THE ROLE OF LATERAL GENE TRANSFER IN PROKARYOTIC EVOLUTION

by

Yan Boucher

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2003

Your file   Votre référence
ISBN: 0-612-83710-6
Our file   Notre référence
ISBN: 0-612-83710-6

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOCHEMISTRY & MOLECULAR BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “The Role of Lateral Gene Transfer in Prokaryotic Evolution” by Yan Boucher in partial fulfillment for the degree of Doctor of Philosophy.

Dated: September 16, 2003

External Examiner: [redacted]

Research Supervisor:

Examining Committee:

Departmental Representative:

DALHOUSIE UNIVERSITY

DATE: September 16, 2003

AUTHOR: Yan Boucher

TITLE: The Role of Lateral Gene Transfer in Prokaryotic Evolution

DEPARTMENT OR SCHOOL: Biochemistry and Molecular Biology

DEGREE: Doctor of Philosophy CONVOCATION: May YEAR: 2004

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

[redacted]

Signature of Author

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

“*Its remains to be seen how far this conception will throw light on the obscure and difficult questions of biological classification, and on those facts of geological succession which are most difficult to reconcile with the usual view of all organisms whatever having originated from a single almost infinitely remote source.*”

Alfred R. Wallace (1872) about spontaneous generation and heterogenesis

Table of Contents

List of Illustrations

List of Tables

Abstract

Lateral gene transfer (LGT) is the propagation of genetic material through ways other than inheritance from a progenitor. The work presented here is part of the recent effort to better understand the importance of LGT in the evolution of the physiological properties harbored by bacteria and archaea. Since LGT is a phenomenon that potentially affects all prokaryotic functions, two main processes were chosen as model systems to study it.

Isoprenoid biosynthesis is a universal and well-characterized metabolic function. The combined genomic, phylogenetic and biochemical data accumulated points toward a critical role for LGT in the evolution of the mevalonate and methylerythritol pathways for the biosynthesis of the isoprenoid precursor isopentenyl diphosphate. The genes encoding the enzymes of pathways involved in the synthesis of archaeal isoprenoid lipids from this precursor were also investigated. This isoprenoid lipid biosynthesis apparatus was found to have evolved through a combination of evolutionary processes, including LGT.

Informational cellular processes (transcription/translation related), to which the ribosome is central, are thought to be more refractory to LGT than metabolic processes. This lead to the choice of ribosomal RNA (rRNA) genes as the second model system to study the importance of LGT. We isolated and sequenced all of the rRNA genes from two species of archaea known to display multiple heterogeneous rRNA gene copies. Clear recombination patterns were indeed found in some of the genes, suggesting a role for LGT in the creation of within species rRNA genes heterogeneity.

During the study of those two model systems for LGT, about 70 different genes were cloned and sequenced (>120kb of DNA). Phylogenetic analysis of those data and data from public databases reveals that LGT has played critical role in evolution of key physiological functions. The vertical inheritance (clonal) model of prokaryotic evolution cannot be easily modified to accommodate recurrent LGT. A new theoretical model of prokaryotic evolution based on LGT as a background force is therefore proposed to replace the old model.

List of Abbreviations and Symbols Used

General

gDNA	genomic DNA
ML	maximum likelihood
NJ	neighbor joining
PCR	polymerase chain reaction
SSH	subtractive (suppressive) hybridization
Homologs	descend from a common ancestor
Orthologs	same gene in different species (divided by speciation)
Paralogs	duplicate genes in the same species (divided by gene duplication)

Isoprenoid biosynthesis

MVA	mevalonate
MEP	2C-methyl-D-erythritol-4-phosphate
HMG	3-hydroxy-3-methylglutaryl
CoA	Coenzyme A
IPP	isopentenyl diphosphate
DMAPP	dimethylallyl diphosphate

rRNA genes

rRNA	ribosomal RNA
SSU	small ribosomal subunit
LSU	large ribosomal subunit
ITS	internal transcribed spacer

Acknowledgements

First of all, I want to thank Ford Doolittle, as none of this would have been possible without him. He has been the perfect supervisor, letting me enough loose so I can be creative but also always being there to answer questions (regardless of these questions being very trivial or very profound). His lab is indeed a special place, where taking a break from work to have philosophical discussions is not only OK, but encouraged.

Many people in the lab provided me with not only technical help but also great company over the years. Early on David Faguy started me up as a graduate student (and even taught me how to do PCR!). John Archibald also gave me a lot of explanations about molecular evolution and without him I could never have understood the concepts of paralogy, orthology and homology.

Just after I arrived in Halifax, I was quite glad when Joel proposed me to move in with him and Mike, providing me with great company when I needed it the most. He has been a great roommate and friend since and I really enjoyed living with him. There is nothing like the two of us smoking some Cubans and drinking port over an evening.

Camilla has been a bit of a second supervisor for me, always having great insights about experimental problems and new ideas and techniques (and aquavit) to share with me. Christophe was my true mentor for anything phylogenetic, being able to answer every single question I ever had. I loved to have the company of a fellow francophone in the lab. There are some things you just can't say in English.

Later in the course of my Ph.D. several great people came to work for Ford. I wish Thane had come earlier to the lab, as he is an extraordinary person, in every sense of the word. Always worth a laugh and always willing to lend a helping hand. Another very special person that came in the lab in the late years is Maureen. I enjoyed very much the discussions I had with her (sometimes non-stop for hours) and her great and undying enthusiasm. Dave Walsh came has the latest addition to the lab and I liked the guy from the first minute I met him. This great Saltspring boy brought some of the west coast spirit with him and has shared my enthusiasm for thinking about microbes ever since he got in Halifax.

Last but not least I want to thank Rebecca, for relieving me with some nice Australian sun in the middle of my last Haligonian winter. She was also a great support in writing this thesis and getting me through the end of my Ph.D.

Introduction

"...if the facts of Archebiosis and heterogenesis are true, and all the lower forms of life are continually being produced de novo, under the influence of unknown laws of development, then we may fairly conclude that, once the earth had arrived at conditions favourable to the production of living organic matter, the process of development would be rapid, and an immense variety of low forms of animals and vegetables would soon people it." Alfred R. Wallace (Wallace, 1872)

Some Darwinists in the late 19th century thought that spontaneous generation (the continuous creation of new microbial life and rapid transformations by heterogenesis) could greatly speed up the rate at which evolutionary change occurs (Wallace, 1872). They envisioned that it could provide an answer to the challenge of William Thompson's claim that earth had not existed in a cooled state long enough to have allowed evolution to occur as described by Darwin in *The Origin of Species* (Farley, 1972a).

Although all biologists now stand by Thomas Huxley's statement that such spontaneous generation only happened in the distant past (Farley, 1972a), we are still amazed by the rapidity with which prokaryotes can adapt to new environments. Simultaneous to the discovery of DNA as the genetic material, although it might not have been fully realized at the time, was the discovery of mechanisms that accelerate the evolution of prokaryotes. Indeed, early experiments by Avery, McLeod and McCarty (1944) on the conversion of non-virulent *Pneumococcus* to a virulent form in the presence of DNA from a virulent strain of the same species, were using the natural capacity of pneumococci to take up DNA from their environment (Avery *et al.*, 1944).

Such uptake of free DNA is now termed transformation and is one of the three main modes by which DNA is exchanged between microorganisms in a non-vertical manner (the others being conjugation and transduction).

This phenomenon of non-vertical DNA transfer, now referred to as lateral or horizontal gene transfer, has come to be recognized as a major force in the evolution of prokaryotes. From being considered a marginal phenomenon appearing only in specific set of circumstances, LGT is now forcing us to reinterpret the way we think about prokaryotes.

This thesis, as a part of the recent movement in trying to understand the real extent and importance of non-vertical genetic exchange among prokaryotes, bears largely on identifying events of LGT and assessing the impact they had on adaptation and evolution of the recipient organisms. This was done mostly through the acquisition and analysis of extensive DNA sequence datasets for genes encoding products involved in a variety of physiological functions.

Chapter 1 is an historical account of the different theories of microbial evolution and how LGT came to take an important place within those theories. This history starts from the very first views on the evolution of prokaryotes (which coincide with theories about spontaneous generation), goes through the current positions and ends proposing further changes in how we view the evolution of Archaea and Bacteria.

Chapter 2 presents phylogenetic analyses of genes whose products take part in key physiological processes in prokaryotes, such as photosynthesis, aerobic respiration, nitrogen fixation, sulfate reduction and methylotrophy. We look at the phylogenetic distribution, genetic organization and biochemistry of those physiological functions to try to understand how LGT affected their evolution.

Chapter 3 presents an in-depth study of the role of LGT in the evolution of one particular metabolic process: the biosynthesis of isoprenoids. It is divided into three sections. The first section concerns the evolution of the two analogous pathways responsible for the synthesis of the universal isoprenoid precursor isopentenyl diphosphate, the mevalonate and methylerythritol pathways. The second section focuses on the evolution of the enzyme catalyzing the first committed step of the mevalonate pathway, HMG-CoA reductase, which has been particularly affected by LGT. The third and last section bears on the origins and evolution of the apparatus responsible for the biosynthesis of a particular class of isoprenoid compounds, the isoprenoid lipids that form the cellular membrane of archaea.

Chapter 4 looks at a specific type of LGT, homologous recombination, and at the role it played in the origins and maintenance of multiple heterogeneous rRNA genes in several taxa of extremely halophilic archaea. This is highly complementary to phylogenetic studies of metabolic genes when looking for LGT, since genes involved in informational processes (related to the replication or transcription of DNA or the translation of its RNA transcript), as opposed to operational (metabolic) processes, are thought to be less frequently involved in LGT (Jain *et al.*, 1999).

Definition of LGT

It is very important to define lateral gene transfer accurately, since the term usually encompasses a range of different phenomena, all having in common that some

DNA fragment from an exogenous source is assimilated in the genome of an organism. It is termed lateral (or horizontal) gene transfer (LGT), as opposed to the vertical transmission of genetic material from a parent cell. **LGT is an outcome**: the fixation of a foreign gene in the genome of an organism.

DNA can be integrated in a genome through homologous recombination. This includes all processes requiring a segment of at least 20 identical base pairs between the recipient DNA and the recombining fragment {such as RecA-mediated homologous recombination, (Smith, 1988)}. The frequency of homologous recombination falls off rapidly as sequence identity decreases between the two DNA molecules to be recombined. Nevertheless, homologous recombination can occur between genes of rather distantly related genomes (different "species", even), especially if they are very slow evolving genes (for example RNA polymerase or rRNA genes).

Similarity of the laterally transferred DNA fragment with the receiving genome is not always required. Genes brought in on stably maintained plasmids by conjugation or fragments integrated by illegitimate recombination (e.g. topoisomerase errors or nicking at a plasmid's origin of replication) are not required to have any similarity to the host's DNA (Michel, 1999). Other processes only require very short stretches of similarity, which are not necessarily part of the acquired gene itself (inverted repeats from transposable elements, site-specific elements of lysogenic bacteriophages or spontaneous rearrangements in chromosomes occurring in short homologous sequences of 3 to 20 bp long).

Processes through which LGT occurs include: (1) **transformation**, transfer of DNA fragments involving free DNA and its penetration in the host, (2) **transduction**, in which the DNA transfer is mediated by a virus and (3) **conjugation**, in which the transfer involves cell-to-cell contact and a conjugative plasmid (or transposon).

Here, I am concerned mostly with LGT as an outcome. Regardless of how a gene was acquired, we can study the impact of its presence in the host or, as will most often be the case here, its continued presence in the descendants of the initial host.

Methods to detect LGT

Many methods are used to detect LGT events. Each of these targets specific types of events (occurring across large or short phylogenetic distances, recent or ancient, involving a particular process or genetic structure). I will here give a short description of the most common methods, along with some of their main characteristics. For more substantial descriptions see Ragan (2001a) and Koonin (2001).

Unusual gene distribution. Gene distribution can be a very strong indicator of LGT. If a gene is only present in two very distant lineages (for example 2Fe-2S ferredoxins are only found in extremely halophilic archaea and cyanobacteria), it is much more likely that it was laterally transferred from one lineage to the other than lost in all other intervening groups. Such gene distribution arguments are, however, based solely on parsimony. Nature doesn't always follow the most parsimonious path, so one has to be careful not to over-interpret evidence for LGT coming from an unusual gene distribution.

Unusual ranking of similarity among homologs. Similarity between a query gene and sequences in public databases is often determined using a BLAST-based tool (Altschul *et al.*, 1997). When searching public databases with BLAST, if a protein or

DNA sequence query "hits" a homolog from a distant taxa with a higher score than it hits homologs from close relatives, there is a possibility that this gene was acquired by LGT. This is a very quick method, but there are several potential problems with it. It is very difficult to determine a BLAST score threshold to retrieve only orthologs of the query (and even harder to find a threshold that will work for multiple gene families). Also, large-scale tests indicate that as many as 40% of best BLASTP matches may target sequences other than the phylogenetically nearest neighbor (Koski and Golding, 2001). This method is therefore not very sensitive and is only effective to detect transfers across large phylogenetic distances (orders or domains).

Atypical composition. Anomalous nucleotide composition (codon usage bias, GC content, trinucleotide frequencies) is widely used to detect LGT events but is only applicable to recent transfers (Ochman *et al.*, 2000). Indeed, genes acquired from other organisms with a different genomic nucleotide composition will be rapidly modified to match the nucleotide composition of their new host (over a time scale of a few million to a few hundred million years). Genes can also have compositional biases for other reasons than an exogenous origin (expression levels, strand biases, etc). Furthermore, genes from compositionally similar donors will certainly be missed.

Presence of an operon structure. According to the "selfish operon" theory of Lawrence and Roth (Lawrence and Roth, 1996), the operon structure would be created and maintained by LGT. This seems an interesting possibility, especially when a particular operon is only found in distant taxa (and more so if the gene order is conserved). However, operons could also be formed as a consequence of selective pressures other than LGT (for example coregulation of the expression of multiple interdependent genes).

Presence of elements associated with particular LGT processes. The identification of site-specific lysogenic phage insertion sequences, transposon insertion elements or plasmid remnants flanking a gene is very direct evidence for an exogenous origin. These elements, however, are quite frequently highly mutated and can be difficult to identify.

Difference in genome content among close relatives. When two closely related prokaryotes differ greatly in their genome content, it is likely that many of the genes found in only one of these neighbors were acquired through LGT. One of the most striking example is *E. coli* 0157:H7, which contains 1387 genes absent from *E. coli* K12, which in turn contains 528 genes absent from the former (Perna *et al.*, 2001).

Incongruence among phylogenetic trees. LGT can result in the anomalous placement of a particular taxon in reference to an "organismal" tree. For example, a bacterial gene grouping strongly amongst archaeal homologs, well away from its bacterial orthologs, is likely to have been acquired by its bacterial host through LGT from an archaeon. Phylogenetic evidence can be very convincing if it comes from a statistically well-supported tree of a single orthologous family. However, methodological artifacts and paralogy cannot always be ruled out and can make the interpretation of a tree fairly difficult (Smith *et al.*, 1992).

I will discuss a large number of potential instances of LGT in this thesis. Doing so, I will often conclude that a gene was transferred from X to Y. This is in fact a shortcut to say that a gene was transferred from a relative of X to an ancestor of Y (as it is extremely unlikely that the organisms from which these genes were sampled were themselves involved in the transfer event). Each inference of LGT in this thesis is an informed decision based on one or more of the lines of evidence mentioned above. If only one element suggesting LGT is found, the case for LGT can still be solid, as long as this one line of evidence is very strong (for example an 80% DNA sequence identity between archaeal and bacterial homologs alone may be sufficient to make a strong case for LGT). When it comes to a phenomenon like LGT, however, it is very difficult to make a completely positive affirmation and most often, assertions about LGT can only be accompanied by a degree of confidence in the particular case being presented. Specifying the direction of a transfer (from X to Y) is also an educated guess. Although disagreement is possible with some of my judgements on the occurrence or direction of particular instances of LGT, it has no special bearing on the overall evidence for LGT, which comes from numerous and varied sources.

Chapter 1: The introduction of lateral gene transfer in theories of microbial evolution

This chapter includes work published in Y. Boucher, C.L. Nesbø and W.F. Doolittle (2001) Microbial genomes: dealing with diversity. *Current Opinion in Microbiology* Jun;4(3):285-9, C.L. Nesbø, Y. Boucher and W.F. Doolittle (2001), Comparative Genomics of Four Archaea: Is there a Core of Non-Transferable Proteins*?* *Journal of Molecular Evolution* Oct-Nov;53(4-5):340-50 and W.F. Doolittle, Y. Boucher, C.L. Nesbø, C.J. Douady, J.O. Andersson and A.J. Roger (2003) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philosophical Transactions of the Royal Society of London B, Series B: Biological Sciences* Dec;358:39-58.

The historical facts mentioned in this chapter come from a number of articles and books on the history of microbiology and the classification of microbes. References to these manuscripts are not always inserted in the text to avoid repetition. Here is a list of the relevant works: (Brock, 1990; Buchanan, 1925; Farley, 1972a, b; Mazumdar, 1995; Stanier and van Niel, 1941; Strick, 2000; Summers, 2000; Winslow *et al.*, 1920).

Theories of prokaryotic evolution in the pre-genomic era

Spontaneous generation

Most great discoveries happen shortly after technical innovations and the field of microbiology is no exception. Microbes were first observed by Antonie Van Leeuwenhoek in 1676, naming them "animalcules" (little animals). He did not extend his observations beyond descriptions. A classification of prokaryotes, essential to formulate evolutionary hypotheses, would have to wait for much later. At the time microorganisms were first observed, the debate on spontaneous generation was raging. In the late 17th century, Francesco Redi and others had quite convincingly demonstrated that it was unlikely that macroscopic animals appeared by spontaneous generation from organic matter (flies and maggots from meat, mice from grain). However, it was still believed that smaller life could arise from organic particles, which possessed some "indestructible vital principle" (in part through the work of John Needham around 1748). It is not before experiments by Louis Pasteur (1861) that many were persuaded that present-day spontaneous generation was impossible (Summers, 2000). After the publication of the *Origin of Species* by Charles Darwin in 1859, which provided a completely naturalistic method of explaining life, the debate on spontaneous generation became more sophisticated. Most of the participants made a crucial distinction between "heterogenesis" and "archebiosis". The term "heterogenesis" was used to describe the process of living things appearing from degenerating material, which itself was derived from previously living things (meat or vegetable infusions). Archebiosis, on the other hand, is the process of living things appearing from inorganic starting materials. Despite finding Pasteur's work highly suggestive, Darwin did still keep an open mind on the possibility of spontaneous generation all through the 1860s and early 1870s (Farley, 1972a). Some Darwinists believed that spontaneous generation was necessary, along with evolution, for any naturalistic science to be consistent and for continuity in nature not to be violated. Alfred Russell Wallace himself was convinced by some arguments presented by one of the most pro-eminent evolutionist supporter of spontaneous generation, H.C. Bastian. In a review of Bastian's best known book *The Beginnings of Life* (1872), Wallace pointed out one of the consequences spontaneous generation and heterogenesis would have on evolution: "*...continuous creation of new microbial life and rapid transformations by heterogenesis could greatly speed up the rate at which evolutionary change occurred.*" This could provide an answer to William Thompson's challenge to Darwinian evolution. Thompson claimed that earth had not existed in a cooled state long enough to allow for the current biological diversity to evolve as described by Darwin in *The Origin of Species.* The incapacity of the spontaneous generation side of the debate to convincingly demonstrate the phenomenon, along with the desire to dissociate Darwinism from the controversy associated with this debate, led Thomas Huxley to define the Darwinian party line position on the origin of life question: spontaneous generation in the present day had been conclusively disproved by Pasteur, but in the earth's distant past scientific naturalism surely called for it to have happened at least one original time. Darwin's famous remark gave a more explicit and concrete basis for this argument:" *It is often said that all the conditions for the first production of living organisms are now present, which could ever have been present. But if (and oh what a big if) one could conceive some **warm little pond** with the right amount of ammonia and phosphoric salts,--light, heat, electricity, etc. present, thus a protein compound was* *chemically formed, ready to undergo itself such complex changes, at the present day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures had formed.*" This view of a world without continuous creation of new organisms but constantly changing through natural processes was a prerequisite to an attempt at a natural (phylogenetic) classification of microbes. Darwin, in the *Origin of Species,* also introduced the concept of a tree to illustrate evolution and phylogenetic relationships: "*limbs divided into great branches… were themselves once, when the tree was small, budding twigs; and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups*". Ernst Haeckel was one of the first to apply the Darwinian analogy to classification of living organisms, depicting their relationship in a tree that included not only two (animals and plants) but also a third kingdom for microorganisms- the Protista (Mazumdar, 1995).

Taxonomic classification based on morphology and physiology

The spontaneous generation debate had not deterred naturalists from recognizing, describing and classifying animalcules on a taxonomical basis. Otto Friderich Müller was one of the first naturalists to interest himself in the matter in the mid-18th century, dividing animalcules into two groups, depending on whether they had external organs or not. Improvement in microscopes allowed for more detailed observations after the 1820's. In 1838, Carl Gustav Ehrenberg published a classification of "Infusoria" (a category that included bacteria, protozoa, rotifers and diatoms) under 22 families, three of which included forms recognizable as bacteria. However, he believed that Infusoria were all animals. In an important work (1854), Ferdinand Cohn suggested that some bacteria should belong to the vegetable kingdom instead {as Mycophyceae or "water fungus", (Summers, 2000)}. While Cohn was arguing for a taxonomy of bacteria which would position them within the known biological realm, Carl Von Nägeli, who was an important figure at the time, maintained that microscopical fungi could arise spontaneously from animal or plant precursors (heterogenesis). There would be little or no constancy in form, a belief that became known as pleomorphism. This belief was derived from observation that many fungi seemed to change form depending on environmental conditions. Pleomorphism was intensely debated and investigated in the late 19th century. The view that each kind of organism had a definite and fixed form became known as monomorphism and was defended by Cohn and his followers. With major developments in morphological methods, pure culture techniques and knowledge of biochemical activities of bacteria, new classification schemes where physiology played a major role were elaborated {Orla-Jensen and Bergey, early 20th century, (Buchanan, 1925)}. Life cycles were also recognized in fungi, putting to rest the debate on pleiomorphism.

Phylogenetic vs. Taxonomic classification

The attention later switched toward the question of whether bacterial classification should be a taxonomy (empirical system) or a phylogeny (natural system). In an article published in the Journal of Bacteriology, Stanier and Van Niel (1941) argued for a natural system, being in disagreement with the taxonomic approach adopted by Bergey and the American Association for Systematic Bacteriology (Winslow *et al.*, 1920). The discovery of DNA as the vehicle for heredity, the development of DNA hybridization techniques (1960's) and of DNA sequencing (1977) would be required to make the option of looking at prokaryotic evolution from a phylogenetic perspective more viable than it previously had been.

The advent of molecular phylogenetics

At this point the fields of molecular biology and biochemistry have become the main players in defining the paradigm in prokaryotic evolution. Zuckerkandl and Pauling first suggested the use of molecular sequences (protein or DNA) to reconstruct organismal phylogeny and argued that they are superior than traditional morphological and physiological characters (Zuckerkandl and Pauling, 1965). They also described what would become the holy grail of molecular systematists, uncovering the true tree of life: "*it will be determined to what extent the phylogenetic tree, as derived from the molecular data in complete independence from the results of organismal biology, coincides with the phylogenetic tree constructed on the basis of organismal biology. If the two phylogenetic trees are mostly in agreement with respect to the topology of branching, the best available single proof of the reality of macro-evolution will be furnished*". This represents well the Darwinian view of prokaryotic evolution at the time, that is, the clonal and tree-like evolution of prokaryotes. Molecular biologists then started trying to define those organismal relationships using at first protein sequences, mostly cytochromes and ferredoxins (1965-1977) (Doolittle, 1999b). After development of DNA sequencing techniques in 1977, DNA sequences started to be used as well. What Zuckerkandl and Pauling did not consider is that different molecules (genes) could yield different trees, and this without the presence of methodological artifacts. This could have been foreseen, as one of the experiments that showed DNA as the genetic material (Avery *et al.*, 1944) relied on transformation, one of the main modes of LGT. Also, in the late 1950s, many investigators observed transfer of antibiotic resistance from *Shigella* to *E. coli*. In 1963, Watanabe suggested that the genetic determinants (RTFs) of this resistance to drugs were in fact plasmids, similar to the F factors identified through their role in the "sexuality" of *E. coli* (Watanabe, 1963). E.S. Andersson understood the possibility that factors such as F and RTF might drastically change prokaryotic evolution. In a *Nature* article in 1966 (Andersson, 1966), he suggested that, if gene transfer through plasmids was found to occur outside of the Enterobacteriaceae, we would need to rethink how we see the evolution of bacteria: *"...if the long-term activity of transfer factors has influenced bacterial evolution, the evolutionary time-scale may have been telescoped into a shorter span than that envisaged purely in terms of the selection of mutants with survival advantages."*

It is interesting to notice the similarity between this statement and the one made by Wallace about Bastian's work on spontaneous generation. It seems that there is an intuitive appeal for mechanisms that would speed-up evolution to be present in nature. Again in the 1970s, some investigators re-iterated the possibility that DNA exchange might be a major player in microbial evolution. One of them in particular, the microbiologist Sorin Sonea, even went as far as claiming that LGT mediated by plasmids and phages made all prokaryotes but a single species (Sonea, 1971, 1988).

These various warnings that LGT could play an important role in prokaryotic biology were at first ignored by molecular phylogeneticists. They could envision the holy grail of systematics, a true tree of life, dreamed of by Darwin and made accessible by the molecular tools developed by Zuckerkandl and Pauling. LGT was then considered by most to be a specialized mechanism affecting only certain sets of genes, such as those responsible for pathogenicity and antibiotic resistance.

