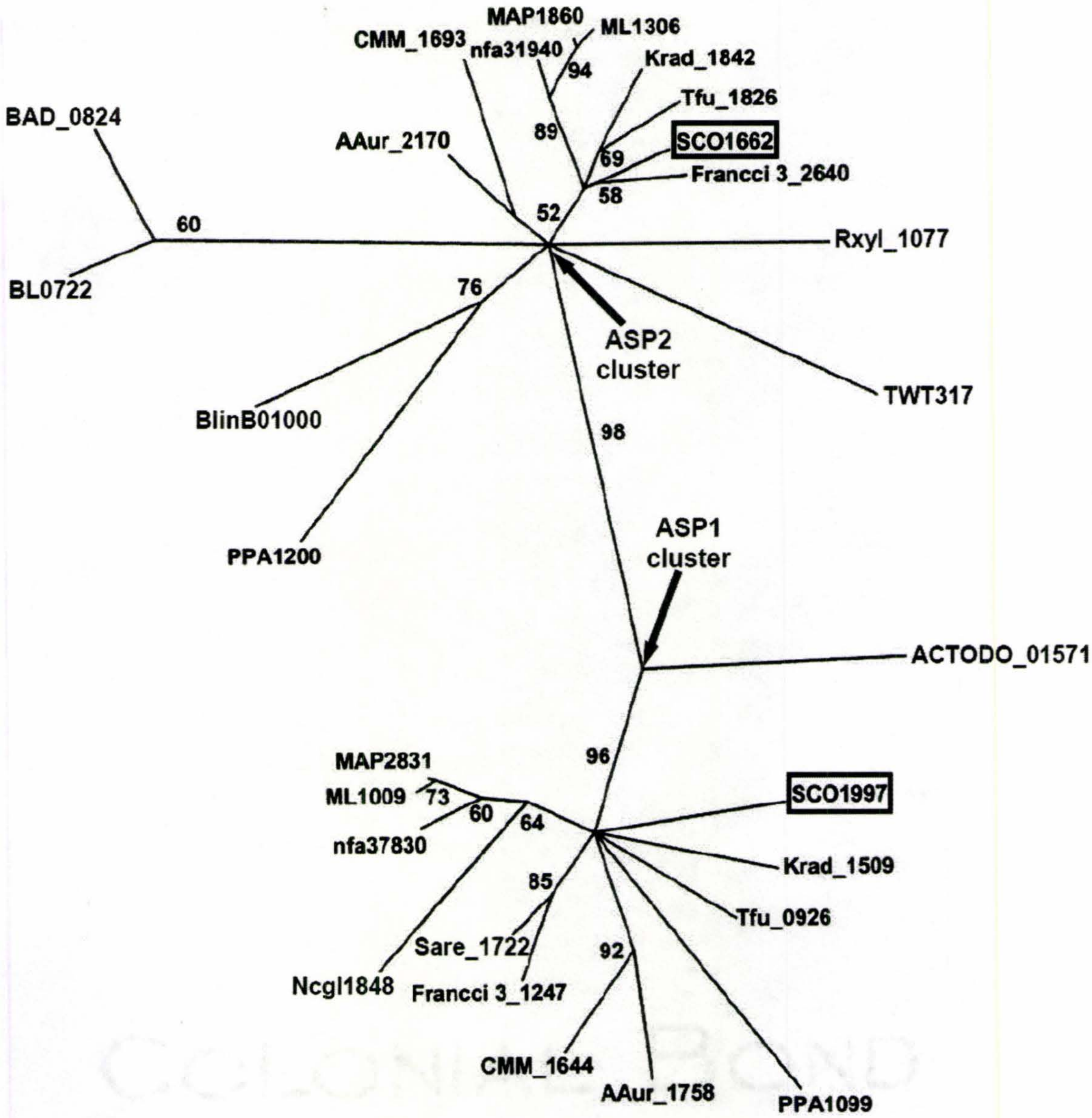


Figure 5.8. Phylogenetic tree of ASP1 and ASP2 homologous genes in different actinobacterial species. The gene ID in blue are from these genomes which only contain one of the two paralogous genes. The bootstrap scores >50% are indicated on various branch points.

CHAPTER 6.

Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis

6.1 Preface

This chapter describes proteins identified from comparative genomic studies that are uniquely shared by either all archaea or the different main groups within archaea. The Archaea-specific protein work has been published (Gao and Gupta, 2007) and the complete reference for this work is as follows: Gao B & Gupta RS. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics. 2007 Mar 29;8(1): 86.

6.2 Introduction

The Archaea are highly diverse in terms of their physiology, metabolism and ecology (Barns et al., 1994; Gribaldo and Brochier-Armanet, 2006; Gupta, 1998a). Presently, very few molecular characteristics are known that are uniquely shared by either all archaea or the different main groups within archaea (Graham et al., 2000; Walsh and Doolittle, 2005). The evolutionary relationships among different groups within the Euryarchaeota branch are also not clearly understood (Ludwig and Klenk, 2001; Brochier et al., 2005a; Gribaldo and Brochier-Armanet, 2006). Comparative studies on limited numbers of archaeal genomes have been carried out by a number of investigators using different criteria. Graham et al. (Graham et al., 2000) analyzed 9 archaeal genomes to identify signature proteins that function uniquely within the Archaea. Their definition of an archaeal signature protein required it to be present in only two different euryarchaeal species and they identified 353 archaeal signature proteins. Makarova and Koonin have analyzed archaeal genomes to identify core sets of genes, which are present in all archaeal species, but which are not restricted to the archaeal species (Makarova et al., 1999; Makarova and Koonin, 2005). Recently, Walsh and Doolittle have analyzed prokaryotic genomes to measure dissimilarity between Archaea and Bacteria (Walsh and Doolittle, 2005). Although it was reported that 28% of the proteins from archaeal genomes are restricted to the Archaea, specific proteins that were present in different groups of archaea were not identified. Other comparative studies using different criteria have been conducted on smaller groups within archaea such as *Pyrococcus, Sulfolobus* and thermoacidophilic organisms (to be discussed later). However, thus far no comprehensive phylogenomics study on different archaeal genomes has been carried out using the same standard criteria to identify proteins or ORFs that are shared by all archaea or its different major lineages. In this study we have carried out comparative analyses of archaeal genomes using uniform criteria to identify CSPs that are uniquely present in archaeal species at different phylogenetic depths (genus or higher) representing all major groups within the Archaea.

6.3 Materials and Methods

6.3.1 Identification of CSPs that are specific to Archaea

To identify proteins which are specific for Archaea or its various subgroups, all proteins in the genomes of *A. pernix* K1 (APE) (Kawarabayasi et al., 1999), *S. acidocaldarius* DSM 639 (Saci) (Chen et al., 2005), *P. aerophilum* str. IM2 (PAE) (Fitz-Gibbon et al., 2002), *P. abyssi* GE5 (PAB) (Fitz-Gibbon et al., 2002), *M. maripaludis* S2 (MMP) (Hendrickson et al., 2004), *M. kandleri* AV19 (MK) (Slesarev et al., 2002), *M. burtonii* DSM 6242 (Mbu) (Allen et al., 2009), *Halobacterium* sp. NRC-1 (VNG) (Ng et al., 2000), *H. walsbyi* DSM 16790 (HQ) (Bolhuis et al., 2006), *T. acidophilum* DSM

1728 (Ta) (Ruepp et al., 2000)and *P. torridus* DSM 9790 (PTO) (Futterer et al., 2004), were analyzed. Protein-protein BLAST searches were carried out on each individual protein using the default parameters, without the low complexity filter, to identify different proteins where all significant hits were from archaea (Altschul et al., 1997). The results of BLAST searches were inspected for sudden increase in Expected values (E-values) from the last archaeal species in the search to the first non-archaeal organism. The proteins that were of interest generally involved a large increase in E-values from the last archaeal hit to the first hit from any other organism. Further, the E-values of these latter hits were expected to be in a range higher than 10^{-4}, which indicates a weak level of similarity that could occur by chance. However, higher E-values are sometimes acceptable for smaller proteins as the magnitude of the E-value depends upon the length of the query sequence (Altschul et al., 1997).

All promising proteins identified by the above criteria were further analyzed using the position-specific iterated (PSI) BLAST program. In the present work, a protein was considered to be archaea-specific if all hits producing significant alignments were from the indicated groups of archaeal species. However, we have also retained a few proteins where 1 or 2 isolated species from other groups (e.g. bacteria or eukaryotes) also had acceptable E-values. We consider these proteins to be also archaea-specific and their presence in isolated unrelated species is very likely due to LGT. For all archaea-specific proteins described here, the protein ID, accession number and their possible functions (also COG or CDD number) are presented in Tables. All proteins indicated in various tables are specific for the Archaea based on these criteria unless otherwise mentioned.

6.3.2 Phylogenetic analyses

Phylogenetic analyses was carried out on a concatenated sequence alignment of 31 universally distributed proteins (Ciccarelli et al., 2006). The information regarding these proteins is provided in the Additional file 10. For each of these proteins sequences from all 29 archaeal species were downloaded and multiple sequence alignments were created using ClustalX 1.83 program (Thompson et al., 1997). A concatenated sequence alignment for all 31 proteins was imported into Gblocks 0.91b (Castresana, 2000) to remove poorly aligned region. The resulting final alignment of 6,252 amino acid sites was used for phylogenetic analyses. A NJ tree based on this dataset was constructed by TREECON 1.3b program with Kimura two-parameter model distance (Van de Peer and De Wachter R, 1994); ML tree were computed under a WAG+F model plus a gamma distribution with four categories by TREE-PUZZLE 5.2 (Schmidt et al., 2002; Whelan and Goldman, 2001); MP tree were obtained by Mega 3.1 package (Kumar et al., 2004). All of the trees were bootstrapped 100 times.

6.4 Results

6.4.1 Phylogenetic analyses of archaeal species

Prior to undertaking comparative studies on archaeal genomes, phylogenetic analysis of sequenced archaeal species was carried out so that the results of phylogenomics analyses could be compared with those obtained by traditional phylogenetic approaches. Phylogenetic trees for the archaeal species based on 16S rRNA as well as concatenated sequences of translation and transcription-related proteins have been published by other investigators (Olsen et al., 1994; Burggraf et al., 1997; Brochier et al., 2005a; Matte-Tailliez et al., 2002). In the present work, we have constructed phylogenetic trees for 29 archaeal species using a set of 31 universally distributed proteins that are involved in a broad range of functions (Ciccarelli et al., 2006). The sequence of *Haloquadratum walsbyi* DSM 16790, which became available afterward, was not included in these studies. Phylogenetic trees based on a concatenated sequence alignment of these proteins were constructed using the NJ, ML and MP methods.

The results of these analyses are presented in Figure 6.1. All three methods gave very similar tree topologies except for the branching positions of *M. kandleri* and *Methanospirillum hungatei*, which were found to be variable. Except for this, the branching pattern of the archaeal species based on our dataset is very similar to that reported by Gribaldo et al. (Brochier et al., 2005a; Gribaldo and Brochier-Armanet, 2006) based on concatenated sequences of translation and transcription-related proteins. In the tree shown, the Crenarchaeota and Euryarchaeota, the two major phyla within Archaea were clearly distinguished from each other. The phylogenetic affinity of *Nanoarchaeum*, which has a long-branch length, was not resolved in this or various other trees (Brochier et al., 2005a; Brochier et al., 2005b). Within Crenarchaeota, *Pyrobaculum* was indicated to be a deeper branch, and *Aeropyrum* branched in between the *Pyrobaculum* and *Sulfolobus*. Within Euryarchaeota, the clades corresponding to *Halobacteria, Thermococci* and *Thermoplasmata* were resolved with high bootstrap scores, but the methanogens were split into 2–3 clusters. One of these clusters that have low bootstrap score consisted of *Methanobacteriales* and *Methanococcales* with *M. kandleri* (*Methanopyrales*) branching in its vicinity (Burggraf et al., 1991; Rivera and Lake, 1996). The second cluster, with higher bootstrap score, showed a grouping of *Methanomicrobiales* and *Methanosarcinales*. These two clusters, which are separated by *Thermoplasmata, Archaeoglobi* and *Halobacteria*, have been referred to as Class I and Class II methanogens by Bapteste et al. (Bapteste et al., 2005).

6.4.2. CSPs specific for all Archaea

Our analyses have identified 1448 CSPs that are unique to different groups of Archaea and for which no homologues are generally found in any bacterial or eukaryotic species. Based on their specificity for different taxonomic groups, these proteins have been divided into a number of different groups (Gao and Gupta, 2007). Among these CSPs, 16 are present in nearly all archaeal species but whose homologues are not found in any Bacteria or Eukaryotes with a single exception (Table 6.1a). Of these, the first 6 proteins in the left column (Table 6.1a) viz. PAB0063, PAB0252, PAB0316, PAB1633, PAB1716 and PAB2291, are present in all sequenced archaeal genomes. The unique presence of these CSPs in all sequenced archaeal genomes indicates that these CSPs could be regarded as distinctive characteristics or molecular signatures for the archaeal domain. The genes for these CSPs likely evolved in a common ancestor of the Archaea and were then vertically acquired by other archaeal species. Makarova and Koonin have also mentioned 6 proteins that are commonly shared by different archaea, but the identity of such proteins was not specified (Makarova and Koonin, 2005). These proteins are likely the same. The remaining 10 proteins in Table 6.1(a) are missing only in *Nanoarcheum equitans*, which is a tiny parasitic organism containing only 536 genes (Waters et al., 2003; Huber et al., 2003). The species distribution pattern of these proteins can be accounted for by one of the following two possibilities. First, it is possible that *N. equitans* is the deepest branching lineage within archaea, as has been suggested (Waters et al., 2003; Huber et al., 2003) and the genes for these 10 proteins evolved in a common ancestor of the other archaea after its divergence (Figure 6.2a). Alternatively, similar to the first 6 proteins, the genes for these 10 proteins evolved in a common ancestor of all archaea, but they were then selectively lost in *N. equitans* (Figure 6.2b) (Makarova and Koonin, 2005; Brochier et al., 2005b; Huber et al., 2003). Based upon our results, one cannot distinguish between these two possibilities. However, in view of the fact that the genome of *N. equitans* has undergone extensive genome shrinkage (only 0.49 Mb) and it is at least 3 times smaller than the next smallest archaeal genome, we favour the latter possibility (Figure 6.2b) (Makarova and Koonin, 2005; Brochier et al., 2005b; Huber et al., 2003).

Of the CSPs that are uniquely present in all archaea, PAB0063 corresponds to tRNA nucleotidyltransferase (CCA-adding enzyme), which builds and repairs the 3' end of tRNA (Cho et al., 2005). Functionally similar enzymes are also present in bacteria and eukaryotes (assigned as Class II), but their sequences share very little homology with the archaeal CCA-adding enzyme (Class I), which explains why no homologs were detected in any bacteria or eukaryotes in blast searches. The main mechanistic difference between class I and class II enzymes is that the tRNA substrate is required to fully define the nucleotide binding site in class I enzyme, whereas class II has a preformed nucleotide binding site that recognizes CTP and ATP in the absence of tRNA (Xiong et al., 2003). Another CSP PAB0316 is assigned as archaeal type DNA primase, which also has its synonymous counterparts in bacterial and eukaryotic species, but shows very little homology to them (Iyer et al., 2005; Ito et al., 2001). In the same way, protein PAB1633 is annotated as a PilT family ATPase, which showed very little similarity to bacterial ATPases involved in type IV pili biogenesis (Ng et al., 2000). Further studies of this protein could provide insights into novel aspects of the archaeal flagellar system. A number of other CSPs viz. PAB1716, PAB0018a, PAB0075, PAB0475 and PAB2104, have also been assigned putative functions based on sequence analysis, but their exact roles in archaeal cells remains to be determined. Interestingly, for protein PAB0075, two gene copies with acceptable E-values are also present in the genomes of *Dehalococcoides ethenogenes* 195, *Dehalococcoides* sp. CBDB1 and *Dehalococcoides* sp. BAV1, which belong to Chloroflexi (Ludwig and Klenk, 2001). Because no homologue of PAB0075 is present in other bacteria, it is likely that this protein was transferred from archaea to the common ancestor of *Dehalococcoides* followed by a gene duplication event.

Table 6.1b lists 20 additional CSPs, which are specific to archaea but missing in a small number of species. Because these CSPs are present in most Euryarchaeota as well as Crenarchaeota species, but not detected in Bacteria or Eukaryotes except one LGT case (PAB2342, see notes in Table 6.1), we consider them also to be distinctive characteristics of most Archaea. Of these CSPs, 11 proteins (viz. PAB0654, PAB0950, PAB1135, PAB1906, PAB7388, PAB0547, PAB0552, PAB0623, PAB1272, PAB1429 and PAB1721) are mainly missing in the 4 *Thermoplasmata* species. *Thermoplasmata* are thermoacidophilic archaea which lack cell envelope (Futterer et al., 2004; Ruepp et al., 2000; Kawashima et al., 2000). Some studies have suggested that high temperature and very low intracellular pH exert selective pressure favouring smaller genomes (Futterer et al., 2004). Thus, it is possible that genes for these CSPs were selectively lost in the *Thermoplasmata* lineage. Most of these CSPs are of unknown function. However, 8 of them have been assigned putative functions with the title of "archaeal type"'. For example, PAB0301 is archaeal sugar kinase, PAB0950 is archaeal transcription factor E α-subunit, PAB1387 is archaeal flagella accessory protein, PAB7094 is archaeal chromatin protein, and PAB0552 is archaeal type Holliday junction resolvase. These CSPs do not show detectable sequence similarity to their counterparts in Bacteria or Eukaryotes, and some studies indicate that they also differ in terms of their structure, function or interaction with other cell components (Daugherty et al., 2001; Aravind et al., 2000).

6.4.3 Proteins that are specific for Crenarchaeota

As mentioned in the section 1.5, the Archaea are divided into 2 main groups, Crenarchaeota and Euryarchaeota, based on 16S rRNA trees as well many other gene trees and characteristics. In comparison to Euryarchaeota, which contain physiologically and metabolically diverse groups of organisms, the Crenarchaeota were thought to be a pure collection of extreme thermophiles and most members metabolize sulfur. However, recent studies indicate that Crenarchaeota are much more diverse in their physiology and ecology than was previously believed (Burggraf et al., 1997). Many species living in the cold ocean also belong to this group based on their branching pattern in 16S rRNA trees, although most of them have not been cultivated (Knittel et al., 2005). Currently, this phylum comprises one single class *Thermoprotei* containing three orders:

Thermoproteales, Desulfurococcales and *Sulfolo*bales. Fortunately, every order has a completely sequenced representative, which provide a platform to explore the characteristics that are unique to crenarchaeal species (Fitz-Gibbon et al., 2002; Chen et al., 2005; Kawarabayasi et al., 2001; She et al., 2001). Comparative genomic surveys have revealed some molecular features that are shared by Crenarchaea but not Euryarchaea, such as the lack of histones, absence of the FtsZ-MinCDE system and distinctive rRNA operon organization (She et al., 2001). Lake et al. have also identified distinctive differences in ribosome structure and an insert in elongation factor EF-G and EF-Tu, which can be used to distinguish Crenarchaeota from Euryarchaeota (Gupta, 1998b; Lake et al., 1984; Rivera and Lake, 1992). However, these features are not unique characteristics of the Crenarchaeota.

BLAST searches on each ORF from the genomes of *A. pernix* and *S. acidocaldarius* DSM 639 (Kawarabayasi et al., 2001; Chen et al., 2005) have identified 11 CSPs which are shared by all five crenarchaeal species, but whose homologs are not found in other archaea, or any bacteria or eukaryotes with only 3 exceptions (see Table 6.2a). The genes for these CSPs likely evolved in a common ancestor of the Crenarchaeota and they provide potential molecular markers for species from this phylum. Additionally, 22 proteins that are listed in Table 6.2b are only found in *A. pernix* and three *Sulfolobus* genomes. These CSPs suggest that *Aeropyrum* and *Sulfolobus* may have shared a common ancestor exclusive of *Pyrobaculum*. However, we have also come across 9 CSPs that are shared by *Aeropyrum* and *Pyrobaculum* (Table 6.2c) and 14 CSPs that are exclusively present in the 3 *Sulfolobus* species and *Pyrobaculum* (see Table

6.2d). Hence, based upon the species distributions of these proteins, the relationships among the *Aeropyrum, Sulfolobales* and *Pyrobaculum* are not entirely clear (Figure 6.2a). In phylogenetic trees *Thermoproteales* (i.e. *Pyrobaculum*) branches consistently earlier than *Desulfurococcales* (i.e. *Aeropyrum*) and *Sulfolobales* (Figure 6.1) (Matte-Tailliez et al., 2002; Brochier et al., 2005a). This observation in conjunction with the fact that *Aeropyrum* and *Sulfolobus* share larger numbers of proteins in common with each other suggests that these two groups likely shared a common ancestor exclusive of *Pyrobaculum* (Figure 6.2b). The proteins that are only found in *Aeropyrum* and *Pyrobaculum,* or in *Sulfolobus* and *Pyrobaculum*, most likely evolved in a common ancestor of the Crenarchaea, but were subsequently lost in either the *Sulfolobales* or *A. pernix* lineages.

In addition to these proteins that are uniquely present in either all sequenced Crenarchaeota genomes or different groups of Crenarchaeota species, these analyses have also identified 264 proteins that are unique for the *Sulfolobales* species (Gao and Gupta, 2007). Of these, 184 proteins are present in all 3 sequenced *Sulfolobus* genomes, whereas the remaining 80 are present in at least two of the three *Sulfolobus* genomes. In this work, since blast analyses were not carried out on all three *Sulfolobus* genomes, it is likely that the numbers of genes or proteins that are uniquely shared by only two *Sulfolobus* genomes is much higher than indicated here. Chen et al. have previously analyzed the genome of *S. acidocaldarius* DSM 639 and indicated the presence of 107 genes that were specific for Crenarchaeota and 866 genes that were specific to Sulfolobus genus (Chen et al., 2005). However, in the present work, relatively few genes that are uniquely shared by various Crenarchaeota species were identified. This difference could be due to more stringent criteria that we have employed for identification of proteins that are specific to different groups.

6.4.4 CSPs specific for Euryarchaeota

The Euryarchaeota, which comprise a majority of the cultured and sequenced archaea, is a morphologically, metabolically and physiologically diverse collection of species as evidenced by the presence in this group of various methanogens, extreme halophiles, cell wall-less archaea and sulfate reducing microbes (Ludwig and Klenk, 2001; Gribaldo and Brochier-Armanet, 2006). No unique biochemical or molecular characteristic that is commonly shared by all of the different lineages is known. The present study has identified 20 CSPs that are only found in Euryarchaeota species with 3 exceptions (see Table 6.3). In this Table, the first 7 proteins (Table 6.3a) are present in most euryarchaeota species. Of these proteins, PAB0082 and PAB2404 were found in all sequenced euryarchaeota species. PAB2404 was also present in *N. equitans*, supporting its placement within the Euryarchaeota (Makarova and Koonin, 2005; Brochier et al., 2005b). The protein PAB0082 is annotated as archaeosine tRNA-ribosyltransferase (ArcTGT), which catalyzes the exchange of guanine with a free 7-cyano-7-deazaguanine (preQ0) base, as the first step in the biosynthesis of an archaea-specific modified base, archaeosine (7-formamidino-7-deazaguanosine) (Ishitani et al., 2002). It should be mentioned that there is another protein PAB0740 in the same genome, which is also annotated and experimentally confirmed as ArcTGT (Aravind and Koonin, 1999). The latter belongs to a family of proteins that are highly conserved in all archaea species (including Crenarchaeota) and some bacteria. It seems that PAB0082 might be involved in RNA modification since it possesses a PUA domain (named after pseudouridine synthase and archaeosine transglycosylase), but its function is likely different from PAB0740. The protein PAB2404, which is annotated as DNA polymerase II large subunit, is highly conserved within Euryarchaeota, but is not found anywhere else except in *Nanoarchaeum*. This enzyme is the major DNA replicase in Euryarchaeota and also a distinctive molecular marker for this group (Shen et al., 2004; Cann and Ishino, 1999). The genes for the above proteins likely evolved in a common ancestor of Euryarchaeota (Figure 6.2) and they provide molecular markers for this diverse group of organisms.

Another 13 CSPs listed in Table 6.3b are found in almost all euryarchaeota, but they are missing in Thermoplasmata. Their distribution suggests that either Thermoplasmata is a deep branching lineage within Euryarchaeota or that the genes for these proteins have been selectively lost from *Thermoplasmata* (Ruepp et al., 2000). Of these proteins, PAB0188 is also present in *N. equitans* supporting its placement with Euryarchaeota. Five other proteins from the first two columns in Table 6.3 (viz. MMP0243, Ta0062, VNG1263c, MMP1287, and VNG2408c) are also not found in the 4 *Thermococci* species. These results can again be explained by either selective loss of these genes from these particular groups or deeper branching of these lineages within the Euryarchaeota species. On the basis of proteins listed in Table 6.3, although one can infer that *Thermoplasmata* and *Thermococci* are deeper branching lineages within Euryarchaeota in comparison to methanogens, their relative branching order cannot be resolved.

Our comparative genomic studies have identified a large number of CSPs that are specific to the major lineages within Euryarchaeota, such as *Thermococci, Halobacteria, Thermoplasmata* and methanogens (Gao and Gupta, 2007). These CSPs constitute unique characteristic for different Euryarchaeota lineages and some of them should play important role in the adaptation of these organisms into extreme environments. Of particular interest, we have identified a number of proteins either specific to all methanogenic archaea or certain subgroups of methanogens. Details about methanogen-specific proteins are described below.

6.4.5 CSPs specific for methanogenic archaea

Currently, the methanogens form the largest group within the Euryarchaeota. They are distinguished from all other prokaryotes by their ability to obtain all or most of their energy via the reduction of CO_2 to methane or by the process of methanogenesis. In Bergey's manual, the methanogenes are divided into 5 distinct orders (viz. *Methanobacteriales, Methanococcales, Methanomicrobiales, Methanosarcinales* and *Methanopyrales*) (Garrity et al., 2001). Some studies have suggested that these organisms possess a set of unique enzymes which are responsible for methanogenesis, such as coenzyme M, Factor 420 and methanopterin (Reeve et al., 1997). However, no systematic study has been carried out thus far to identify proteins that are uniquely present in different methanogens. Our blast searches of proteins from different methanogens have led to identification of 31 proteins, which are uniquely found in various methanogenic archaea. Twenty of these 31 proteins are present in all sequenced methanogens, while 11 proteins are missing only in *M. stadtmanae*, which is a human intestinal inhabitant (see notes in Table 6.4). This archaeon generate methane by reduction of methanol with H_2 and lacks many proteins present in the genomes of other methanogens (Vandewijngaard et al., 1991; Fricke et al., 2006). Thus, it is highly likely that the 11 proteins missing in *M. stadtmanae* were selectively lost from this species. Therefore, it is very likely that the genes for these 31 proteins that are commonly shared by virtually all methanogens (Table 6.4a) evolved in a common ancestor of all methanogens.

These analyses have also identified 10 proteins that are uniquely shared by various methanogens as well as *A. fulgidus* (see Table 6.4b). The genes for these proteins likely evolved in a common ancestor of *A. fulgidus* and various methanogenic archaea and they point to a close relationship between these two groups of organisms (Figure 6.3). Ten additional proteins are present in *A. fulgidus* as well as various *Methanosarcinales* and *M. hungatei* (Methanomicrobiales) (Table 6.4c). It is likely that the genes for these proteins also evolved in a common ancestor of *A. fulgidus* and various methanogenic archaea, but they were selectively lost in other methanogens. Of the proteins that are commonly shared by *A. fulgidus* and various methanogenic archaea, MMP0607 is reported to be a novel repressor of nif and glnA genes, which are involved in nitrogen assimilation (Lie and Leigh, 2003). Interestingly, 2 homologs of this protein are also found in 3 *Dehalococcoides* species, but nowhere else, which are very likely due to LGT. Protein MMP0984 is the ε-subunit of carbon-monoxide dehydrogenase complex, which is made up of five subunits in different methanogens (Murakami and Ragsdale, 2000). The epsilon subunits are required for the reversible oxidation of CO to CO_2 [81]. All of the other components could be found in a few bacterial species, while the ε-subunit is restricted to methanogenic archaea and *A. fulgidus* (Lindahl and Chang, 2001; Klenk et al., 1997). Protein MMP1499 is identified as a transcriptional regulator with a Helix-turn-helix (HTH) motif, but its exact role has not been reported.

Among the genes that are uniquely shared by various methanogenic archaea (or these archaea plus *A. fulgidus*), two large gene clusters responsible for methanogenesis are found. The proteins MMP1346, MMP1560–MMP1564 and MMP1566–MMP1567 (Table 6.4) are parts of an eight-component complex, coenzyme M methyltransferase (Mtr), which catalyzes an energy-conserving, sodium-ion-translocating step in methanogenesis from H_2 and CO_2 (Harms et al., 1995). *M. maripaludis* contains all of the known Mtr subunits, but the gene coding for MtrF is fused into the N-terminal region of MtrA (Hendrickson et al., 2004). All other methanogenic archaeal genomes contain complete set of *mtr* genes. It is of interest to note that for the protein MMP1567 (MtrH), homologues with low E-values are also found in two *Desulfitobacterium hafniense* strains as well as in three *Rhizobiales* species (*Aminobacter lissarensis, Methylobacterium chloromethanicum,* and *Hyphomicrobium chloromethanicum*; α-proteobacteria) (see note in Table 6.4). These three rhizobiae species can use methyl halides as a sole source of carbon and energy, and all of them possess a set of *cmu* genes which are essential for methyl chloride degradation (Warner et al., 2005). In particular, the CmuB protein which is homologous to MMP1567 transfers a methyl group to methylcobalamin:H_4 folate (H4F), which is analogous to the reverse of the reaction catalyzed by MtrH in archaea (McAnulla et al., 2001). In view of the sequence and functional similarity between MtrH and CmuB proteins, it is likely that the mtrH gene was laterally transferred from a methanogenic archaeon to the common ancestor of the above three rhizobiae species to serve the new functional role. The function of the laterally transferred mtrH related gene in *D. hafniense* is not known at present.

The proteins MMP1555–MMP1559 in Table 6.4 form another gene cluster, encoding the subunits of Methyl-coenzyme M reductase (MCR). This complex catalyzes the final reaction of the energy conserving pathway in which methylcoenzyme M and coenzyme B are converted to methane and the heterodisulfide CoM-S-S-CoB (Grabarse et al., 2001; Ermler et al., 1997). Except for these proteins, the other proteins listed in Table 6.4 are of putative or unknown functions. It is likely that these proteins are involved in some aspects of methanogenesis or other unknown pathways unique to methanogenic archaea. These proteins provide molecular markers for methanogens, which can be used for identification of new archaeal species capable of methane production.

The BLAST searches of the *M. maripaludis* and *M. kandleri* genomes have identified 10 proteins that are uniquely shared by all of the following species belonging to the orders *Methanobacteriales* (*M. thermoautotrophicus*), Methanococcales (*M. jannaschii*, *M. maripaludis*) and Methanopyrales (*M. kandleri*) (Gao and Gupta, 2007). Of these, only 2 proteins are present in *M. stadtmanae*, which is also a *Methanobacteriales* that has lost most of its genes due to its adaptation to the human intestine (Fricke et al., 2006). The genes for these 10 proteins likely evolved in a common ancestor of the above groups of methanogens (Figure 6.3), which corresponds to the cluster of methanogenic archaea referred to as "Class I methanogens" (Gribaldo and

Brochier-Armanet, 2006). Interestingly, these studies have also identified 10 proteins that are uniquely shared by these methanogenic orders and *M. hungatei*, which branches distantly in phylogenetic trees (Gribaldo and Brochier-Armanet, 2006). The unique presence of these proteins in these methanogens suggests that species from these groups shared a common ancestor exclusive of other methanogenic archaea (Figure 6.3).

Fifteen additional proteins discovered in this work are uniquely present in *M. kandleri* and various *Methanobacteriales* indicating that these two groups are more closely related to each other than the Methanococcales (Figure 6.3) (Gao and Gupta, 2007). We have also come across 7 proteins that are uniquely shared by *Methanococcales* and *Methanobacteriales*, and 4 proteins that are only present in *Methanococcales* and *Methanopyrales*. The most likely explanation to account for the species distributions of these latter proteins is that their genes also originated in a common ancestor of the above three groups of methanogens, but were selectively lost in either the *Methanobacteriales* or *Methanopyrales* lineages. These analyses have also identified 14 additional proteins that are uniquely present in all 5 *Methanosarcinales* species, as well as 7 proteins that are only found in various *Methanosarcinales* and *M. hungatei*. Lastly, these studies have also identified 55 proteins that are uniquely present in *M. maripaludis* and *M. jannaschii* and 68 proteins that are only present in *M. burtonii* and 3 *Methanosarcina* species, all belonging to the *Methanosarcinaceae* family (Figure 6.3) indicating that they are likely distinctive characteristics of species from these groups.

Of the proteins that are uniquely found in *Methanococcales, Methanobacteriales, Methanopyrales* and *Methanomicrobiales*, 12 proteins viz. MMP1448–MMP1454, MMP1456, MMP1458–MMP1460 and MMP1467 are from a big gene cluster eha, which encodes the multisubunit membrane-bound [Ni-Fe] hydrogenase (Tersteegen and Hedderich, 1999). Two of these proteins, MMP1456 and MMP1458, are only found in *Methanococcales* (Gao and Gupta, 2007). The whole *eha* operon is composed of 20 ORFs in the genome of *M. thermoautotrophicus* and of these only these 12 proteins are restricted to these methanogens while the other subunits have counterparts in bacteria. The precise roles of these 12 proteins, which are predicted to be integral membrane proteins in the hydrogenase complex, have not been determined (Tersteegen and Hedderich, 1999). Among the other proteins that are specific for these groups of methanogens, MMP0127 and MMP1716 are Hmd homologs, which catalyze the reversible dehydrogenation of N5, N10-methylenetetrahydromethanopterin (Hartmann et al., 1996). In the proteins that are specific for the *Methanococcales*, one large gene cluster (MMP0233–MMP0240) is found, but no information is available concerning its possible function. Except for these proteins, all other proteins that are specific for these methanogenic archaea are of unknown or putative function.

6.4.6 Proteins restricted to several archaeal lineages or showing sporadic distribution

In addition to the above proteins that are restricted to specific lineages of archaea, we have also identified 63 proteins, which are shared by several archaeal groups (see Table 6.5). The distribution pattern of these proteins could provide useful insights concerning the phylogenetic relationship between different groups. However, their distribution patterns could also be explained by gene losses in specific lineages or LGT between particular groups. Table 6.5 shows many proteins that are uniquely shared by various methanogenic archaea, *Archaeoglobus* and *Thermococci*. The first 5 proteins in Table 6.5a (PAB0076, PAB0138, PAB0965, PAB1927 and PAB1994) are present in all of the *Thermococci* and most of the methanogens. Four of these proteins are also present in *A. fulgidus*. The next 13 proteins in this Table are also uniquely found in most of the *Thermococci* as well as a number of methanogens and also in many cases in *A. fulgidus*. In addition, 6 proteins listed in Table 6.5b are only found in various *Thermococci* and *A. fulgidus*. These results suggest a closer relationship between the methanogenic archaea, *A. fulgidus* and *Thermococci* within the Euryarchaeota lineage. In conjunction with our earlier inference that *A. fulgidus* forms an outgroup of the methanogenic archaea, these results suggest that the above three groups are related in the following manner: *Thermococci* → *A. fulgidus* → Methanogens.

Although the relationship suggested above is the most likely explanation for the observed results, we have also come across three proteins (VNG1263c, MMP11287 and VNG2408c) that are uniquely present in various *Halobacteria*, *A. fulgidus* and different methanogens. To account for their species distribution, one has to postulate that their genes have been selectively lost from the *Thermococci*. In addition, 9 proteins are only found in various *Halobacteria* as well as *Methanosarcinales* and *Methanomicrobiales* (Table 6.5c). Their distribution requires again either selective gene losses from other lineages or LGT from *Halobacteria* to these methanogens.

Our analyses have also uncovered 30 proteins that are uniquely shared by species from *Thermoplasmata* and *Sulfolobus* (see Table 6.5). Among these proteins, 7 are present in all *Thermoplasmata* and *Sulfolobus* species for which sequence information is available, while the remainder are missing in 1 or more species. It has been reported that there has been much lateral gene transfer between *T. acidophilum* and *S. solfataricus*, both of which inhabit the same environment (Ruepp et al., 2000). However, the shared presence of these proteins in these two groups could also result from a unique shared ancestry of these thermo-acidophilic archaea.

Another 43 Archaea-specific proteins are sporadically present in different archaeal species (data not shown)(Gao and Gupta, 2007). A number of proteins in this group are present in a limited number (between 3 to 6) of archaeal species belonging to different groups. There are 2 possible explanations that can account for their sporadic distribution: First, it is possible that some of these genes are the remnants of sequences that also originated in an ancestral lineage of Archaea but they have been selectively lost in many species because they are not required for growth. Second, the sporadic presence of these genes in a number of archaeal species can also be explained if some of these genes originally evolved in a particular group or species of archaea and then transferred to other archaea by LGT (Bapteste et al., 2005). However, in view of the observed specificity of these genes/proteins for archaea, the LGTs in these cases need to be selective and limited to within archaea.