During the 1980s, the first conflicts between trees built from different genes were noticed. Ambler, Meyer, Kamen and collaborators (Ambler *et al.*, 1979a; Ambler *et al.*, 1979b) noted that trees for purple non-sulfur photosynthetic bacteria (we would now call them proteobacteria) made with different proteins (cytochrome c-551 and c', high potential iron-sulfur proteins) had different topologies. Furthermore, these molecular trees disagreed with traditional classifications based on cell physiology, biochemistry and morphology. Although acknowledging several possible causes, they concluded that "*the simplest (if most difficult to prove) explanations for all the anomalies described is that the proteins concerned have reached some of the organisms by processes of interspecific gene transfer*" (Ambler *et al.*, 1979a).

LGT as an evolutionary force: evidence from anecdotal cases

The inconsistency found between phylogenetic trees of cytochrome proteins was only the first of numerous examples of conflict between trees that would be uncovered during the 1980s, as DNA sequencing became more common and more affordable. Several of these early claims were in fact poorly supported and unlikely, perhaps as a result of the enthusiasm about new methods of phylogenetic analysis coupled with a lack of knowledge on their limitations and proper use. Russell Doolittle's group (Smith *et al.*, 1992) and later Michael Syvanen (who had always been a strong proponent of LGT), suggested caution regarding identification of LGT events: "*because so many examples* *have been offered and then retracted, and because there remains considerable skepticism as to the occurrence of horizontal transfer, one should be obliged to consider the rigor of the various criteria used for making this judgment*" (Syvanen, 1994). Improvements in tree reconstruction algorithms, in amino acid and nucleotide substitution models, and in phylogenetic practices subsequently lead to number of convincing and well-supported cases for LGT. Tables 1.1 and 1.2 present a sampling of such strong anecdotal cases of LGT, but the list is far from exhaustive.

It is not before the advent of genomics, however, that we would see the true extent of how prevalent LGT is among prokaryotes. Anecdotal cases are very revealing about the evolutionary history of specific genes or systems, but do not convey much information on the evolution of an organism as a whole.

Table 1.1 Representative cases of lateral gene transfer for prokaryotes.

Gene	Functional category	Type of LGT[a]	Direction of LGT[b]	Detection method	Reference
Aminoacyl-tRNA synthetases	Translation, ribosomal structure and biogenesis	GA, OR or HR	Mulitple within and between all domains	PD	(Woese, Olsen et al. 2000) (Wolf, Aravind *et al.* 1999)
Ribosomal protein RpS14	Translation, ribosomal structure and biogenesis	GA, OR	Multiple within Eubacteria	PD, GC	(Brochier, Philippe *et al.* 2000)
Ribosomal proteins RpL31, RpL33, RpS14, RpS18, RpL32, RpL28	Translation, ribosomal structure and biogenesis	GA, OR	Multiple within Eubacteria	PD, GC	(Makarova, Ponomarev *et al.* 2001)
Elongation factor Tu (tufB)	Translation, ribosomal structure and biogenesis	GA	Streptococcaceae to Enterococcaceae	PD, SM	(Ke, Boissinot *et al.* 2000)
rRNA operons	Translation, ribosomal structure and biogenesis	GA or OR	*Thermobispora bispora* to *Thermomonospora chromogena*	PD	(Yap, Zhang *et al.* 1999)
Recombination/repair system (mutS, mutL, mutH, mutU)	DNA replication, recombination and repair	HR	Multiple between *E. coli* strains	PD	(Denamur, Lecointre *et al.* 2000)
Walker-type chromosome partitioning ATPase (parA)	Cell division and chromosome partitioning	GA, OR or HR	Multiple between Archaea and Eubacteria	PD	(Gerdes, Moller-Jensen *et al.* 2000)
Nodulation genes (nodA, nodB)	Signal transduction mechanisms	OR	Multiple inter- and intra-specific within Rhizobiaceae	PD	(Wernegreen and Riley 1999)
Two-component signal transduction systems	Signal transduction mechanisms	GA, OR	Multiple from Eubacteria to Archaea and Eukarya	PD, DP	(Koretke, Lupas *et al.* 2000)
Lysine biosynthesis gene cluster (lys20, hacA, hacB, rimK, argC, argB)	Amino acid transport and metabolism	GA, OR or HR	Between *Pyrococcus horikoshii* and *Thermus thermophilus*	PD, DP	(Nishida, Nishiyama *et al.* 1999)
Histidine biosynthesis genes (HisD, HisG)	Amino acid transport and metabolism	GA or OR	Multiple between Eubacteria and Archaea	PD	(Bond and Francklyn 2000)
Glutamine synthetase (GSI)	Amino acid transport and metabolism	OR	Euryarchaea to low G+C Gram positives	PD	(Pesole, Gissi *et al.* 1995)
Glutamine synthetase (GSII)	Amino acid transport and metabolism	OR	Multiple inter- and intra-generic within Rhizobiaceae	PD	(Turner and Young 2000)
Glutamate synthase	Amino acid transport and metabolism	OR	Multiple within and between Archaea and Eubacteria	PD	(Nesbø, L'Haridon *et al.* 2001)
Myo-inositol 1P synthase	Carbohydrate transport and metabolism	GA, OR	Multiple within and between Archaea and Eubacteria	PD	(Nesbø, L'Haridon *et al.* 2001)

Table 1.1 Representative cases of lateral gene transfer for prokaryotes (continued).

Gene	Functional category	Type of LGT[a]	Direction of LGT[b]	Detection method	Reference
Glyceraldehyde-3-phosphate dehydrogenase (Gap1,Gap2)	Carbohydrate transport and metabolism	GA or OR	β-proteobacteria to CFB Eubacteria	PD	(Figge, Schubert *et al.* 1999)
ABC transport system (maltose, trehalose)	Carbohydrate transport and metabolism	GA	Between *Pyrococcus furiosus* and *Thermococcus litoralis*	DP, SM	(Diruggiero, Dunn *et al.* 2000)
Rubisco (form I, rbcL)	Carbohydrate transport and metabolism	GA or OR	Multiple within Eubacteria and from Eubacteria to Eukarya	PD	(Palmer 1995)
HMG-CoA reductase	Lipid metabolism	OR	Multiple between all domains	PD	(Boucher and Doolittle 2000)
MEP cytidyltransferase	Lipid metabolism	GA	Eubacteria to *Pyrococcus horikoshii*	PD, DP	(Lange, Rujan *et al.* 2000)
A/V ATPases (A and B subunits)	Energy production and conversion	GA	Multiple from Acrhaea to Eubacteria	PD, DP	(Olendzenski, Liu *et al.* 2000)
Photosynthesis gene cluster (crtEF-bchCXYZ-puf and bchFNBHLM-lhaA-puhA)	Energy production and conversion	GA or OR	α-proteobacteria to δ-proteobacteria	PD	(Igarashi, Harada *et al.* 2001)
HSP70 (dnaK)	Posttranslational modification, protein turnover, chaperones	GA	Multiple from Eubacteria to Archaea	PD	(Gribaldo, Lumia *et al.* 1999)
Catalase-peroxidase	Inorganic ion transport and metabolism	GA or OR	Archaea to δ-proteobacteria	PD	(Faguy and Doolittle 2000)
Nitrogenase (nifH)	Inorganic ion transport and metabolism	GA or OR	α-proteobacteria to β-proteobacteria	PD	(Hurek, Egener *et al.* 1997)

[a] Type of LGT: GA = Gene acquisition, HR = Homologous recombination, OR = Orthologous replacement

[b] Detection method: PD = Phylogenetic discordance, SM = Overall sequence or motifs similarity, GC = Genomic context, DP = Distribution pattern

Table 1.2 Representative cases of lateral gene transfer for eukaryotes.

Gene	Functional category	Type of LGT[a]	Direction of LGT[b]	Detection method	Reference
Glutamate synthase (small subunit)	Amino acid transport and metabolism	OR	Low G+C Gram pos. to Eukaryotes	PD	(Andersson and Roger 2002)
Glyceraldehyde-3-phosphate dehydrogenase	Carbohydrate transport and metabolism	OR	Eubacteria to kinetoplastids and Eubacteria to diplonemids	PD	(Qian and Keeling 2001)
Phosphoenolpyruvate carboxykinase	Carbohydrate transport and metabolism	GA or OR	Archaea to *Giardia*	PD	(Suguri, Henze *et al.* 2001)
N-acetylneuraminate lyase	Carbohydrate transport and metabolism	GA	Eubacteria to *Trichomonas*	PD, DP	(de Koning, Brinkman *et al.* 2000)
5-monophosphate-dehydrogenase	Nucleotide metabolism	OR	proteobacteria to *Cryptosporidium*	PD	(Striepen, White *et al.* 2002)
Sulfide dehydrogenase	Sulfur metabolism	GA	Anaerobic prokaryote to diplomonads	PD, DP	(Andersson and Roger 2002)
Malic enzyme	Energy production and conversion	OR	Archaea to *Entamoeba*	PD, DP	(Field, Rosenthal *et al.* 2000)
NADH oxidase	Energy production and conversion	GA	Eubacteria to *Giardia* and Eubacteria to *Entamoeba*	PD, DP	(Nixon, Wang *et al.* 2002)

[a]Type of LGT: GA = Gene acquisition, HR = Homologous recombination, OR = Orthologous replacement

[b]Detection method: PD = Phylogenetic discordance, SM = Overall sequence or motifs similarity, GC = Genomic context, DP = Distribution pattern

LGT as an evolutionary force: evidence from genomic studies

Genome content variability

Comparisons of complete and partial genome sequences. Genome sequences of multiple closely related microorganisms first became available for a variety of pathogenic bacteria. The first to be compared were two *Helicobacter pylori* strains from the human gut. Similar genomic organizations and gene orders were found in the two strains, although 6-7% of open reading frames (ORFs) in each strain were missing from the other (Alm *et al.*, 1999; Alm and Trust, 1999). Most of that difference was found in a single hypervariable region. The comparison of closely related *Chlamydia trachomatis* genomes revealed similar zones of plasticity, these two strains otherwise showing a very high degree of conservation of gene order and content (Kalman *et al.*, 1999; Read *et al.*, 2000). When these strains were compared to the more distantly related *Chlamydia pneumoniae* genome, 214 ORFs specific to the latter organism were found while 70 ORFs specific to *C. trachomatis* could be identified (Kalman *et al.*, 1999; Read *et al.*, 2000). Three strains of another common pathogen, *Neisseiria meningitidis*, have also been compared (Parkhill *et al.*, 2000; Tettelin *et al.*, 2000). *N. meningitidis* MC58 contains 239 ORFs (11.1%) that are absent from *N. meningitidis* Z2491, which itself harbours 208 (9.8%) ORFs missing from MC58 (Lan and Reeves, 2000). The majority of these 208 ORFs coded for proteins contributing to virulence and resided in three large pathogenicity islands, which were suspected to be of foreign origin on the basis of their GC content, different from the bulk of the chromosome.

Important variation in genome content has also been observed between closely related microorganisms that are not pathogenic. The alkaliphile *Bacillus halodurans* has a similar genome size to *Bacillus subtilis* and SSU rRNA phylogeny identifies them as close relatives (Takami *et al.*, 2000). About a third of the ORFs identified in *B. halodurans* did not have a match in *B. subtilis*. These proteins specific to *B. halodurans* included 10 sigma (σ) factors (which might have played a role in its adaptation to alkaline environments) and large numbers of transposases and recombinases (112 in *B. halodurans* compared to 12 in *B. subtilis*). Eleven of these transposases and recombinases were found within AT-rich or GC-rich islands, likely to have played a role in the introduction of foreign genes in the *B. halodurans* genome. Important genome content variability between close relatives can also be observed in members of the domain Archaea. Comparison of the genomes of the hyperthermophiles *Pyrococcus furiosus* and *Pyrococcus horikoshii* shows that the former is about 170kb (10%) larger than the latter, which is missing *trp*, *his*, *aro*, *leu-ile-val*, *arg*, *pro*, *cys*, *thr* and *mal* operons (Maeder *et al.*, 1999). Within this 170kb is a 16kb region that is 99% identical to a region of the same length in *Thermococcus litoralis*. This region, which encodes an ABC transport system for maltose and trehalose, is most certainly a transposon, as it is flanked by insertion elements (Diruggiero *et al.*, 2000).

Partial genome sequences can also be compared *in silico*. One method for such comparisons, developed by McClelland *et al.* (McClelland *et al.*, 2000), has been applied to compare the complete *Escherichia coli* K12 genome with the nearly completely sequenced genomes of other Enterobacteriaceae (*Klebsiella pneumoniae, Salmonella enterica* serovars Typhimurium, Typhi and Paratyphi A, and *Yersinia pestis*). 1350 (30.6%) of K12's genes, in 394 locations, were absent in all *Salmonella* genomes and 1165 of these were absent from *K. pneumoniae* as well. Most of the missing genes are found within 28 10-40kb regions of the K12 genome. However, according to McClelland *et al.*, there would be about 100 sites where a single gene is absent from representatives of the *Salmonella* genera but present in *E. coli* K12 and flanked by genes found in both genomes. Studying a few large pathogenicity islands would therefore not be sufficient to understand the differences between *Salmonella* and *E. coli.* The complete genome sequence of another *E. coli*, the pathogenic strain O157:H7 (Perna *et al.*, 2001), presents one of the most extreme examples of within-species variation observed to date. The average nucleotide identity between orthologous sequences from K12 and O157:H7 is 98.4%, showing that they diverged very recently. However, K12 has 528 genes that are not found in O157:H7 and the latter strain has 1387 genes absent from K12. The genes present in only one of the two genomes are found scattered in 1770 O157:H7 specific islands (1.34Mb of DNA) and 234 K12 specific islands (0.53Mb), most of these islands, which range in size from 50 to 88bp, have a base composition suggesting a foreign origin.

Microbial genomes can also be compared using techniques that do not involve the complete sequencing of all of the genomes being compared. DNA microarrays based on a completely sequenced genome have been used to compare the latter to closely related genomes for which no sequence information is available. Using this approach, the genome of *Mycobacterium tuberculosis* H37Rv was compared to those of *Mycobacterium bovis* and *M. bovis* attenuated BCG vaccine strains (Behr *et al.*, 1999). Eleven regions (91 ORFs) of the H37Rv genome were absent from *M. bovis* and another five (38 ORFs) were missing from the BCG strains (Salamon *et al.*, 2000). Using microarrays, the completely sequenced genomes of *Helicobacter pylori* (strains 26695 and J99) were compared to 15 strains of *H. pylori* of variable virulence (Salama *et al.*, 2000). A full 22% of *H. pylori* genes were found to be dispensable in one or more strains. Similar results were obtained by Ochman and Jones when an *E. coli* K12 microarray was used to compare K12 to five other strains of *E. coli* (Ochman and Jones, 2000). Between 82 and 392 ORFs were missing in one or another strain. Independent measures of genome size found that these five strains contain between 65 and 1183kb (25%) of DNA that is not present in K12.

Genomic subtraction – hybridization of DNA from two genomes and removal of common sequences – specifically targets the parts of two genomes that are different. If a substantial number of clones from a genomic subtraction library are sequenced and one of the genomes compared has been completely sequenced, it is possible to estimate by how many genes the two genomes differ. Such a comparison has been done between the completely sequenced genome of the hyperthermophilic bacterium *Thermotoga maritima* MSB8 and the closely related strain *Thermotoga sp.* RQ2 {99.7% identity between their SSU rRNA sequences, (Nesbo *et al.*, 2002)}. This study estimated that RQ2 contained 350-400 genes that are not found in MSB8 (20% of the RQ2 genome). A similar approach found that *Salmonella enterica* serovar Typhimurium contained about 140kb of DNA not found in *S. enterica* serovar Typhi (3% of its genome) (Emmerth *et al.*, 1999). When genomic subtraction was performed between the prototype virulent strain of *Listeria monocytogenes* and the prototype epidemic strain of the same species, the two genomes were found to differ by 150 to 190 kb of DNA (DNA present in only one of the two strains) (Herd and Kocks, 2001).

Quantification of LGT within a genome

In the last few years, it has become standard for publications reporting the complete genomic sequence of a prokaryotic species to estimate the fraction of genes that could be of "foreign" origin. The articles describing the genome of the hyperthermophilic bacterium *Thermotoga maritima* and the methanogenic archaeon *Methanosarcina mazei* illustrate this very well. For *Thermotoga*, BLAST-based similarity searches led to an estimation that 24% of this bacterium's ORFs had been acquired by its ancestors through LGT from archaea (Nelson *et al.*, 1999). The archaeal origin of a limited subset of these genes was confirmed by actual phylogenetic analysis. Using similar methods, the genome of another hyperthermophilic bacterium, *Aquifex aeolicus*, was also estimated to harbour a large number of archaeal genes (Deckert *et al.*, 1998). Since these bacteria both thrive in hot temperature environments dominated by archaea, it has been suggested that they acquired numerous archaeal genes through their adaptation to life at higher temperatures (Nesbo *et al.*, 2001).

Among archaea of which the complete genome has been sequenced, *Methanosarcina mazei* and *Methanosarcina acetivorans* have both the largest genomes and the highest number of genes with their most similar homolog in a bacterial genome – 1043 (30%) and 945 (21%) of their respective totals of 3,371 and 4,524 ORFs (Deppenmeier *et al.*, 2002; Galagan *et al.*, 2002). For *M. mazei*, half of these ORFs have no match in archaea and are therefore extremely likely to be transfers. Using the same BLAST-based similarity searches, the next largest archaeal genome, that of the crenarchaeote *Sulfolobus solfataricus*, is found to have one of the lowest bacterial gene contents (12.0% of 2,977 genes) (Deppenmeier *et al.*, 2002). However, about 17.0% of the genes in its genome have their closest match to the acidophilic euryarchaeotes *Thermoplasma volcanium* and *Thermoplasma acidophilum*. Like *Thermotoga* and hyperthermophilic archaea, *Thermoplasma* and *Sulfolobus* share a common niche: they are both acidophilic scavengers. This environmental proximity might have facilitated genetic exchange between these archaea belonging to different phyla.

Most completed prokaryotic genome sequences show some proportion of "foreign" genes, even though it is usually lower than the extremes mentioned above. Going further than estimating how much LGT affected a given genome, attempts have been made at systematically calculating the amount of laterally transferred genes in each and every available genome sequence.

Ochman and Lawrence performed systematic analyses of most prokaryotic genomes using a methodology detecting laterally acquired genes through their atypical GC content and codon biases (which match the donors base composition rather than the host's) (Garcia-Vallve *et al.*, 2000; Ochman *et al.*, 2000). They have reported foreign gene contents as low as zero (for the intracellular parasites *Mycoplasma genitalium* and *Rickettsia prowazekii*) and as high as 16.6% (for the cyanobacterium *Synechococcus* PCC6803). This technique has been largely criticized for allowing many false positives (genes with biased base composition due to their expression profile) and false negatives (genes laterally transferred from a donor with a genomic base composition matching the host's) (Hooper and Berg, 2002; Koski *et al.*, 2001; Ragan, 2001a, b). In any instance, this technique is also limited by its inability to detect transfers older than a few hundred million years (Lawrence and Ochman, 1997).

BLAST tools have been extensively used by Koonin and his collaborators to assess the extent of LGT in prokaryotic genomes, defining as foreign-in-origin genes that are more similar to homologs found in distant rather than closely related genomes (Koonin *et al.*, 2001). They call this method "unexpected ranking of sequence similarity among homologs". Considering only exchanges between domains (bacteria to archaea and vice-versa), their estimates range from 0% (the obligate aphid intracellular symbiont *Buchnera sp.*) to 15.6% (the extremely halophilic archaeon *Halobacterium sp.* NRC-1). Such values do not seem very impressive: the average fraction of archaeal genes in bacterial genomes was only 3% and the values for bacterial genes in archaeal genomes ranged from 4 to 8%. However, their method excludes transfers within domains, likely to be much more common than those occurring between domains, as well as early inter-domain transfers, because they accept only genes whose within-domain hits are all significantly less than a between-domains hit. If the same method is used to detect exchange between phyla within Bacteria, higher estimates of LGT are indeed obtained (again 0% for *Buchnera*, but 32.6% for *Treponema pallidum* and 28.8% for *Bacillus halodurans*).

Added together transfers between distant and closely related organisms could mean that for some genomes, very few genes actually have the same evolutionary history, back to the last universal common ancestor. None of the *in silico* methods to detect LGT among complete genome sequences, however, has the capacity to detect the most frequent type of event, gene recombination between members of a species or genus.

LGT as an evolutionary force: evidence from MLEE and MLST studies

The fact that prokaryotes reproduce by binary fission lead to the assumption that their evolution took place entirely by asexual processes. Such assumptions seemed to be confirmed by studies of *E. coli* using multilocus enzyme electrophoresis (MLEE) (this technique allows calculation of the level of linkage disequilibrium present in a population: strong linkage indicates non-random association of alleles, i.e. low levels of recombination). The finding that *E. coli* populations seemed to be clonal was rapidly extended as a paradigm applicable to all prokaryotic populations (Maynard Smith, 1999).

However, more recent studies have shown that low linkage disequilibrium can also appear in bacterial populations where recombination is frequent (Maynard Smith, 1999). Also, when MLEE was applied to other bacterial populations, it soon became apparent that a range of clonality is possible, from a fully sexual populations (*Neisseria gonorrhoeae*), through sexual at the fine scale but clonal between more distant populations (*Rhizobium*) to apparently clonal at all levels (*Salmonella*) (Maynard Smith, 1999). Direct analysis of nulceotide sequences from *E. coli* isolates (four loci from 12 isolates) also brought into question the interpretation of MLEE data. Indeed, Guttman and Dykhuizen (1994) estimated a rate of recombination of $5x10^{-9}$ changes per nucleotide per generation (50-fold higher than the mutation rate).

Analysis of large amounts of sequence data for the detection of recombination became possible with the development of multiolocus sequence typing (MLST), a scaled-up version of the technique used by Guttman *et al.* a decade earlier (Guttman and Dykhuizen, 1994). MLST involves the sequencing of several housekeeping genes (well separated from one another on the chromosome) from a very large number of isolates.

From these data, recombination rate can be estimated by selecting clusters of isolates that have identical allelic profiles (clones) and identifying minor variants that differ from the typical allelic profiles of these clones at only one of the multiple housekeeping loci sequenced (Feil *et al.*, 2000). Each of these variants either arose by point mutation (single-nucleotide difference) or recombination (multiple-nucleotide differences). This type of analysis, when performed with *Neisseria meningitidis*, *Streptococcus pneumoniae* and *Staphylococcus aureus*, yields r/m ratios (relative probability that an individual nucleotide site changed by recombination or mutation) of 100:1, 61:1 and 24:1, respectively. These numbers do not take into account highly deleterious point mutations and recombination events between identical sequences, as these events will not be observed (Feil *et al.*, 2001).

Estimating the r/m ratio requires a large amount of data from several closely related isolates. The three species mentioned above are the only ones for which such an extensive dataset is available. However, for species with more limited MLST datasets, the importance of recombination can be estimated indirectly by assessing the degree of incongruence between phylogenetic trees of different housekeeping genes within these datasets (Feil *et al.*, 2001). Using a maximum likelihood approach, a phylogenetic tree is computed for each gene from a species and its difference in log likelihood with the trees of other genes from that species is compiled. These values are compared to a null distribution generated using random tree topologies. High rates of recombination should result in incongruent trees for different genes from the same species: trees that are significantly different from each other (no more similar to each other than they would be to randomly generated trees). Important incongruence was found in the three species for which r/m ratios were calculated, *N. meningitidis*, *S. pneumoniae* and *S. aureus*, as well as *Streptococcus pyogenes* (Feil *et al.*, 2001). Less incongruence was found in *E. coli* and *Haemophilus influenzae,* suggesting that the rate of recombination might be lower in these species. However, clear mosaicism can be found in some of *H. influenzae* sequences and there is evidence from nucleotide sequences for recombination in *E. coli* (Milkman and Bridges, 1993). There is therefore apparent conflict between the congruence among gene trees and the direct observation of recombination in these species. Either recombination occurs occasionally but not at a sufficient frequency to disrupt the clonal frame or these species encompass distinct biological or ecological clusters (Feil *et al.*, 2001).

Rates of recombination and point mutation can also vary greatly through a prokaryotic population. Mismatch repair (MMR) systems control the fidelity of recombination by recognizing mispaired and unpaired bases in heteroduplex regions and block the strand transfer catalyzed by RecA. Inactivation of this system can increase the rate of mutation 10^2- to 10^3-fold, and natural populations of several γ-proteobacteria contain MMR-deficient mutants at a frequency exceeding 1% (Matic *et al.*, 1997). Denamur *et al.* (Denamur *et al.*, 2000) detected significant sequence mosaicism in MMR genes of *E. coli* isolates, suggesting that genes of this system could be frequently inactivated and reactivated by recombination. The SOS response system can also lead to increased rates of recombination, mostly through the overproduction of RecA (Matic *et al.*, 1995).

Although the sample of prokaryotic species for which we have information about recombination rates is very biased (pathogenic bacteria belonging to the firmicutes or the proteobacteria), a few trends seem to be emerging. Recombination rates, when they can be calculated, can often be higher than point mutation rates. These rates also seem to vary widely among species, with the capacity to be naturally transformable playing a role, but not being determinant. Some transformable species indeed show greater congruence between phylogenies of housekeeping genes (*H. influenzae*) than species that are not (*S. aureus*, *S. pyogenes*).

Is a paradigm shift required in theories of prokaryotic evolution?