6.5 Discussion

Comparative analyses of sequenced archaeal genomes presented here have led to identification of large numbers of proteins that are distinctive characteristics of either all archaea or its different main groups. Based upon these proteins, all of the main groups within Archaea (e.g. Crenarchaeota, Euryarchaeota, *Halobacteria, Thermococci*, *Thermoplasmata,* Methanogens) and their subgroups can now be clearly distinguished in molecular terms. The species distribution of these signature proteins strongly suggests that their genes have evolved or originated at various stages in the evolution of archaea, but once evolved, they are indicated to be generally stably retained in various descendents of these lineages with minimal gene loss or LGTs. Based upon the species distributions of these proteins, the evolutionary stages where the genes for these proteins have likely evolved are shown in Figure 6.4. The evolutionary relationships among archaea have thus far been mainly inferred on the basis of their branching in phylogenetic trees based on 16S rRNA and certain protein sequences (Ludwig and Klenk, 2001; Olsen et al., 1994; Gribaldo and Brochier-Armanet, 2006; Olsen and Woese, 1997; Brown and Doolittle, 1997; Brendel et al., 1997). The results of our analyses although they support many inferences reached based on phylogenetic trees (viz. identification of all of the main clades in phylogenetic trees in molecular terms) (Figure 6.1), they also differ from them in important regards. In particular, our results shed important light on certain phylogenetic relationships that were very puzzling or were not resolved based on earlier studies. Some of these novel inferences are discussed below.

In phylogenetic trees based on 16S rRNA and various proteins sequences, the methanogenic archaea form at least two distinct clusters (see Figure 6.1) (Gribaldo and Brochier-Armanet, 2006; Bapteste et al., 2005; Brochier et al., 2004; Slesarev et al., 2002). In addition, in many of these trees, *M. kandleri* branches distinctly from all other methanogenic archaea (Gribaldo and Brochier-Armanet, 2006; Brochier et al., 2004; Rivera and Lake, 1996). The methanogenic archaea in these trees are interspersed by other groups of non-methanogenic archaea such as *Halobacteriales, Archaeoglobus, Thermoplasmatales* and *Thermococcales* (see Figure 6.1) (Gribaldo and Brochier-Armanet, 2006; Brochier et al., 2004; Rivera and Lake, 1996). This has led to important questions concerning the origin of methanogenesis i.e. whether it evolved only once and its absence in the intervening lineages (Gribaldo and Brochier-Armanet, 2006; Bapteste et al., 2005; Makarova and Koonin, 2005; Reeve et al., 1997). To account for these results, it has been suggested that methanogenesis evolved once in a common ancestor of the above groups, i.e. different methanogenic archaea, *Halobacteriales, Archaeoglobus, Thermoplasmatales* and also possibly *Thermococcales*, comprising virtually all euryarchaeota, but that the various genes involved in this process were subsequently lost from different groups except the methanogens (Gribaldo and Brochier-Armanet, 2006; Bapteste et al., 2005; Slesarev et al., 2002). This scenario, in essence, proposes that the common ancestor of different physiologically and metabolically distinct groups within euryarchaeota was a methanogen and this capability was independently lost in all other lineages.

In contrast to this proposal, our phylogenomics analyses have identified 31 proteins that are uniquely present in virtually all methanogens, as well as many proteins that are specifically shared by different subgroups of methanogens. Of these proteins only about 1/3 are indicated to be directly involved in methanogenesis and the cellular functions of others are presently not known. The unique presence of such large numbers of proteins by nearly all methanogens, but none of the above groups of archaea, strongly indicates that the genes for these proteins evolved in a common ancestor of various methanogens. These results strongly suggest that all methanogenic archaea form a mononphyletic lineage exclusive of all other groups of archaea (Figure 6.4). Importantly, these studies have also identified 10 proteins that are uniquely shared by all methanogens as well as by *A. fulgidus*. In contrast, we have not come across any protein that various methanogenic archaea uniquely share with any of the *Halobacterales* or *Thermoplasmatales*. These observations are highly significant because they strongly suggest that Archaeoglobus and all of the methanogens shared a common ancestor exclusive of all other archaea. In other words, the ancestral lineage that led to the origin of methanogenesis very likely evolved from the *Archaeoglobus* lineage (Figure 6.4). It is also significant that of the proteins that are uniquely shared by *Archaeoglobus* and methanogens, several form part of complexes that are important for nitrogen assimilation and methanogenesis. These results support the view that these characteristics have their origin within the Archaeoglobus lineage.

The present work also provides clarification regarding the phylogenetic position of *M. kandleri*. In phylogenetic trees based on 16S rRNA or different protein sequences, the branching of this species is highly variable and it often forms the deepest branch within the Euryarchaeota (Gribaldo and Brochier-Armanet, 2006; Brochier et al., 2004; Burggraf et al., 1991; Rivera and Lake, 1996). In the present work, we have identified 31 proteins that are uniquely shared by all methanogens including *M. kandleri*, as well as 10 proteins that *M. kandleri* specifically shares with various *Methanobacteriales* and *Methanococcales*, and 15 additional proteins that are only found in *M. kandleri* and the two *Methanobacteriales* species (*M. thermoautotrophicus* and *M. stadtmanae*). These observations reliably place *M. kandleri* with other methanogenic archaea with the *Methanobacteriales* as its closest relatives (Figure 6.4). Our results also suggest a closer relationship of the *Thermococcales* to the *Archaeoglobus* and methanogenic archaea, although this relationship is not as strongly supported as between *Archaeoglobus* and Methanogens.

The observed differences in the evolutionary relationships among methanogens based upon phylogenomics analyses versus those by traditional phylogenetic methods can in principle be accounted for by three explanations. First, it is possible that the branching patterns of various clades in phylogenetic trees are misleading and they have been affected by factors such as long branch attraction effect (Philippe et al., 2005b; Felsenstein, 1981). Second, the polyphyletic branching of methanogens can also be explained (as indicated earlier) if the genes uniquely shared by all methanogens evolved in an early branching lineage such as *M. kandleri*, but subsequently they were either completely or partially lost from various non-methanogenic (viz. *Halobacteriales*, *Thermoplasmatales* and *Archaeoglobus*) groups that lie in between the two methanogenic clusters (Figure 6.1). Third, lateral transfer of these genes from one methanogenic archaea to all others can also explain these results. Of these possibilities, we favour the first explanation, as the last two require extensive gene loss or LGT from (or into) multiple independent lineages.

The present work also supports the placement of *N. equitans* within the Euryarchaeota lineage. *N. equitans* has a very small genome (only 0.49 Mb), which is at least 3 times smaller than any other archaeal genome. Due to its very small size, there are only 6 genes that *N. equitans* uniquely shares with all other archaea. However, our analysis indicates that whereas *N. equitans* shares a few genes (PAB2404 and PAB 0188) with most of the Euryarchaeota, it does not share any gene uniquely with most of the Crenarchaeota species, indicating its closer affinity for the former lineage. Although our analysis of the *N. equitans* genome has not revealed any strong signals indicating its specific affinity for any of the Euryarchaeota groups, the shared presence of some proteins by *N. equitans* and *Thermococci* (and in some cases also *A. fulgidus* and methanogens) suggest that it may be related to the *Thermococci*. However, because of the extensive gene losses that have occurred in this genome, we are not able to draw any reliable inference in this regard. Therefore, although we have depicted *N. equitans* as a deep branching lineage within Euryarchaeota (Figure 6.4), based upon our analysis, its placement within Euryarchaeota is not resolved.

The present work also suggests that Thermoplasmatales might be a deeper branching lineage within Euryarchaeota in comparison to the *Thermococcales, Halobacteriales, Archaoglobous* and Methanogens. This inference is suggested by the observation that a number of proteins that are uniquely present in almost all other Euryarcheota species are missing in the *Thermoplasmatales*. Although the absence of these proteins in the *Thermoplasmatales* can be explained by specific gene loss, the possibility that the genes for at least some of these proteins have evolved after the branching of *Thermoplasmatales* deserves serious consideration. The deeper branching of the *Thermoplasmatales* within the Euryarchaeota will place it closer to the

Crenarchaeota. Such a placement could prove helpful in understanding why so many genes (i.e. 30) are uniquely shared by various *Thermoplasmatales* and the *Sulfolobales*.

For the archaea-specific proteins identified in the present work, sequence information at present is available from only a limited number of archaeal species. Hence, it is important to obtain information for these genes/proteins from other archaeal species to confirm whether these proteins are distinctive characteristics of the specified groups or a subgroup of such species. These proteins in addition to their utility for phylogenetic and taxonomic studies also provide valuable means for understanding archaeal biology (Makarova and Koonin, 2005; Galperin and Koonin, 2004). The cellular functions of most of these proteins are not known and further studies in this regard should prove very helpful in the discovery of novel biochemical and physiological characteristics that are unique to either all or different groups of archaea (Galperin and Koonin, 2004). Lastly, the primary sequences of many of these genes/proteins are also highly conserved and they provide novel means for identification of different groups of archaea in various environmental settings by means of PCR amplification and other molecular biological and immunological methods.

6.6 Figures 1-4 and Tables 1-5

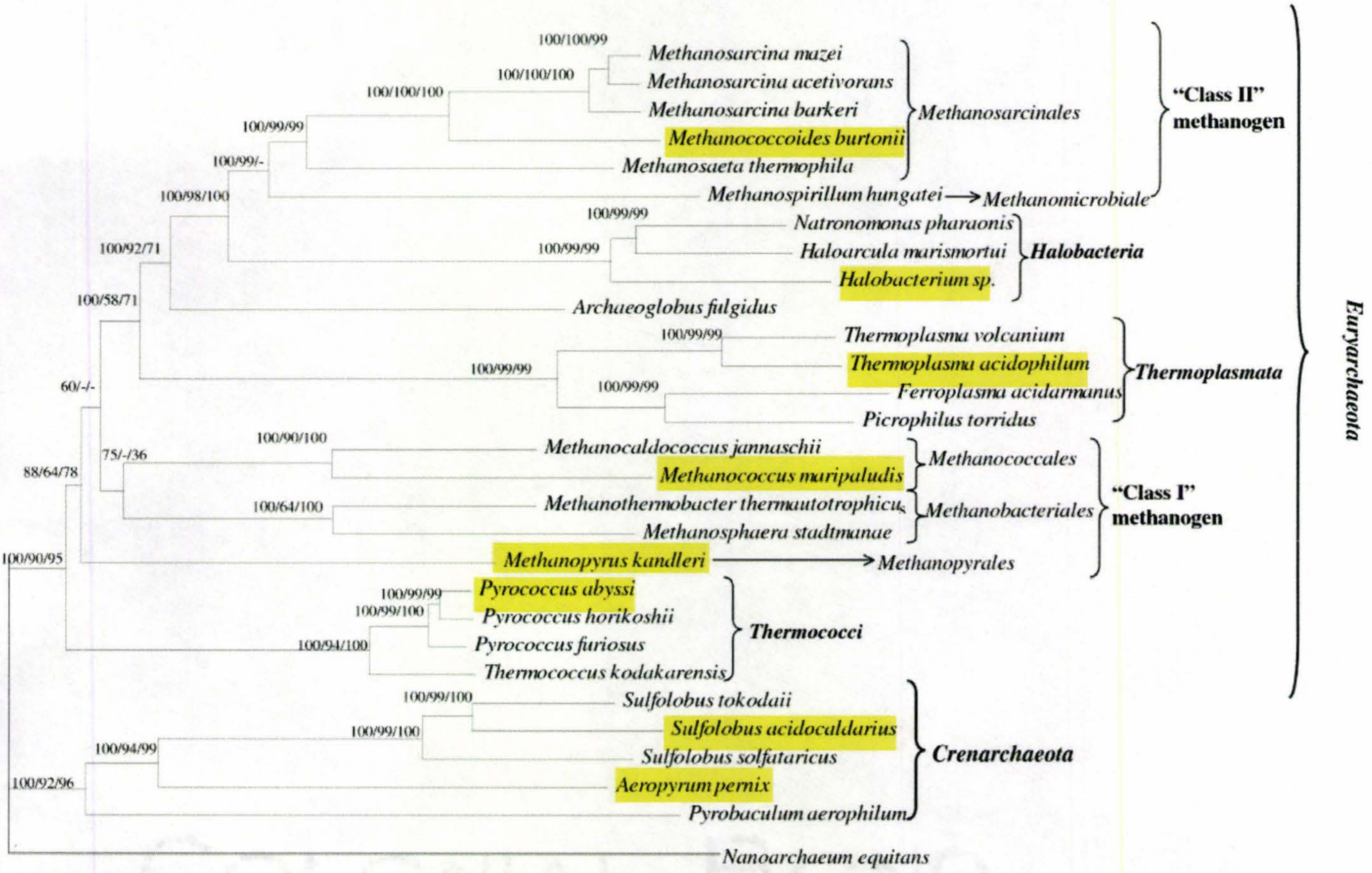


Figure 6.1. A neighbour-joining distance tree based on a concatenated sequence alignment for 31 widely distributed proteins. The numbers on the nodes indicate bootstrap scores observed in NJ/ML/MP analyses. The species shaded in yellow were selected as the query genomes for blast searches.

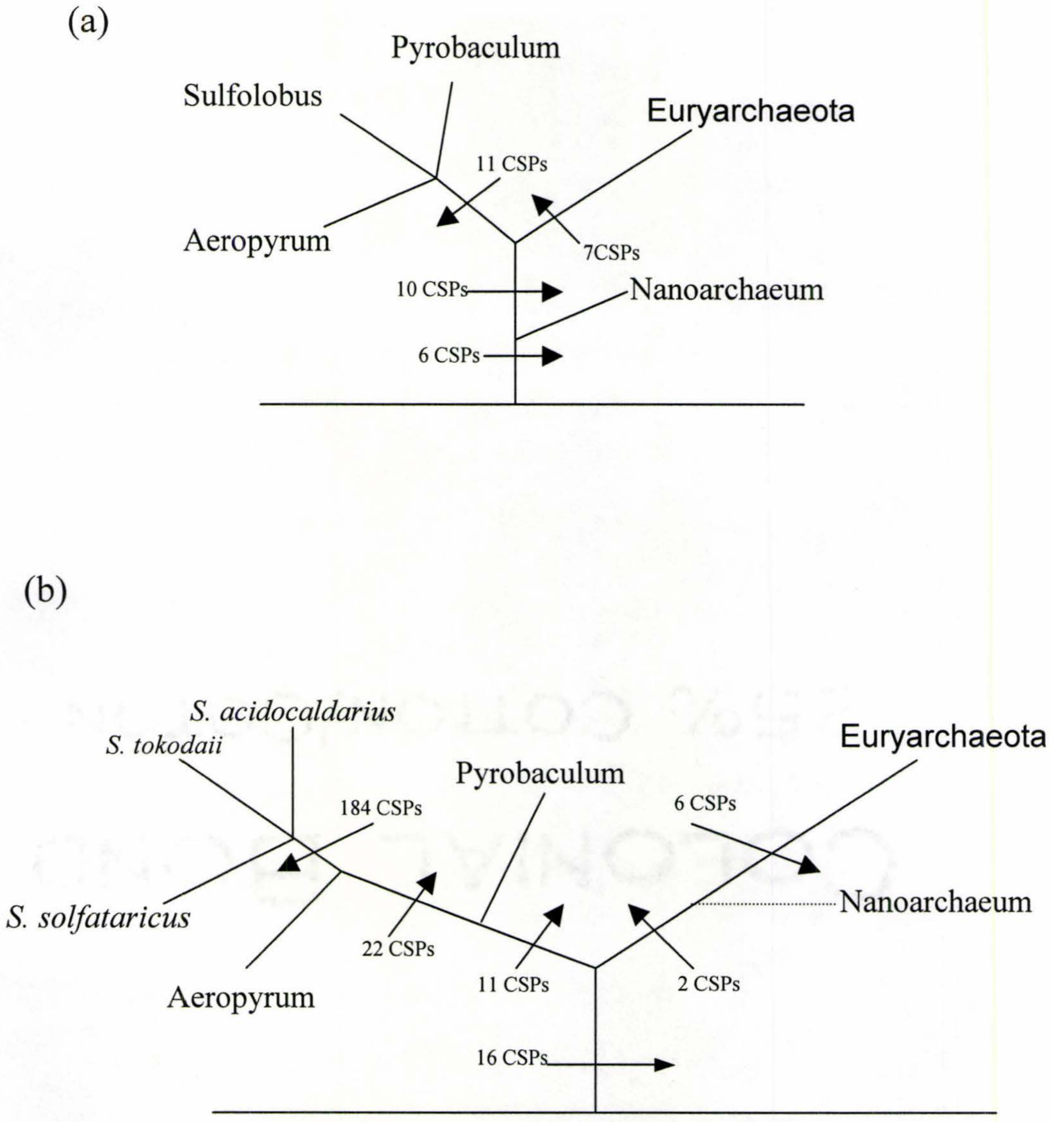


Figure 6.2. Interpretive diagrams showing the suggested evolutionary stages where genes for the CSPs that are specific for the Crenarchaeota and Euryarchaeota as well as some of the Crenarchaeota subgroups, likely originated. The top diagram (A) indicates the evolutionary interpretation of the CSPs in the absence of any other information, whereas that below (B) indicates our interpretation of this data taking into consideration other relevant information discussed in the text. The branching pattern shown here is unrooted and the CSPs that are shared by all archaea were introduced in a common ancestor of all archaea. The dotted line for *N. equitans* in (B) indicates that its placement within Euryarchaeota lineage is uncertain. Edited from (Gao and Gupta, 2007).

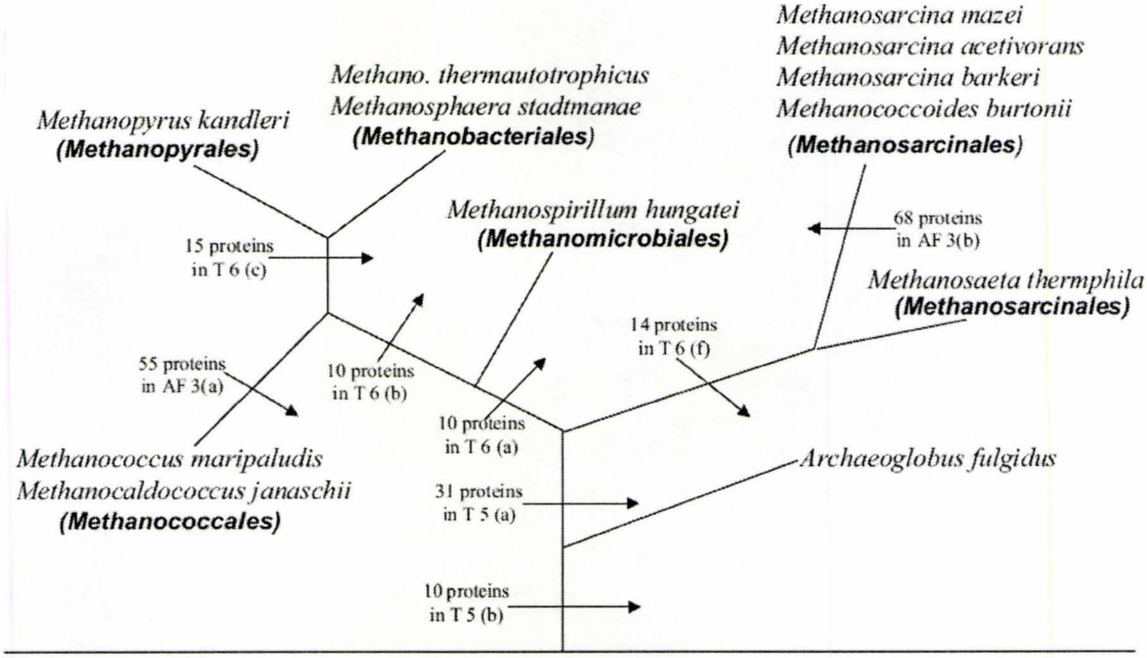


Figure 6.3. An interpretive diagram showing the evolutionary stages where genes for different CSPs that are specific for methanogenic archaea likely originated. The 10 CSPs that are uniquely shared by *A. fulgidus* and various methanogenic archaea indicate that this lineage is the closest ancestor of all methanogens.

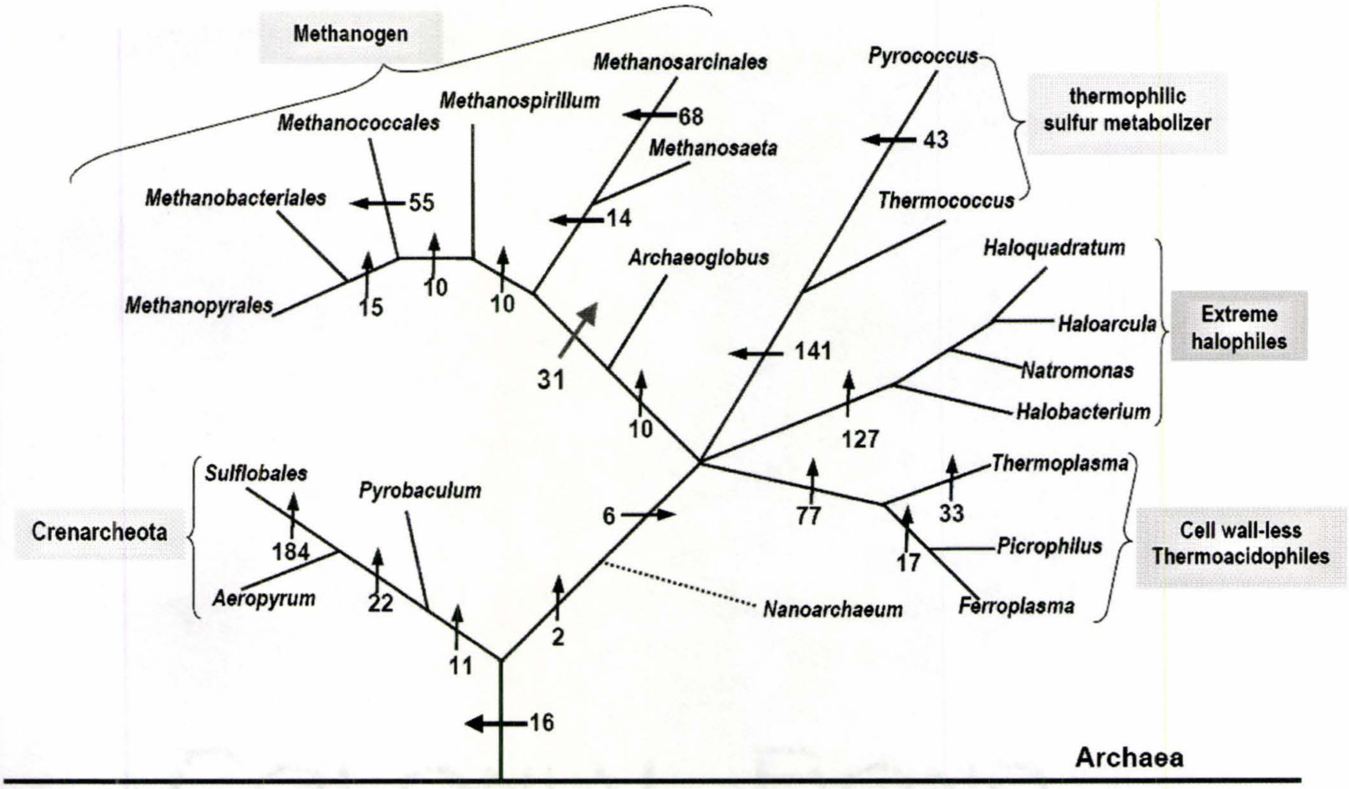


Figure 6.4. A summary diagram showing the species distribution patterns of various Archaea-specific CSPs. The arrows mark the suggested evolutionary stages where proteins that are uniquely shared by the indicated groups were introduced. The branching pattern shown here is based upon the species distribution patterns of these proteins and it is unrooted. The dotted line for Nanoarchaeum indicates that its placement within Euryarchaeota is uncertain. Modified from (Gao and Gupta, 2007).

Table 6.1. CSPs specific to all Archaea

(a) CSPs specific to all Archaea				(b) Archaea-specific CSPs with gene loss in few species			
PAB0063	[NP_125796]	Cca	COG1746	PAB0301	[NP_126142]	SK	COG1685
PAB0247	[NP_126062]	DNA binding	COG1571	PAB0654	[NP_126650]		CDD8168
PAB0252	[NP_126069]	RNA-binding	CDD16214	PAB0950	[NP_127106]	TFIIE	CDD480
PAB0316	[NP_126166]	DNA primase	COG0358	PAB1112	[NP_127373]		CDD5727
PAB1716	[NP_126666]	NMD3	CDD16276	PAB1135	[NP_127406]		CDD8168
PAB2291	[NP_125771]		CDD6629	PAB1241	[NP_127355]		CDD9682
PAB0018a	[NP_125721]	RNA binding	COG2888	PAB1387	[NP_127161]	flaJ	COG1955
PAB0075[1]	[NP_125817]	dehydratase	CDD23288	PAB1715	[NP_126667]		CDD9801
PAB0439	[NP_126328]		COG1308	PAB1906	[NP_126377]		CDD2531
PAB0475	[NP_126376]	regulator	COG1709	PAB7094	[NP_126085]	Alba	CDD25844
PAB1040	[NP_127251]	SpoU	CDD6631	PAB7388	[NP_127197]	Ribosomal_LX	CDD2437
PAB1106	[NP_127361]		CDD9578	PAB0469.1n	[NP_877631]		CDD8674
PAB1706	[NP_126677]		COG1634	PAB0547	[NP_126484]		COG1759
PAB2062	[NP_126118]		CDD16190	PAB0552	[NP_126501]	Hjr	CDD29957
PAB2104	[NP_126058]	HTH	COG1395	PAB0623	[NP_126611]		CDD9586
				PAB1272	[NP_127310]		COG1759
				PAB1429	[NP_127105]		COG2433
				PAB1721	[NP_126657]		COG2248
				PAB2342[2]	[NP_125707]		CDD15774
				PAB7309	[NP_126897]		CDD2523

These CSPs were identified by BLASTP searches and their specificity is further confirmed by PSI-BLAST searches. For details, see method section. The protein ID number starting with PAB represents query protein from the genome of *P. abyssi* GE5, which was used as probe to perform the blast search. Accession numbers for these proteins are shown in square brackets. The possible cellular functions and COG or CDD number of some proteins are noted. For other proteins, the cellular functions are not known.

Note [1]. Two low-scoring homologs to PAB0075 were also found in *Dehalococcoides ethenogenes* 195 (*Chloroflexi*) and *Dehalococcoides* sp. CBDB1.

Note [2]. A homolog to PAB2342 is also found in *Oenococcus oeni* PSU-1, *Leuconostoc mesenteroides* subsp. mesenteroides ATCC 8293 and *Clostridium perfringens* str. 13.

Table 6.2. CSPs specific to Crenarchaeota

(a) CSPs specific to Crenarchaeota	(c) CSPs specific to *Sulfolobus* and *Pyrobaculum*	(d) CSPs specific to *Aeropyrum* and *Sulfolobus*
APES019 [NP_147243] ribonuclease p3	Saci_0004 [YP_254727]	APE0143 [NP_146996] COG5491
APE0488 [NP_147273] COG4914	Saci_0005 [YP_254728]	APE0145 [NP_146997]
APE0503 [NP_147284] COG4755	Saci_0035 [YP_254758]	APE0168 [NP_147017]
APE0505[1] [NP_147285] CDD26165	Saci_0223 [YP_254935] PaREP8	APE0238 [NP_147072]
APE0623 [NP_147373] COG4888	CDD46009	APE0429 [NP_147222]
APE0975 [NP_147640] COG4879	Saci_0224 [YP_254936] = Saci_0223	APE0663 [NP_147399] COG5431
APE1561 [NP_148025] COG4900	Saci_0660 [YP_255337]	APE0902 [NP_147588]
APE1627[2] [NP_148064] CDD26669	Saci_0857 [YP_255517]	APE1113 [NP_147720]
APE1644 [BAA80645]	Saci_1129 [YP_255774]	APE1364 [NP_147897]
	Saci_1813 [YP_256412] COG4113	APE1626 [NP_148063]
(b) CSPs specific to *Aeropyrum* and *Pyrobaculum*	Saci_1883 [YP_256481] = Saci_1813	APE1817 [NP_148186] COG5399
	Saci_2070 [YP_256657]	APE1848 [NP_148210] COG1259
APE0106 [NP_146969]	Saci_2080 [YP_256667] = Saci_1813	APE1936 [BAA80945]
APE0730 [NP_147451]	Saci_2195 [YP_256774] = Saci_0223	APE1966 [NP_148294]
APE0874 [NP_147564]	Saci_2357 [YP_256931] = Saci_0223	APE1996 [NP_148313]
APE1194 [NP_147776] HTH_luxr COG5625		APE2102 [NP_148384]
APE1228 [NP_147804]		APE2195 [NP_148451] COG2083
APE1230 [NP_147806]		APE2325 [NP_148539]
APE1236 [NP_147812]		APE2340 [NP_148552]
APE2409 [NP_148589]		APE2435 [NP_148607] COG4920
APE2602 [NP_148718]		APE2454 [BAA81469]
		APE2463 [NP_148628]

The protein ID number starting with APE and Saci represents query protein from the genome of *A. pernix* K1 and *S. acidocaldarius* DSM 639. "=" means paralogous genes.

Note [1]. A low scoring homolog to APE0505 is also found in *Ferroplasma acidarmanus* Fer1.
Note [2]. A low scoring homolog to APE1627 is also found in *Aquifex aeolicus* VF5.

Table 6.3. CSPs specific to Euryarchaeota

(a) CSPs specific to almost all Euryarchaeota	(b) CSPs specific to Euryarchaeota except Thermoplasmata	(c) CSPs specific to Euryarchaeota except Thermoplasmata and Halobacteria
PAB0082 [NP_125825] Tgt COG1549	PAB0161 [NP_125931] COG1326	PAB0076 [NP_125818] CDD15620
MMP0243* [NP_987363] CDD9595	PAB0172 [NP_125944] ATPase COG2117	PAB0138 [NP_125896] CDD9576
PAB1089 [NP_127334] COG2150	PAB0188[1] [NP_125970] CDD8172	PAB0965 [NP_127127] CDD15705
PAB2404[1] [NP_125813] Pol II COG1933	PAB0951 [NP_127107] COG4044	PAB1927[3] [NP_126347] CDD29323
PAB2435 [NP_126297] CDD25834	PAB1055 [NP_127280] COG4743	PAB1994 [NP_126245] CDD9568
PAB0315 [NP_126165] CDD29150	PAB1284 [NP_127297] RecJ COG1107	
Ta0062* [NP_393541] CDD26662	MMP1287* [NP_988407] CDD2419	
	PAB1338 [NP_127222] CDD9842	
	PAB1517 [NP_126975] COG1356	
	PAB1804 [NP_126517] CDD15772	
	PAB2224 [NP_125887] CDD5728	
	VNG1263c* [AAG19620] CDD2419	
	VNG2408c* [AAG20496] COG3365	

The protein ID number starting with MMP, Ta and VNG represents query protein from the genome of *M. maripaludis* S2, *T. acidophilum* and *Halobacterium* sp. NRC-1. * means protein is missing in the genomes of 4 *Thermococci* species.

Note [1]. Homologs to PAB2404 and PAB0188 are also found in in *Nanoarchaeum equitans* Kin4-M.
Note [2]**.** Homolog to PAB1055 is also found in *Dehalococcoides sp.* CBDB1 and *D. ethenogenes* 195.
Note [3]**.** Homolog to PAB1927 is also found in *Rubrobacter xylanophilus* DSM 9941.

Table 6.4. CSPs specific to Methanoarchaeota

(a) CSPs specific to methanogen+*A. fulgidus*				(b) CSPs specific to Methanoarchaeota						
MMP0372	[NP_987492]	MTD	CDD2518	MMP0001	[NP_987121]	COG4014	MMP1346	[NP_988466]	MtrX	COG4002
MMP0400[1]	[NP_987520]		COG1707	MMP0021[5]	[NP_987141]	COG4079	MMP1555	[NP_988675]	MCR_B	CDD25889
MMP0499[5]	[NP_987619]	ArsR	CDD28947	MMP0143	[NP_987263]	COG4069	MMP1556	[NP_988676]	MCR_D	CDD3015
MMP0607[2]	[NP_987727]	NrpR	COG1693	MMP0154	[NP_987274]	COG4070	MMP1557	[NP_988677]	MCR_C	CDD15906
MMP0961[5]	[NP_988081]		CDD15263	MMP0311[5]	[NP_987431]	COG4048	MMP1558	[NP_988678]	MCR_G	CDD29638
MMP0962	[NP_988082]		COG4855	MMP0312	[NP_987432]	COG4050	MMP1559	[NP_988679]	MCR_A	CDD8362
MMP0976[5]	[NP_988096]		COG1810	MMP0337	[NP_987457]	COG4029	MMP1560	[NP_988680]	MtrE	CDD9765
MMP0984[5]	[NP_988104]	CO_dh	CDD3060	MMP0421	[NP_987541]	COG4052	MMP1561	[NP_988681]	MtrD	CDD9766
MMP1499[5]	[NP_988619]	HTH	COG4800	MMP0563[5]	[NP_987683]	COG4090	MMP1562	[NP_988682]	MtrC	CDD17461
MMP1567[3]	[NP_988687]	MtrH	CDD25859	MMP0642	[NP_987762]	COG4020	MMP1563	[NP_988683]	MtrB	CDD23666
				MMP0656	[NP_987776]	COG4051	MMP1564[4]	[NP_988684]	MtrA	COG4063
				MMP0665	[NP_987785]	COG4066	MMP1566	[NP_988686]	MtrG	CDD9769
				MMP0698[5]	[NP_987818]	COG4033	MMP1593	[NP_988713]		COG1571
				MMP0701[5]	[NP_987821]	COG4081	MMP1644	[NP_988764]		COG4022
				MMP1223	[NP_988343]	COG4065	MMP1704	[NP_988824]		COG4008
				MMP1309[5]	[NP_988429]	COG4073				

Note [1]. A homolog to MMP0400 is found in *Solibacter usitatus* Ellin6076 and *Rubrobacter xylanophilus* DSM 9941;

Note [2]. A homolog to MMP0607 is found in *Dehalococcoides sp.* CBDB1 and *D. ethenogenes* 195;

Note [3]. A homolog to MMP1567 is found in 2 *Desulfitobacterium hafniense* strains (*Firmicutes*), and the CmuB protein from 3 sepecies belonging to *Rhizobiales* of α-proteobacteria also show great similarity with MtrH;

Note [4]. A homolog to MMP1564 is also found in *Dechloromonas aromatica* RCB;

Note [5]. These 10 proteins are absent in the genome of *Methanosphaera stadtmanae* DSM 3091.

Table 6.5. CSPs restricted to several archaeal lineages

(a) CSPs only found in Thermococci, *Archaeoglobus* and Class I methanogens		(b) CSPs mainly shared by Halobacteria and Class II methanogens
PAB0036 [NP_125764]	PAB1251 [NP_127332] endonuclease COG3780	VNG0240C[2] [AAG18840] COG4031
PAB0054 [NP_125787] CDD41919	PAB1779 [NP_126559] CDD43950	VNG1236C [AAG19598]
PAB0176 [NP_125948] CDD43579	PAB1806[1] [NP_126515] CDD43599	VNG1611C [AAG19875] COG4749
PAB1127 [NP_127394] CDD30177	PAB2413 [NP_126288] COG1710	VNG1670C [AAG19921] COG3612
PAB1291 [NP_127284] CDD41906		VNG1891H [AAG20086]
PAB1584 [NP_126876] COG4072		VNG2315H [AAG20425]
PAB1860 [NP_126440]		VNG2508C [AAG20570] MC1 CDD45747
PAB0813 [NP_126902] COG1630		VNG2524H [AAG20585]
PAB0853 [NP_126970]		VNG2669G [AAG20696] Cyo COG4083
(c) CSPs mainly shared by Thermoplasmata and *Sulfolobus*		
Ta0035 [NP_393514] COG5592	Ta0793a [NP_394256]	Saci_0055 [YP_254778]
Ta0164 [NP_393642]	Ta0938 [NP_394396]	Saci_0322 [YP_255031]
Ta0165 [NP_393643]	Ta0939 [NP_394397] PQQC CDD45213	Saci_0323 [YP_255032]
Ta0267 [NP_393747] CDD43623	Ta1156 [NP_394612]	Saci_0979 [YP_255633] sdhD
Ta0308 [NP_393788]	Ta1345 [NP_394801]	Saci_1065 [YP_255715]
Ta0347 [NP_393826] TauA CDD31059	Ta1440 [NP_394894]	Saci_1491 [YP_256105] CDD40171
Ta0547 [NP_394021]	Ta1453 [NP_394906]	Saci_1560 [YP_256166]
Ta0548m[3] [NP_394022]	Ta1507 [NP_394957] CDD29645	Saci_1747 [YP_256346] SoxE CDD46414
Ta0583 [NP_394007]	Saci_0040 [YP_254763]	Saci_1952 [YP_256548]
Ta0759 [NP_394223]	Saci_0054 [YP_254777]	Saci_2078 [YP_256665]

Note [1]. A homolog to PAB1806 is also found in *Aquifex aeolicus* VF5;
Note [2]. A homolog to VNG0240c is also found in *Methanopyrus kandleri*;
Note [3]. Two low-scoring homolog for Ta0548 is also found in *Gloeobacter violaceus* PCC 7421.

CHAPTER 7.

Conclusions

7.1 Research Summary

In order to develop a reliable understanding of microbial systematics and phylogeny, it is necessary to develop new well-defined (molecular or biochemical) criteria for identifying all of the main groups or divisions within Prokaryotes. These new criteria or properties should be such that they should enable identification and circumscription of all of the major taxa (at various taxonomic levels) in clear molecular and/or biochemical terms (Gupta and Griffiths, 2002). Furthermore, it is also of central importance to understand how different groups are related to each other and have branched off from a common ancestor (Gupta and Gao, 2010; Gupta, 2000a; Gupta, 2001). The significance of rare genomic changes lies in their ability to provide novel means to resolve these important issues in phylogeny (Gupta, 1998b; Gupta and Griffiths, 2002). Based on simply the absence/presence of various identified RGCs, the main groups within Prokaryotes can now be defined and distinguished from each other in clear molecular terms (Gupta and Gao, 2010; Gupta, 2000a).