The standard account of prokaryotic evolution has been couched in Darwinian terms of vertical inheritance patterns best represented in the form of trees. Lately, however, a rapidly expanding body of evidence for lateral gene transfer has been complicating that modeling strategy. LGT is generally accepted by all evolutionary biologists, but it is frequently conceived of as a special case within an all-encompassing model of generation-to-generation inheritance. From this point of view, lateral gene transfer may somewhat obscure vertical patterns (the tree-like structure), but the dispute is mostly the extent at which the obscuring occurs – a lot or a little, and at which points in evolutionary history (Gogarten *et al.*, 2002). If the messing-up occurs a lot, a tree would not be an adequate analogy to describe evolution of prokaryotes (Doolittle, 1999b). However, the tree analogy might still hold for certain genes that are much less frequently exchanged than others (Brochier *et al.*, 2002). Also, the fact that gene trees are, most of the time, only incongruent on some branches is often taken as evidence that some vertical pattern exists and is conserved over long evolutionary time. Others are of the opinion that the tree-like pattern we often observe in evolutionary reconstructions is actually created by a high rate of recombination and LGT among closely related prokaryotes and a low rate between more distantly related microbes (Gogarten *et al.*, 2002). A web analogy has also been suggested as an alternative to the tree and is even used by certain evolutionary reconstruction methods (Doolittle, 1999a).

Is there a paradigm under which we could work that would accommodate both observations of vertical patterns and of the scrambling of these patterns? The old paradigm of vertical inheritance has clearly been overthrown, at least in its pure form (Doolittle, 2000b). But even allowing for exceptions (some LGT) does not save the concept, because it requires an *ad hoc* decision about how much scrambling is allowed. How many genes with congruent phylogenies do we need to say we recovered the organismal tree? Which genes will these be? At which level do we look for congruence: strains, species, genus, order? It is clear that both vertical and lateral (horizontal) inheritance are major players in prokaryotic evolution. But the current evolutionary paradigm, whether it includes a little or a lot of exchange, begs for quantification, which we might never be able to provide, and most certainly varies from one part of the tree to the other.

I am going to argue for a new paradigm that is broader than that of vertical inheritance, in its original or modified versions accounting for some level of LGT. We will simply assume that exchange forms the background instead of vertical inheritance and that we are trying to discover vertical patterns in a sea of transfers. This can be expressed as a set of four interconnected claims.

Claim 1. DNA Recombination and Replication are interdependent processes. Similarly to replication, recombination is essential for DNA to be an efficient medium for genetic information. The structure of DNA itself suggests the existence of this mechanism, just as it suggested the mode of replication to Watson and Crick (Cox, 2001). In prokaryotes, an important function of homologous recombination is the repair of stalled or collapsed replication forks. DNA replication under standard growth conditions is therefore dependent on homologous recombination and is often termed recombination dependent DNA replication (Cox, 2001). Indeed unrepaired DNA breaks are lethal for any cell or its progeny (Kowalczykowski, 2000). Such breaks happen naturally, as a result of the encounter of a replication fork with "normal" discountinuities in the template, with a frequency that can be close to one such event per cell division (Kowalczykowski, 2000). Some enzymes, such as RecA (or RecA-like proteins), are essential for homologous recombination to occur (Kogoma, 1997; Smith, 1988). This could readily explain the high lethality (50-70%) of RecA mutants, since at least 50% of the cells will encounter at least one lesion that causes replication fork collapse and require recombination function for resumption (Kowalczykowski, 2000). The RecA enzyme, like many others proteins involved in recombination, is highly conserved among all domains of life and even found in viruses (Lehman, 2003). This suggests an ancient origin for those systems, possibly tightly linked with DNA replication systems. Furthermore, in several prokaryotes, recombination seems to be the source of more genetic variability than point mutations (which include replication errors) (Feil *et al.*, 2001), making both processes important in the generation of the genetic diversity required for adaptation and evolution.

Claim 2. All genes can be exchanged. All genes can be transferred or recombined, and have been at some point in their history. There are, however, different propensities for transfer exhibited by different genes. Recombination does become more difficult as organisms become more distantly related, since divergent DNA is not as readily recombined by the RecA system (Majewski *et al.*, 2000). Even though those barriers can be reduced {in mismatch repair deficient mutants or during SOS response, (Denamur *et al.*, 2000; Matic *et al.*, 2000; Radman, 1999)}, extant cells from two very distant lineages might not be exchanging much DNA with each other. Recombination might be less frequent across large evolutionary distances, but it is still possible. Even if two sequences were so divergent that they found it impossible to recombine because of the biochemical properties of the enzymatic recombination machinery, other processes – such as illegitimate recombination or recombination through bacteriophage site-specific recognition sequences or insertion elements of transoposons – could take over.

Claim 3. All prokaryotes are susceptible to LGT. Even if a prokaryote is not naturally competent, some conditions in nature might allow it to be transformed with exogenous DNA, as was demonstrated with *E. coli* in freshwater (Baur *et al.*, 1996; Demaneche *et al.*, 2001). Plasmids and conjugative transposons can have wide phylogenetic ranges (Amabile-Cuevas and Chicurel, 1992; Rawlings and Tietze, 2001) and most prokayotes are thought to have specific viruses (Bamford, 2003; Campbell, 2003; Filee *et al.*, 2003). Many microbes also use DNA as a food source, leaving the possibility for the penetration of some DNA fragments in such cells (Redfield, 1993). Environmental stress (or cellular damage) might also make possible the penetration of DNA inside cells that are normally impervious to it. The combination of all these factors makes it likely that the vast majority of prokaryotes can, under a particular set of circumstances, uptake DNA from an exogenous source.

Claim 4. Not all exchanges have detectable consequences. Although transfer (recombination) is part of the history of a gene, it does not necessarily have meaningful effects on its phylogenetic signal. A lot of transfer may have no detectable consequences, in that it involved identical sequences or created differences that did not affect the lineage's evolutionary history at a significant or detectable level. Most likely, the vast majority of exchange happens by recombination of identical DNA, which makes no difference at all to the cells receiving it or, consequently, to any phylogenetic reconstruction of their history. Consequential or meaningful exchange comes in only with non-identical exchange, which means that the question of exchange has two levels of relevance. Asking whether a gene has been exchanged at some point in its history is irrelevant, because the answer is obviously a trivial 'Yes'. The second question, however, which asks whether there are consequential gene transfers that influence phylogenetic reconstruction, is still highly pertinent to evolutionary inquiry.

If these four claims hold, it seems advantageous to consider LGT as the evolutionary background rather than vertical inheritance. Current evidence seems to point in favor of these claims. However, further data coming from large-scale MLST analyses of a much wider diversity of prokaryotes than is actually available will be required to verify the importance of recombination in creating genetic diversity. More knowledge about the role of viruses in prokaryotic biology, mostly concerning their frequency, host range and diversity would be needed to better understand the importance of transduction for microbial evolution.

The consequences of considering exchange as a background would be numerous in our search for vertical patterns in evolution. First, we would have to evaluate the possibility of the presence of recombination in our datasets. For moderately conserved genes on a large phylogenetic scale, such recombination is unlikely to be detectable or to pose a problem for phylogenetic inference. However, phylogenies on smaller scales (species level) should systematically be screened for such events {for example using recombination-detection software or phylogenetic programs taking into account this possibility such as split decomposition (Bandelt and Dress, 1992)}. Phylogeny of genes that are conserved across large phylogenetic ranges (such as SSU rRNA) should also be checked for recombination, which can then occur between more distantly related organisms (genus level).

Adopting this view of LGT does not preclude reconstruction of the phylogeny of important groups of genes (such as translational apparatus proteins) covering a large timescale, but simply requires that the investigator test for the possibility of LGT influencing the result of the reconstruction.

Chapter 2: Lateral gene transfer of physiological processes and the origins of prokaryotic groups

This chapter includes work that will be published as Y. Boucher, C.J. Douady, T.R. Papke, D.A. Walsh, E.M. Boudreau, C.L. Nesbø, R.J. Case and W.F. Doolittle (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics* 37: 283-328.

Introduction

Prokaryotic organisms are often classified according to their most striking physiological properties: green sulfur (photosynthesis and sulfur oxidation) and green non-sulfur bacteria (photosynthesis), purple sulfur bacteria (photosynthesis and sulfur oxidation), methylotrophs (oxidation of C_1 compounds), methanogens (methane production). Phylogenetically distant groups, however, often share these characteristics. Photosynthesis, sulfate reduction, methylotrophy are all properties found in a number of distantly related organisms (Figure 2.1). Two decades ago, when one of these particular physiological properties was revealed to be homologous rather then analogous between two microorganisms, evolutionists would claim that this feature was present in their common ancestor. With the now widely recognized prevalence of lateral gene transfer (LGT) in prokaryotes, such an inference has to be more carefully considered (Doolittle, 1999a).

With the patchy distribution observed for most metabolic processes and their presence in both bacteria and archaea, refusing the idea that some groups acquired these processes by LGT quickly leads to a *totipotent universal ancestor*. This ancestor would have had a proteome considerably more complex than that of any modern prokaryote, being capable of almost the full range of autotrophic, heterotrophic, anaerobic and aerobic biochemistries found in the diverse extent prokaryotes (Doolittle, 2000a). Furthermore, for pathways or genes that are sparsely distributed, it is quite often more parsimonious to assume LGT than independent losses in a multitude of lineages. Although nature doesn't always behave parsimoniously, it is the only logical guide we have. Also, in many instances, orthologs from two divergent organisms will be very similar to each other, so similar that claiming that this does not result from LGT would require both many losses in intermediate lineages AND an incredible shift of the evolutionary rate in those organisms that retain the trait.

Literature describing the evolution of different physiological processes has been steadily accumulating in the last decade, propelled by the wealth of information provided by genome sequencing and genome comparison. A major role for LGT has been frequently invoked to explain the dispersal of important phy siological functions in evolutionarily distant groups (Chistoserdova *et al.*, 1998; Klein *et al.*, 2001; Pereira *et al.*, 2001). However, a thorough comparison of the evolution of these different functions, to identify common underlying principles, has not been done. In this chapter, I aim to describe a number of these processes, summarize what is known about their evolution, emphasizing on the potential role LGT might have played in individual cases and look for broad similarities between them. This is by no means an exhaustive description of the evolution of all physiological processes where LGT is suspected to have played some role, but rather a selection of cases for which significant information is available. The cases were also selected to cover a diversity of processes: sulfate reduction, C_1 compounds oxidation, aerobic respiration, nitrogen fixation and photosynthesis. Comparison will allow a better understanding of the evolutionary dynamic of important physiological processes and of how much this dynamic differs from one process to another.

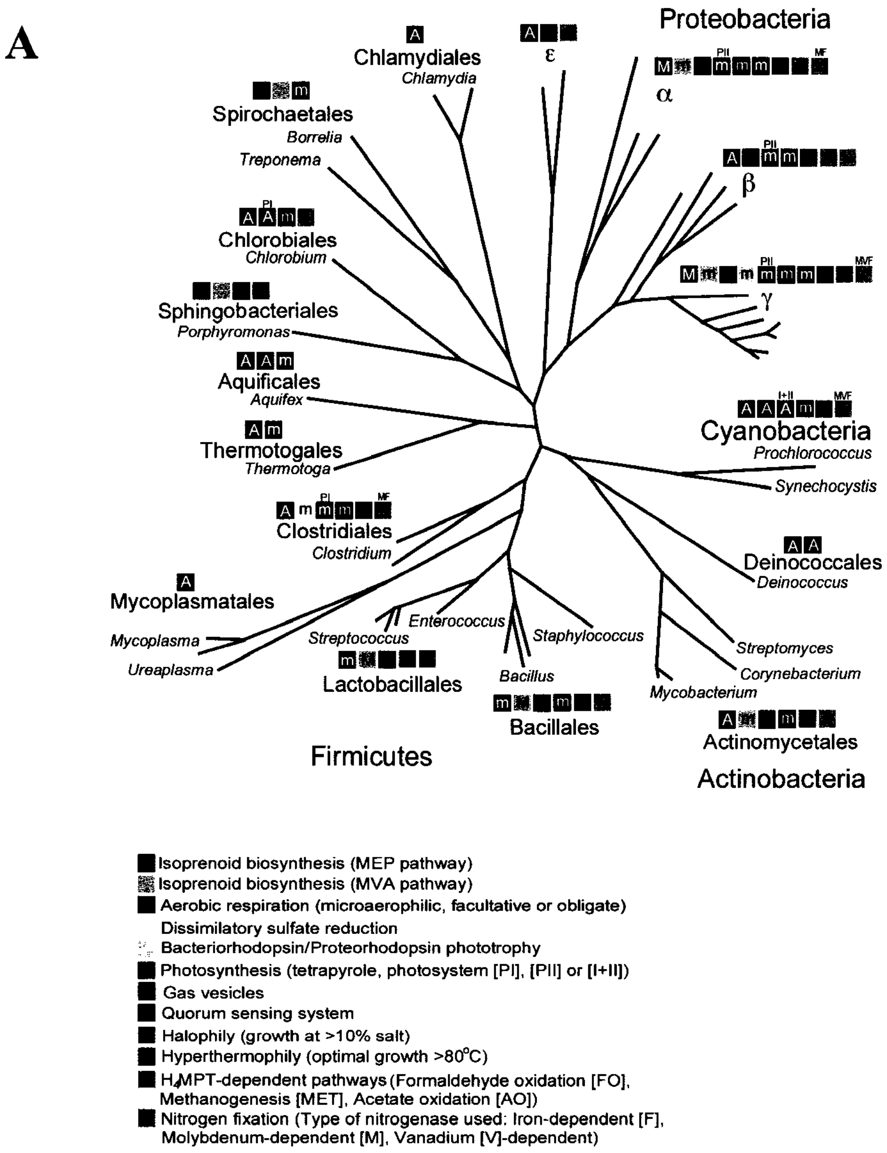

A
Proteobacteria
α
β
γ
ε
PII
MF
MVF
Chlamydiales
Chlamydia
Spirochaetales
Borrelia
Treponema
PI
Chlorobiales
Chlorobium
Sphingobacteriales
Porphyromonas
Aquificales
Aquifex
Thermotogales
Thermotoga
Clostridiales
Clostridium
Mycoplasmatales
Mycoplasma
Ureaplasma
Enterococcus
Streptococcus
Lactobacillales
Staphylococcus
Bacillus
Bacillales
Firmicutes
I+II
Cyanobacteria
Prochlorococcus
Synechocystis
Deinococcales
Deinococcus
Streptomyces
Corynebacterium
Mycobacterium
Actinomycetales
Actinobacteria
Isoprenoid biosynthesis (MEP pathway)
Isoprenoid biosynthesis (MVA pathway)
Aerobic respiration (microaerophilic, facultative or obligate)
Dissimilatory sulfate reduction
Bacteriorhodopsin/Proteorhodopsin phototrophy
Photosynthesis (tetrapyrole, photosystem [PI], [PII] or [I+II])
Gas vesicles
Quorum sensing system
Halophily (growth at >10% salt)
Hyperthermophily (optimal growth >80°C)
H4MPT-dependent pathways (Formaldehyde oxidation [FO], Methanogenesis [MET], Acetate oxidation [AO])
Nitrogen fixation (Type of nitrogenase used: Iron-dependent [F], Molybdenum-dependent [M], Vanadium [V]-dependent)

Figure 2.1 Distribution of the physiological properties reviewed here among the different prokaryotic orders. The properties are mapped on translational apparatus proteins phylogenetic trees of both (A) Bacteria (Brochier *et al.*, 2002) and (B) Archaea (Matte-Tailliez *et al.*, 2002). Each physiological property is represented by a colored square and the letters found within these squares illustrate how widespread a property is within an order; (A) potentially present in all (or practically all) taxa, (M) present in a majority (>50%) of taxa and (m) present in a minority of taxa (<50%) (the absence of a letter indicates that a property is present in a given lineage but that it is difficult to evaluate if a majority or a minority of members of this lineage harbor it).

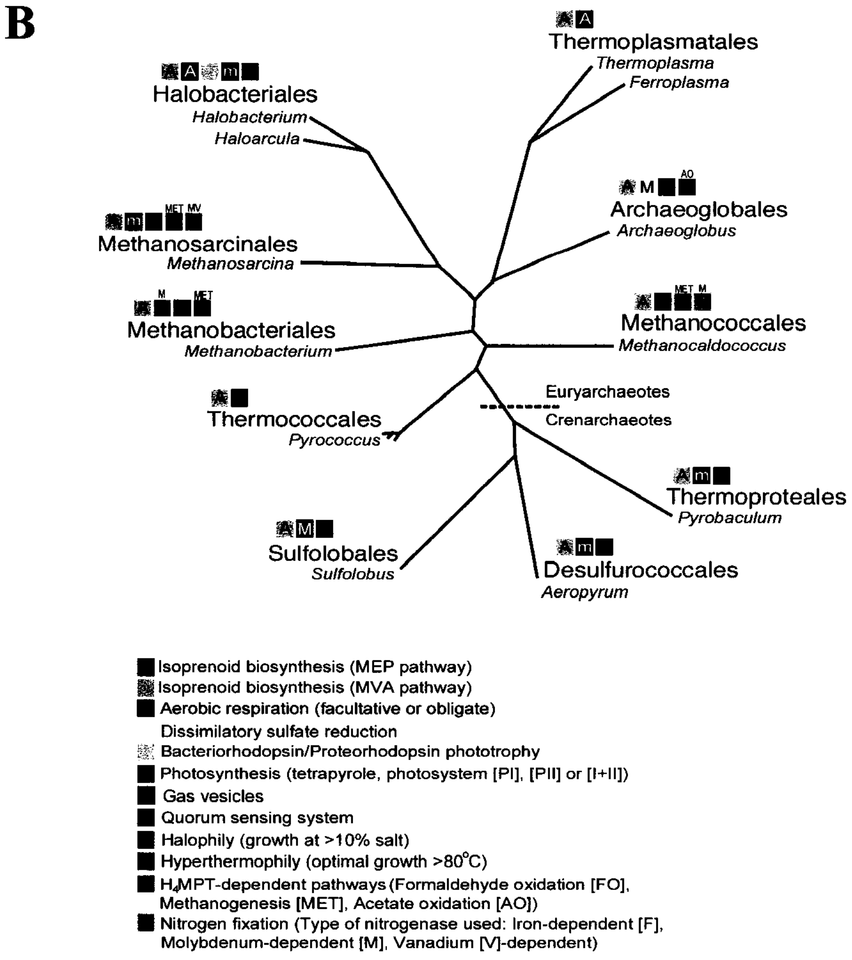

B
Thermoplasmatales
Thermoplasma
Ferroplasma
A
Halobacteriales
Halobacterium
Haloarcula
A m
Archaeoglobales
Archaeoglobus
M AO
Methanosarcinales
Methanosarcina
m MET MV
Methanobacteriales
Methanobacterium
M MET
Methanococcales
Methanocaldococcus
MET M
Thermococcales
Pyrococcus
Euryarchaeotes
Crenarchaeotes
Thermoproteales
Pyrobaculum
m
Sulfolobales
Sulfolobus
M
Desulfurococcales
Aeropyrum
m
■ Isoprenoid biosynthesis (MEP pathway)
▨ Isoprenoid biosynthesis (MVA pathway)
■ Aerobic respiration (facultative or obligate)
Dissimilatory sulfate reduction
▨ Bacteriorhodopsin/Proteorhodopsin phototrophy
■ Photosynthesis (tetrapyrole, photosystem [PI], [PII] or [I+II])
■ Gas vesicles
■ Quorum sensing system
■ Halophily (growth at >10% salt)
■ Hyperthermophily (optimal growth >80°C)
■ H4MPT-dependent pathways (Formaldehyde oxidation [FO], Methanogenesis [MET], Acetate oxidation [AO])
■ Nitrogen fixation (Type of nitrogenase used: Iron-dependent [F], Molybdenum-dependent [M], Vanadium [V]-dependent)

Figure 2.1 Distribution of the physiological properties reviewed here among the different prokaryotic orders (continued).

Materials and Methods

Gene alignment construction

Amino acid sequences of all isoprenoid biosynthesis genes were obtained from the NCBI web site (http://www.ncbi.org). For each orthologous gene family, one protein sequence of known function (biochemically characterized) was used as a query to probe the database (non-redundant GenBank) using BLASTP (Altschul *et al.*, 1997). All significant matches to the query which represented orthologous genes were retrieved (orthology was judged on the basis of sequence similarity to the query and functional annotation in the database). The amino acid sequences were then aligned using CLUSTALW with the default settings (Thompson *et al.*, 1994). The alignments were subsequently edited manually in MacClade (Maddison and Maddison, 1989) to remove gaps and ambiguous characters.

Phylogenetic analysis

The phylogenetic analysis of genes involved in the physiological processes studied here was performed at the amino acid level. PROML (Felsenstein, 1993) was used for maximum likelihood tree reconstructions, with the JTT amino acid substitution matrix, a rate heterogeneity model with gamma-distributed rates over four categories with the α parameter estimated using TREE-PUZZLE, global rearrangements and randomized input order of sequences (10 jumbles). Bootstrap support values represent a consensus (obtained using CONSENSE) of 100 Fitch-Margoliash distance trees (obtained using PUZZLEBOOT and FITCH) from pseudo-replicates of the original alignment (obtained using SEQBOOT). The settings of PUZZLEBOOT were the same as those used for

PROML, except that no global rearrangements and randomized input order of sequences are available in this program. PROML, CONSENSE, FITCH and SEQBOOT are from the version 3.6a of the PHYLIP package (http://evolution.genetics. washington.edu/phylip.html). TREE-PUZZLE and PUZZLEBLOOT can be obtained from the programs website (http://www.tree-puzzle.de).

Results

Photosynthesis

Tetrapyrole-based photosynthesis (as opposed to bacteriorhodopsin/ proteorhodopsin-based phototrophy and henceforth referred to simply as photosynthesis) is responsible for generating a large portion of the energy required by organisms on Earth and is arguably the biosphere's most important function/process. Photosynthetic organisms generate biochemical energy from light energy via a complex arrangement of light-absorbing pigments (e.g. chlorophyll, bacteriochlorophyll, carotenoids and phycobillins) and membrane-integral proteins (e.g. light-harvesting and reaction center proteins). Because there are many similarities among the different photosynthetic systems, photosynthesis is thought to have evolved only once (Baymann *et al.*, 2001; Schubert *et al.*, 1998). However, its scattered distribution among different taxa of bacteria has been difficult to explain.

Photosynthesis can occur either oxygenically or anoxygenically (with or without generation of oxygen). At the center of each of these photosynthetic processes are light-absorbing tetrapyrole molecules termed chlorophyll (oxygenic photosynthesis) or bacteriochlorophyll (anoxygenic photosynthesis), similar in structure to heme, but containing an Mg atom instead of a Fe atom. Photosynthetic processes can also be classified according to the type of terminal electron acceptor present in the photosynthetic reaction center. Type I photosynthetic reaction centers have iron-sulfur clusters and type II photosynthetic reaction centers are linked to quinones. Type I reaction centers are found in Heliobacteria and Green Sulfur Bacteria (Chlorobiales) while type II reaction centers are found in Purple Bacteria (Proteobacteria) and Green Non-Sulfur Bacteria (Chloroflexales) (Table 2.1). When type I or type II photosynthetic reaction centers occur individually in bacteria, they perform anoxygenic photosynthesis. The oxygenic photosynthesis of cyanobacteria (and chloroplasts) employs type I and type II reaction centers, which are found physically associated in cellular membranes. However, under anoxic conditions and when reducible substrates are obtainable from the environment, some algae and cyanobacteria are capable of performing photosynthesis using only the type I reaction center, and do so without generating oxygen.

One of the interesting and unexpected early outcomes of SSU rRNA phylogeny was that photosynthetic organisms have a patchy distribution among the different lineages. Although Woese (Woese, 1987) first interpreted this in terms of multiple losses of photosynthetic capacity, lateral gene transfer now seems a more reasonable and likely hypothesis. Regarding the photosynthetic reaction centers, for instance, there are (at least) two different lateral gene transfer theories that attempt to explain how type I and II came to have their observed organismal distribution (Figure 2.1).

Table 2.1 Characteristics of the photosynthetic apparatus of various groups of bacteria

Bacterial groups	**Example genus**	**Reaction Center Type**	**Oxygenic (O) or Anoxygenic (A)**	**Tetrapyroles**
Heliobacteria (Heliobacteriaceae)	*Heliobacillus*	I	A	Bchl g
Green Sulfur Bacteria (Chlorobiales)	*Chlorobium*	I	A	Bchl a + c,d,e
Green Non-Sulfur bacteria (Chloroflexales)	*Chloroflexus*	II	A	Bchl a + c
Purple bacteria (Proteobacteria)	*Rhodobacter*	II	A	Bchl a or b
Blue-green bacteria (Cyanobacteria)	*Synechococcus*	I and II	O	Chl a

Bchl: Bacteriochlorophyll
Chl: Chlorophyll

In the 1990's, a "fusion" hypothesis was proposed by both Mathis (Mathis, 1990) and Blankenship (Blankenship, 1992) and a modified fusion theory, the "heterologous fusion model" was proposed by Xiong *et al.* (Xiong *et al.*, 1998). The fusion models propose that, *via* gene duplications, a single ancient type I-like photosynthetic reaction center evolved into the two anoxygenic types currently known. The theories also suggest that oxygenic photosynthesis found in Cyanobacteria (and subsequently plants and eukaryotic algae) resulted from a "fusion" event in which the two types of anoxygenic photosynthetic system merged in a single organism, an ancestral cyanobacterium. The acquisition of a second photosynthetic reaction center, required for this fusion event, would have required lateral transfer.

Baymann *et al.* (Baymann *et al.*, 2001) suggest that an "export" model can also explain the distribution of photosynthetic reaction centers. They propose that the type I reaction centers evolved first and that a gene duplication event generated two type I reaction centers in an ancestor of the Cyanobacteria. The cyanobacterial lineage would have subsequently evolved the duplicated reaction centers into the currently observed oxygenic type I and type II reaction centers. Under the export model, the anoxygenic type II reaction center found in Green Non-Sulfur and Purple Bacteria is the result of a transfer from an ancestral cyanobacterium.

Differential loss theories, such as that proposed by Olson and Pierson (Olson and Pierson, 1987), explain the observed distribution of photosynthetic reaction centers by the loss of either the type I or type II reaction centers in the anoxygenic lineages. This type of explanation eliminates the need for LGT to play a role in explaining the distribution but also require cyanobacterial/oxygenic photosynthesis to have been the first to evolve.

Proponents of the transfer (fusion and export) models consider this unlikely, as phylogenetic analyses display photosynthesis genes from anoxygenic photosynthetic orgasnisms as ancestral to those from oxygenic phototsynthetic organisms and organelles (cyanobacteria and chloroplasts) (Baymann *et al.*, 2001; Xiong *et al.*, 2000).

Phylogenetic analysis of genes involved in Mg-tetrapyrole synthesis (e.g. chlorophyll and bacteriochlorophyll) also suggests that lateral gene transfer has been integral to the evolution of photosynthetic organisms. Comparison of these genes performed by Xiong *et al.* (Xiong *et al.*, 2000), showed *Chlorobium tepidum* and *Chloroflexus aurantiacus* as each other's closest relatives among the photosynthetic organisms. Interestingly, these bacteria have type I and type II photosynthetic reaction centers, respectively. Additionally, the authors pointed out that phylogenetic trees of reaction center apoproteins "are incongruent with the pigment biosynthesis protein trees." Analyses of bacteriochorophyll biosynthesis genes encoded on a super-operon belonging to the β-proteobacterium (purple bacterium) *Rubrivivax gelatinosus* revealed that these genes were more closely related to their α-proteobacterial homologs, while genes not involved in photosynthesis flanking the gene cluster supported the SSU rRNA phylogeny (i.e. β-proteobacterial origin) (Igarashi *et al.*, 2001). Furthermore, it was noted that nearly all of the 31 photosynthesis genes of this super-operon had their closest sequence identity to genes found in *Rhodospeudomonas palustris*, a species belonging to the α-Proteobacteria, suggesting that possibly the entire gene cluster was transferred in a single event. These data in sum suggest that in some cases the separate components of photosynthesis (e.g. reaction centers and bacteriochlorophyll biosynthesis pathways) each have discrete evolutionary histories and in some cases the entire photosynthetic complex has been transferred as a unit.

Aerobic respiration

Electron transport chains, in both aerobic and anaerobic respiration, are coupled via electrochemical proton potential to ADP phosphorylation with inorganic phosphate, which is catalysed by ATP synthase. The genes coding for subunits of this enzyme complex are some of the most studied regarding LGT, being a paradigmatic example of essential genes that experience frequent exchange across vast phylogenetic distances (Hilario and Gogarten, 1993). Evidence has been accumulating over the last decade that respiratory chains are rather dynamic systems that display great variability in their components. Aerobic respiration, where oxygen is used as the terminal electron acceptor, has been particularly well studied, both at the biochemical and genetic levels. Several of the proteins involved in this process in Archaea have been independently suggested to be of bacterial origin (Kennedy *et al.*, 2001; Lemos *et al.*, 2002; Osborne and Gennis, 1999; Pereira *et al.*, 2001). Aerobic respiration is also very patchily distributed among both Bacteria and Archaea, as a single order can contain both obligate aerobes and strict anaerobes (i.e. Sulfolobales).

The aerobic respiratory chain (Figure 2.2) is generally described as five complexes. Complexes I and II are the two main dehydrogenases involved in aerobic respiration and transfer of electrons from NADH and succinate, respectively, to the quinone pool (Lancaster, 2002). The main component of this pool varies greatly among prokaryotes, but is often ubiquinone or menaquinone (Collins and Jones, 1981). The

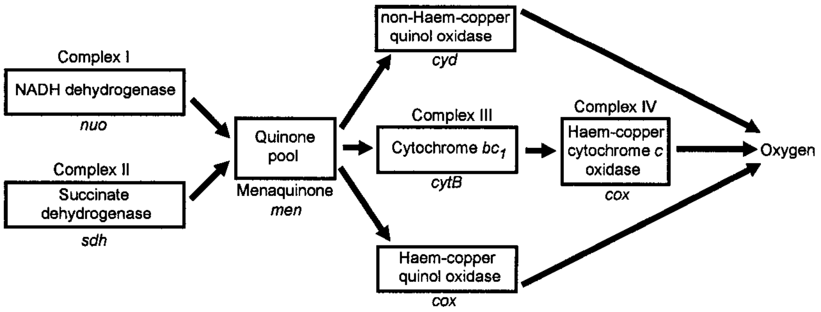


Figure 2.2 Diagram of the aerobic electron transport chain in prokaryotes. Elements of this transport chain present in a given organism differ from one organism to another. Only the most widespread and well-studied proteins are shown here: many organisms harbor other electron carriers in their transport chains. Boxes contain the name of the enzyme complexes forming the respiratory chain and the name of the genes encoding major components of these complexes are indicated underneath.

quinones can be oxidized by haem or non-haem-copper quinol oxidases and by cytochrome *bc*$_1$ (complex III). Quinol oxidases transfer electrons directly to oxygen. Cytochrome *bc*$_1$ is subsequently oxidized by haem-copper cytochrome *c* oxidase (complex IV), which then relays the electrons to oxygen (Lubben, 1995). The proton gradient created by the respiratory chain is then used by ATP synthase (complex V) to produce ATP. Other electron carriers, usually small cytoplasmic molecules, are also involved in the transport chains of some prokaryotes (i.e. Rieske proteins, ferredoxins and blue copper proteins).

Gene duplication is frequent for many components of aerobic respiratory chains, which makes it hard to identify cases of LGT: they can be confounded with differential loss of paralogs. Furthermore, most of these proteins are either small or have highly variable membrane bound regions, which greatly reduces the phylogenetic information they contain. However, gene transfers over large phylogenetic distances (inter-domain) have been reliably identified using phylogeny in conjunction with distribution patterns. Updated versions of aerobic respiration protein phylogenies from earlier literature, with more recent sequence data included, are presented in Figure 2.3.

The genes coding for the subunits of NADH dehydrogenase are frequently found as an operon in prokaryotes (10 to 14 subunits encoded by *nuo* genes). This is the case for the *nuo* genes of *Halobacterium sp.* NRC-1, which have recently been suggested to be of bacterial origin based on sequence similarity and phylogenetic analysis of the different subunits (Kennedy *et al.*, 2001). It is also clear, with the addition of recently sequenced genomes to phylogenetic analyses, that the strictly anaerobic hyperthermophile *Archaeoglobus fulgidus* acquired several homologs of *nuo* subunit genes from a relative

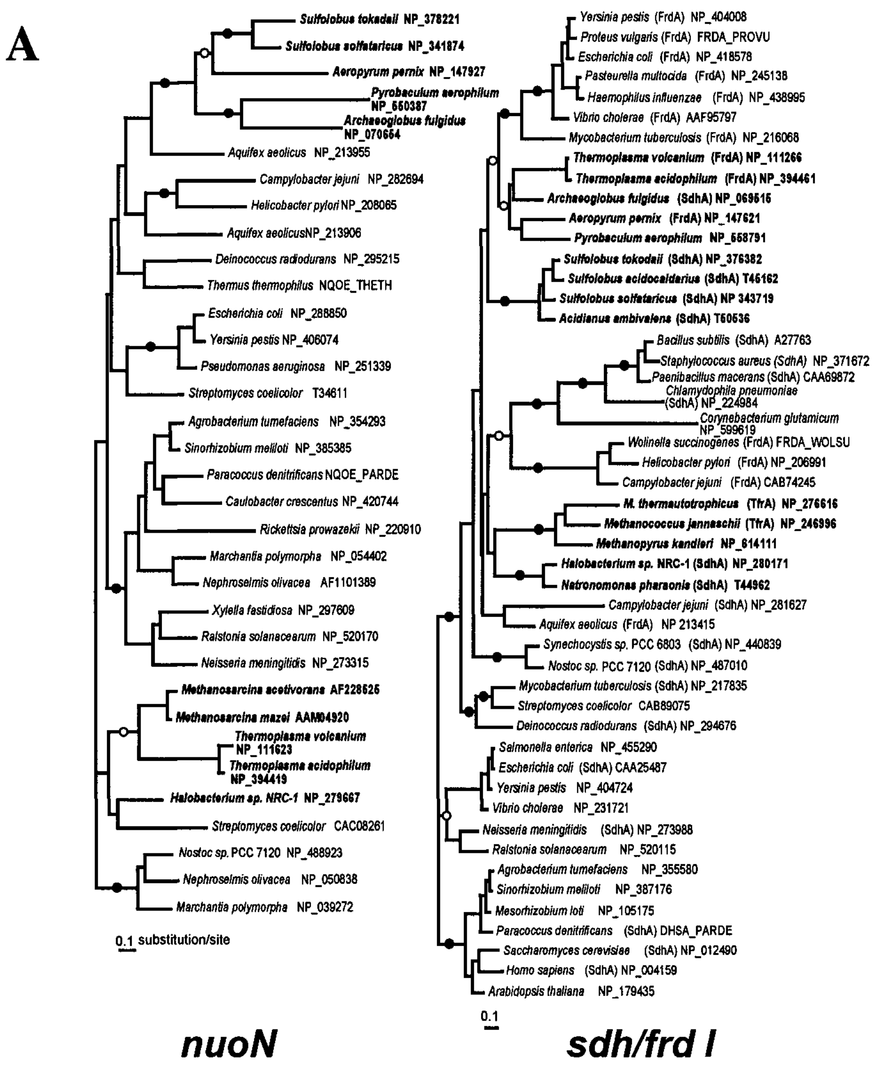

A
Sulfolobus tokadaii NP_378221
Sulfolobus solfataricus NP_341874
Aeropyrum pernix NP_147927
Pyrobaculum aerophilum NP_550387
Archaeoglobus fulgidus NP_070664
Aquifex aeolicus NP_213955
Campylobacter jejuni NP_282694
Helicobacter pylori NP_208065
Aquifex aeolicusNP_213906
Deinococcus radiodurans NP_295215
Thermus thermophilus NQOE_THETH
Escherichia coli NP_288850
Yersinia pestis NP_406074
Pseudomonas aeruginosa NP_251339
Streptomyces coelicolor T34611
Agrobacterium tumefaciens NP_354293
Sinorhizobium meliloti NP_385385
Paracoccus denitrificans NQOE_PARDE
Caulobacter crescentus NP_420744
Rickettsia prowazekii NP_220910
Marchantia polymorpha NP_054402
Nephroselmis olivacea AF1101389
Xylella fastidiosa NP_297609
Ralstonia solanacearum NP_520170
Neisseria meningitidis NP_273315
Methanosarcina acetivorans AF228525
Methanosarcina mazei AAM04920
Thermoplasma volcanium NP_111623
Thermoplasma acidophilum NP_394419
Halobacterium sp. NRC-1 NP_279667
Streptomyces coelicolor CAC08261
Nostoc sp. PCC 7120 NP_488923
Nephroselmis olivacea NP_050838
Marchantia polymorpha NP_039272
0.1 substitution/site
nuoN
Yersinia pestis (FrdA) NP_404008
Proteus vulgaris (FrdA) FRDA_PROVU
Escherichia coli (FrdA) NP_418578
Pasteurella multocida (FrdA) NP_245138
Haemophilus influenzae (FrdA) NP_438995
Vibrio cholerae (FrdA) AAF95797
Mycobacterium tuberculosis (FrdA) NP_216068
Thermoplasma volcanium (FrdA) NP_111266
Thermoplasma acidophilum (FrdA) NP_394461
Archaeoglobus fulgidus (SdhA) NP_069515
Aeropyrum pernix (FrdA) NP_147621
Pyrobaculum aerophilum NP_558791
Sulfolobus tokodaii (SdhA) NP_376382
Sulfolobus acidocaldarius (SdhA) T46162
Sulfolobus solfataricus (SdhA) NP 343719
Acidianus ambivalens (SdhA) T50536
Bacillus subtilis (SdhA) A27763
Staphylococcus aureus (SdhA) NP_371672
Paenibacillus macerans (SdhA) CAA69872
Chlamydophila pneumoniae (SdhA) NP_224984
Corynebacterium glutamicum NP_599619
Wolinella succinogenes (FrdA) FRDA_WOLSU
Helicobacter pylori (FrdA) NP_206991
Campylobacter jejuni (FrdA) CAB74245
M. thermautotrophicus (TfrA) NP_276616
Methanococcus jannaschii (TfrA) NP_246996
Methanopyrus kandleri NP_614111
Halobacterium sp. NRC-1 (SdhA) NP_280171
Natronomonas pharaonis (SdhA) T44962
Campylobacter jejuni (SdhA) NP_281627
Aquifex aeolicus (FrdA) NP 213415
Synechocystis sp. PCC 6803 (SdhA) NP_440839
Nostoc sp. PCC 7120 (SdhA) NP_487010
Mycobacterium tuberculosis (SdhA) NP_217835
Streptomyces coelicolor CAB89075
Deinococcus radiodurans (SdhA) NP_294676
Salmonella enterica NP_455290
Escherichia coli (SdhA) CAA25487
Yersinia pestis NP_404724
Vibrio cholerae NP_231721
Neisseria meningitidis (SdhA) NP_273988
Ralstonia solanacearum NP_520115
Agrobacterium tumefaciens NP_355580
Sinorhizobium meliloti NP_387176
Mesorhizobium loti NP_105175
Paracoccus denitrificans (SdhA) DHSA_PARDE
Saccharomyces cerevisiae (SdhA) NP_012490
Homo sapiens (SdhA) NP_004159
Arabidopsis thaliana NP_179435
0.1
sdh/frd I

Figure 2.3 Best maximum likelihood phylogenetic trees of aerobic respiratory chain proteins. Black dots indicate maximum likelihood distances bootstrap support over 95% and white dots support over 80%. Archaea are highlighted in bold. (A) subunit N of NADH dhydrogenase (*nuo*N) and subunit I of succinate dehydrogenase/fumarate reductase (*sdh*I/*frd*I). (B) menaquinone biosynthesis enzymes 1,4-dihydroxy-2-naphtoic acid prenyltransferase (*menA*) and synthase (*menB*) and subunit I of cytochrome *bd* oxidase (*cyd*I). (C) cytochrome *b* (*cyt*B) and cytochrome b_6 + subunit IV (split *cyt*B) and subunit I of cytochrome *c* oxidase (*cox*I).

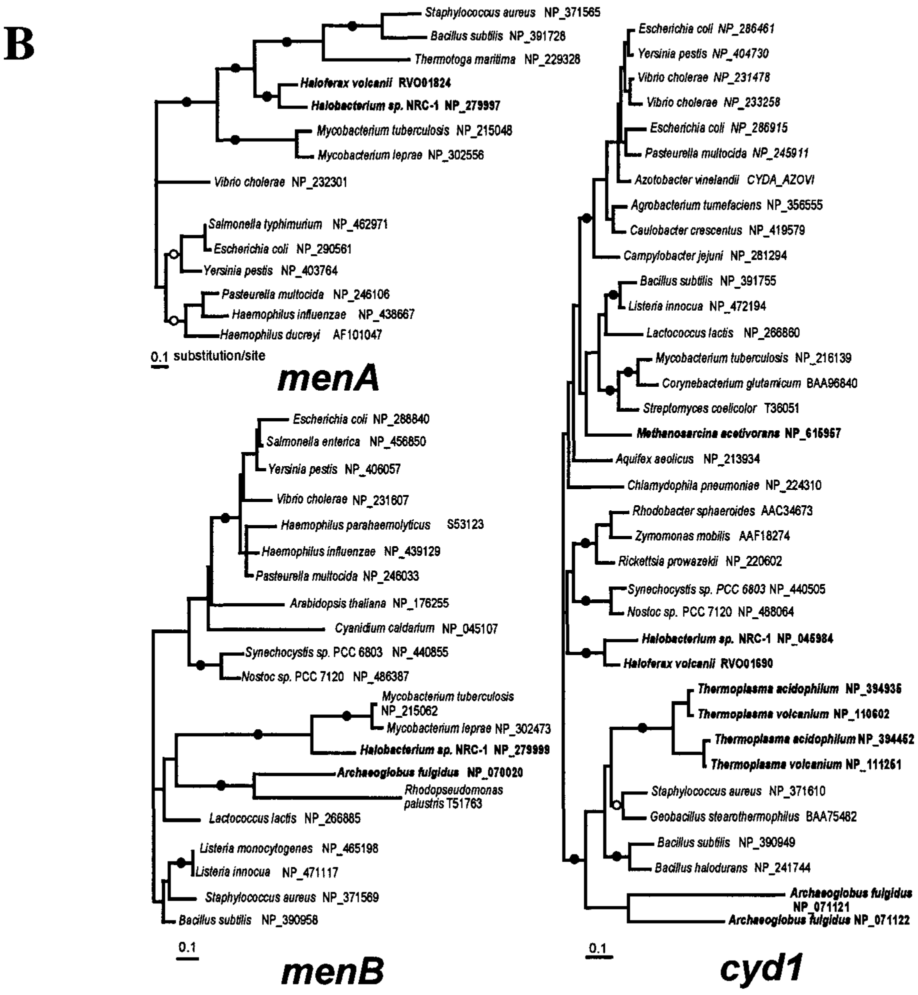

B
Staphylococcus aureus NP_371565
Bacillus subtilis NP_391728
Thermotoga maritima NP_229328
Haloferax volcanii RVO01824
Halobacterium sp. NRC-1 NP_279997
Mycobacterium tuberculosis NP_215048
Mycobacterium leprae NP_302556
Vibrio cholerae NP_232301
Salmonella typhimurium NP_462971
Escherichia coli NP_290561
Yersinia pestis NP_403764
Pasteurella multocida NP_246106
Haemophilus influenzae NP_438667
Haemophilus ducreyi AF101047
0.1 substitution/site
menA
Escherichia coli NP_288840
Salmonella enterica NP_456850
Yersinia pestis NP_406057
Vibrio cholerae NP_231607
Haemophilus parahaemolyticus S53123
Haemophilus influenzae NP_439129
Pasteurella multocida NP_246033
Arabidopsis thaliana NP_176255
Cyanidium caldarium NP_045107
Synechocystis sp. PCC 6803 NP_440855
Nostoc sp. PCC 7120 NP_486387
Mycobacterium tuberculosis NP_215062
Mycobacterium leprae NP_302473
Halobacterium sp. NRC-1 NP_279999
Archaeoglobus fulgidus NP_070020
Rhodopseudomonas palustris T51763
Lactococcus lactis NP_266885
Listeria monocytogenes NP_465198
Listeria innocua NP_471117
Staphylococcus aureus NP_371569
Bacillus subtilis NP_390958
0.1
menB
Escherichia coli NP_286461
Yersinia pestis NP_404730
Vibrio cholerae NP_231478
Vibrio cholerae NP_233258
Escherichia coli NP_286915
Pasteurella multocida NP_245911
Azotobacter vinelandii CYDA_AZOVI
Agrobacterium tumefaciens NP_356555
Caulobacter crescentus NP_419579
Campylobacter jejuni NP_281294
Bacillus subtilis NP_391755
Listeria innocua NP_472194
Lactococcus lactis NP_266860
Mycobacterium tuberculosis NP_216139
Corynebacterium glutamicum BAA96840
Streptomyces coelicolor T36051
Methanosarcina acetivorans NP_615957
Aquifex aeolicus NP_213934
Chlamydophila pneumoniae NP_224310
Rhodobacter sphaeroides AAC34673
Zymomonas mobilis AAF18274
Rickettsia prowazekii NP_220602
Synechocystis sp. PCC 6803 NP_440505
Nostoc sp. PCC 7120 NP_488064
Halobacterium sp. NRC-1 NP_045984
Haloferax volcanii RVO01690
Thermoplasma acidophilum NP_394935
Thermoplasma volcanium NP_110602
Thermoplasma acidophilum NP_394452
Thermoplasma volcanium NP_111261
Staphylococcus aureus NP_371610
Geobacillus stearothermophilus BAA75482
Bacillus subtilis NP_390949
Bacillus halodurans NP_241744
Archaeoglobus fulgidus NP_071121
Archaeoglobus fulgidus NP_071122
0.1
cyd1

Figure 2.3 Best maximum likelihood phylogenetic trees of aerobic respiratory chain proteins (continued).

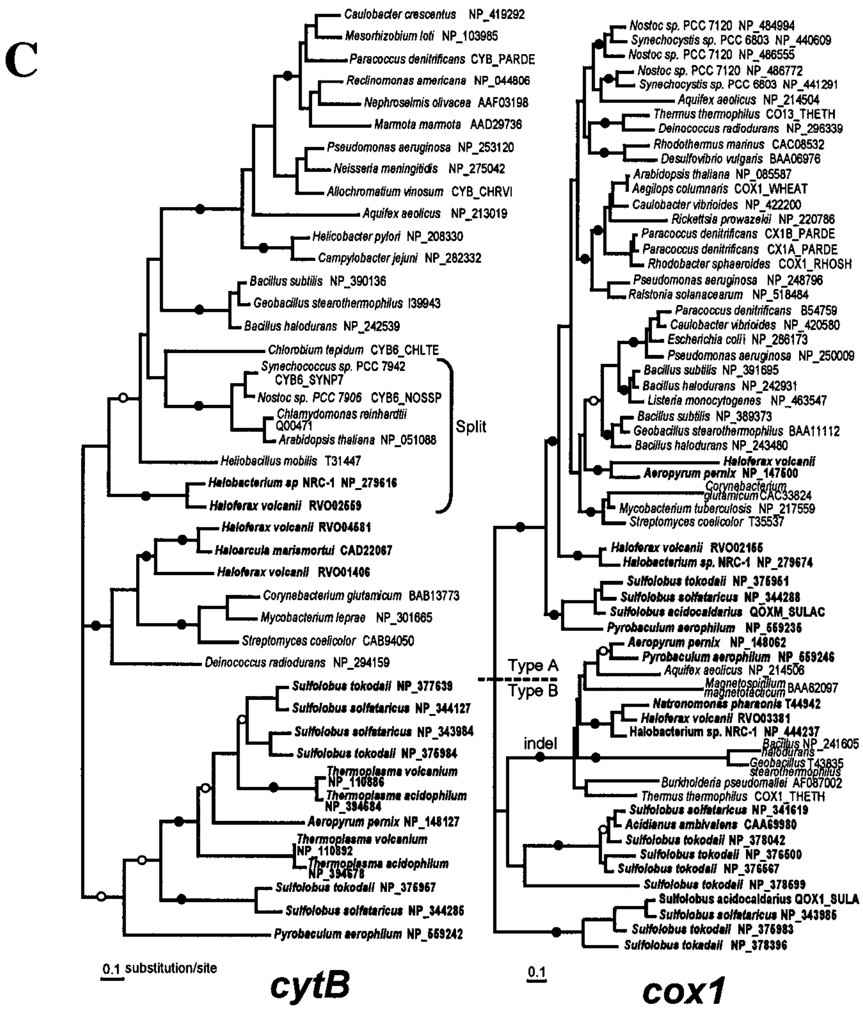

C
Caulobacter crescentus NP_419292
Mesorhizobium loti NP_103985
Paracoccus denitrificans CYB_PARDE
Reclinomonas americana NP_044806
Nephroselmis olivacea AAF03198
Marmota marmota AAD29736
Pseudomonas aeruginosa NP_253120
Neisseria meningitidis NP_275042
Allochromatium vinosum CYB_CHRVI
Aquifex aeolicus NP_213019
Helicobacter pylori NP_208330
Campylobacter jejuni NP_282332
Bacillus subtilis NP_390136
Geobacillus stearothermophilus I39943
Bacillus halodurans NP_242539
Chlorobium tepidum CYB6_CHLTE
Synechococcus sp. PCC 7942 CYB6_SYNP7
Nostoc sp. PCC 7906 CYB6_NOSSP
Chlamydomonas reinhardtii Q00471
Arabidopsis thaliana NP_051088
Heliobacillus mobilis T31447
Halobacterium sp NRC-1 NP_279616
Haloferax volcanii RVO02559
Split
Haloferax volcanii RVO04581
Haloarcula marismortui CAD22067
Haloferax volcanii RVO01406
Corynebacterium glutamicum BAB13773
Mycobacterium leprae NP_301665
Streptomyces coelicolor CAB94050
Deinococcus radiodurans NP_294159
Sulfolobus tokodaii NP_377639
Sulfolobus solfataricus NP_344127
Sulfolobus solfataricus NP_343984
Sulfolobus tokodaii NP_375984
Thermoplasma volcanium NP_110886
Thermoplasma acidophilum NP_394684
Aeropyrum pernix NP_148127
Thermoplasma volcanium NP_110892
Thermoplasma acidophilum NP_394678
Sulfolobus tokodaii NP_375967
Sulfolobus solfataricus NP_344285
Pyrobaculum aerophilum NP_559242
0.1 substitution/site
cytB
Nostoc sp. PCC 7120 NP_484994
Synechocystis sp. PCC 6803 NP_440609
Nostoc sp. PCC 7120 NP_486555
Nostoc sp. PCC 7120 NP_486772
Synechocystis sp. PCC 6803 NP_441291
Aquifex aeolicus NP_214504
Thermus thermophilus CO13_THETH
Deinococcus radiodurans NP_296339
Rhodothermus marinus CAC08532
Desulfovibrio vulgaris BAA06976
Arabidopsis thaliana NP_085587
Aegilops columnaris COX1_WHEAT
Caulobacter vibrioides NP_422200
Rickettsia prowazekii NP_220786
Paracoccus denitrificans CX1B_PARDE
Paracoccus denitrificans CX1A_PARDE
Rhodobacter sphaeroides COX1_RHOSH
Pseudomonas aeruginosa NP_248796
Ralstonia solanacearum NP_518484
Paracoccus denitrificans B54759
Caulobacter vibrioides NP_420580
Escherichia coli NP_286173
Pseudomonas aeruginosa NP_250009
Bacillus subtilis NP_391695
Bacillus halodurans NP_242931
Listeria monocytogenes NP_463547
Bacillus subtilis NP_389373
Geobacillus stearothermophilus BAA11112
Bacillus halodurans NP_243480
Haloferax volcanii
Aeropyrum pernix NP_147600
Corynebacterium glutamicum CAC33824
Mycobacterium tuberculosis NP_217559
Streptomyces coelicolor T35537
Haloferax volcanii RVO02155
Halobacterium sp. NRC-1 NP_279674
Sulfolobus tokodaii NP_375951
Sulfolobus solfataricus NP_344288
Sulfolobus acidocaldarius QOXM_SULAC
Pyrobaculum aerophilum NP_559236
Type A
Type B
Aeropyrum pernix NP_148062
Pyrobaculum aerophilum NP_559246
Aquifex aeolicus NP_214506
Magnetospirillum magnetotacticum BAA62097
Natronomonas pharaonis T44942
Haloferax volcanii RVO03381
Halobacterium sp. NRC-1 NP_444237
indel
Bacillus halodurans NP_241605
Geobacillus stearothermophilus T43835
Burkholderia pseudomallei AF087002
Thermus thermophilus COX1_THETH
Sulfolobus solfataricus NP_341619
Acidianus ambivalens CAA69980
Sulfolobus tokodaii NP_378042
Sulfolobus tokodaii NP_376500
Sulfolobus tokodaii NP_376567
Sulfolobus tokodaii NP_378699
Sulfolobus acidocaldarius QOX1_SULA
Sulfolobus solfataricus NP_343986
Sulfolobus tokodaii NP_375983
Sulfolobus tokadaii NP_378396
0.1
cox1

Figure 2.3 Best maximum likelihood phylogenetic trees of aerobic respiratory chain proteins (continued).

of the facultatively aerobic crenarchaeote *Pyrobaculum aerophilum* (Figure 2.3A). These *nuo* subunits homologs encode an $F_{420}H_2$ dehydrogenase. This enzyme is involved in reoxidation of the reduced cofactor $F_{420}H_2$, which plays a role in lactate oxidation in *Archaeoglobus* (Kunow *et al.*, 1994).