The primary focus of my research project is phylogenomics studies of Actinobacteria. Sequence alignments of various proteins from different bacterial species have identified a number of CSIs that are specific for either all actinobacteria or certain subgroups within this phylum (Gao and Gupta, 2005). CSIs that are specific for the entire Actinobacteria phylum are found in various proteins, including: CoxI, GluRS, CTPsyn, Gft, GlyRS, TrmD, Gyrase A, SahH and SHMT. When this work was initiated, the sequence information for most of these proteins was only available for species whose genomes have been sequenced. Thus, the actinobacteria-specificity of several of these

CSIs (viz. CoxI, GluRS and CTPsyn) and a large insert in 23S rRNA has been examined by sequencing fragments of these genes from 23 actinobacterial species, covering many different families. All of these gene fragments, except two in GluRS, were found to contain these CSIs providing strong evidence that they are distinctive characteristics of the entire phylum (Gao and Gupta, 2005). In view of their actinobacteria-specificity, these CSIs provide good molecular markers for circumscribing the Actinobacteria phylum and distinguishing species of this group from all other bacteria. In addition to these CSIs, which are specific for all Actinobacteria, we have identified many other CSIs that are specific for certain subgroups within this large phylum. Some of these CSIs also provide information regarding the interrelationships among different subgroups. Based on the shared CSIs (viz. HolB, S3 and DnaK), two large clades within Actinobacteria phylum *Corynebacterineae/Pseudonocardineae/Micromonosporineae* and *Micrococcineae/Actinomycineae/Bifidobacteriaceae* are revealed. This inference is also supported by the clustering of these subgroups in phylogenetic trees based on combined sequences for multiple proteins.

Besides identifying CSIs that are specific for Actinobacteria, comprehensive BLAST searches were carried out on each ORF from 12 actinobacterial genomes to identify CSPs that are specific for Actinobacteria or its various subgroups. In this study, a large number of actinobacteria-specific CSPs were identified, homologs of which are not found in any other bacteria (Gao et al., 2006). Based on their species distribution pattern, these CSPs could be grouped as actinobacterial phylum-specific, suborder-specific, family-specific, genus-specific and so on. For example, 4 CSPs are shared by all actinobacteria, including *Rubrobacter*, which is the deepest branch within this phylum; 4 CSPs are specific for the suborder *Corynebacterineae*; and 32 CSPs are unique to the genus *Mycobacterium*. These proteins provide novel molecular means for defining and circumscribing the Actinobacteria phylum and its different subgroups (Gao et al., 2006). Because of their specificity, these group-specific CSPs likely play important role in the cell that could confer unique biochemical or physiological characteristics to Actinobacteria.

In this work, I also carried out functional studies of one of the 4 actinobacteria-specific proteins, ASP1 (Gene ID: SCO1997) from *Streptomyces coelicolor*. The X-ray crystal structure of ASP1 was determined at 2.2 Å. The overall structure of ASP1 contains a similar fold as the large NP-1 family of nucleoside phosphorylase enzymes; however, detailed structural comparison suggests that their functions are not related. Further comparative analysis revealed two regions expected to be important for protein function: a central, divalent metal ion binding pore, and a highly conserved elbow shaped helical region at the C-terminus. Sequence analyses revealed that ASP1 is paralogous to another actinobacteria-specific protein ASP2 (SCO1662 from *S. coelicolor*) and that both proteins likely carry out similar function. Our structural data in combination with sequence analysis supports the idea that ASP1 and ASP2 mediate similar function. This function is predicted to be novel since both the structures and sequences of these proteins do not match any known protein with or without known function. Our results suggest that this function could involve divalent metal ion binding/transport. It will be intriguing to determine what contribution, if any, this highly conserved 'pore' region makes toward

ASP1 function. Future genetic and biochemical studies of these proteins is therefore of great interest in linking the conservation of the biology of actinobacteria and their unique genes.

The second prokaryotic group that I have worked on is Archaea, which are highly diverse in terms of their physiology, metabolism and ecology. Presently, very few molecular characteristics are known that are uniquely shared by either all archaea or the different main groups within archaea. The evolutionary relationships among different groups within the Euryarchaeota branch are also not clearly understood. We have carried out comprehensive analyses on each ORF in the genomes of 11 archaea to search for proteins that are unique to either all Archaea or for its main subgroups (Gao and Gupta, 2007). These studies have identified 1448 CSPs or ORFs that are distinctive characteristics of Archaea and its various subgroups and whose homologues are not found in other organisms. Six of these CSPs are unique to all Archaea, 10 others are only missing in *Nanoarchaeum equitans* and a large number of CSPs are specific for various main groups within Archaea (e.g. Crenarchaeota, Euryarchaeota, *Sulfolobales* and *Desulfurococcales, Halobacteriales, Thermococci, Thermoplasmata*, all methanogenic archaea or particular groups of methanogens). Of particular importance is the observation that 31 CSPs are uniquely present in virtually all methanogens (including *M. kandleri*) and 10 additional CSPs are only found in different methanogens as well as *A. fulgidus*. In contrast, no CSP was exclusively shared by various methanogens and any of the *Halobacteriales* or *Thermoplasmatales*. These results strongly indicate that all methanogenic archaea form a monophyletic group exclusive of other archaea and that this lineage likely evolved from *Archaeoglobus*. In addition, 15 CSPs that are uniquely shared by *M. kandleri* and *Methanobacteriales* suggest a close evolutionary relationship between them. In contrast to the phylogenomics studies, a monophyletic grouping of methanogenic archaea is not supported by phylogenetic analyses based on protein sequences. The identified archaea-specific CSPs provide novel molecular markers or signature proteins that are distinctive characteristics of Archaea and all of its major subgroups. The species distributions of these CSPs provide novel insights into the evolutionary relationships among different groups within Archaea, particularly regarding the origin of methanogenesis. Most of these CSPs are of unknown function and further studies should lead to the discovery of novel biochemical and physiological characteristics that are unique to either all Archaea or its different subgroups.

In addition to the work that has been described in Chapters 2-6, I also carried out similar phylogenomics studies on two additional bacterial groups, and the references for the published work are as follows: (1) Gao B, Mohan R, Gupta RS. Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. Int J Syst Evol Microbiol. 2009 Feb;59(Pt 2):234-47. (2) Gupta RS & Gao B. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus *Clostridium* sensu stricto (cluster I). Int J Syst Evol Microbiol. 2009 Feb;59(Pt 2):285-94. A brief summary of these two studies is given below.

The class Gammaproteobacteria, which forms one of the largest groups within bacteria, is currently distinguished from other bacteria solely on the basis of its branching in phylogenetic trees. No molecular or biochemical characteristic is known that is unique to the class Gammaproteobacteria or its different subgroups (orders). The relationship among different orders of gammaproteobacteria is also not clear. In this study, we carried out detailed phylogenomic and comparative genomic analyses on gammaproteobacteria that clarify some of these issues (Gao et al., 2009a). Phylogenetic trees based on concatenated sequences for 13 and 36 universally distributed proteins were constructed for 45 members of the class Gammaproteobacteria covering 13 of its 14 orders. In these trees, species from a number of the subgroups formed distinct clades and their relative branching order was indicated as follows (from the most recent to the earliest diverging): *Enterobacteriales→Pasteurellales→Vibrionales→Aeromonadales→Alteromonadales→ Oceanospirillales, Pseudomonadales→Chromatiales, Legionellales, Methylococcales, Xanthomonadales, Cardiobacteriales, Thiotrichales*. Four conserved indels in 4 widely distributed proteins that are specific for gammaproteobacteria are also described. A 2-aa indel in 59-phosphoribosyl-5-aminoimidazole-4-carboxamide transformylase (AICAR transformylase; PurH) was a distinctive characteristic of all gammaproteobacteria (except Francisella tularensis). Two other conserved indels (a 4-aa indel in RNA polymerase b-subunit and a 1-aa indel in ribosomal protein L16) were found uniquely in various species of the orders *Enterobacteriales, Pasteurellales, Vibrionales, Aeromonadales* and *Alteromonadales*, but were not found in other gammaproteobacteria. Lastly, a 2-aa indel in leucyl-tRNA synthetase was commonly present in the above orders of the class Gammaproteobacteria and also in some members of the order *Oceanospirillales*. The presence of conserved indels in these gammaproteobacterial orders indicates that species from these orders shared a common ancestor that was separate from other bacteria, a suggestion that is supported by phylogenetic studies. Systematic BLASTP searches were also conducted on various ORF in the genome of *Escherichia coli* K- 12. These analyses identified 75 proteins that were unique to most members of the class Gammaproteobacteria or were restricted to species from some of its main orders. The genes for these proteins have evolved at various stages during the evolution of gammaproteobacteria and their species distribution pattern, in conjunction with other results presented here, provide valuable information regarding the evolutionary relationships among these bacteria.

The species of *Clostridium* comprise a very heterogeneous assemblage of bacteria that do not form a phylogenetically coherent group. It has been proposed previously that only a subset of the species of *Clostridium* that form a distinct cluster in the 16S rRNA tree (cluster I) should be regarded as true representatives of the genus *Clostridium* (i.e. *Clostridium* sensu stricto). However, this cluster is presently defined only in phylogenetic terms, and no biochemical, molecular or phenotypic characteristic is known that is unique to species from this cluster. We carried out phylogenomic and comparative analyses based on sequenced clostridial genomes in an attempt to bridge this gap and to clarify the evolutionary relationships among species of clostridia. In phylogenetic trees for species of clostridia based on concatenated sequences for 37 highly conserved proteins, the species of *Clostridium* cluster I formed a strongly supported clade that was separated from all other clostridia by a long branch. Several other *Clostridium* species that are not part of this cluster, grouped reliably with other species of clostridia in a number of well-resolved clades. Our comparative genomic analyses have identified 3 conserved indels in 3 highly conserved proteins (a 4-aa indel in DNA gyrase A, a 1-aa indel in ATP synthase beta subunit and a 1-aa indel in ribosomal protein S2) that are unique to the species of *Clostridium* cluster I and are not found in any other bacteria. BLASTP searches on various proteins in the genomes of *Clostridium tetani* E88 and *Clostridium perfringens* SM101 have also identified more than 10 proteins that are found uniquely in the cluster I species. These results provide evidence that the species of *Clostridium* cluster I are not only phylogenetically distinct but also share many unique molecular characteristics. These newly identified molecular markers provide useful tools to define and circumscribe the genus *Clostridium* sensu stricto in more definitive terms. We have also identified a 7–9 aa conserved insert in the enzyme phosphoglycerate dehydrogenase that is uniquely found in the *Clostridium thermocellum, Thermoanaerobacter pseudethanolicus, Thermoanaerobacter tengcogensis* and *Caldicellulosiruptor saccharolyticus* homologues, and is absent from all other bacteria. These species form a well-defined clade in the phylogenetic trees and this indel provides a potential molecular marker for this clostridial cluster.

7.2 Future Work

The CSIs, CSPs and phylogenetic analyses that are presented here provide exciting prospects for future work and applications. First, the resolving power of CSIs and CSPs for inferring relationship among higher taxonomic ranks makes them valuable for defining or evaluating the taxonomy structure of Prokaryotes. In the future, the studies on prokaryotic systematics and taxonomy should incorporate analysis from genomic comparison instead of being solely based on the 16S rRNA or even earlier ambiguous morphological characteristics. Second, group-specific CSIs and CSPs can be used to test the ecological diversity and richness of differerent prokaryotic groups in metagenomics or clinical samples (Hugenholtz and Tyson, 2008; Gill et al., 2006; Ahmad et al., 2003). PCR primers could be constructed for gene fragments that contain useful CSIs or genes for group-specfic CSPs. Then after sequencing, we can detect the presence of certain prokaryotic lineages based on the presence or absence of the CSIs and CSPs. Furthermore, if the function of the detected CSPs are known, we can probably infer the contribution or function of the prokaryotic lineage in the community (Gianoulis et al., 2009; Turnbaugh and Gordon, 2008).

Third, the functional significance of all CSIs and most CSPs are unknown. Most of the discovered CSIs are present in universal proteins (viz. CoxI, DnaK, CTP synthase, Gyrase A and B subunits, etc.) that are involved in essential functions. The primary biochemical functions of these proteins are vital for cell survival and they are expected to remain the same in all organisms. Hence, the question arises: what is the functional significance of these evolutionarily preserved CSIs that are specific for different bacterial lineages? Our recent work on two CSIs in GroEL and DnaK demonstrated that knocking out the indel region will result in bacterial death (Singh and Gupta, 2009). Our lab is currently working on understanding why these indels are essential for protein function. Unlike the CSIs, which are found in universal proteins, virtually all of the CSPs are of unknown function (Griffiths et al., 2006; Gao et al., 2006; Kainth and Gupta, 2005). The discovery of these proteins points to our lack of knowledge regarding many fundamental aspects of cells, particularly functions that are unique to different bacterial groups. Hence, an important challenge is to understand the cellular functions of these genes/proteins (Danchin, 1999; Galperin and Koonin, 2004; Roberts, 2004). However, it is not an easy task to study unknown proteins, which will involve multi-aspect information, such as gene expression, knockout, cellular localization, protein-protein interaction, structure, enzymatic assays, etc.

Beyond their application in phylogenetic analysis and their potential for functional studies, there is more to explore of CSIs and CSPs, such as their origin and fate in the cellular network. We know that not all genes of a bacterial genome come from the last universal common ancestor (Koonin, 2003). Most genes such as these lineage-specific CSPs and also CSIs arise throughout the evolutionary process of prokaryotes, and that is why they can be identified at different phylogenetic depth (Daubin and Ochman, 2004; Lerat et al., 2005). Although their origin is still mysterious, studies have shown that indels might result from replication loops, and the new genes might come from gene duplications or gene transfers that occurred at different evolutionary stages (Ochman et al., 2000; Lawrence and Hendrickson, 2005). The reason that they evolved in the progenitor cell and were retained by every daughter lineage is that they confer adaptive advantage to the bacteria to fight with the changing environment (Kuo and Ochman, 2009; Raskin et al., 2006). Later on, these new genes gradually build up their interaction with other proteins in the existing cellular network and become a required component in the network (Narra et al., 2008; Abby and Daubin, 2007).

Unlike CSPs, CSIs are not independent functional units. We found that most of these lineage-specific indels are located in the surface loop region of the available protein structures (Shah et al., 2009). A recent study showed that the indels in surface loops play an important role in determining domain-domain, or protein-protein interactions (Akiva et al., 2008). Furthermore, since some CSPs share the same lineage-specificity with CSIs, it is likely that the CSIs act as docking surface and pull the protein to form the interaction. Thus, we hypothesize about the evolution of cellular networks. As shown in Figure 7.1, in a universal protein, group A bacteria all have a CSI at the same position, which form a direct interaction with a group A-specific CSP to function together. While group-B bacteria has a different CSI either at the same position or at different position and interact with a CSP that is unique to group B. In this way, new genes become incorporated into the cellular network and function in the cellular network to make the bacteria fit the environment. This hypothesis needs more experimental evidence such as the interaction of CSIs and CSPs, and the function of these two markers.

7.3 Closing Remarks

The availability of genome sequence data is enabling identification of numerous rare genomic changes consisting of CSIs and CSPs that are specific for different groups of prokaryotes at various taxonomic levels. The discovery of these new molecular markers is proving very useful in understanding some of the critical issues in prokaryotic phylogeny and systematics. In particular, based on these molecular markers, it is now possible to identify and circumscribe most of the major phyla, as well as many of their subgroups, within Bacteria and Archaea in definitive molecular terms. Additionally, these molecular markers are also providing means to logically infer the evolutionary relationship within each phylum. Because of their taxa specificity, functional studies on these newly discovered molecular markers hold much promise for discovering novel biological characteristics that are distinctive characteristics of different groups of prokaryotes.

7.4 Figure

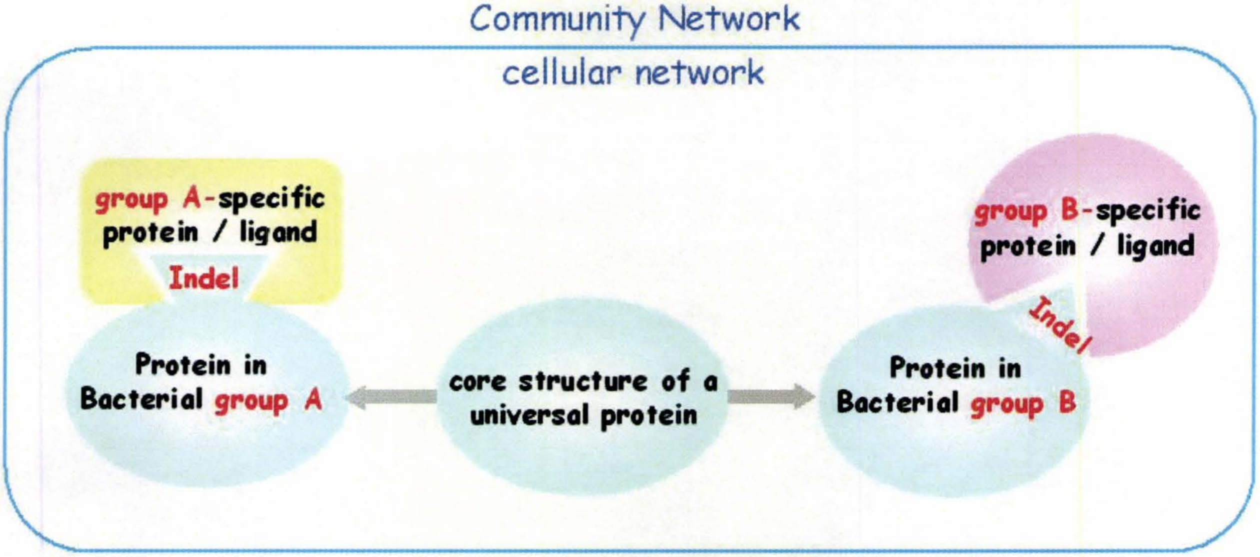

Community Network
cellular network
group A-specific protein / ligand
Indel
Protein in Bacterial group A
core structure of a universal protein
Protein in Bacterial group B
group B-specific protein / ligand
Indel

Figure 7.1 Cartoon illustrating the basic concept concerning the cellular functions of various CSIs and CSPs. Most of the CSIs are present in essential proteins for which their sequence are highly conserved and their core functions remain the same in all lineages. It is postulated that the CSIs located on protein surfaces are involved in binding of CSPs or ligands, which share the same lineage-specificity.

BIBLIOGRAPHY

Abby,S. and Daubin,V. (2007). Comparative genomics and the evolution of prokaryotes. Trends Microbiol *15*, 135-141.

Adams,P.D., Grosse-Kunstleve,R.W., Hung,L.W., Ioerger,T.R., Mccoy,A.J., Moriarty,N.W., Read,R.J., Sacchettini,J.C., Sauter,N.K., and Terwilliger,T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr *58*, 1948-1954.

Ahmad,S., Itani,L.Y., and Araj,G.F. (2003). Molecular fingerprinting of multidrug-resistant *Mycobacterium tuberculosis* strains in Beirut reveals genetic diversity and father to daughter transmission. J Med Liban *51*, 4-8.

Ahmad,S., Selvapandiyan,A., and Bhatnagar,R.K. (2000). Phylogenetic analysis of Gram-positive bacteria based on grpE, encoded by the dnaK operon. Int J Syst Evol Microbiol *50*, 1761-1766.

Akiva,E., Itzhaki,Z., and Margalit,H. (2008). Built-in loops allow versatility in domain-domain interactions: Lessons from self-interacting domains. Proc Natl Acad Sci U S A *105*, 13292-13297.

Allen,M.A., Lauro,F.M., Williams,T.J., Burg,D., Siddiqui,K.S., De Francisci,D., Chong,K.W., Pilak,O., Chew,H.H., De Maere,M.Z., Ting,L., Katrib,M., Ng,C., Sowers,K.R., Galperin,M.Y., Anderson,I.J., Ivanova,N., Dalin,E., Martinez,M., Lapidus,A., Hauser,L., Land,M., Thomas,T., and Cavicchioli,R. (2009). The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold adaptation. ISME J *Epub ahead of print*.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W., and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Aravind,L. and Koonin,E.V. (1999). Novel predicted RNA-binding domains associated with the translation machinery. J Mol Evol *48*, 291-302.

Aravind,L., Makarova,K.S., and Koonin,E.V. (2000). Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. Nucleic Acids Res *28*, 3417-3432.

Aris-Brosou,S. (2005). Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol *22*, 200-209.

Bagwell,C.E., Bhat,S., Hawkins,G.M., Smith,B.W., Biswas,T., Hoover,T.R., Saunders,E., Han,C.S., Tsodikov,O.V., and Shimkets,L.J. (2008). Survival in nuclear waste, extreme resistance, and potential applications gleaned from the genome sequence of *Kineococcus radiotolerans* SRS30216. PLoS One *3*, e3878.

Bapteste,E., Brochier,C., and Boucher,Y. (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea *1*, 353-363.

Barns,S.M., Fundyga,R.E., Jeffries,M.W., and Pace,N.R. (1994). Remarkable Archaeal Diversity Detected in A Yellowstone-National-Park Hot-Spring Environment. Proc Natl Acad Sci U S A *91*, 1609-1613.

Bazylinski,D.A. and Frankel,R.B. (2004). Magnetosome formation in prokaryotes. Nat Rev Microbiol *2*, 217-230.

Beiko,R.G., Harlow,T.J., and Ragan,M.A. (2005). Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A *102*, 14332-14337.

Belanger,A.E., Besra,G.S., Ford,M.E., Mikusova,K., Belisle,J.T., Brennan,P.J., and Inamine,J.M. (1996). The embAB genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc Natl Acad Sci USA *93*, 11919-11924.

Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H., Harper,D., Bateman,A., Brown,S., Chandra,G., Chen,C.W., Collins,M., Cronin,A., Fraser,A., Goble,A., Hidalgo,J., Hornsby,T., Howarth,S., Huang,C.H., Kieser,T., Larke,L., Murphy,L., Oliver,K., O'Neil,S., Rabbinowitsch,E., Rajandream,M.A., Rutherford,K., Rutter,S., Seeger,K., Saunders,D., Sharp,S., Squares,R., Squares,S., Taylor,K., Warren,T., Wietzorrek,A., Woodward,J., Barrell,B.G., Parkhill,J., and Hopwood,D.A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature *417*, 141-147.

Bentley,S.D., Corton,C., Brown,S.E., Barron,A., Clark,L., Doggett,J., Harris,B., Ormond,D., Quail,M.A., May,G., Francis,D., Knudson,D., Parkhill,J., and Ishimaru,C.A. (2008). Genome of the actinomycete plant pathogen *Clavibacter michiganensis* subsp sepedonicus suggests recent niche adaptation. J Bacteriol *190*, 2150-2160.

Bentley,S.D., Maiwald,M., Murphy,L.D., Pallen,M.J., Yeats,C.A., Dover,L.G., Norbertczak,H.T., Besra,G.S., Quail,M.A., Harris,D.E., von Herbay,A., Goble,A., Rutter,S., Squares,R., Squares,S., Barrell,B.G., Parkhill,J., and Relman,D.A. (2003). Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. Lancet *361*, 637-644.

Berg,S., Starbuck,J., Torrelles,J.B., Vissa,V.D., Crick,D.C., Chatterjee,D., and Brennan,P.J. (2005). Roles of conserved proline and glycosyltransferase motifs of embC in biosynthesis of lipoarabinomannan. J Biol Chem *280*, 5651-5663.

Bergsten,J. (2005). A review of long-branch attraction. Cladistics *21*, 163-193.

Bolhuis,H., Palm,P., Wende,A., Falb,M., Rampp,M., Rodriguez-Valera,F., Pfeiffer,F., and Oesterhelt,D. (2006). The genome of the square archaeon *Haloquadratum walsbyi* : life at the limits of water activity. BMC Genomics *7*.

Boucher,Y., Douady,C.J., Papke,R.T., Walsh,D.A., Boudreau,M.E.R., Nesbo,C.L., Case,R.J., and Doolittle,W.F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet *37*, 283-328.

Brendel,V., Brocchieri,L., Sandler,S.J., Clark,A.J., and Karlin,S. (1997). Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. J Mol Evol *44*, 528-541.

Brennan,P.J. (2003). Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. Tuberculosis *83*, 91-97.

Brochier,C., Forterre,P., and Gribaldo,S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. Genome Biol *5*.

Brochier,C., Forterre,P., and Gribaldo,S. (2005a). An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol *5*.

Brochier,C., Gribaldo,S., Zivanovic,Y., Confalonieri,F., and Forterre,P. (2005b). Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol *6*.

Brochier-Armanet,C., Boussau,B., Gribaldo,S., and Forterre,P. (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol *6*, 245-252.

Brosch,R., Gordon,S.V., Garnier,T., Eiglmeier,K., Frigui,W., Valenti,P., Dos Santos,S., Duthoy,S., Lacroix,C., Garcia-Pelayo,C., Inwald,J.K., Golby,P., Garcia,J.N., Hewinson,R.G., Behr,M.A., Quail,M.A., Churcher,C., Barrell,B.G., Parkhill,J., and Cole,S.T. (2007). Genome plasticity of BCG and impact on vaccine efficacy. Proc Natl Acad Sci U S A *104*, 5596-5601.

Brown,J.R. and Doolittle,W.F. (1997). Archaea and the prokaryote-to-eukaryote transition. Microbiol Rev *61*, 456-502.

Bruggemann,H., Henne,A., Hoster,F., Liesegang,H., Wiezer,A., Strittmatter,A., Hujer,S., Durre,P., and Gottschalk,G. (2004). The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin. Science *305*, 671-673.

Brunger,A.T., Adams,P.D., Clore,G.M., Delano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M., Pannu,N.S., Read,R.J., Rice,L.M., Simonson,T., and Warren,G.L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr *54*, 905-921.

Bull,A.T. and Stach,J.E. (2007). Marine actinobacteria: new opportunities for natural product search and discovery. Trends Microbiol *15*, 491-499.

Burggraf,S., Huber,H., and Stetter,K.O. (1997). Reclassification of the crenarchaeal orders and families in accordance with 16S rRNA sequence data. Int J Syst Bacteriol *47*, 657-660.

Burggraf,S., Stetter,K.O., Rouviere,P., and Woese,C.R. (1991). *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. Syst Appl Microbiol *14*, 346-351.

Cann,I.K.O. and Ishino,Y. (1999). Archaeal DNA replication: Identifying the pieces to solve a puzzle. Genetics *152*, 1249-1267.

Castresana,J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

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.

Cerdeno-Tarraga,A.M., Efstratiou,A., Dover,L.G., Holden,M.T.G., Pallen,M., Bentley,S.D., Besra,G.S., Churcher,C., James,K.D., De Zoysa,A., Chillingworth,T., Cronin,A., Dowd,L., Feltwell,T., Hamlin,N., Holroyd,S., Jagels,K., Moule,S., Quail,M.A., Rabbinowitsch,E., Rutherford,K.M., Thomson,N.R., Unwin,L., Whitehead,S., Barrell,B.G., and Parkhill,J. (2003). The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res *31*, 6516-6523.

Chater,K.F. and Chandra,G. (2006). The evolution of development in Streptomyces analysed by genome comparisons. FEMS Microbiol Rev *30*, 651-672.

Chen,L.M., Brugger,K., Skovgaard,M., Redder,P., She,Q.X., Torarinsson,E., Greve,B., Awayez,M., Zibat,A., Klenk,H.P., and Garrett,R.A. (2005). The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota. J Bacteriol *187*, 4992-4999.

Cho,H.D., Verlinde,C.L., and Weiner,A.M. (2005). Archaeal CCA-adding enzymes - Central role of a highly conserved beta-turn motif in RNA polymerization without translocation. J Biol Chem *280*, 9555-9566.

Ciaramella,M., Napoli,A., and Rossi,M. (2005). Another extreme genome: how to live at pH 0. Trends Microbiol *13*, 49-51.

Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B., and Bork,P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M., and Tiedje,J.M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res *37*, D141-D145.

Cole,S.T. (2002). Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. Microbiology-Sgm *148*, 2919-2928.

Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D., Mungall,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R.M., Devlin,K., Duthoy,S., Feltwell,T., Fraser,A., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Lacroix,C., Maclean,J., Moule,S., Murphy,L., Oliver,K., Quail,M.A., Rajandream,M.A., Rutherford,K.M., Rutter,S., Seeger,K., Simon,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Taylor,K., Whitehead,S., Woodward,J.R., and Barrell,B.G. (2001). Massive gene decay in the leprosy bacillus. Nature *409*, 1007-1011.

Daffe,M. and Draper,P. (1998). The envelope layers of mycobacteria with reference to their pathogenicity. Adv Microb Physiol *39*, 131-203.

Danchin,A. (1999). From function to sequence, an integrated view of the genome texts. Physica A *273*, 92-98.

Daubin,V., Gouy,M., and Perriere,G. (2002). A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Res *12*, 1080-1090.

Daubin,V. and Ochman,H. (2004). Bacterial Genomes as new gene homes: The genealogy of ORFans in *E. co*li. Genome Res. *14*, 1036-1042.

Daugherty,M., Vonstein,V., Overbeek,R., and Osterman,A. (2001). Archaeal shikimate kinase, a new member of the GHMP-kinase family. J Bacteriol *183*, 292-300.

Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M., Huber,R., Feldman,R.A., Short,J.M., Olsen,G.J., and Swanson,R.V. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature *392*, 353-358.

Delano,W.L. (2002). The PyMOL User's Manual. Palo Alto, CA: DeLano Scientific).

Delsuc,F., Brinkmann,H., and Philippe,H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet *6*, 361-375.

Doerks,T., von Mering,C., and Bork,P. (2004). Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. Nucleic Acids Res *32*, 6321-6326.

Domenech,P., Barry,C.E., and Cole,S.T. (2001). *Mycobacterium tuberculosis* in the post-genomic age. Curr Opin Microbiol *4*, 28-34.

Doolittle,W.F. (1999). Phylogenetic classification and the universal tree. Science *284*, 2124-2128.

Dutilh,B.E., Snel,B., Ettema,T.J.G., and Huynen,M.A. (2008). Signature genes as a phylogenomic tool. Mol Biol Evol *25*, 1659-1667.

Embley,T.M. and Stackebrandt,E. (1994). The Molecular Phylogeny and Systematics of the Actinomycetes. Annu Rev Microbiol *48*, 257-289.

Emsley,P. and Cowtan,K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126-2132.

Endrizzi,J.A., Kim,H.S., Anderson,P.M., and Baldwin,E.P. (2004). Crystal structure of *Escherichia coli* cytidine triphosphate synthetase, a nucleotide-regulated glutamine amidotransferase/ATP-dependent amidoligase fusion protein and homologue of anticancer and antiparasitic drug targets. Biochemistry *43*, 6447-6463.

Ermler,U., Grabarse,W., Shima,S., Goubeaud,M., and Thauer,R.K. (1997). Crystal structure of methyl coenzyme M reductase: The key enzyme of biological methane formation. Science *278*, 1457-1462.

Farrar,M.D., Ingham,E., and Holland,K.T. (2000). Heat shock proteins and inflammatory acne vulgaris: molecular cloning, overexpression and purification of a *Propionibacterium acnes* GroEL and DnaK homologue. FEMS Microbiol Lett *191*, 183-186.

Felsenstein,J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol *17*, 368-376.

Fitz-Gibbon,S.T., Ladner,H., Kim,U.J., Stetter,K.O., Simon,M.I., and Miller,J.H. (2002). Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc Natl Acad Sci U S A *99*, 984-989.

Forterre,P. and Philippe,H. (1999). Where is the root or the universal tree of life? Bioessays *21*, 871-879.

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.

Freilich,S., Kreimer,A., Borenstein,E., Yosef,N., Sharan,R., Gophna,U., and Ruppin,E. (2009). Metabolic-network-driven analysis of bacterial ecological strategies. Genome Biol *10*, R61.

Fricke,W.F., Seedorf,H., Henne,A., Kruer,M., Liesegang,H., Hedderich,R., Gottschalk,G., and Thauer,R.K. (2006). The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H-2 for methane formation and ATP synthesis. J Bacteriol *188*, 642-658.

Froula,J.L. and Francino,M.P. (2007). Selection against spurious promoter motifs correlates with translational efficiency across bacteria. PLoS One *2*, e745.

Futterer,O., Angelov,A., Liesegang,H., Gottschalk,G., Schleper,C., Schepers,B., Dock,C., Antranikian,G., and Liebl,W. (2004). Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. Proc Natl Acad Sci U S A *101*, 9091-9096.

Galperin,M.Y. and Koonin,E.V. (2000). Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol *18*, 609-613.

Galperin,M.Y. and Koonin,E.V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res *32*, 5452-5463.

Gao,B. and Gupta,R.S. (2005). Conserved indels in protein sequences that are characteristic of the phylum Actinobacteria. Int J Syst Evol Microbiol *55*, 2401-2412.

Gao,B. and Gupta,R.S. (2007). Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics *8*.

Gao,B., Mohan,R., and Gupta,R.S. (2009a). Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. Int J Syst Evol Microbiol *59*, 234-247.

Gao,B., Paramanathan,R., and Gupta,R.S. (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. Antonie Van Leeuwenhoek *90*, 69-91.

Gao,B., Sugiman-Marangos,S., Junop,M.S., and Gupta,R.S. (2009b). Structural and Phylogenetic Analysis of a Conserved Actinobacteria-Specific Protein (ASP1; SCO1997) from *Streptomyces coelicolor*. BMC Struct Biol *9*, 40.

Garrity,G.M., Bell,J.A., and liburn,T.G. (2001). Taxonomic outline of the prokaryotes. In Bergey's Manual of Systematic Bacteriology, D.R.Boone, R.W.Castenholz, and G.M.Garrity, eds. (Berlin: Springer-Verlag).

Garrity,G.M., Bell,J.A., and Lilburn,T. (2005). The revised road map to the manual. In Bergey's Manual of Systematic Bacteriology, D.J.Brenner, N.R.Krieg, J.T.Staley, and G.M.Garrity, eds. Springer US), pp. 159-187.

Garrity,G.M. and Holt,J.G. (2001). The road map to the manual. In Bergey's Manual of Systematic Bacteriology, D.R.Boone and R.W.Castenholz, eds. (Berlin: Springer-Verlag), pp. 119-166.

George,R.A., Spriggs,R.V., Bartlett,G.J., Gutteridge,A., MacArthur,M.W., Porter,C.T., Al Lazikani,B., Thornton,J.M., and Swindells,M.B. (2005). Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci U S A *102*, 12299-12304.

Gianoulis,T.A., Raes,J., Patel,P.V., Bjornson,R., Korbel,J.O., Letunic,I., Yamada,T., Paccanaro,A., Jensen,L.J., Snyder,M., Bork,P., and Gerstein,M.B. (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A *106*, 1374-1379.

Gill,S.R., Pop,M., Deboy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M., and Nelson,K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355-1359.

Gogarten,J.P., Doolittle,W.F., and Lawrence,J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol Biol Evol *19*, 2226-2238.

Gogarten,J.P. and Townsend,J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol *3*, 679-687.

Gonzalez,J.M., Sheckells,D., Viebahn,M., Krupatkina,D., Borges,K.M., and Robb,F.T. (1999). *Thermococcus waiotapuensis* sp. nov., an extremely thermophilic archaeon isolated from a freshwater hot spring. Arch Microbiol *172*, 95-101.

Goodfellow,M. and Williams,S.T. (1983). Ecology of Actinomycetes. Annu Rev Microbiol *37*, 189-216.

Grabarse,W., Mahlert,F., Duin,E.C., Goubeaud,M., Shima,S., Thauer,R.K., Lamzin,V., and Ermler,U. (2001). On the mechanism of biological methane formation: Structural evidence for conformational changes in methyl-coenzyme M reductase upon substrate binding. J Mol Biol *309*, 315-330.

Graham,D.E., Overbeek,R., Olsen,G.J., and Woese,C.R. (2000). An archaeal genomic signature. Proc Natl Acad Sci U S A *97*, 3304-3308.

Gribaldo,S. and Brochier-Armanet,C. (2006). The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci *361*, 1007-1022.

Griffiths,E. and Gupta,R.S. (2001). The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division. Microbiology-Sgm *147*, 2611-2622.

Griffiths,E. and Gupta,R.S. (2004a). Distinctive protein signatures provide molecular markers and evidence for the monophyletic nature of the Deinococcus-thermus phylum. J Bacteriol *186*, 3097-3107.

Griffiths,E. and Gupta,R.S. (2004b). Signature sequences in diverse proteins provide evidence for the late divergence of the order Aquificales. Int Microbiol *7*, 41-52.

Griffiths,E. and Gupta,R.S. (2006). Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. Int J Syst Evol Microbiol *56*, 99-107.

Griffiths,E. and Gupta,R.S. (2007). Identification of signature proteins that are distinctive of the Deinococcus-Thermus phylum. Int Microbiol *10*, 201-208.

Griffiths,E., Petrich,A.K., and Gupta,R.S. (2005). Conserved indels in essential proteins that are distinctive characteristics of Chlamydiales and provide novel means for their identification. Microbiology-Sgm *151*, 2647-2657.

Griffiths,E., Ventresca,M.S., and Gupta,R.S. (2006). BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species. BMC Genomics *7*, 14.

Grosse-Kunstleve,R.W. and Adams,P.D. (2003). Substructure search procedures for macromolecular structures. Acta Crystallogr D Biol Crystallogr *59*, 1966-1973.

Gupta,R.S. (1998a). Life's third domain (Archaea): An established fact or an endangered paradigm? A new proposal for classification of organisms based on protein sequences and cell structure. Theor Popul Biol *54*, 91-104.

Gupta,R.S. (1998b). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev. *62*, 1435-1491.

Gupta,R.S. (2000a). The natural evolutionary relationships among prokaryotes. Crit Rev Microbiol. *26*, 111-131.

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

Gupta,R.S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int Microbiol *4*, 187-202.

Gupta,R.S. (2003). Evolutionary relationships among photosynthetic bacteria. Photosynth Res *76*, 173-183.

Gupta,R.S. (2004). The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Crit Rev Microbiol *30*, 123-143.

Gupta,R.S. (2006). Molecular signatures (unique proteins and conserved indels) that are specific for the epsilon proteobacteria (Campylobacterales). BMC Genomics *7*.