The analysis of the *Halobacterium sp.* NRC-1 genome also revealed a menaquinone biosynthesis operon of *men* genes very similar to bacterial homologs (Kennedy *et al.*, 2001). Several of these genes indeed cluster very strongly with Gram positive bacteria in phylogenies, confirming their bacterial origin (Figure 2.3B). *Archaeoglobus* also harbors a *menB* gene, which is very similar to its ortholog from the α-proteobacterium *Rhodopseudomonas palustris*, thus likely acquired from the latter by the archaeon.

Succinate dehydrogenase (*sdh*) and fumarate dehydrogenase (*frd*), respectively involved in aerobic and anaerobic respiration, cannot presently be distinguished at the structure level (both are composed of two subunits, A and B) or even the sequence level. The patchy distribution of these enzymes within Archaea suggests that they do not descend from a common ancestor, but rather have been acquired from Bacteria through mutltiple LGT events (Lemos *et al.*, 2002). This is apparent in a phylogenetic tree of subunit A, where Archaea are polyphyletic (Figure 2.3A). The tree also suggests more LGT within the Archaea, between crenarchaeotes and euryarchaeotes.

Only one case of lateral transfer of cytochrome *b* (*cyt*B) has so far been reported, from Proteobacteria to *Aquifex* (Schutz *et al.*, 2000). With several new archaeal genome sequences available, it becomes clear that the only two euryarchaeal groups harbouring a CytB (*Thermoplasma* and extremely halophilic archaea) also acquired it by LGT.

Thermoplasma cytB genes cluster strongly with crenarchaeotes (Figure 2.3C). The two different *cytB* genes found in extremely halophilic archaea, one of the usual unsplit type and the other of the cytochrome b_6 + subunit IV split type, most likely originated from two different bacterial sources (Figure 2.3C).

Phylogeny of subunit I of non-haem-copper quinol oxidase (*cyd*, composed of subunits I and II, also known as cytochrome *bd* oxidase) suggested that *Archaeoglobus fulgidus* and *Halobacterium sp.* NRC-1 possess a Cyd protein of bacterial origin (Osborne and Gennis, 1999). More recently, subunit I *cyd* genes have also been found in the genomes of *Methanosarcina* and *Thermoplasma*. Addition of these gene sequences to the subunit I phylogeny produces a polyphyletic Archaea, signifying multiple independent acquisitions by LGT (Figure 2.3B). *Halobacterium* and *Methanosarcina* are the only archaea known to also have a subunit II, whereas this subunit is present in a vast range of bacteria, a distribution pattern suggesting a bacterial origin.

The homologous haem-copper quinol oxidase and haem-copper cytochrome *c* oxidase (*cox*) of the two most widespread types, A and B, are composed of two subunits, I and II (Pereira *et al.*, 2001). Archaeal haem-copper oxidases generally have a higher amino acid identity with bacterial homologs rather than homologs from other archaeal groups. An origin among Gram positive bacteria (Firmicutes and Actinobacteria) has also been proposed for these enzymes (Pereira *et al.*, 2001). Furthermore, an updated phylogeny confirms a polyphyletic pattern for archaeal *cox* genes among bacterial homologs (Figure 2.3C). The *cox* gene tree also suggests LGT among archaea (between *Aeropyrum pernix* and *Haloferax volcanii*) and supports the initial claim of transfer from Firmicutes to Proteobacteria by Castresana *et al.* (Castresana *et al.*, 1994). There is therefore evidence for LGT of *cox* genes both within and between the two prokaryotic domains.

Phylogenetic analysis is difficult for several electron carrier proteins (blue copper proteins, Rieske proteins, ferredoxins), because of their short size and rapid rate of evolution. However, the patchy distribution pattern of these proteins in Archaea suggests that they are frequently lost in different lineages and regained by LGT. For example, ferredoxins of the 2Fe-2S type are found solely in cyanobacteria and extremely halophilic archaea, making it very unlikely that they evolved in the common ancestor of Bacteria and Archaea and were lost afterwards in all lineages but these two (Pfeifer *et al.*, 1993). The same reasoning can apply to blue copper proteins, which are found in several bacterial lineages but only in a few select groups of archaea (*Methanoarcina*, *Sulfolobus*, *Thermoplasma* and extremely halophilic archaea) (Van Driessche *et al.*, 1999). Rieske proteins are more widespread but still have a very patchy distribution, both in Bacteria and Archaea (Henninger *et al.*, 1999). Usually peripheral to the respiratory system, these proteins have certainly been lost many times. However, explaining their presence in a few distant lineages without invoking LGT would not only require an enormous number of loss events, but also begs for an explanation as to how they were maintained throughout evolution. LGT reduces drastically the number of loss events required to explain their distribution and allows these proteins to escape evolutionary elimination by dispersing them to other lineages.

Nitrogen Fixation

Nitrogen fixation, the conversion of N_2 (dinitrogen) to NH_4^+ (ammonium), is, like many steps in the global nitrogen cycle, restricted to the prokaryotic domains. In both

Bacteria and Archaea, it is accomplished through the nitrogenase complex (Evans and Burris, 1992). There are two components within this complex, dinitrogenase (component 1) and dinitrogenase reductase (component 2). The dinitrogenase (encoded by *nifD* and *nifK*) contains an iron-sulfur center and an iron-molybdenum co-factor. The dinitrogenase reductase (encoded by *nifH*) contains a single iron-sulfur center. Variant forms, described as "alternative nitrogenases", are also found in some bacteria and archaea. These are encoded by homologs of *nifD, K* and *H* designated *vnfD, K* and *H* or *anfD, K* and *H*, and contain an iron-vanadium (*vnfD,K,H*) or iron-only (*anfD,K,H*) cofactor, in place of iron-molybdenum in component 1. The dinitrogenase component also contains a third subunit in alternative nitrogenases (encoded by *vnf*G or *anf*G), which appears to play a role in cofactor processing and function. The multiple subunits of nitrogenase complexes are almost always encoded within one operon or gene cluster, often along with other genes involved in the regulation of nitrogen fixation or related processes (Kessler *et al.*, 1998). Some prokaryotes harbor two or all three of the different nitrogenase complexes (*nif, anf* and *vnf*) (Bishop and Premakumar, 1992; Chien *et al.*, 2000).

Nitrogen-fixing organisms are very widespread among Bacteria. Members of groups as diverse as the Proteobacteria (all subdivisions), Cyanobacteria, Firmicutes, Actinobacteria, Fusobacteria, Chlorobia and the Spirochaetes all possess this biochemical activity and the homologs of the *nifD*, *nifK* and *nifH* genes required to perform it (Young, 1992). The *nifH* gene, which is very conserved in sequence among different orthologs, has been used as a molecular marker to detect and identify nitrogen-fixing organisms in a variety of environments (Mehta *et al.*, 2003). However, in phylogenetic studies of this gene, multiple species of the β-proteobacterial genus *Azoarcus* cluster strongly with *nifH* from α-proteobacteria, instead of other species of *Azoarcus* (Hurek *et al.*, 1997), suggesting LGT of *nifH* among proteobacteria. Phylogenetic analyses to assess the possible lateral transfer of the *nif* genes among bacteria were also performed by Hirsch *et al.* (Hirsch *et al.*, 1995). Their *nifK* phylogeny displayed a clustering of the orthologs from the cyanobacterium *Anabaena* and the actinobacterium *Frankia* with proteobacterial sequences. More recent phylogenetic analyses of the *nifH* gene and its homologs (including *vnfH* and *anfH*), with a greatly increased organismal sampling, also recover these specific relationships (Ohkuma *et al.*, 1999). Our phylogeny of the *nifD* gene, updated from Chien *et al.* (Chien *et al.*, 2000), is congruent with both *nifK* and *nifH* phylogenies, grouping together molybdenum-dependent nitrogenases from cyanobacteria, *Frankia* and proteobacteria (Cluster I, Figure 2.4).

On a smaller phylogenetic scale, Qian *et al.* (Qian *et al.*, 2003) have detected conflict between *nifD* and SSU rRNA phylogenies of multiple strains from the Rhizobiales genus *Bradyrhizobium.* They interpreted this incongruence between phylogenies as lateral transfer of the *nifD* gene within this genus. LGT could indeed be a major mode of dissemination for the genes necessary for nitrogen fixation within the α-proteobacterial order Rhizobiales. This is suggested by the fact that proteins involved in this process are often encoded on what are known as symbiotic, or *sym*, plasmids (Finan *et al.*, 2001). These plasmids can carry a variety of genes for N_2 fixation (nitrogenase genes) and metabolism as well as genes for the colonization of host plants and

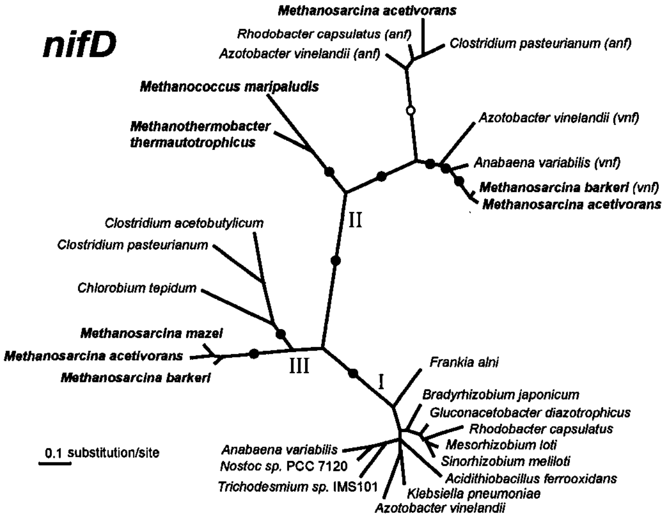

nifD
Methanosarcina acetivorans
Rhodobacter capsulatus (anf)
Azotobacter vinelandii (anf)
Clostridium pasteurianum (anf)
Methanococcus maripaludis
Methanothermobacter thermautotrophicus
Azotobacter vinelandii (vnf)
Anabaena variabilis (vnf)
Methanosarcina barkeri (vnf)
Methanosarcina acetivorans
II
Clostridium acetobutylicum
Clostridium pasteurianum
Chlorobium tepidum
Methanosarcina mazei
Methanosarcina acetivorans
Methanosarcina barkeri
III
I
Frankia alni
Bradyrhizobium japonicum
Gluconacetobacter diazotrophicus
Rhodobacter capsulatus
Mesorhizobium loti
Sinorhizobium meliloti
Acidithiobacillus ferrooxidans
Klebsiella pneumoniae
Azotobacter vinelandii
Anabaena variabilis
Nostoc sp. PCC 7120
Trichodesmium sp. IMS101
0.1 substitution/site

Figure 2.4 Best maximum likelihood phylogenetic tree of the dinitrogenase (component 1) α subunit (*nifD*). Black dots indicate maximum likelihood distances bootstrap support over 95% and white dots support over 80%. Archaea are highlighted in bold. The three major *nif* clusters are recovered (Chien *et al.*, 2000). Cluster I includes conventional molybdenum-dependent nitrogenases from Proteobacteria, *Frankia* and Cyanobacteria. Cluster II includes molybdenum-dependent nitrogenases from methanogenic archaea, molybdenum- and vanadium-independent alternative nitrogenases (*anf*) and vanadium-dependent alternative nitrogenases (*vnf*). Cluster III includes molybdenum-dependent enzymes from *Clostridium* and *Methanosarcina* and an uncharacterized enzyme from *Chlorobium.*

nodulation. For example, *Sinorhizobium meliloti* contains two megaplasmids in addition to its chromosome. The larger plasmid (called pSymB, pExo or pRmeSU47B), carries several genes required for nodulation, while the smaller (pSymA, pNod-Nif or pRmeSU47A) contains both nodulation and nitrogen fixation genes (Finan *et al.*, 2001). Another member of the Rhizobiales, *Rhizobium etli* CFN42, contains six plasmids. One of these (the symbiotic plasmid) carries the majority of genes required for nitrogen fixation and another is required for the mobilization of the former (Tun-Garrido *et al.*, 2003).

Within Archaea, the ability to fix nitrogen is so far found only in methanogenic archaea. Although the methanogens are not a monophyletic group of archaea, homologs of the *nif* genes and/or nitrogenase activity are found in members of all orders with this metabolic property (Methanopyrales, Methanococcales, Methanobacteriales, Methanomicrobiales and Methanosarcinales). The most studied of the methanogens regarding nitrogen fixation, *Methanocaldococcus maripaludis*, contains a *nif* gene operon, with a gene order similar to that found in bacterial operons (Kessler *et al.*, 1998) and which is part of a larger regulon with other nitrogen fixation genes (Kessler and Leigh, 1999). Similar operons have been found in all characterized methanogenic archaea (Kessler *et al.*, 1998). Some methanogens possess more than one cluster of nitrogenase genes. This was demonstrated for *Methanosarcina barkeri* 227, which harbors two separate clusters, most likely originating from different sources. The first cluster contains a complete set of nitrogenase structural genes *(nifHDK2)*; these show high similarity to gene products from *Clostridium pasteurianum* (Chien and Zinder, 1996) and *Chlorobium tepidum*, to the exclusion of homologs from other methanogenic archaea (Figure 2.4). The second cluster shows high similarity to bacterial vanadium-dependent alternative nitrogenase genes (*vnf*). The latter cluster also includes a *vnfG* homolog, encoding for the distinct subunit present only in alternative nitrogenases, so far found only in the genus *Methanosarcina* among archaea (Chien *et al.*, 2000). The *vnf*D gene of this cluster and its homolog from *Anabaena variabilis* show one of the highest degrees of identity ever found between bacterial and archaeal gene products (80% amino acid identity) and therefore most likely originates from a recent LGT between relatives of these species (Chien *et al.*, 2000).

Overall, nitrogenase complexes seem to have a very dynamic evolution in prokaryotes. LGT has been detected at various phylogenetic levels for the genes encoding nitrogenase subunits: among members of the genera *Bradyrhizobium*, between proteobacterial classes for some species of the genus *Azoarcus*, between bacterial phyla for Proteobacteria, Cyanobacteria and *Frankia* and between domains for *Methanosarcina* and *Anabaena*. The fact that nitrogen fixation genes are usually found within one cluster, both in Bacteria and Archaea, and that they have been identified on plasmids (at least within the Rhizobiales) make lateral transfer of this function all the more plausible.

Sulfate reduction

Biological sulfate reduction is an essential component of the biosphere's basal machinery (Barton and Tomei, 1995; Fauque, 1995; Klenk *et al.*, 1997; Shen *et al.*, 2001; Stetter *et al.*, 1987; Wagner *et al.*, 1998; Widdel, 1986, 1988). Not only does it have a great impact in the global sulfur cycle, but it may also be the dominant respiratory process in marine or freshwater anaerobic ecosystems (Castro *et al.*, 2000; Cooling *et al.*, 1996; Klenk *et al.*, 1997). Since the early sixties, two sulfate reduction pathways have been recognized in prokaryotes (Peck, 1961). These two pathways mostly differ in the amount of sulfate reduced, of H_2S produced and in the substrate they use.

The assimilatory pathway (Postgate, 1952) corresponds to the preponderant mechanism by which reduced sulfate is incorporated into organic compounds (Kredich, 1996) whereas the dissimilatory pathway (Postgate, 1952) uses sulfate as a final electron acceptor. Figure 2.5 presents a comparative view of these two pathways. Along the assimilatory route, inorganic sulfate is activated to adenosine-5'-phosphosulfate (APS). Subsequently, the APS is transformed to 3'phosphoadenosine-5'-phosphosulfate (PAPS) and the PAPS is itself reduced to sulfite. Finally, sulfite is reduced to sulfide, allowing it to react with *O*-acetyl-L-serine to form L-cysteine (Kredich, 1996; Peck, 1961). Unlike its counterpart, the dissimilatory pathway requires only three steps, the APS being directly transformed to sulfite. The assimilatory pathway is widespread across all domains of life: it does not show the patchy distribution we take as the hallmark of gene transfer. The dissimilatory pathway, on the other hand, clearly represents such an example and will be the only one considered here.

Five or six genes are thought to be involved in the dissimilatory pathway. Sulfate activation is catalyzed by a homotrimeric sulfate adenylyltransferase (Dahl and Truper, 2001) encoded by the *sat* gene (Figure 2.5). This enzyme is also known to act during the first step of the assimilatory pathway (catalyzed by a heterodimeric sulfate adenylyltransferase in *E. coli*) and to facilitate the reverse reaction in sulfur oxidizers (Foster *et al.*, 1994; Renosto *et al.*, 1991). The second step, corresponding to the transformation of APS in sulfite, is catalyzed by APS reductase, encoded by the *aps* α and *aps* β genes. The smallest active form of this enzyme is still debated, but an *aps* $\alpha_2\beta_2$

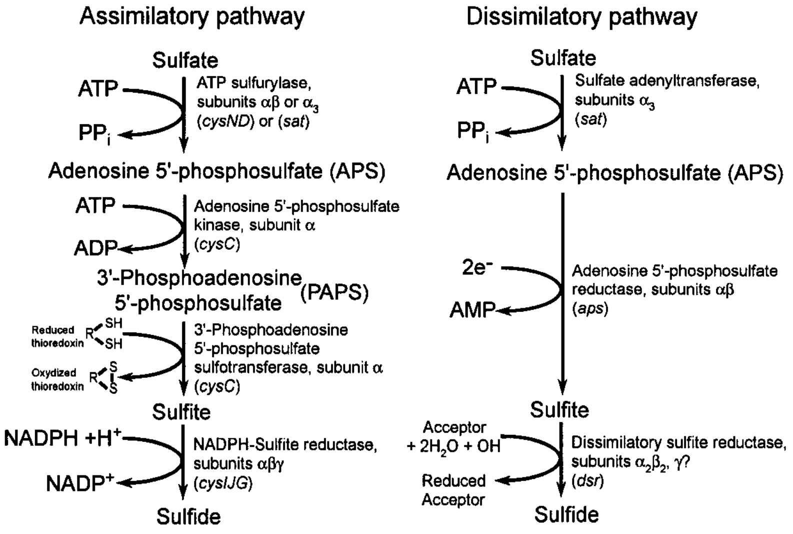

Assimilatory pathway
Sulfate
ATP
PPi
ATP sulfurylase, subunits αβ or α3 (cysND) or (sat)
Adenosine 5'-phosphosulfate (APS)
ATP
ADP
Adenosine 5'-phosphosulfate kinase, subunit α (cysC)
3'-Phosphoadenosine 5'-phosphosulfate (PAPS)
Reduced thioredoxin R SH SH
Oxydized thioredoxin R S S
3'-Phosphoadenosine 5'-phosphosulfate sulfotransferase, subunit α (cysC)
Sulfite
NADPH +H+
NADP+
NADPH-Sulfite reductase, subunits αβγ (cysIJG)
Sulfide
Dissimilatory pathway
Sulfate
ATP
PPi
Sulfate adenyltransferase, subunits α3 (sat)
Adenosine 5'-phosphosulfate (APS)
2e-
AMP
Adenosine 5'-phosphosulfate reductase, subunits αβ (aps)
Sulfite
Acceptor + 2H2O + OH-
Reduced Acceptor
Dissimilatory sulfite reductase, subunits α2β2, γ?
(dsr)
Sulfide

Figure 2.5 Schematic representation of the two pathways for the reduction of sulfate. Sulfate is reduced through the assimilatory pathway to be incorporated into organic compound. The dissimilatory pathway uses sulfate as a final electron acceptor.

heterotetrameric structure if often suggested (Dahl and Truper, 2001). The third and last enzyme of this pathway is encoded by two or three genes: *dsr* α, *dsr* β and a putative γ subunit that is encoded by the gene *dsvC* (*Archaeoglobus fulgidus* nomenclature).

When comparing the chromosomal organization of these genes between the three complete genomes that display the entire set of genes (i.e., the euryarchaeote *Archaeoglobus fulgidus*, the crenarchaeote *Pyrobaculum aerophilum*, and the green sulfur bacterium *Chlorobium tepidum*), the gene order seems rather flexible (Figure 2.6). Indeed, only the two subunits of the APS reductase and the *dsr* α and β subunit genes seem to always be co-expressed. However, as shown in Figure 2.6, all these genes tend to reside in one or two islands. Sperling *et al.* (Sperling *et al.*, 1998) showed that the *sat* gene and the two APS reductase subunit genes (*aps* α and *aps* β) are part of the same operon in *Archaeoglobus fulgidus.*

Dissimilatory sulfate reduction occurs rarely and patchily among prokaryotes (Friedrich, 2002; Larsen *et al.*, 1999; Shen *et al.*, 2001; Wagner *et al.*, 1998). Interestingly, all but one of the five lineages of sulfate reducers are bacterial (the archaeon *Archaeoglobus* vs. four bacterial groups: Thermodesulfobacteria, Nitrospira, δ-Proteobacteria and Firmicutes) (Friedrich, 2002; Shen *et al.*, 2001). This odd distribution as well as phylogenetic considerations have led to the postulate that *Archaeoglobus* sulfate reduction is of bacterial origin (Friedrich, 2002; Klein *et al.*, 2001; Larsen *et al.*, 1999). We review data bearing on each of the three steps of the dissimilatory pathway in turn.

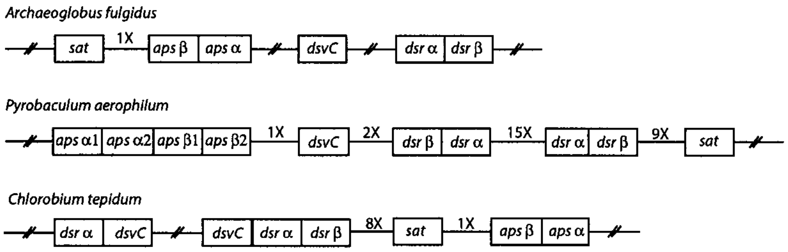


Figure 2.6 Genetic arrangement of the genes encoding for enzymes involved in sulfate reduction for the euryarchaeote *Archaeoglobus fulgidus*, the crenarchaeote *Pyrobaculum aerophilum* and the green sulfur bacteria *Chlorobium tepidum*. 1X represents one open reading frame found between two sulfate reduction genes.

Sulfate activation. Schwenn (Schwenn, 1997), showed that the phylogenetic topology inferred from *sat* genes with assimilatory function was in agreement with the generally accepted taxonomy (bacteria, plant, fungi and animals falling in four independent clades). Sperling (Sperling *et al.*, 1998) and more recently Dahl and Trüper (Dahl and Truper, 2001) included *sat* sequences from both dissimilatory reducers and assimilatory oxidizing organisms in phylogenies and obtained a distinct clade for each function. For these authors, this was an indication of an ancestral duplication of the *sat* gene, leading to assimilatory and dissimilatory paralogs. When including newly available sequences in a phylogenetic analysis, however, this deep paralogy scenario is inconsistent with topology of the tree obtained (Figure 2.7). For instance *Archaeoglobus* and *Pyrobaculum aerophilum sat* genes cluster very strongly together (98% bootstrap support) and with sequences from *Sulfolobus solfataricus* and *S. tokodaii* (both sulfide oxidizers) and the actinobacterium *Mycobacterium* (100% bootstrap supprot). These results not only refute the paralogy hypothesis but also show multiple LGT events for *sat*. Effectively, not only do *Archaeoglobus* and *Pyrobaculum aerophilum* form a strong clade, but *Pyroccocus* also clusters with *Aeropyrum pernix*. This arrangement differs greatly from phylogeny based on the SSU rRNA gene, where the crenarchaeotes (*Pyrobaculum, Sulfolobus, Aeropyrum*) and the euryarchaeotes (*Archaeoglobus, Pyroccocus*) form distinct clades. A similar picture is also observed among bacteria and between eukaryotes and bacteria. That is, we find high support (I) for a clade consisting of *Entamoeba* and the δ-proteobacterium *Desulfovibrio* to the exclusion of other eukaryotes and proteobacteria and (II) for a clade including most eukaryotic sequences plus *Aquifex* and two proteobacteria to the exclusion of other proteobacteria (Figure 2.7).

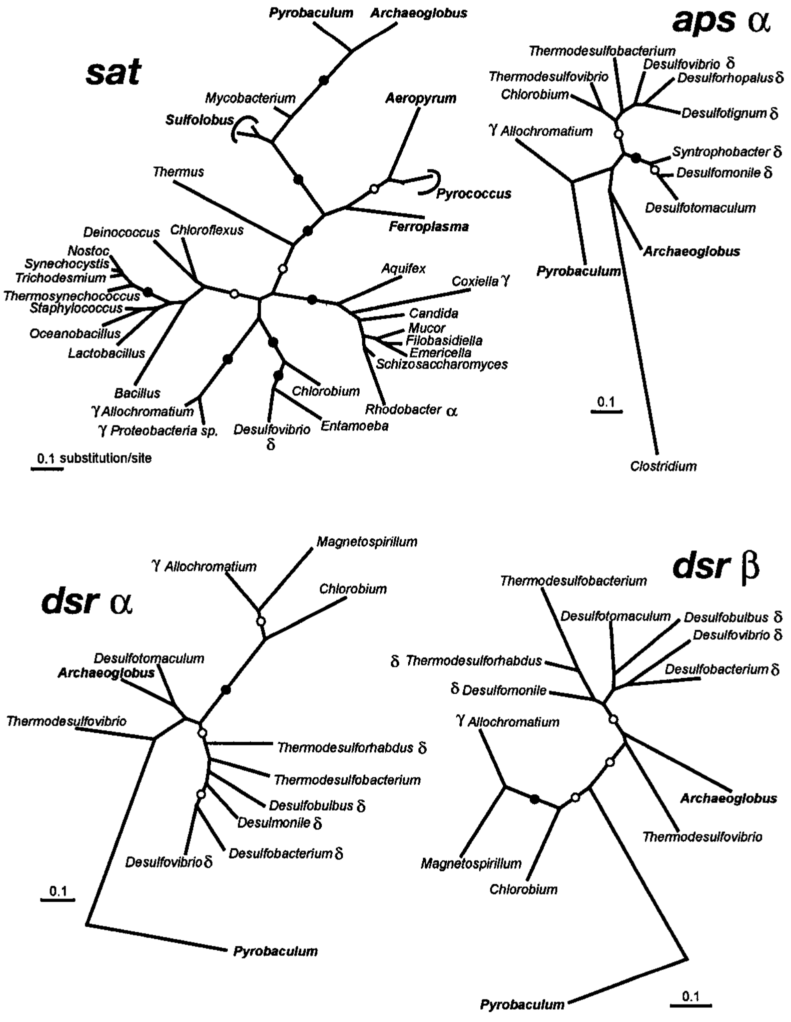


Figure 2.7 Best maximum likelihood phylogenetic trees of the enzymes of the dissimilatory pathway for sulfate reduction. Trees were constructed for sulfate adenyltransferase (*sat*), adenosine 5'-phosphosulfate (*aps*) and subunits α and β of dissimilatory sulfite reductase (*dsr* α and *dsr* β). Phylogenies were constructed using the PROML algorithm. Black dots indicate maximum likelihood distances bootstrap support over 95% and white dots support over 80%. Archaea are highlighted in bold.

APS to sulfite reduction. Hipp and co-workers (Hipp *et al.*, 1997) showed that sequences of *aps* genes from *Archaeoglobus fulgidus* were as or less distant from *Desulfovibrio vulgaris* (sulfate-reducing d-proteobacterium) than from *Allochromatium vinosum* (sulfur-oxidizing γ-proteobacterium). This statement led them to formulate two hypotheses: the presence of these genes in *Archaeoglogus fulgidus* is due to LGT, or *aps* α and *aps* β both experienced an ancestral duplication (two different functions catalyzed by two paralogous APS reductases). Even if they did not fully refute the LGT hypothesis, Hipp *et al.* (Hipp *et al.*, 1997) clearly favored the duplication scenario. More recently, Friedrich (2002) (Friedrich, 2002) significantly improved the sampling of the *aps α* gene and alternatively favored the LGT hypothesis. Effectively, the Archaeoglobales *aps α* genes are more similar to their proteobacterial homologs than is expected from the SSU rRNA marker.

Our reanalysis of *aps α* (Figure 2.7) cannot rule out an ancestral duplication but it clearly corroborates Friedrich's view that at least some transfer did occur. For example *Syntrophobacter* and *Desulfomonile* (both δ-proteobacteria) form a strong cluster with the firmicute *Desulfotomaculum* and not with the three other representatives of the δ-Proteobacteria (*Desulfovibro, Desulforhopalus* and *Desulfotignum*). The limited sampling of *aps β* makes it difficult to infer its evolutionary past. However, we observed (results not shown) that, similarly to *aps α*, the *aps β* from *Pyrobaculum* clusters with the oxidizing enzyme from *Allochromatium vinosum.* Like *aps α*, the Archaeoglobales *aps β* genes are more similar to their proteobacterial homologs than is expected from the SSU rRNA phylogeny.

Reduction of sulfite to sulfide. Over the last decade, dissimilatory sulfate reductase has been extensively studied (Dahl *et al.*, 1993; Dahl and Truper, 2001; Hipp *et al.*, 1997; Klein *et al.*, 2001; Larsen *et al.*, 1999; Molitor *et al.*, 1998; Stahl *et al.*, 2002; Wagner *et al.*, 1998). These investigations clearly demonstrated that *dsr* α and β are the product of an ancestral duplication (Hipp *et al.*, 1997; Molitor *et al.*, 1998). Among dissimilatory sulfate reducers, numerous and compelling examples of lateral gene transfer of *dsr* have been found (Klein *et al.*, 2001; Larsen *et al.*, 1999). For example, at the phylum level, some members of the firmicute genus *Desulfotomaculum* (*dsr* β) and representatives of the phylum *Thermodesulfobacteria* (*dsr* α and β) nest within the δ-Proteobacteria. Also, *Archaeoglobus* very likely represents an inter-domain LGT event as it is, with *Pyrobaculum*, the only member of the Archaea to possess *dsr* genes. Furthermore, those genes are more similar in sequence to their bacterial homologs than is expected from the divergence of their SSU rRNA genes. Klein *et al.* (Klein *et al.*, 2001) also found further evidence for the bacterial origin of *Archaeoglobus's dsr*. Indeed, using reciprocal rooting analyses (*dsr* α used to root *dsr* β and *vice versa*) *Thermodesulfovibrio* was found as the deepest branch (Klein *et al.*, 2001), not *Archaeoglobus*, as would be expected for an archaeal enzyme in a bacterial phylogeny.

Overall, it seems that the distribution of enzymes of the sulfate reduction dissimilatory pathway has been multiply affected by LGT. Evidence of transfer between different phyla of Bacteria has been observed in all three steps of the pathways, mostly involving proteobacteria and firmicutes. At a higher level it is likely that both *aps* and *dsr* genes were transferred between bacteria and the archaeon *Archaeoglobus*. Along the same lines, the close relationship of, for example, *Entamoeba* and *Desulfovibrio sat* genes, seems to indicate that this initial step was also affected by inter-domain LGT. Reasons for these multiples occurrences are not easy to guess, but the operon structure of the subunits of the enzymes catalyzing each step and the island distribution of the pathway have surely favored these dispersals.

Methylotrophy

Aerobic methylotrophs are bacteria capable of growth using C_1 compounds more reduced than formic acid as their carbon and energy source (Maden, 2000; Vorholt, 2002). In this process, termed methylotrophic metabolism, formaldehyde is formed as a central intermediate. Two main types of pathway have been described for the oxidation of formaldehyde to CO_2 in methylotrophic bacteria: (1) a cyclic pathway initiated by the condensation of formaldehyde with ribulose monophosphate, and (2) linear pathways that involve a dye-linked formaldehyde dehydrogenase or C_1 unit conversion bound to the cofactors tetrahydrofolate (H_4F), tetrahydromethanopterin (H_4MPT), gluthathione (GSH) or mycothiol (MySH). The pathways involving cofactors have certain steps in common: (1) spontaneous or enzyme-catalyzed condensation of formaldehyde with the cofactor, (2) oxidation of the cofactor-bound C_1 unit and its conversion to formate and (3) the oxidation of formate to CO_2 using formate dehydrogenase.

The H_4F-dependent pathway (Figure 2.8) is found in a number of bacteria, some archaea and eukaryotes. H_4F itself (as a C_1 fragment carrier) is involved in various essential biosynthetic processes, like purine and thymidylate synthesis, and is therefore a widespread cofactor. The thiol (glutathione/mycothiol)-dependent pathways are also widespread among bacteria, in addition to being found in plants, mammals and yeasts.

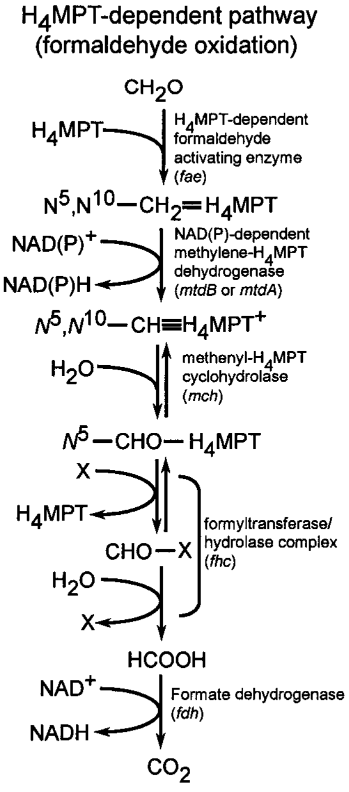

H4MPT-dependent pathway
(formaldehyde oxidation)
CH2O
H4MPT
H4MPT-dependent formaldehyde activating enzyme (fae)
N5,N10—CH2=H4MPT
NAD(P)+
NAD(P)H
NAD(P)-dependent methylene-H4MPT dehydrogenase (mtdB or mtdA)
N5,N10—CH≡H4MPT+
H2O
methenyl-H4MPT cyclohydrolase (mch)
N5—CHO—H4MPT
X
H4MPT
CHO—X
H2O
X
formyltransferase/ hydrolase complex (fhc)
HCOOH
NAD+
NADH
Formate dehydrogenase (fdh)
CO2

Figure 2.8 Schematic representation of the H_4MPT (tetrahydromethanopterin) dependent pathway for the oxidation of formaldehyde in methylotrophic metabolism. X represents an unknown cofactor, assumed to be an analogue to methanofuran from methanogenic archaea.

Reactions dependent on H_4MPT however, were initially thought to be restricted to methanogenic and sulfate-reducing archaea, where they play a central part in energy metabolism. In sulfate-reducers like *Archaeoglobus*, H_4MPT is used to mediate reactions involved in acetate oxidation. In methanogens, H_4MPT takes part in the reduction of CO_2 to methane.

In 1998, a methylotrophic bacterium, *Methylobacterium extroquens*, was discovered to harbour a cluster of genes homologous to known genes encoding for H_4MPT-dependent enzymes in methanogens and sulfate-reducing archaea (Chistoserdova *et al.*, 1998). Most enzymes encoded by genes from this cluster have been characterized and found to catalyze different steps of the oxidation of formaldehyde to formate (Figure 2.8). The distribution of one of these enzymes, methenyl-H_4MPT cyclohydrolase (Mch), has been thoroughly studied in prokaryotes. The *mch* gene was found in a variety of methylotrophic genera of Proteobacteria from the α, β and γ subdivisions (in addition to the α-proteobacterium *Methylobacterium extorquens* and methanogenic and sulfate-reducing archaea). Phylogenetic analysis of this gene revealed that the archaeal and bacterial homologs clearly form two distinct but related groups (Vorholt *et al.*, 1999). The sparse distribution of *mch* gene among the Bacteria in general and Proteobacteria in particular is hard to explain through differential loss. It is more likely, given the ubiquity of this gene and of the H_4MPT cofactor in methanogenic archaea (Thauer, 1998), that *mch* and functionally related genes (encoding for H_4MPT-dependent enzymes) have been laterally transferred from Archaea to Bacteria in an ancient event (Vorholt *et al.*, 1999).

One of the multiple analogous pathways for methylotrophic metabolism found in Proteobacteria is therefore likely to have been acquired in one or multiple LGT events from methanogenic or sulfate-reducing archaea. The H_4MPT-dependent pathway has yet to be found in bacteria other than Proteobacteria, even though some methylotrophs are known to exist within the Actinobacteria and the Firmicutes. This restricted distribution is surprising, given that the genes of this pathway are located in a single cluster in *M. extorquens*, making it easy to transfer as a unit. This cluster structure also suggests that the pathway was disseminated among Proteobacteria through LGT, a process thought to be involved in the formation and maintenance of operons (Lawrence and Roth, 1996). The history of methylotrophy needs much more detailed elaboration, and this capacity does not define a monophyletic prokaryotic group. It is however the central pathway of carbon and energy metabolism for many species, a further illustration of the role of LGT in adaptation at the most fundamental physiological level.

Discussion

The various physiological systems reviewed share some similarities regarding their evolution: (1) they are patchily distributed in prokaryotes; (2) LGT can be detected at multiple phylogenetic levels, either for entire operons/clusters or for single genes; (3) at least part of the system is encoded within a cluster or an operon. The implications of each of these shared characteristics for our understanding of the evolution of prokaryotic metabolism are numerous. I consider some of them here.

Differential loss vs. LGT

There is no pattern of gene distribution explicable by LGT that cannot also be explained by differential loss of paralogs from a more gene-rich ancestor, and many authors have used such reasoning, for instance Woese (1987) to explain the distribution of photosynthesis, and Castresana (2001) to rationalize phylogenetic patterns of bioenergetic proteins with the SSU rRNA tree. But often such differential loss scenarios will be wildly unparsimonious when other phylogenetic information is taken into account.

Take for instance the distribution of *menA* genes shown in Figure 2.3B. One *could* explain the "*(Haloferax/Halobacterium)(Thermotoga/Bacillus/Staphylococcus)*" cluster through differential loss in two ways. First, the last common ancestor of Bacteria and Archaea (the universal ancestor) contained at least two *menA*s. One has been lost in all but these four genomes, while the other was the ancestor of all other *menA*s. By this scenario, we must place the root of the universal prokaryotic tree between the five taxa and *Mycobacterium*. That is implausible on many other grounds, however, so we should examine the second differential loss scenario. This would root the organismal tree in the usual place, separating all bacteria from all archaea, and requires that the ancestor of the *menA* gene found in *Haloferax, Halobacterium*, *Thermotoga, Bacillus* and *Staphylococcus* was present in the genome of the universal ancestor. But if that is correct, then each of the branches below this corresponds to a paralog-generating gene duplication that occurred *prior to the time of the universal ancestral cell*, and each of these paralogs was also present in that cell's genome. There would be at least five such paralogs in this instance, but in many cases of more common but patchily distributed genes, the required number of paralogs could be dozens. So there are really two kinds of unparsimoniousness in differential loss scenarios: the invocation of multiple losses all along the line leading to a modern taxon with the rare paralog, and the evocation of a universal ancestor with more genes and more physiological abilities than any known contemporary prokaryote.

Such theoretical arguments can be applied to most of the examples discussed in this review. For example, H_4MPT-dependent methylotrophy is only found in Proteobacteria in the bacterial domain and it is highly unlikely to have been present in the bacterial ancestor and lost in all lineages but Proteobacteria. Since homologous genes are also present in some archaea (i.e. methanogens), this trait must have been transferred from one group to the other. Given that H_4MPT-dependent enzymes are an integral part of methanogenic metabolism and only one of multiple pathways for C_1 compounds oxidation in Proteobacteria, this particular set of enzymes most likely evolved in Archaea and was later transferred to Bacteria. The same goes for vanadium-dependent alternative nitrogenases, which are found in multiple bacterial lineages, but so far only in one group of methanogenic archaea, the Methanosarcinales. Dissimilatory sulfate reduction is another example. This metabolic capacity is present in a few groups of bacteria (Proteobacteria, Thermodesulfobacteria, Clostridia) and in one archaeal group (Archaeoglobales). Since the distribution of the enzymes for dissimilatory sulfate reduction is fairly sparse in both Bacteria and Archaea, it is very hard to identify their origin. However, there is little doubt that sulfate reduction enzymes genes have been transferred between domains, unless one endorses a scenario of numerous losses in both domains. Aerobic respiration also seems to be a derived feature in Archaea. Most obligate or facultative aerobes among archaea are found within clades of anaerobes. Indeed, Halobacteriales and Thermoplasmatales are the only aerobic euryarchaeotes, and are late-branching lineages thought to have evolved from anaerobic ancestors (Forterre *et* *al.*, 2002). Among the four obligately or facultatively aerobic crenarchaeotes whose genomes have been sequenced, *Pyrobaculum aerophilum* is the only species of its genus that is not a strict anaerobe, *Aeropyrum pernix*'s closest cultured relative *Thermodiscus maritimus* is a strict anaerobe and *Sulfolobus solfataricus* and *Sulfolobus tokadaii* (both strict aerobes) belong to the order Sulfolobales, which also contains strict anaerobes like *Stygioglobus*. The respiratory chain proteins required for aerobic metabolism are, for the most part, only found in aerobic archaea. Phylogenetic analyses of most of these enzymes reveal archaeal sequences as polyphyletic among bacterial relatives. Also, archaeal respiratory enzymes frequently have higher similarity to bacterial homologs. Taken together, these observations suggest that aerobic respiration evolved secondarily in archaea, by acquisition of genes from bacteria.

It seems that the lateral transfer of important physiological properties, such as dissimilatory sulfate reduction, nitrogen fixation and aerobic respiration crosses domains boundary and might have had major impact on the evolution of a number of bacterial and archaeal groups. In the case of the acquisition of the H_4MPT-dependent formaldehyde oxidation pathway, it is not the function itself that was transferred between Archaea and Bacteria, but rather a set of genes that has been co-opted for a different function after LGT.

LGT occurs at all phylogenetic levels

LGT of physiological processes can be observed at multiple phylogenetic levels. Most of the functions reviewed here, at least for some of their component genes, seem to have been transferred at the domain level: aerobic respiration proteins from Bacteria to Archaea, the H_4MPT-dependent formaldehyde oxidation pathway from methanogenic archaea to Proteobacteria, vanadium-dependent alternative nitrogenase complexes from Bacteria to *Methanosarcina*, dissimilatory sulfate reduction enzymes from Bacteria to Archaeoglobales. Evidence is also present for transfer within domains: between bacterial phyla for the genes of both isoprenoid biosynthesis pathways and dissimilatory sulfate reduction enzymes and between archaeal orders for the *cox* and *nuo* respiratory genes. When the required sampling is present, LGT is also observed between closely related organisms (within a class or a genus): the nitrogenase subunit gene *nifD* between strains of the genus *Bradyrhizobium*, the *nifH* gene from α-Proteobacteria to some species of the β-proteobacterial genus *Azoarcus*. The occurrence of LGT at different phylogenetic levels indicates not only that a function can be transferred between highly divergent groups, but also that LGT can play a role in the subsequent dissemination and evolution of this function within the group by which it was acquired.

The role of operons and gene clusters in LGT

The “Selfish operon” model of Lawrence and Roth (Lawrence and Roth, 1996) imagines that operons and gene clusters are created and maintained by LGT. If multiple gene products are required for a function, only acquisition of all these genes will be beneficial to the naïve host. Thus only organisms in which the genes are linked can easily serve as donors. From the gene's-eye perspective, linkage permits escape from evolutionary elimination by invasion of new genomes, making these clusters or operons “selfish” (Lawrence and Roth, 1996). Clusters can nonetheless confer highly beneficial functions that can be of long-term use to their new host (e.g. an alternative pathway for C_1 compounds oxidation, the capacity for dissimilatory sulfate reduction, etc.) that could not have been provided by single genes. Under this model, the very existence of operons and clusters of genes can be considered as evidence for LGT. In any circumstances, it is clear that the presence of these genetic structures facilitates lateral transfer of complex functions (i.e. requiring multiple gene products). The fact that all physiological systems reviewed here are usually encoded by one or two operon(s)/cluster(s) can be seen as indirect evidence that they have been disseminated by LGT.

Limits to dissemination by LGT

Tetrapyrole-based photosynthesis is found only within Bacteria. Methanogenesis, for its part, is found in several archaeal orders but not in Bacteria. A legitimate question to ask is why are these processes limited to particular groups? For methanogenesis and terapyrole-based photosynthesis, an explanation for their limited distribution might be that they are very complex processes that are encoded by a large number of genes that are not co-localized in the genome of their host. Methanogenic reduction of CO_2 is a process that minimally requires eight enzymes and six coenzymes, which are encoded by genes that are not clustered in the genomes of methanogens (Graham and White, 2002). Lateral transfer of methanogenic metabolism *in toto* would therefore be very difficult. Part of the system, however, could conceivably be transferred. It is indeed likely to be the case, as we discussed earlier, for the methanogenesis enzymes homologs found in the H_4MPT-dependent pathway of proteobacterial methylotrophs.

Tetrapyrole-based photosynthesis, a potentially dispensable but nevertheless complex process requiring numerous pigments and proteins (chlorophyll, bacteriochlorophyll, carotenoids, phycobillins, light-harvesting and reaction center proteins), seems to have experienced transfer both across large and small phylogenetic distances (transfer of genes encoding for photosynthetic reaction centers between bacterial phyla and of more-or-less the entire photosynthesis apparatus between proteobacterial classes). These transfers are made possible by the fact that most genes involved in photosynthesis can be found as a large cluster (or super-operon) in several bacterial groups (for example the proteobacterium *Rubrivivax gelatinosus* or the Heliobacterium *Heliobacillus mobilis*) (Igarashi *et al.*, 2001; Xiong *et al.*, 1998). Therefore, the absence of tetrapryole-based photosynthesis in the archaeal domain cannot be explained by the complexity or essentiality of this process alone. A biased sampling of diversity could also lead to overlooking the presence of a physiological process in certain phylogenetic groups. For example, another type of phototrophy, which is based on bacteriohodopsin or proteorhodopsin, was earlier thought to be present only in archaea but has recently been found to occur in bacteria as well, through cultivation-independent methods (Beja *et al.*, 2000). According to Hugenholtz (2002), 53% of archaeal phyla and 37% of bacterial phyla are known only from molecular data and have no cultured representatives. This shows how little we know of prokaryotic diversity, and suggests that the absence of certain physiological processes from a domain of life could simply derive from our limited sampling of diversity.

Also difficult to explain is the limited distribution of the H_4MPT-dependent pathway for oxidation of formaldehyde, which is exclusively found in Proteobacteria. This process is sparsely distributed within Proteobacteria but appears to have been disseminated by LGT within this group. If it was able to spread within Proteobacteria, what prevented this system from invading other microbial groups? The answer to this limited distribution might reside in the need for the presence of particular structures that are not easily transferred. For example, the activity of monoxygenases, essential for methylotrophy and ammonium oxidation, is dependent on the presence of internal membrane systems (Bedard and Knowles, 1989).

Prokaryotic genomes may be extensively mosaic, but each living prokaryotic cell is a functioning whole. All combinations of biochemical processes cannot proceed simultaneously in a single cell and some may require special structural accommodations to do so (nitrogen fixation with respiration, for instance). Archaeal membranes, with their unique isoprenyl-ether lipids, may be incompatible with some bacterial enzyme systems. There will also be many other such taxon-specific cellular limitations to exchange. By establishing the range of transfer of taxon-defining traits, we can determine the factors limiting dissemination by lateral transfer.

LGT and the structure of the prokaryotic domains

LGT was first discovered in the context of the spread of antibiotic resistance among pathogens, and the strongest evidence for its importance still comes from comparison of sequences of closely-related pathogens (Welch *et al.*, 2002). However, there are ample individual cases of transfers that must have occurred long before there were any animals for bacteria to infect, and estimates as high as 30% for the number of genes in the genomes of individual non-pathogenic (or "environmental") bacteria that have been received from archaea, and *vice versa* {for review, see (Doolittle *et al.*, 2003)}. Such transfers can effect major transformations in the biology of a recipient, and (as we have argued) sometimes must have been foundational events in the appearance of groups we now recognize as coherent "classes" or even "phyla". The traits transferred are often fundamental to cell physiology and "lifestyle", and are of the sort that microbiologists traditionally used to define such prokaryotic taxa, or employed in deducing phylogenetic relationships between them, especially before SSU rRNA achieved its current dominant role in microbial systematics.

Chapter 3: Origins and Evolution of Isoprenoid Biosynthesis

This chapter includes work published in Y. Boucher and W.F. Doolittle (2000) The Role of Lateral Gene Transfer in the Evolution of Isoprenoid Biosynthesis Pathways. *Molecular Microbiology* Aug;37(4): 703-716, Y. Boucher, H. Huber, S. L'Haridon, K.O. Stetter and W.F. Doolittle (2001) Bacterial Origin for the Isoprenoid Biosynthesis Enzyme HMG-CoA Reductase of the Archaeal Orders Thermoplasmatales and Archaeoglobales. *Molecular Biology and Evolution* Jul;18(7):1378-88. and Y. Boucher, M. Kamekura and W.F. Doolittle (2003) Origins and evolution of isoprenoid lipid biosynthesis in archaea. *Molecular microbiology* (submitted).

New sequences have been deposited in Genbank under accession numbers Y10011, AF247973, AJ566212, AJ564466 to AJ564495 and AJ299203 to AJ299219.

Introduction

More than 22,000 naturally-occurring isoprenoids are known (Conolly and Hill, 1992). Isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) are together the universal precursors of all isoprenoids, which occupy an important place in the biochemistry of all cells (Table 3.1). Sterols and dolichols for example, are essential to eukaryotic life, acting as membrane stabilizers and carbohydrate carriers, respectively. In the unique cell membranes of archaea, isopranyl glycerol ethers make up the hydrophobic core, and the isoprenoid moieties of these ether lipids are essential for

Table 3.1 Main functions of major isoprenoid compounds produced in the different domains of life.