Gupta,R.S. (2009). Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol *[Epub ahead of print]*.

Gupta,R.S. and Gao,B. (2009). Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I). Int J Syst Evol Microbiol *59*, 285-294.

Gupta,R.S. and Gao,B. (2010). Recent Advances in Understanding Microbial Systematics. In Microbial Population Genetics, J.P.Xu, ed. Caister Academic Press).

Gupta,R.S. and Golding,G.B. (1996). The origin of the eukaryotic cell. Trends Biochem Sci *21*, 166-171.

Gupta,R.S. and Griffiths,E. (2002). Critical issues in bacterial phylogeny. Theor Popul Biol *61*, 423-434.

Gupta,R.S. and Griffiths,E. (2006). Chlamydiae-specific proteins and indels: novel tools for studies. Trends Microbiol. *14*, 527-535.

Gupta,R.S. and Lorenzini,E. (2007). Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. BMC Evol Biol *7*.

Gupta,R.S., Mark,P.I., Chandrasekera,C., and Johari,V. (2003). Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues. Int J Syst Evol Microbiol *53*, 1833-1842.

Gupta,R.S. and Mok,A. (2007). Phylogenomics and signature proteins for the alpha Proteobacteria and its main groups. BMC Microbiol *7*.

Hao,W. and Golding,G.B. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res *16*, 636-643.

Harms,U., Weiss,D.S., Gartner,P., Linder,D., and Thauer,R.K. (1995). The energy conserving N5-methyltetrahydromethanopterin:coenzyme M methyltransferase complex from *Methanobacterium thermoautotrophicum* is composed of eight different subunits. Eur J Biochem *228*, 640-648.

Hartmann,G.C., Klein,A.R., Linder,M., and Thauer,R.K. (1996). Purification, properties and primary structure of H-2-forming N-5,N(10)methylenetetrahydromethanopterin dehydrogenase from *Methanococcus thermolithotrophicus*. Arch Microbiol *165*, 187-193.

Hendrickson,E.L., Kaul,R., Zhou,Y., Bovee,D., Chapman,P., Chung,J., de Macario,E.C., Dodsworth,J.A., Gillett,W., Graham,D.E., Hackett,M., Haydock,A.K., Kang,A., Land,M.L., Levy,R., Lie,T.J., Major,T.A., Moore,B.C., Porat,I., Palmeiri,A., Rouse,G., Saenphimmachak,C., Soll,D., Van Dien,S., Wang,T., Whitman,W.B., Xia,Q., Zhang,Y., Larimer,F.W., Olson,M.V., and Leigh,J.A. (2004). Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J Bacteriol *186*, 6956-6969.

Hendrickson,W.A., Horton,J.R., and Lemaster,D.M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO Journal *9*, 1665-1672.

Holm,L., Kaariainen,S., Rosenstrom,P., and Schenkel,A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics *24*, 2780-2781.

Horan,K.L., Freeman,R., Weigel,K., Semret,M., Pfaller,S., Covert,T.C., van Soolingen,D., Leao,S.C., Behr,M.A., and Cangelosi,G.A. (2006). Isolation of the genome sequence strain *Mycobacterium avium* 104 from multiple patients over a 17-year period. J Clin Microbiol *44*, 783-789.

House,C.H. and Fitz-Gibbon,S.T. (2002). Using homolog groups to create a whole-genomic tree of free-living organisms: An update. J Mol Evol *54*, 539-547.

Huber,H., Hohn,M.J., Stetter,K.O., and Rachel,R. (2003). The phylum Nanoarchaeota: Present knowledge and future perspectives of a unique form of life. Res Microbiol *154*, 165-171.

Hugenholtz,P. and Tyson,G.W. (2008). Microbiology - Metagenomics. Nature *455*, 481-483.

Hutchings,M.I., Palmer,T., Harrington,D.J., and Sutcliffe,I.C. (2009). Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. Trends Microbiol *17*, 13-21.

Huynen,M.A. and Bork,P. (1998). Measuring genome evolution. Proc Natl Acad Sci U S A *95*, 5849-5856.

Ikeda,H., Ishikawa,J., Hanamoto,A., Shinose,M., Kikuchi,H., Shiba,T., Sakaki,Y., Hattori,M., and Omura,S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol *21*, 526-531.

Ishitani,R., Nureki,O., Fukai,S., Kijimoto,T., Nameki,N., Watanabe,M., Kondo,H., Sekine,M., Okada,N., Nishimura,S., and Yokoyama,S. (2002). Crystal structure of archaeosine tRNA-guanine transglycosylase. J Mol Biol *318*, 665-677.

Ito,N., Nureki,O., Shirouzu,M., Yokoyama,S., and Hanaoka,F. (2001). Crystallization and preliminary X-ray analysis of a DNA primase from hyperthermophilic archaeon *Pyrococcus horikoshii*. J Biochem (Tokyo) *130*, 727-730.

Iyer,L.M., Koonin,E.V., Leipe,D.D., and Aravind,L. (2005). Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. Nucleic Acids Res *33*, 3875-3896.

Jones,W.J., Nagle,D.P., and Whitman,W.B. (1987). Methanogens and the Diversity of Archaebacteria. Microbiol Rev *51*, 135-177.

Joshi,S.M., Pandey,A.K., Capite,N., Fortune,S.M., Rubin,E.J., and Sassetti,C.M. (2006). Characterization of mycobacterial virulence genes through genetic interaction mapping. Proc Natl Acad Sci U S A *103*, 11760-11765.

Kainth,P. and Gupta,R.S. (2005). Signature proteins that are distinctive of alpha proteobacteria. BMC Genomics *6*.

Kalinowski,J., Bathe,B., Bartels,D., Bischoff,N., Bott,M., Burkovski,A., Dusch,N., Eggeling,L., Eikmanns,B.J., Gaigalat,L., Goesmann,A., Hartmann,M., Huthmacher,K., Kramer,R., Linke,B., McHardy,A.C., Meyer,F., Mockel,B., Pfefferle,W., Puhler,A., Rey,D.A., Ruckert,C., Rupp,O., Sahm,H., Wendisch,V.F., Wiegrabe,I., and Tauch,A. (2003). The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J Biotechnol *104*, 5-25.

Karlin,S., Mrazek,J., and Gentles,A.J. (2003). Genome comparisons and analysis. Curr Opin Struct Biol *13*, 344-352.

Kasting,J.F. and Siefert,J.L. (2002). Life and the evolution of Earth's atmosphere. Science *296*, 1066-1068.

Kawarabayasi,Y., Hino,Y., Horikawa,H., Jin-no,K., Takahashi,M., Sekine,M., Baba,S., Ankai,A., Kosugi,H., Hosoyama,A., Fukui,S., Nagai,Y., Nishijima,K., Otsuka,R., Nakazawa,H., Takamiya,M., Kato,Y., Yoshizawa,T., Tanaka,T., Kudoh,Y., Yamazaki,J., Kushida,N., Oguchi,A., Aoki,K., Masuda,S., Yanagii,M., Nishimura,M., Yamagishi,A., Oshima,T., and Kikuchi,H. (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res *8*, 123-140.

Kawarabayasi,Y., Hino,Y., Horikawa,H., Yamazaki,S., and Haikawa,Y. (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res *6*, 83-101.

Kawashima,T., Amano,N., Koike,H., Makino,S., Higuchi,S., Kawashima-Ohya,Y., Watanabe,K., Yamazaki,M., Kanehori,K., Kawamoto,T., Nunoshiba,T., Yamamoto,Y., Aramaki,H., Makino,K., and Suzuki,M. (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc Natl Acad Sci U S A *97*, 14257-14262.

Kennedy,S.P., Ng,W.V., Salzberg,S.L., Hood,L., and DasSarma,S. (2001). Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res *11*, 1641-1650.

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

Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D., Richardson,D.L., Kerlavage,A.R., Graham,D.E., Kyrpides,N.C., Fleischmann,R.D., Quackenbush,J., Lee,N.H., Sutton,G.G., Gill,S., Kirkness,E.F., Dougherty,B.A., McKenney,K., Adams,M.D., Loftus,B., Peterson,S., Reich,C.I., Mcneil,L.K., Badger,J.H., Glodek,A., Zhou,L.X., Overbeek,R., Gocayne,J.D., Weidman,J.F., McDonald,L., Utterback,T., Cotton,M.D., Spriggs,T., Artiach,P., Kaine,B.P., Sykes,S.M., Sadow,P.W., DAndrea,K.P., Bowman,C., Fujii,C., Garland,S.A., Mason,T.M., Olsen,G.J., Fraser,C.M., Smith,H.O., Woese,C.R., and Venter,J.C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature *390*, 364-&.

Knittel,K., Losekann,T., Boetius,A., Kort,R., and Amann,R. (2005). Diversity and distribution of methanotrophic archaea at cold seeps. Appl Environ Microbiol *71*, 467-479.

Knoll,A.H. (1999). Paleontology - A new molecular window on early life. Science *285*, 1025-1026.

Koonin,E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol *1*, 127-136.

Koonin,E.V. (2009a). Darwinian evolution in the light of genomics. Nucleic Acids Res *37*, 1011-1034.

Koonin,E.V. (2009b). Evolution of genome architecture. Int J Biochem Cell Biol *41*, 298-306.

Koonin,E.V., Makarova,K.S., and Aravind,L. (2001). Horizontal gene transfer in prokaryotes: Quantification and classification. Annu Rev Microbiol *55*, 709-742.

Koonin,E.V. and Wolf,Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res *36*, 6688-6719.

Korbel,J.O., Snel,B., Huynen,M.A., and Bork,P. (2002). SHOT: a web server for the construction of genome phylogenies. Trends Genet *18*, 158-162.

Koski,L.B. and Golding,G.B. (2001). The closest BLAST hit is often not the nearest neighbor. J Mol Evol *52*, 540-542.

Koski,L.B., Morton,R.A., and Golding,G.B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol *18*, 404-412.

Krissinel,E. and Henrick,K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*, 774-797.

Kumar,S., Tamura,K., and Nei,M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform *5*, 150-163.

Kunin,V., Goldovsky,L., Darzentas,N., and Ouzounis,C.A. (2005). The net of life: Reconstructing the microbial phylogenetic network. Genome Res *15*, 954-959.

Kunin,V. and Ouzounis,C.A. (2003). The balance of driving forces during genome evolution in prokaryotes. Genome Res *13*, 1589-1594.

Kunisawa,T. (2007). Gene arrangements characteristic of the phylum Actinobacteria. Antonie Van Leeuwenhoek *92*, 359-365.

Kuo,C.H. and Ochman,H. (2009). The fate of new bacterial genes. FEMS Microbiol Rev *33*, 38-43.

Lake,J.A., Henderson,E., Oakes,M., and Clark,M.W. (1984). Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci U S A *81*, 3786-3790.

Lange,M. and Ahring,B.K. (2001). A comprehensive study into the molecular methodology and molecular biology of methanogenic Archaea. FEMS Microbiol Lett *25*, 553-571.

Lathe,W.C., Snel,B., and Bork,P. (2000). Gene context conservation of a higher order than operons. Trends Biochem Sci *25*, 474-479.

Lawrence,J.G. (2003). Gene organization: Selection, selfishness, and serendipity. Annu Rev Microbiol *57*, 419-440.

Lawrence,J.G. and Hendrickson,H. (2005). Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol *8*, 572-578.

Lawrence,J.G. and Ochman,H. (1997). Amelioration of bacterial genomes: Rates of change and exchange. J Mol Evol *44*, 383-397.

Lechevalier,H.A. and Lechevalier,M.P. (1967). Biology of actinomycetes. Annu Rev Microbiol *21*, 71-100.

Lee,J.H., Karamychev,V.N., Kozyavkin,S.A., Mills,D., Pavlov,A.R., Pavlova,N.V., Polouchine,N.N., Richardson,P.M., Shakhova,V.V., Slesarev,A.I., Weimer,B., and O'Sullivan,D.J. (2008). Comparative genomic analysis of the gut bacterium Bifidobacterium longum reveals loci susceptible to deletion during pure culture growth. BMC Genomics *9*.

Lerat,E., Daubin,V., Ochman,H., and Moran,N.A. (2005). Evolutionary origins of genomic repertoires in bacteria. PLoS Biol *3*, 807-814.

Lichtarge,O., Bourne,H.R., and Cohen,F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lie,T.J. and Leigh,J.A. (2003). A novel repressor of nif and glnA expression in the methanogenic archaeon *Methanococcus maripaludis*. Mol Microbiol *47*, 235-246.

Lin,J. and Gerstein,M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. Genome Res *10*, 808-818.

Lindahl,P.A. and Chang,B. (2001). The evolution of acetyl-CoA synthase. Orig Life Evol Biosph *31*, 403-434.

Livingstone,C.D. and Barton,G.J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci *9*, 745-756.

Ludwig,W. and Klenk,H.P. (2001). Overview:A phylogenetic backbone and taxonomic framework for prokaryotic systamatics. In Bergey's Manual of Systematic Bacteriology, Boone D.R. and Castenholz R.W., eds. (Berlin: Springer-Verlag), pp. 49-65.

Lunin,V.V., Dobrovetsky,E., Khutoreskaya,G., Zhang,R.G., Joachimiak,A., Doyle,D.A., Bochkarev,A., Maguire,M.E., Edwards,A.M., and Koth,C.M. (2006). Crystal structure of the CorA Mg^{2+} transporter. Nature *440*, 833-837.

Lykidis,A., Mavromatis,K., Ivanova,N., Anderson,I., Land,M., DiBartolo,G., Martinez,M., Lapidus,A., Lucas,S., Copeland,A., Richardson,P., Wilson,D.B., and Kyrpides,N. (2007). Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. J Bacteriol *189*, 2477-2486.

Maguire,M.E. (2006). The structure of CorA: a Mg^{2+}-selective channel. Current Opinion in Structural Biology *16*, 432-438.

Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,C.T., Saxman,P.R., Farris,R.J., Garrity,G.M., Olsen,G.J., Schmidt,T.M., and Tiedje,J.M. (2001). The RDP-II (Ribosomal Database Project). Nucleic Acids Res *29*, 173-174.

Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I., and Koonin,E.V. (1999). Comparative genomics of the archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. Genome Res *9*, 608-628.

Makarova,K.S. and Koonin,E.V. (2005). Evolutionary and functional genomics of the Archaea. Curr Opin Microbiol *8*, 586-594.

Mao,C., Cook,W.J., Zhou,M., Koszalka,G.W., Krenitsky,T.A., and Ealick,S.E. (1997). The crystal structure of *Escherichia coli* purine nucleoside phosphorylase: a comparison with the human enzyme reveals a conserved topology. Structure *5*, 1373-1383.

Marri,P.R., Bannantine,J.P., and Golding,G.B. (2006). Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol Rev *30*, 906-925.

Marri,P.R. and Golding,G.B. (2008). Gene amelioration demonstrated: the journey of nascent genes in bacteria. Genome *51*, 164-168.

Matsunaga,T., Okamura,Y., Fukuda,Y., Wahyudi,A.T., Murase,Y., and Takeyama,H. (2005). Complete genome sequence of the facultative anaerobic magnetotactic bacterium *Magnetospirillum* sp strain AMB-1. DNA Res *12*, 157-166.

Matte-Tailliez,O., Brochier,C., Forterre,P., and Philippe,H. (2002). Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol *19*, 631-639.

McAlpine,J.B., Bachmann,B.O., Piraee,M., Tremblay,S., Alarco,A.M., Zazopoulos,E., and Farnet,C.M. (2005). Microbial Genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. J Nat Prod *68*, 493-496.

McAnulla,C., Woodall,C.A., McDonald,I.R., Studer,A., Vuilleumier,S., Leisinger,T., and Murrell,J.C. (2001). Chloromethane utilization gene cluster from *Hyphomicrobium chloromethanicum* strain CM2(T) and development of functional gene probes to detect halomethane-degrading bacteria. Appl Environ Microbiol *67*, 307-316.

Mcleod,M.P., Warren,R.L., Hsiao,W.W.L., Araki,N., Myhre,M., Fernandes,C., Miyazawa,D., Wong,W., Lillquist,A.L., Wang,D., Dosanjh,M., Hara,H., Petrescu,A., Morin,R.D., Yang,G., Stott,J.M., Schein,J.E., Shin,H., Smailus,D., Siddiqui,A.S., Marra,M.A., Jones,S.J.M., Holt,R., Brinkman,F.S.L., Miyauchi,K., Fukuda,M., Davies,J.E., Mohn,W.W., and Eltis,L.D. (2006). The complete genome of *Rhodococcus* sp RHA1 provides insights into a catabolic powerhouse. Proc Natl Acad Sci U S A *103*, 15582-15587.

Michel,H., Behr,J., Harrenga,A., and Kannt,A. (1998). Cytochrome C oxidase: Structure and spectroscopy. Annu Rev Biophys Biomol Struct *27*, 329-356.

Moran,N.A. and Wernegreen,J.J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol *15*, 321-326.

Morris,R.P., Nguyen,L., Gatfield,J., Visconti,K., Nguyen,K., Schnappinger,D., Ehrt,S., Liu,Y., Heifets,L., Pieters,J., Schoolnik,G., and Thompson,C.J. (2005). Ancestral antibiotic resistance in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A *102*, 12200-12205.

Murakami,E. and Ragsdale,S.W. (2000). Evidence for intersubunit communication during acetyl-CoA cleavage by the multienzyme CO dehydrogenase/acetyl-CoA synthase complex from *Methanosarcina thermophila* - Evidence that the beta subunit catalyzes C-C and C-S bond cleavage. J Biol Chem *275*, 4699-4707.

Nakamura,Y., Itoh,T., Matsuda,H., and Gojobori,T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet *36*, 760-766.

Narra,H.P., Cordes,M.H.J., and Ochman,H. (2008). Structural features and the persistence of acquired proteins. Proteomics *8*, 4772-4781.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,L.D., Nelson,W.C., Ketchum,K.A., McDonald,L., Utterback,T.R., Malek,J.A., Linher,K.D., Garrett,M.M., Stewart,A.M., Cotton,M.D., Pratt,M.S., Phillips,C.A., Richardson,D., Heidelberg,J., Sutton,G.G., Fleischmann,R.D., Eisen,J.A., White,O., Salzberg,S.L., Smith,H.O., Venter,J.C., and Fraser,C.M. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature *399*, 323-329.

Ng,W.V., Kennedy,S.P., Mahairas,G.G., Berquist,B., Pan,M., Shukla,H.D., Lasky,S.R., Baliga,N.S., Thorsson,V., Sbrogna,J., Swartzell,S., Weir,D., Hall,J., Dahl,T.A., Welti,R., Goo,Y.A., Leithauser,B., Keller,K., Cruz,R., Danson,M.J., Hough,D.W., Maddocks,D.G., Jablonski,P.E., Krebs,M.P., Angevine,C.M., Dale,H., Isenbarger,T.A., Peck,R.F., Pohlschroder,M., Spudich,J.L., Jung,K.H., Alam,M., Freitas,T., Hou,S.B., Daniels,C.J., Dennis,P.P., Omer,A.D., Ebhardt,H., Lowe,T.M., Liang,R., Riley,M., Hood,L., and DasSarma,S. (2000). Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci U S A *97*, 12176-12181.