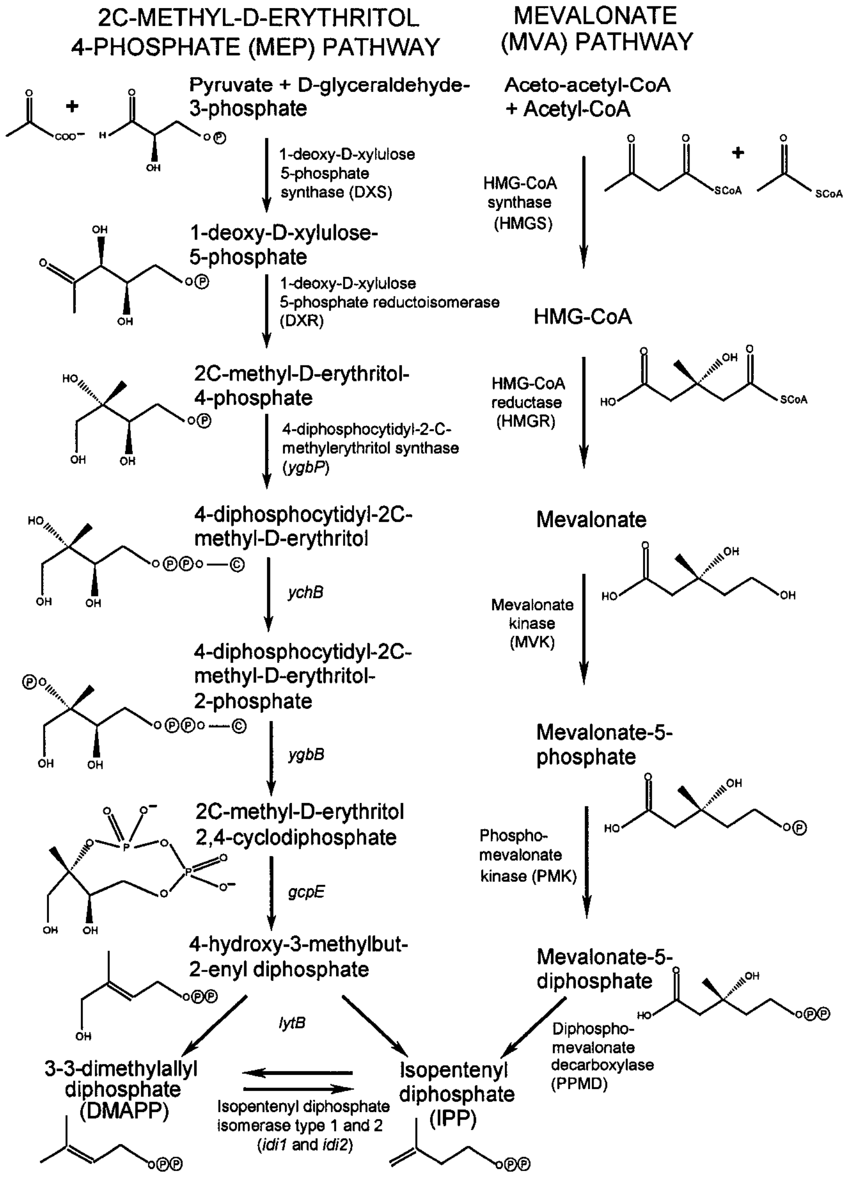
Organisms	Isoprenoids produced	Main function	Reference
Bacteria			
ubiquitous in eubacteria	ubiquinone, menaquinone	respiratory chain oxidizable pool	Gennis and Stewart, 1996
wide taxonomic distribution	hopanoids	membrane stabilizers	Rohmer *et al.,* 1984
ubiquitous in eubacteria	bactoprenols	carbohydrate carriers in biosynthesis of peptidoglycan	Putra *et al.,* 1998
Chloroflexus	verrucosanol	modulator of membrane fluidity	Hefter *et al.,* 1993
Synechocystis	carotenoids, phytol	photosynthesis	Lichtentaler, 1998
Streptomyces	antibiotics	competitive advantage	Seto *et al.,* 1996
Flavobacterium	zeaxanthin	photosynthesis	Britton *et al.,* 1979
Archaea			
ubiquitous in archaea	isoprenoid ethers	cell membrane essential components	De Rosa *et al.,* 1986
Eukaryotes			
plants (plastid)	carotenoids, phytol, plastoquinone	photosynthetic machinery	Eisenreich *et al.,* 1998
	monoterpenoids, diterpenoids	components of essential oils and resins	Rohmer, 1999
			Lichtentaler, 1998
plants (cytoplasm)	sesquiterpenes	components of essential oils and resins	Eisenreich *et al.,* 1998
	sterols	membrane stabilizers	Rohmer, 1999
	ubiquinone	quinones of electron transport chains	Lichtentaler, 1998
animals	sterols	membrane stabilizers, steroid hormones and bile acids precursors in vertebrates	Rohmer, 1999 Goldstein and Brown, 1990
	ubiquinone	quinones of electron transport chains	
	dolichols	carbohydrates carriers	

membrane integrity (De Rosa *et al.*, 1980; De Rosa *et al.*, 1986; Moldoveanu and Kates, 1988). Bacteria use the isoprenoid-containing compounds ubiquinone and menaquinone in their electron transport chains (Gennis and Stewart, 1996).

Two pathways are responsible for the biosynthesis of IPP and DMAPP. These pathways are named after their main intermediate product: the mevalonate (MVA) pathway and the 2-*C*-methyl-D-erythritol-4-phosphate (MEP) pathway (Figure 3.1). Although their products are identical, these pathways proceed through different chemical steps, and use non-homologous enzymes. The better described of the two, the MVA pathway, was for decades thought to be the only route for the synthesis of isoprenoid precursors (Banthorpe *et al.*, 1972; Beytia and Porter, 1976; De Rosa *et al.*, 1986; Goldstein and Brown, 1990; Qureshi and Porter, 1981; Rohmer, 1999). In this pathway, aceto-acetyl-CoA and acetyl-CoA are first converted to HMG-CoA (3-hydroxy-3-methylglutaryl coenzyme A), which is then reductively deacylated to mevalonate. After two subsequent phosphorylation steps, the mevalonate is converted by decarboxylation to the final product of the MVA pathway, IPP. The other precursor required for isoprenoid biosynthesis, DMAPP, is obtained through conversion of its isomer IPP by isopentenyl diphosphate isomerase (IDI), which can be of either of two analogous types (IDI1 or IDI2) (Kaneda *et al.*, 2001). The MEP pathway, for its part, starts with pyruvate and D-glyceraldehyde-3-phosphate, derived from glycolysis. Through several biochemical steps, these precursors are converted to 2-*C*-methyl-D-erythritol-4-phosphate (MEP), which is further processed to 4-hydroxy-3-methylbut-2-enyl diphosphate (Hoeffler *et al.*, 2002). In the final step of the MEP pathway, the latter compound is simultaneously converted to IPP and DMAPP (making isomerases like IDI1 or IDI2 non-essential for organisms using this pathway). Section 1 of this chapter describes the evolution of the MVA and MEP pathways in prokaryotes, with an emphasis on the role of lateral gene transfer.

As mentioned above, isoprenoids are particularly important in archaea, as their entire cell membrane is composed of phosphoglycerol ether-linked to isoprenoid side-chains. Following the synthesis of IPP by the MVA pathway, several other biochemical steps are required to obtain these membrane lipids. Evolutionary analyses were extended to include all characterized enzymes involved in the synthesis of these compounds from IPP: isopentenyl diphosphate isomerases (conversion of IPP to its isomer DMAPP, which is required to initiate the biosynthesis of isoprenoids side chains), short and medium chain prenyltransferases (synthesizing isoprenoid side chains), *sn*-glycerol-1-phosphate dehydrogenase (synthesizing the stereospecific phosphoglycerol backbone) and geranylgeranylglyceryl phosphate synthase (which links the first isoprenoid side-chain to the phosphoglycerol backbone). All analyses bearing on the evolution of isoprenoid lipid biosynthesis in archaea are described in section 2.

Among all isoprenoid biosynthesis enzymes submitted to evolutionary analysis, one of the most interesting case of LGT discovered concerned the key enzyme catalyzing the first committed step of the MVA pathway, 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMG-CoA reductase or HMGR). The genome sequence of *Archaeoglobus fulgidus* revealed that this archaeon harbors a version of this enzyme more closely related to bacterial HMGRs than to orthologs from other archaea. Section 3 describes a further investigation of this case, presenting the results of a survey of the HMGR gene in close relatives of *A. fulgidus* and other archaea that live in similar environments, along with a detailed evolutionary analysis of the sequence data obtained.



2C-METHYL-D-ERYTHRITOL 4-PHOSPHATE (MEP) PATHWAY
MEVALONATE (MVA) PATHWAY
Pyruvate + D-glyceraldehyde-3-phosphate
Aceto-acetyl-CoA + Acetyl-CoA
1-deoxy-D-xylulose 5-phosphate synthase (DXS)
HMG-CoA synthase (HMGS)
1-deoxy-D-xylulose-5-phosphate
1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR)
HMG-CoA
2C-methyl-D-erythritol-4-phosphate
4-diphosphocytidyl-2-C-methylerythritol synthase (ygbP)
HMG-CoA reductase (HMGR)
4-diphosphocytidyl-2C-methyl-D-erythritol
Mevalonate
ychB
Mevalonate kinase (MVK)
4-diphosphocytidyl-2C-methyl-D-erythritol-2-phosphate
Mevalonate-5-phosphate
ygbB
2C-methyl-D-erythritol 2,4-cyclodiphosphate
Phospho-mevalonate kinase (PMK)
gcpE
4-hydroxy-3-methylbut-2-enyl diphosphate
Mevalonate-5-diphosphate
lytB
Diphospho-mevalonate decarboxylase (PPMD)
3-3-dimethylallyl diphosphate (DMAPP)
Isopentenyl diphosphate (IPP)
Isopentenyl diphosphate isomerase type 1 and 2 (idi1 and idi2)

Figure 3.1 Schematic representation of the two pathways for the biosynthesis of the universal precursors of isoprenoids IPP and DMAPP. The circled P indicates a phosphate moiety and the circled C a cytidyl moiety.

Section I: The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways

Eukaryotes, with the exception of photosynthetic eukaryotes (Lichtentaler, 1998), solely use the MVA pathway (Figure 3.2). Most plants and eukaryotic algae, in addition to the MVA pathway operating in their cytoplasm, harbour activity of the MEP pathway in their chloroplasts. Like non-photosynthetic eukaryotes, archaea derive their isoprenoids from MVA, as shown by tracer studies more than two decades ago (Kates and Kushwaha, 1978). However, only orthologs of the enzymes catalyzing the first three steps of the MVA pathway found in bacteria and eukaryotes are present in archaea. The genes coding for the enzymes that catalyze the last two steps of the pathway (phosphomevalonate kinase, or PPMK, and diphosphomevalonate decarboxylase, or PPMD) have no match in most archaeal genomes (Smit and Mushegian, 2000). Since the corresponding activities are essential for cell membrane synthesis in archaea, some other enzymes (analogs) must perform equivalent functions to allow IPP production.

Complete and partial genome sequences indicate that a complex evolutionary story is required to account for the distribution of the IPP biosynthesis pathways in the bacterial domain (Table 3.2). Some bacteria appear to use one or the other pathway exclusively, some use both pathways and others may have one complete pathway and elements of the other. The distribution of the pathways is not strongly correlated with phylogenetic position based on ribosomal RNA.

Table 3.2 Distribution of genes of the MEP and MVA pathways for isoprenoid biosynthesis.

	MEP pathway							MVA pathway						
BACTERIA	DXS	DXR	ygbB	ygbP	ychB	gcpE	lytB	HMGR	HMGS	MK	PMK	DPMD	IDI1	IDI2
AQUIFICALES														
Aquifex aeolicus	+	+	+	+	+	+	+	-	-	-	-	-	-	-
THERMOTOGALES														
Thermotoga maritima	+	+	+	+	+	+	+	-	-	-	-	-	-	-
GREEN NON-SULFUR														
Chloroflexus aurantiacus								l+/+	l+/+	l+	l+	l+/+		
DEINOCOCCUS/THERMUS														
Deinococcus radiodurans	+	+	+	+	+	+	+	-	-	-	-	-	-	-
CYTOPHAGA/BACTEROIDES														
Porphyromonas gingivalis	+	+	+	+	+	+	+							
Flavobacterium sp.								l+	l+	l+	l+	l+		
HIGH G+C GRAM POSITIVES														
Mycobacterium tuberculosis	+	+	+	+	+	+	+	-	-	-	-	-	+	-
Mycobacterium leprae	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Mycobacterium phlei	l+	l+	l+	l+	l+	l+	l+							
Mycobacterium smegmatis	l+	l+	l+	l+	l+	l+	l+							
Mycobacterium avium	+	+	+	+	+	+	+							
Mycobacterium bovis	+	+	+	+	+	+	+							
Corynebacterium ammoniagenes	l+	l+	l+	l+	l+	l+	l+							
Corynebacterium diphteriae	+	+	+	+	+	+	+							
Actinoplanes sp. A40644	l+	l+	l+	l+	l+	l+	l+	l+	l+	l+	l+	l+		
Streptomyces coelicolor	+	+	+	+	+	+							+	
Streptomyces aeriouvifer CL190	l+	l+	l+	l+	l+	l+	l+	l+/+	l+/+	l+/+	l+/+	l+/+		+
Streptomyces niveus	l+	l+	l+	l+	l+	l+	l+							
Kitasatospora griseola	+							l+/+	l+/+	l+/+	l+/+	l+/+		+
Streptomyces exfoliatus	l+	l+	l+	l+	l+	l+	l+							
Streptomyces ruber								l+	l+	l+	l+	l+		
Streptomyces spheroides	l+	l+	l+	l+	l+	l+	l+							
Streptomyces ghanaensis	l+	l+	l+	l+	l+	l+	l+							
Streptomyces blastmycetium	l+	l+	l+	l+	l+	l+	l+							
LOW G+C GRAM POSITIVES														
Bacillus subtilis	+	+	+	+	+	+	+	-	-	-	-	-	-	+
Bacillus halodurans	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Listeria monocytogenes	+	+	+	+	+	+	+	+	+	+	+	+	-	+
Listeria innocua	+	+	+	+	+	-	-	+	+	+	+	+	-	+
Staphylococcus aureus	-	-	-	-	+	-	-	+	+	+	+	+	-	+
Streptococcus pneumoniae	-	-	-	-	-	-	-	+	+	+	+	+	-	+
Lactococcus lactis	+	-	-	-	-	-	-	+	+	+	+	+	-	+
Oceanobacillus iheyensis	-	-	-	-	+	-	-	+	+	+	+	+	-	+
Staphylococcus carnosus								l+	l+	l+	l+	l+		
Streptococcus mutans								l+/+	l+	l+	l+	l+		
Streptococcus pyogenes								+	+	+	+	+		+
Enterococcus faecalis								+	+	+	+	+		
Enterococcus faecium	+				+			+	+	+	+	+		
Lactobacillus plantarum								l+	l+	l+	l+	l+		
Clostridium difficile	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Clostridium acetobutylicum	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Alicyclobacillus acidoterrestris	l+	l+	l+	l+	l+	l+	l+							
Ureaplasma urealyticum	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mycoplasma pneumoniae	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mycoplasma pulmonis	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mycoplasma genitalium	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CYANOBACTERIA														
Nostoc sp.	+	+	+	+	+	+	+	-	-	-	-	-	-	+
Synechocystis sp. PCC6803	+	+	+	+	+	+	+	-	-	-	-	-	-	+
GREEN SULFUR														
Chlorobium tepidum	+	+	+											
SPIROCHAETES														
Borrelia burgdorferi	-	-	-	-	-	-	-	+	+	+	+	+	-	+
Treponema pallidum	+	+	+	+	+	+	+	-	-	-	-	-	-	-
CHLAMYDIALES														
Chlamydia trachomatis	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Chlamydia muridarum	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Chlamydophila pneumoniae	+	+	+	+	+	+	+	-	-	-	-	-	-	-

Table 3.2 Distribution of genes of the MEP and MVA pathways for isoprenoid biosynthesis (continued).

	MEP pathway							MVA pathway						
BACTERIA	DXS	DXR	ygbB	ygbP	ychB	gcpE	lytB	HMGR	HMGS	MK	PMK	DPMD	IDI1	IDI2
ALPHA-PROTEOBACTERIA														
Agrobacterium tumefasciens	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Mesorhizobium loti	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Sinorhizobium meliloti	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Zymomonas mobilis	I+	I+/+	I+/+	I+/+	I+	I+	I+							
Methylobacterium organophilum	I+	I+	I+	I+	I+	I+	I+							
Methylobacterium fujisawaense	I+	I+	I+	I+	I+	I+	I+							
Rhodobacter capsulatus	+	+	+										+	
Rhodopseudomonas acidophila	I+	I+	I+	I+	I+	I+	I+							
Rhodopseudomonas palustris	I+	I+	I+	I+	I+	I+	I+							
Caulobacter crescentus	+	+	+	+	+	+	+	-	+	-	-	-	-	-
Rickettsia conorii	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Paracoccus zeaxanthinifaciens								+	+	+	+	+	+	
Rickettsia prowazekii	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BETA-PROTEOBACTERIA														
Neisseria gonorrhea	+	+	+	+	+	+	+							
Neisseria meningitidis	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Bordetella pertussis	+	+	+	+	+	+	+							
Burkholderia gladioli	I+	I+	I+	I+	I+	I+	I+							
Burkholderia caryophylli	I+	I+	I+	I+	I+	I+	I+							
Ralstonia solanacearum	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Ralstonia pickettii	I+	I+	I+	I+	I+	I+	I+							
GAMMA-PROTEOBACTERIA														
Actinobacillus actinomycetemcomitans	+	+	+	+	+	+	+							
Vibrio cholerae	+	+	+	+	+	+	+	+	-	-	-	-	-	-
Escherichia coli	+	+	+	+	+	+	+	-	-	-	-	-	+	-
Haemophilus influenzae	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Pseudomonas mevalonii								+						
Pseudomonas aeruginosa	I+/+	I+/+	I+/+	I+/+	I+/+	I+/+	I+/+		+					
Pseudomonas fluorescens	I+	I+	I+	I+	I+	I+	I+		+					
Pasteurella multocida	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Shewanella putrefaciens	+	+	+	+	+	+	+							
Legionella pneumophila								+		+		+	+	
Salmonella typhi	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Salmonella typhimurium	I+/+	I+/+	I+/+	I+/+	I+/+	I+/+	I+/+	-	-	-	-	-	+	-
Thiobacillus ferrooxidans	+	+	+	+	+	+	+							
Yersinia pestis	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Acinetobacter calcoaceticus	I+	I+	I+	I+	I+	I+	I+							
Citrobacter freundii	I+	I+	I+	I+	I+	I+	I+							
Xylella fastidiosa	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Erwinia carotovora	I+	I+	I+	I+	I+	I+	I+							
DELTA-PROTEOBACTERIA														
Myxococcus fulvus								I+	I+	I+	I+	I+		
EPSILON-PROTEOBACTERIA														
Campylobacter jejuni	+	+	+	+	+	+	+	-	-	-	-	-	-	-
Helicobacter pylori	+	+	+	+	+	+	+	-	-	-	-	-	-	-

Table 3.2 Distribution of genes of the MEP and MVA pathways for isoprenoid biosynthesis (continued).

	MEP pathway							MVA pathway						
ARCHAEA	DXS	DXR	ygbB	ygbP	ychB	gcpE	lytB	HMGR	HMGS	MK	PMK	DPMD	IDI1	IDI2
DESULFUROCOCCALES														
Aeropyrum pernix	-	-	-	-	-	-	-	+	+	+	-	-	-	+
SULFOLOBALES														
Sulfolobus solfataricus	-	-	-	-	-	-	-	+	+	+	-	-	-	+
Sulfolobus tokodaii	-	-	-	-	-	-	-	+	+	+	-	-	-	+
THERMOPROTEALES														
Pyrobaculum aerophilum	-	-	-	-	-	-	-	+	+	+	-	-	-	+
ARCHAEOGLOBALES														
Archaeoglobus fulgidus	-	-	-	-	-	-	-	+	+	+	-	-	-	+
HALOBACTERIALES														
Halobacterium sp. NRC-1	-	-	-	-	-	-	-	I+/+	I+/+	I+/+	-	+	+	+
Haloferax	-	-	-	-	-	-	-	+				+	+	+
Haloarcula	-	-	-	-	-	-	-	+					+	
Halococcus	-	-	-	-	-	-	-						+	
Halorubrum	-	-	-	-	-	-	-					+	+	
Halorhabdus	-	-	-	-	-	-	-						+	+
Haloterrigena	-	-	-	-	-	-	-					+	+	
Natrialba	-	-	-	-	-	-	-						+	
Natrinema	-	-	-	-	-	-	-						+	
Natronobacterium	-	-	-	-	-	-	-					+	+	+
Natronomonas	-	-	-	-	-	-	-					+	+	+
Natronorubrum	-	-	-	-	-	-	-					+	+	+
METHANOBACTERIALES														
Methanothermobacter thermoautotrophicus	-	-	-	-	-	-	-	+	+	+	-	-	-	+
METHANOCOCCALES														
Methanococcus jannaschii	-	-	-	-	-	-	-	+	+	+	-	-	-	+
THERMOCOCCALES														
Pyrococcus abyssii	-	-	-	-	-	-	-	+	+	+	-	-	-	+
Pyrococcus horikoshii	-	-	-	+	-	-	-	+	+	+	-	-	-	+
THERMOPLASMATALES														
Thermoplasma acidophilum	-	-	-	-	-	-	-	+	+	+	-	+	-	+
Thermoplasma volcanii	-	-	-	-	-	-	-	+	+	+	-	+	-	+

	MEP pathway							MVA pathway						
EUKARYOTES	DXS	DXR	ygbB	ygbP	ychB	gcpE	lytB	HMGR	HMGS	MK	PMK	DPMD	IDI1	IDI2
STREPTOPHYTA														
Arabidopsis thaliana	+	+	+	+	+	+	+	+	+	+	+	+	+	-
RHODOPHYTA														
Cyanidium caldarium	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+		
CHLOROPHYTA														
Scenedesmus obliquus	I+	I+	I+	I+	I+	I+	I+	I-	I-	I-	I-	I-		
Chlorella fusca	I+	I+	I+	I+	I+	I+	I+	I-	I-	I-	I-	I-		
Chlamydomonas reinhardtii	I+	I+	I+	I+	I+	I+	I+	I-	I-	I-	I-	I-		
HETEROKONTA														
Ochromonas danica	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+	I+		
APICOMPLEXA														
Plasmodium falciparum	I+/+	I+/+	I+/+	I+	I+	I+/+	I+/+	I-	I-	I-	I-	I-		
EUGLENOZOA														
Euglena gracilis								I+	I+	I+	I+	I+		
Trypanozoma cruzi								I+	I+	I+	I+	I+		
Leishmania major								I+	I+	I+	I+	I+		+
METAZOA														
Homo sapiens	-	-	-	-	-	-	-	+	+	+	+	+	+	-
Drosophila melanogaster	-	-	-	-	-	-	-	+	+	+	+	+	+	-
Caenorhabditis elegans	-	-	-	-	-	-	-	+	+	+	+	+	+	-
FUNGI														
Saccharomyces cerevisiae	-	-	-	-	-	-	-	+	+	+	+	+	+	-
Rhodoturula glutinis	I-	I-	I-	I-	I-	I-	I-	I+	I+	I+	I+	I+		
Aschersonia aleyrodis	I-	I-	I-	I-	I-	I-	I-	I+	I+	I+	I+	I+		
DIPLOMONADIDA														
Giardia lamblia								+	+	+	+	+		

A blank means no information about the presence/absence of this gene is available

- Gene absent from complete genome

+ Gene has been sequenced for this organism

I+ Tracer studies have shown biosynthetic activity for either pathway (feeding experiments with isotopically labeled precursors)

A minority of bacteria seem to use only the MVA pathway. Of the 96 complete bacterial genome sequences currently available, only the spirochaete *Borrelia burgdorferi* and firmicutes of the genera *Staphylococcus, Enterococcus, Streptococcus, Lactococcus* and *Listeria* have the complete MVA pathway. Indeed, the firmicutes *Staphylococcus carnosus*, *Staphylococcus aureus*, *Streptococcus mutans* and *Lactobacillus plantarum* have been shown to use the MVA pathway to make their IPP, either through detection of enzymatic activities or feeding experiments involving labeled precursor incorporation in the synthesis of mevalonate (tracer studies) (Gough *et al.*, 1970; Horbach *et al.*, 1993; Kleinig, 1975; Moshier and Chapman, 1973). Very few bacteria outside the firmicutes have been shown to use the MVA pathway in the absence of the MEP pathway: *Flavobacterium sp.* and *Myxococcus fulvus* use it to synthesize carotenoids and *Chloroflexus aurantiacus* to make the membrane diterpene verrucosanol (Table 3.1).

The MEP pathway is by far the more common in bacteria. It is found either alone, with some genes of the MVA pathway or in addition to the complete MVA pathway (Table 3.2). Of all the characterized genes of the MEP pathway, only *ygbP* and *ygbB* are usually found physically close in the genome (either in an operon or fused in a single protein) (Herz *et al.*, 2000; Rohdich *et al.*, 1999).

Figure 3.2 summarizes available information on the presence of the IPP biosynthesis pathways among bacteria. The resulting distribution pattern resembles those of thermophily or photosynthetic capacity among bacteria: unrelated bacteria can possess the same pathway, and related ones different pathways. Another interesting insight coming from the observed distribution is that the MEP pathway seems to be older than the MVA pathway, at least in bacterial groups for which several genera have been studied. Indeed, the MEP pathway is by far the more prevalent, especially amongst Proteobacteria and Actinobacteria, for which most data is available. Also, whenever the MEP pathway is present, it assumes a primary metabolic role and the MVA pathway (or a few of its genes) fulfill a secondary role (for example in *S. aeriouvifer, Actinoplanes sp.* and *P. mevalonii*, see next section). The MVA pathway assumes a primary role only when found alone.