Normand,P., Lapierre,P., Tisa,L.S., Gogarten,J.P., Alloisio,N., Bagnarol,E., Bassi,C.A., Berry,A.M., Bickhart,D.M., Choisne,N., Couloux,A., Cournoyer,B., Cruveiller,S., Daubin,V., Demange,N., Francino,M.P., Goltsman,E., Huang,Y., Kopp,O.R., Labarre,L., Lapidus,A., Lavire,C., Marechal,J., Martinez,M., Mastronunzio,J.E., Mullin,B.C., Niemann,J., Pujic,P., Rawnsley,T., Rouy,Z., Schenowitz,C., Sellstedt,A., Tavares,F., Tomkins,J.P., Vallenet,D., Valverde,C., Wall,L.G., Wang,Y., Medigue,C., and Benson,D.R. (2007). Genome characteristics of facultatively symbiotic *Frankia* sp strains reflect host range and host plant biogeography. Genome Res *17*, 7-15.

Novichkov,P.S., Omelchenko,M.V., Gelfand,M.S., Mironov,A.A., Wolf,Y.I., and Koonin,E.V. (2004). Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol *186*, 6575-6585.

Novichkov,P.S., Wolf,Y.I., Dubchak,I., and Koonin,E.V. (2009). Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol *191*, 65-73.

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.

Ochiai,K., Yamanaka,T., Kimura,K., and Sawada,O. (1959). Studies on inheritance of drug resistance between Shigella strains and *Escherichia coli* strains. Nihon Iji Shimpo 34-46.

Ochman,H. (2005). Genomes on the shrink. Proc Natl Acad Sci U S A *102*, 11959-11960.

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

Ohno,M., Shiratori,H., Park,M.J., Saitoh,Y., Kumon,Y., Yamashita,N., Hirata,A., Nishida,H., Ueda,K., and Beppu,T. (2000). *Symbiobacterium thermophilum* gen. nov., sp nov., a symbiotic thermophile that depends on co-culture with a Bacillus strain for growth. Int J Syst Evol Microbiol *50*, 1829-1832.

Oliynyk,M., Samborskyy,M., Lester,J.B., Mironenko,T., Scott,N., Dickens,S., Haydock,S.F., and Leadlay,P.F. (2007). Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. Nat Biotechnol *25*, 447-453.

Olsen,G.J. and Woese,C.R. (1997). Archaeal genomics: An overview. Cell *89*, 991-994.

Olsen,G.J., Woese,C.R., and Overbeek,R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol *176*, 1-6.

Oren,A. (2004). Prokaryote diversity and taxonomy: current status and future challenges. Philos Trans R Soc Lond B Biol Sci *359*, 623-638.

Pace,N.R. (1997). A molecular view of microbial diversity and the biosphere. Science *276*, 734-740.

Payandeh,J. and Pai,E.F. (2006). A structural basis for Mg2+ homeostasis and the CorA translocation cycle. EMBO J *25*, 3762-3773.

Pflugrath,J.W. (1999). The finer things in X-ray diffraction data collection. Acta Crystallogr D Biol Crystallogr *55*, 1718-1725.

Philippe,H., Delsuc,F., Brinkmann,H., and Lartillot,N. (2005a). Phylogenomics. Annu Rev Ecol Syst *36*, 541-562.

Philippe,H., Zhou,Y., Brinkmann,H., Rodrigue,N., and Delsuc,F. (2005b). Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol *5*.

Pugmire,M.J. and Ealick,S.E. (2002). Biochem J. Biochem J *361*, 1-25.

Ragan,M.A. (2001). Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev *11*, 620-626.

Ragan,M.A. and Beiko,R.G. (2009). Lateral genetic transfer: open issues. Phil. Trans. R. Soc. B *364*, 2241-2251.

Rainey,F.A., Ward-Rainey,N.L., Janssen,P.H., Hippe,H., and Stackebrandt,E. (1996). *Clostridium paradoxum* DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences. Microbiology *142*, 2087-2095.

Ranea,J.A. (2006). Genome evolution: Micro(be)-economics. Heredity *96*, 337-338.

Raoult,D., Ogata,H., Audic,S., Robert,C., Suhre,K., Drancourt,M., and Claverie,J.M. (2003). *Tropheryma whipplei* twist: A human pathogenic Actinobacteria with a reduced genome. Genome Res *13*, 1800-1809.

Raskin,D.M., Seshadri,R., Pukatzki,S.U., and Mekalanosl,J.J. (2006). Bacterial genomics and pathogen evolution. Cell *124*, 703-714.

Reeve,J.N., Nolling,J., Morgan,R.M., and Smith,D.R. (1997). Methanogenesis: Genes, genomes, and who's on first? J Bacteriol *179*, 5975-5986.

Rivera,M.C. and Lake,J.A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science *257*, 74-76.

Rivera,M.C. and Lake,J.A. (1996). The phylogeny of *Methanopyrus kandleri*. Int J Syst Bacteriol *46*, 348-351.

Roberts,R.J. (2004). Identifying protein function - A call for community action. PLoS Biol *2*, 293-294.

Rokas,A. and Holland,P.W.H. (2000). Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol *15*, 454-459.

Roller,C., Ludwig,W., and Schleifer,K.H. (1992). Gram-positive bacteria with a high DNA G+C content are characterized by a common insertion within their 23S rRNA genes. J Gen Microbiol *138*, 167-175.

Ruepp,A., Graml,W., Santos-Martinez,M.L., Koretle,K.K., Volker,C., Mewes,H.W., Frishman,D., Stocker,S., Lupas,A.N., and Baumeister,W. (2000). The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature *407*, 508-513.

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

Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V., and Altschul,S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res *29*, 2994-3005.

Schell,M.A., Karmirantzou,M., Snel,B., Vilanova,D., Berger,B., Pessi,G., Zwahlen,M.C., Desiere,F., Bork,P., Delley,M., Pridmore,R.D., and Arigoni,F. (2002). The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. Proc Natl Acad Sci U S A *99*, 14422-14427.

Schleper,C., Jurgens,G., and Jonuscheit,M. (2005). Genomic studies of uncultivated archaea. Nat Rev Microbiol. *3*, 479-488.

Schmidt,H.A., Strimmer,K., Vingron,M., and von Haeseler,A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502-504.

Schopf,J.W. (1978). Evolution of earliest cells. Sci Am *239*, 110-112.

Schrempf,H. (2001). Recognition and degradation of chitin by streptomycetes. Antonie Van Leeuwenhoek *79*, 285-289.

Schueler-Furman,O. and Baker,D. (2003). Conserved residue clustering and protein structure prediction. Proteins *52*, 225-235.

Schumann,P., Kampfer,P., Busse,H.J., and Evtushenko,L.I. (2009). Proposed minimal standards for describing new genera and species of the suborder Micrococcineae. Int J Syst Evol Microbiol *59*, 1823-1849.

Shah,H.N., Olsen,I., Bernard,K., Finegold,S.M., Gharbia,S., and Gupta,R.S. (2009). Approaches to the study of the systematics of anaerobic, Gram-negative, non-sporeforming rods: Current status and perspectives. Anaerobe *Epub ahead of print*.

She,Q., Singh,R.K., Confalonieri,F., Zivanovic,Y., Allard,G., Awayez,M.J., Chan-Weiher,C.C.Y., Clausen,I.G., Curtis,B.A., De Moors,A., Erauso,G., Fletcher,C., Gordon,P.M.K., Heikamp-de Jong,I., Jeffries,A.C., Kozera,C.J., Medina,N., Peng,X., Thi-Ngoc,H.P., Redder,P., Schenk,M.E., Theriault,C., Tolstrup,N., Charlebois,R.L., Doolittle,W.F., Duguet,M., Gaasterland,T., Garrett,R.A., Ragan,M.A., Sensen,C.W., and Van der Oost,J. (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci U S A *98*, 7835-7840.

Shen,Y., Tang,X.F., Matsui,E., and Matsui,I. (2004). Subunit interaction and regulation of activity through terminal domains of the family D DNA polymerase from *Pyrococcus horikoshii.* Biochem Soc Trans *32*, 245-249.

Siew,N., Azaria,Y., and Fischer,D. (2004). The ORFanage: an ORFan database. Nucleic Acids Res *32*, D281-D283.

Siew,N. and Fischer,D. (2003). Twenty thousand ORFan microbial protein families for the biologist? Structure. (Camb.) *11*, 7-9.

Singh,B. and Gupta,R.S. (2009). Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol Genet Genomics *281*, 361-373.

Slesarev,A.I., Mezhevaya,K.V., Makarova,K.S., Polushin,N.N., Shcherbinina,O.V., Shakhova,V.V., Belova,G.I., Aravind,L., Natale,D.A., Rogozin,I.B., Tatusov,R.L., Wolf,Y.I., Stetter,K.O., Malykh,A.G., Koonin,E.V., and Kozyavkin,S.A. (2002). The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A *99*, 4644-4649.

Smith,I. (2003). *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev *16*, 463-496.

Snel,B., Bork,P., and Huynen,M.A. (1999). Genome phylogeny based on gene content. Nat Genet *21*, 108-110.

Snel,B., Bork,P., and Huynen,M.A. (2002). Genomes in flux: The evolution of archaeal and proteobacterial gene content. Genome Res *12*, 17-25.

Snel,B., Huynen,M.A., and Dutilh,B.E. (2005). Genome trees and the nature of genome evolution. Annu Rev Microbiol *59*, 191-209.

Soliveri,J.A., Gomez,J., Bishai,W.R., and Chater,K.F. (2000). Multiple paralogous genes related to the *Streptomyces coelicolor* developmental regulatory gene whiB are present in Streptomyces and other actinomycetes. Microbiology-Uk *146*, 333-343.

Stackebrandt,E. (2006). Defining Taxonomic Ranks. In The prokaryotes, M.Dworkin, S.Falkow, E.Rosenberg, K.-H.Schleifer, and E.Stackebrandt, eds. Springer), pp. 29-57.

Stackebrandt,E., Rainey,F.A., and WardRainey,N.L. (1997). Proposal for a new hierarchic classification system, Actinobacteria classis nov. Int J Syst Bacteriol. *47*, 479-491.

Stackebrandt,E. and Schumann,P. (2006). Introduction to the taxonomy of actinobacteria. In Prokaryotes, M.Dworkin, ed. Springer New York), pp. 297-321.

Steel,M. and Penny,D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol *17*, 839-850.

Stinear,T.P., Seemann,T., Harrison,P.F., Jenkin,G.A., Davies,J.K., Johnson,P.D.R., Abdellah,Z., Arrowsmith,C., Chillingworth,T., Churcher,C., Clarke,K., Cronin,A., Davis,P., Goodhead,I., Holroyd,N., Jagels,K., Lord,A., Moule,S., Mungall,K., Norbertczak,H., Quail,M.A., Rabbinowitsch,E., Walker,D., White,B., Whitehead,S., Small,P.L.C., Brosch,R., Ramakrishnan,L., Fischbach,M.A., Parkhill,J., and Cole,S.T. (2008). Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of Mycobacterium tuberculosis. Genome Res *18*, 729-741.

Strong,M., Sawaya,M.R., Wang,S.S., Phillips,M., Cascio,D., and Eisenberg,D. (2006). Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A *103*, 8060-8065.

Sutcliffe,I.C. and Harrington,D.J. (2004). Lipoproteins of *Mycobacterium tuberculosis*: an abundant and functionally diverse class of cell envelope components. FEMS Microbiol Rev *28*, 645-659.

Syvanen,M. (1985). Cross-Species Gene-Transfer - Implications for A New Theory of Evolution. J Theor Biol *112*, 333-343.

Takarada,H., Sekine,M., Kosugi,H., Matsuo,Y., Fujisawa,T., Omata,S., Kishi,E., Shimizu,A., Tsukatani,N., Tanikawa,S., Fujita,N., and Harayama,S. (2008). Complete genome sequence of the soil actinomycete *Kocuria rhizophila*. J Bacteriol *190*, 4139-4146.

Tamura,K., Dudley,J., Nei,M., and Kumar,S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol *24*, 1596-1599.

Teichmann,S.A. and Mitchison,G. (1999). Is there a phylogenetic signal in prokaryote proteins? J Mol Evol *49*, 98-107.

Tersteegen,A. and Hedderich,R. (1999). *Methanobacterium thermoautotrophicum* encodes two multisubunit membrane-bound [NiFe] hydrogenases - Transcription of the operons and sequence analysis of the deduced proteins. Eur J Biochem *264*, 930-943.

Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S., Deboy,R.T., Davidsen,T.M., Mora,M., Scarselli,M., Ros,I.M.Y., Peterson,J.D., Hauser,C.R., Sundaram,J.P., Nelson,W.C., Madupu,R., Brinkac,L.M., Dodson,R.J., Rosovitz,M.J., Sullivan,S.A., Daugherty,S.C., Haft,D.H., Selengut,J., Gwinn,M.L., Zhou,L.W., Zafar,N., Khouri,H., Radune,D., Dimitrov,G., Watkins,K., O'Connor,K.J.B., Smith,S., Utterback,T.R., White,O., Rubens,C.E., Grandi,G., Madoff,L.C., Kasper,D.L., Telford,J.L., Wessels,M.R., Rappuoli,R., and Fraser,C.M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". Proc Natl Acad Sci U S A *102*, 13950-13955.

Thomas,C.M. and Nielsen,K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol *3*, 711-721.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F., and Higgins,D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res *25*, 4876-4882.

Turnbaugh,P.J. and Gordon,J.I. (2008). An invitation to the marriage of metagenomics and metabolomics. Cell *134*, 708-713.

Ueda,K., Yamashita,A., Ishikawa,J., Shimada,M., Watsuji,T., Morimura,K., Ikeda,H., Hattori,M., and Beppu,T. (2004). Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res *32*, 4937-4944.

Vagin,A.A., Steiner,R.A., Lebedev,A.A., Potterton,L., McNicholas,S., Long,F., and Murshudov,G.N. (2004). REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D Biol Crystallogr *60*, 2184-2195.

Van de Peer,Y. and De Wachter R (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci *10*, 569-570.

van Passel,M.W.J., Marri,P.R., and Ochman,H. (2008). The emergence and fate of horizontally acquired genes in Escherichia coli. PLoS Comput Biol *4*.

van de Wijngaard,W.M.H., Creemers,J., Vogels,G.D., and Vanderdrift,C. (1991). Methanogenic pathways in *Methanosphaera stadtmanae*. FEMS Microbiol Lett *80*, 207-212.

Ventura,M., Canchaya,C., Del Casale,A., Dellaglio,F., Neviani,E., Fitzgerald,G.F., and van Sinderen,D. (2006a). Analysis of bifidobacterial evolution using a multilocus approach. Int J Syst Evol Microbiol *56*, 2783-2792.

Ventura,M., Canchaya,C., Fitzgerald,G.F., Gupta,R.S., and van Sinderen,D. (2006b). Genomics as a means to understand bacterial phylogeny and ecological adaptation: the case of bifidobacteria. Antonie Van Leeuwenhoek.

Ventura,M., Canchaya,C., Tauch,A., Chandra,G., Fitzgerald,G.F., Chater,K.F., and van Sinderen,D. (2007). Genomics of Actinobacteria: Tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev *71*, 495-548.

Ventura,M., van Sinderen,D., Fitzgerald,G.F., and Zink,R. (2004). Insights into the taxonomy, genetics and physiology of bifidobacteria. Antonie Van Leeuwenhoek *86*, 205-223.

Walsh,D.A. and Doolittle,W.F. (2005). The real 'domains' of life. Curr Biol *15*, R237-R240.

Ward,A.C. and Bora,N. (2006). Diversity and biogeography of marine actinobacteria. Curr Opin Microbiol *9*, 279-286.

Warner,K.L., Larkin,M.J., Harper,D.B., Murrell,J.C., and McDonald,I.R. (2005). Analysis of genes involved in methyl halide degradation in *Aminobacter lissarensis* CC495. FEMS Microbiol Lett *251*, 45-51.

Warren,R., Hsiao,W.W.L., Kudo,H., Myhre,M., Dosanjh,M., Petrescu,A., Kobayashi,H., Shimizu,S., Miyauchi,K., Masai,E., Yang,G., Stott,J.M., Schein,J.E., Shin,H., Khattra,J., Smailus,D., Butterfield,Y.S., Siddiqui,A., Holt,R., Marra,M.A., Jones,S.J.M., Mohn,W.W., Brinkman,F.S.L., Fukuda,M., Davies,J., and Eltis,L.D. (2004). Functional characterization of a catabolic plasmid from polychlorinated-biphenyl-degrading *Rhodococcus* sp strain RHA1. J Bacteriol *186*, 7783-7795.

Waters,E., Hohn,M.J., Ahel,I., Graham,D.E., Adams,M.D., Barnstead,M., Beeson,K.Y., Bibbs,L., Bolanos,R., Keller,M., Kretz,K., Lin,X.Y., Mathur,E., Ni,J.W., Podar,M., Richardson,T., Sutton,G.G., Simon,M., Soll,D., Stetter,K.O., Short,J.M., and Noordewier,M. (2003). The genome of *Nanoarchaeum equitans*: Insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A *100*, 12984-12988.

Welch,R.A., Burland,V., Plunkett,G., Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J., Stroud,D., Mayhew,G.F., Rose,D.J., Zhou,S., Schwartz,D.C., Perna,N.T., Mobley,H.L.T., Donnenberg,M.S., and Blattner,F.R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A *99*, 17020-17024.

Whelan,S. and Goldman,N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol *18*, 691-699.

Woese,C.R. (1987). Bacterial Evolution. Microbiol Rev *51*, 221-266.

Woese,C.R. (2006). How We Do, Don't and Should Look at Bacteria and Bacteriology. In The Prokaryotes, M.Dworkin, S.Falkow, E.Rosenberg, K.-H.Schleifer, and E.Stackebrandt, eds. Springer), pp. 3-23.

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 U S A *87*, 4576-4579.

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

Wolf,Y.I., Rogozin,I.B., Grishin,N.V., and Koonin,E.V. (2002). Genome trees and the Tree of Life. Trends Genet *18*, 472-479.

Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S., and Koonin,E.V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res *11*, 356-372.

Xiong,J. (2006). Photosynthesis: what color was its origin? Genome Biol *7*.

Xiong,Y., Li,F., Wang,J.M., Weiner,A.M., and Steitz,T.A. (2003). Crvstal structures of an archaeal class ICCA-adding enzyme and its nucleotide complexes. Mol Cell *12*, 1165-1172.

Yang,S., Doolittle,R.F., and Bourne,P.E. (2005). Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A *102*, 373-378.

Yang,Z.H. and Nielsen,R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol *17*, 32-43.

Zazopoulos,E., Huang,K.X., Staffa,A., Liu,W., Bachmann,B.O., Nonaka,K., Ahlert,J., Thorson,J.S., Shen,B., and Farnet,C.M. (2003). A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat Biotechnol *21*, 187-190.

Zhi,X., Li,W., and Stackebrandt,E. (2009). An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class Actinobacteria, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int J Syst Evol Microbiol *59*, 589-608.