No scheme consistent with simple vertical inheritance of genetic information can parsimoniously describe this patchy distribution of the MVA and MEP biosynthetic pathways among prokaryotes. Database surveys and phylogenetic analysis of all genes of the two isoprenoid biosynthesis pathways were performed to verify to what extent LGT contributed to the evolution of this metabolic process in prokaryotes.

Figure 3.2 Schematic representation of SSU rRNA phylogenetic analysis of the bacterial domain based on Hugenholtz *et al.* (1998). All bacteria known to display activity or the presence of genes from either MVA or MEP pathways are included. Green and red highlights mean that the organism most likely uses the MEP or MVA pathway, respectively, for the biosynthesis of primary metabolites. We infer that an organism uses a pathway biosynthetically if (1) DXS , DXR , *ygbP, ygbB, ychB, gcpE* and *lytB* are encoded in its genome for the MEP pathway; HMGR, HMGS, MVK, PMK and PPMD genes for the MVA pathway or (2) tracer studies have shown that appropriate primary metabolites (essential isoprenoids) are synthesized by the MVA or the MEP pathway. The presence of only some genes of the MVA or the MEP pathways is indicated by an asterisk. Biosynthetic activity known to lead to secondary metabolites is represented by two asterisks. The asterisks are green when they concern the presence of genes of the MEP pathway and red for genes of the MVA pathway. For species whose names are displayed in black, either no pathway is present or the pathway used is undetermined. Numbers in parenthesis represent the number of species of the specified genus displaying similar metabolism with regard to MVA or MEP pathway activity. Complete genome sequence is available for underlined species.

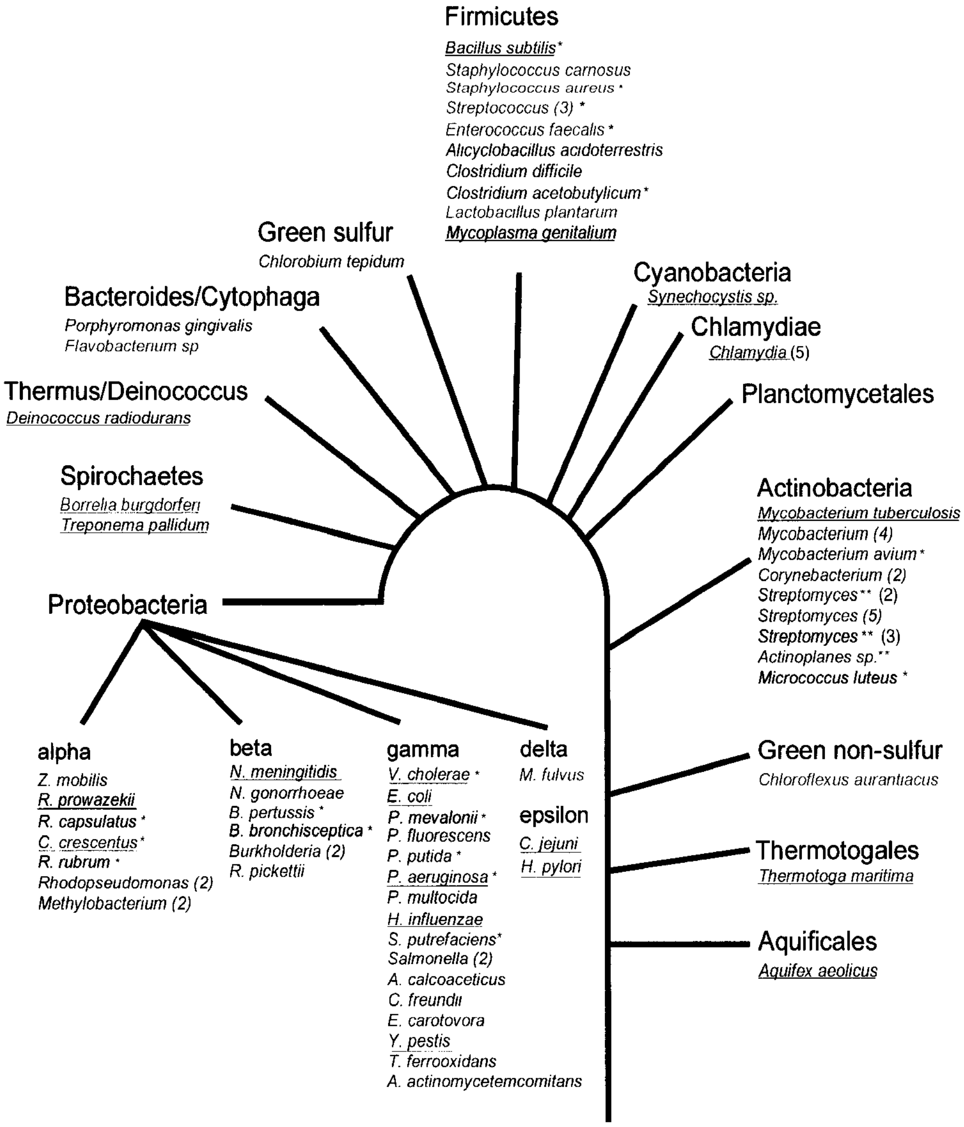


Figure 3.2 Schematic representation of SSU rRNA phylogenetic analysis of the bacterial domain based on Hugenholtz *et al.* (1998).

Materials and methods

Genomic DNA extraction

Cell pellets of the Chloroflexales bacterium *Chloroflexus aurantiacus* j-10 were given by Dr. Reidun Sirevåg (University of Oslo, Oslo, Norway). Genomic DNA was isolated from these cell pellets as described in Wilson (1994). *Giardia lamblia* (WB, ATCC#30957) total genomic DNA was kindly provided by J.M. Archibald.

Polymerase chain reaction (PCR), cloning and sequencing

Preliminary sequence data for the HMG-CoA reductase gene of *Giardia lamblia* was obtained from the genome Project website at the Marine Biological Laboratory (www.mbl.edu/LABS/JBPC/). Only the 5' and 3' ends of the HMGR gene were available for *G. lamblia*. These partial sequences were used to design PCR primers to each end of the gene (forward 5'-ATCCAGGCTCTTGATACA ATG-3', reverse 5'-GTGCTCCATCTCCAGGCTCTT-3'). The primers were used to amplify 1170 bp of the HMGR gene from *G. lamblia*. PCR conditions were standard (10 ng template DNA, 1X PCR buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTP, 1 μM of each primer and 1 U Taq-polymerase), with 35 amplification cycles (94°C for 1 minute, 45°C for 1 minute, and 72 °C for 1 minute). The PCR product was subsequently cloned with the invitrogen TOPO TA cloning kit and sequenced by LI-COR automated sequencing.

A portion of the *Chloroflexus aurantiacus* HMGR gene (522 bp) was amplified from pure genomic DNA using universal class 2 HMGR degenerate primers. PCR conditions were the same as for the amplification of *Giardia*'s HMGR gene. The primers (class 2 HMGR universal forward 5'-TTAGCTACCGARGARCCNTC-3’, class 2 HMGR universal reverse 5’-CCGTTCATGATNCCYTTRTT-3’) were designed based on an amino acid alignment of all HMGR gene sequences available in GenBank.

Gene alignment construction

The gene alignments were constructed and edited as describe in chapter 2, with the following modifications. Regions corresponding to the PCR primers in novel sequences (*Giardia lamblia* and *Chloroflexus aurantiacus* HMGR genes) were removed from the alignments during the manual editing in MacClade (Maddison and Maddison, 1989). 5’ and 3’ regions absent from the novel sequences (that are only partial) were kept in the alignment.

Phylogenetic analysis

All isoprenoid biosynthesis genes of the MVA and MEP pathways were subjected to phylogenetic analysis at the amino acid level. Maximum likelihood phylogenetic analyses and distance maximum likelihood bootstrapping were performed as described in chapter 2.

Results

Simultaneous presence of the two isoprenoid biosynthesis pathways in a single organism

Biosynthesis of secondary metabolites by actinomycetes.

Streptomyces are actinomycetes, high G+C Gram positive bacteria (Actinobacteria) known to produce antibiotics and other secondary metabolites, several of them containing isoprenoid moieties. Most *Streptomyces* species use exclusively the MEP pathway for the biosynthesis of IPP (Orihara *et al.*, 1998), which is used in primary metabolic functions like menaquinone synthesis, as well as some secondary metabolic functions like antibiotic production (Table 3.3). However, some *Streptomyces* and other actinomycetes harbour both complete IPP biosynthesis pathways. Among those, *Streptomyces aeriouvifer* CL190 has been shown to use the MEP pathway at the beginning of its exponential growth cycle to synthesize menaquinone and then, at stationary phase, to switch to the MVA pathway to produce IPP for synthesis of the antibiotic naphterpin (Seto *et al.*, 1996; Seto *et al.*, 1998). Furthermore, pravastatin (an HMGR inhibitor) can suppress production of naphterpin by *S. aeriouvifer* without affecting growth. *Actinoplanes sp.* is similar to *S. aeriouvifer* in its use of the IPP biosynthesis pathways, synthesizing menaquinone *via* the MEP pathway in early growth, and then employing the MVA pathway for synthesis of BE-406441 (a human thioredoxin system inhibitor) in late growth (Seto *et al.*, 1996; Seto *et al.*, 1998). It is important to notice that, in both of these cases, the MEP pathway is used for a primary metabolic

Table 3.3 IPP biosynthesis pathways employed by different actinomycetes to synthesize antibiotics and other isoprenoid derivatives

Species	MVA pathway	MEP pathway
Streptomyces ruber	napyradiomycin	
Streptomyces sp. KO-3988	furaquinocin	
Kitasatospora griseola	terpentecin	
Streptomyces aeriouvifer CL190	naphterpin	menaquinone
Actionoplanes sp.	BE-406441	menaquinone
Streptomyces sp. UC5319		pentalenolactone
Streptomyces spheroides		novobiocin
Streptomyces exfoliatus		carquinostatin
Streptomyces niveus		novobiocin
Streptomyces ghanaensis		moenomycin
Streptomyces blastmycetium		teleocidin

function and the MVA pathway fulfills a non-essential role, synthesizing secondary metabolites.

Its primary role in metabolite synthesis may suggest that the MEP pathway is ancestral and that some *Streptomyces* subsequently acquired the MVA pathway, which is dispensable and absent from most *Streptomyces* species (Seto *et al.*, 1996; Seto *et al.*, 1998). The MVA pathway is also missing from other studied Actinobacteria lineages (*Mycobacterium*, *Corynebacterium*, see Table 3.2), and the assumption that it is an ancestral feature would require its independent loss from many lineages. More compelling than this parsimony argument, however, is the fact that, by phylogenetic analysis, unequivocal transfer of HMGR (from archaea) has clearly occurred within *Streptomyces*.

Indeed, the HMGRs found in *Kitasatospora griseola* (previously known as *Streptomyces griseolosporeus*) (BAA74565), *Streptomyces sp.* KO-3988 (BAA74566) and *Streptomyces aeriouvifer* CL190 (BAA70975) (Dairi *et al.*, 2000; Takahashi *et al.*, 1999) are of class 1, and thus of eukaryotic or archaeal origin. Our phylogenetic analysis supports a class 1 HMGR for these actinomycetes (Figure 3.3). A specific relationship of the *Streptomyces/Kitasatospora* HMGRs to their archaeal orthologs is suggested by the presence of a shared insertion (Figure 3.4). These HMGRs, acquired by LGT, play a role in the MVA pathway that synthesizes antibiotics. Indeed, several terpenoid and hemiterpenoid antibiotics are produced by the MVA pathway in different actinomycetes species (Table 3.3).

Figure 3.3 Best maximum likelihood phylogenetic trees for the enzymes of the MVA pathway. Trees are arbitrarily rooted and only relevant nodes with support values over 50% are displayed. Archaeal taxa are highlighted in bold. A) HMG-CoA synthase (HMGS). B) HMG-CoA reductase (HMGR) rooted with class 2 enzymes. C) Phosphomevalonate kinase (PMK) rooted with Mevalonate kinase (MK). D) Diphosphomevalonate decarboxylase (PPMD).

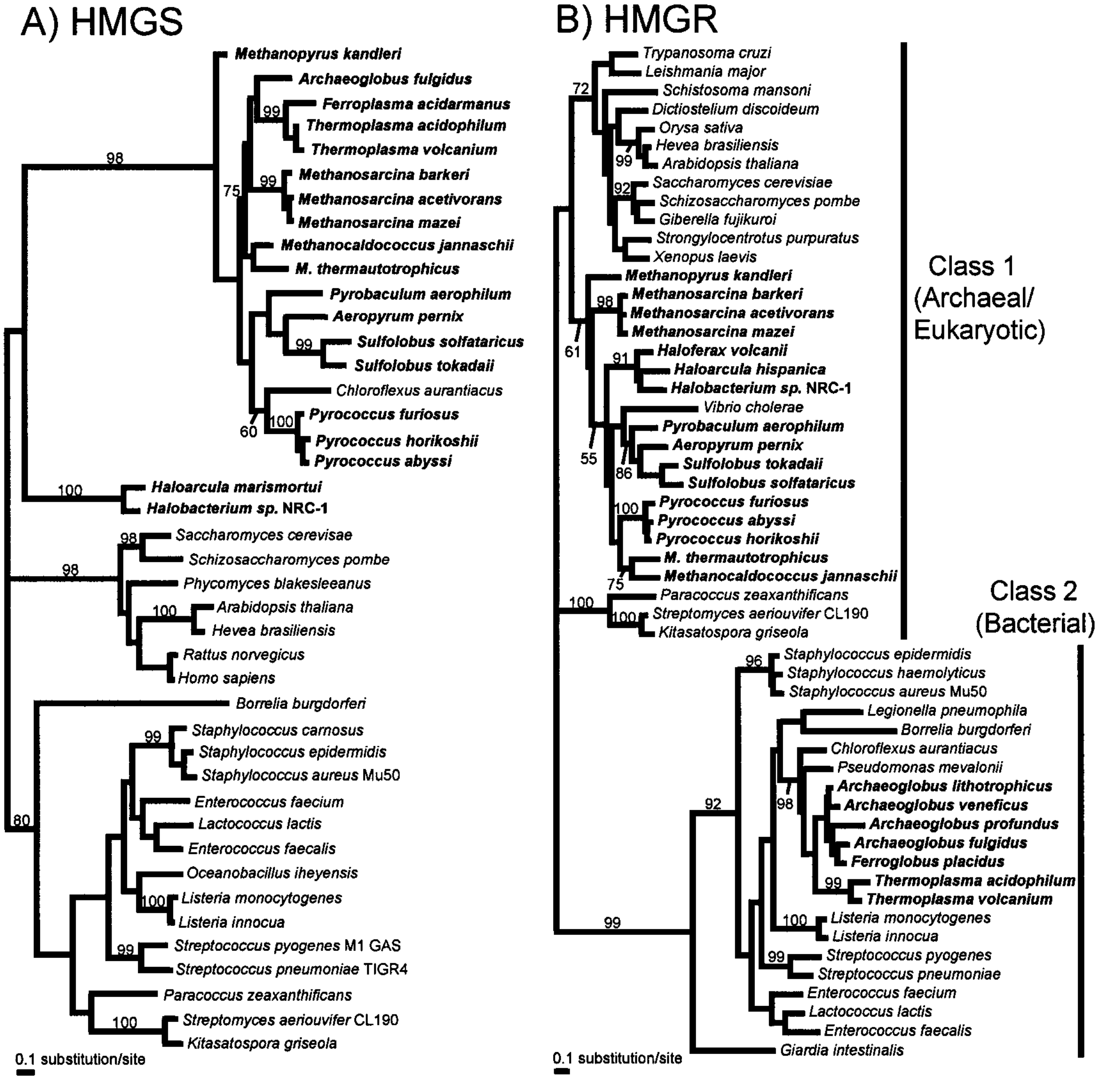


Figure 3.3 Best maximum likelihood phylogenetic trees for the enzymes of the MVA pathway.

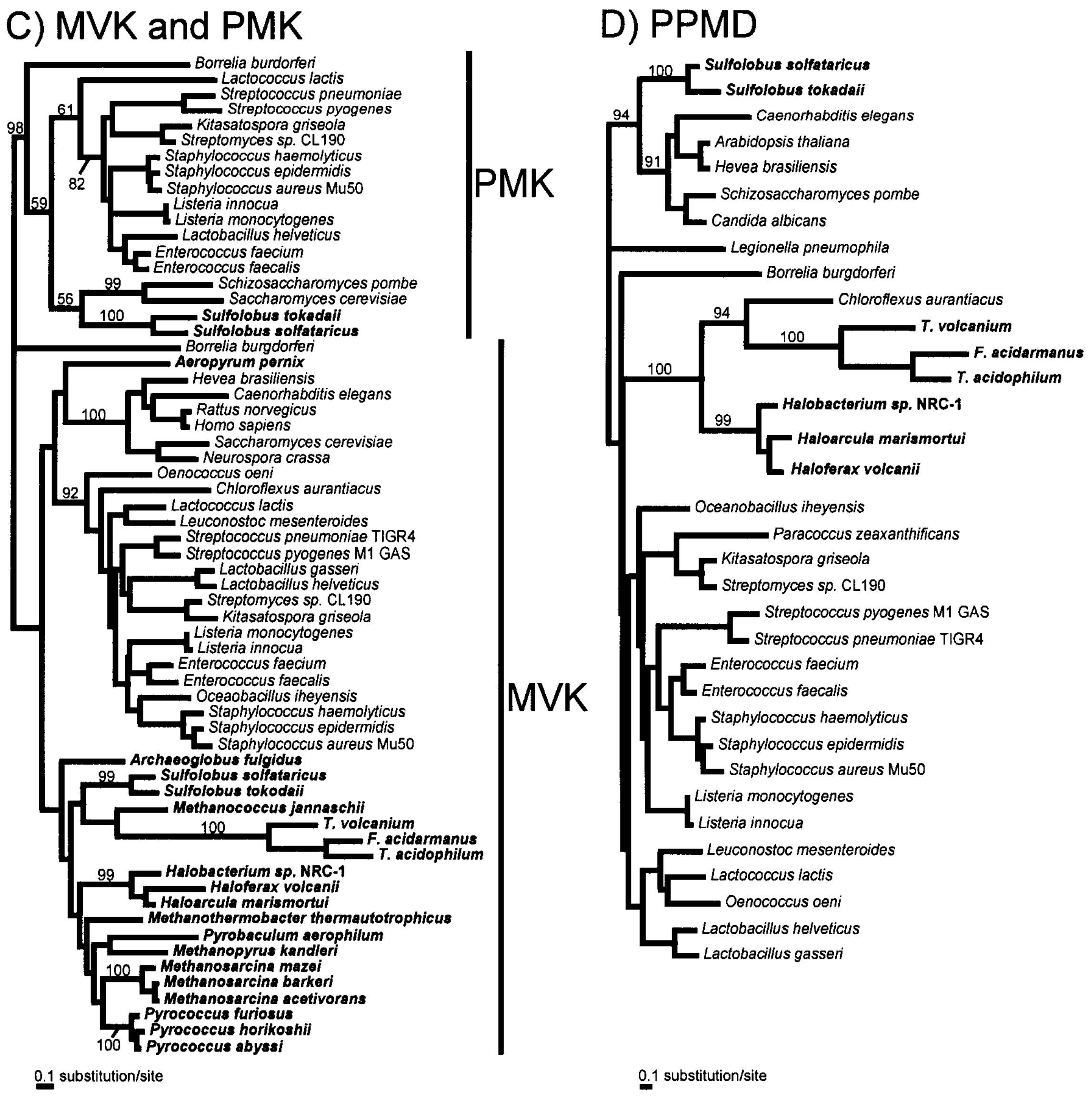


Figure 3.3 Best maximum likelihood phylogenetic trees for the enzymes of the MVA pathway (continued).

Figure 3.4 Protein sequence alignment of representative class 1 and class 2 HMG-CoA reductases. All known protein sequences, including isozymes, were aligned using ClustalW and the alignment was edited by eye. A subset of this alignment is presented here, extending from two residues before the HMG-CoA binding ENVIG motif (underlined) to eighteen residues after the active site glutamate (asterisk). Residues conserved among both classes in at least 10 of 12 taxa are highlighted in black, including conservative substitutions {as defined in the PIMA algorithm (Smith and Smith, 1992)}. The highlights in two shades of gray indicate residues conserved for either class 1 or class 2 HMGRs. The boxed four amino-acids insertion is shared only by archaea plus *V. cholerae* and three actinomycetes species (including *Kitasatospora griseola* presented here). The amino-acid positions in the original protein sequences are indicated at each ends of the alignment, except for *G. lamblia*, which is missing the 5' end.

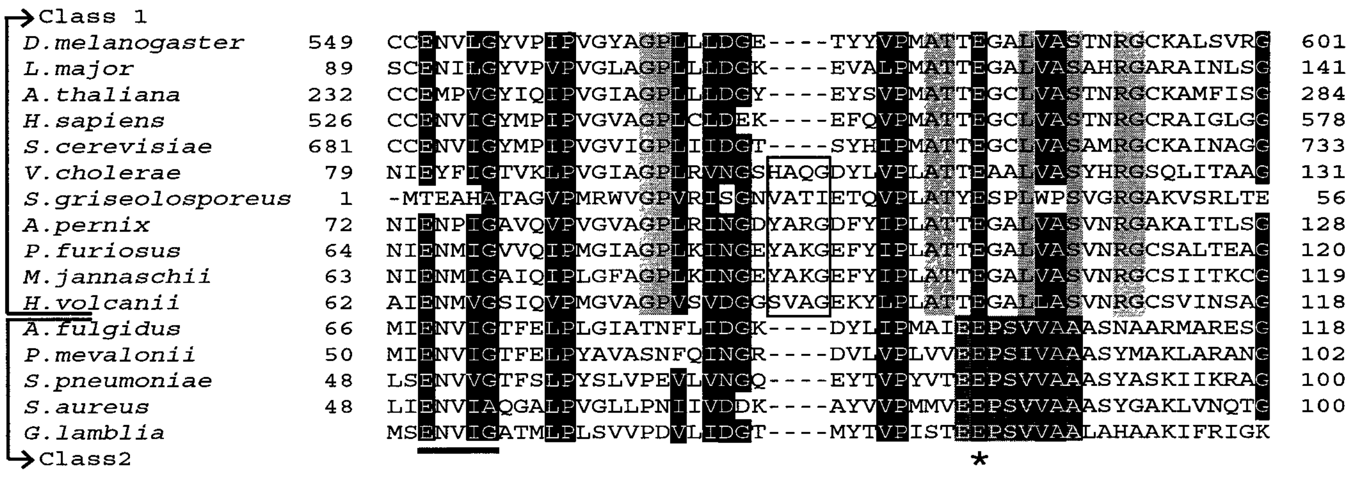

Class 1
D.melanogaster 549 CCENVLGYVPIPVGYAGPLLLDGE----TYYVPMATTEGALVASTNRGCKALSVRG 601
L.major 89 SCENILGYVPVPVGLAGPLLLDGK----EVALPMATTEGALVASAHRGARAINLSG 141
A.thaliana 232 CCEMPVGYIQIPVGIAGPLLLDGY----EYSVPMATTEGCLVASTNRGCKAMFISG 284
H.sapiens 526 CCENVIGYMPIPVGVAGPLCLDEK----EFQVPMATTEGCLVASTNRGCRAIGLGG 578
S.cerevisiae 681 CCENVIGYMPIPVGVIGPLIIDGT----SYHIPMATTEGCLVASAMRGCKAINAGG 733
V.cholerae 79 NIEYFIGTVKLPVGIAGPLRVNGSHAQGDYLVPLATTEAALVASYHRGSQLITAAG 131
S.griseolosporeus 1 -MTEAHATAGVPMRWVGPVRISGNVATIETQVPLATYESPLWPSVGRGAKVSRLTE 56
A.pernix 72 NIENPIGAVQVPVGVAGPLRINGDYARGDFYIPLATTEGALVASVNRGAKAITLSG 128
P.furiosus 64 NIENMIGVVQIPMGIAGPLKINGEYAKGEFYIPLATTEGALVASVNRGCSALTEAG 120
M.jannaschii 63 NIENMIGAIQIPLGFAGPLKINGEYAKGEFYIPLATTEGALVASVNRGCSIITKCG 119
H.volcanii 62 AIENMVGSIQVPMGVAGPVSVDGGSVAGEKYLPLATTEGALLASVNRGCSVINSAG 118
A.fulgidus 66 MIENVIGTFELPLGIATNFLIDGK----DYLIPMAIEEPSVVAAASNAARMARESG 118
P.mevalonii 50 MIENVIGTFELPYAVASNFQINGR----DVLVPLVVEEPSIVAAASYMAKLARANG 102
S.pneumoniae 48 LSENVVGTFSLPYSLVPEVLVNGQ----EYTVPYVTEEPSVVAAASYASKIIKRAG 100
S.aureus 48 LIENVIAQGALPVGLLPNIIVDDK----AYVVPMMVEEPSVVAAASYGAKLVNQTG 100
G.lamblia MSENVIGATMLPLSVVPDVLIDGT----MYTVPISTEEPSVVAALAHAAKIFRIGK
Class2
*

Figure 3.4 Protein sequence alignment of representative class 1 and class 2 HMG-CoA reductases.

The complete MVA pathway of *Streptomyces aeriouvifer* CL190 was recently obtained by Kaneda *et al.* (Kaneda *et al.*, 2001) and found to be encoded within a single operon, along with a type 2 IDI. This operon structure bolsters the suggestion that the MVA pathway was disseminated within the actinomycetes by LGT.

Pathway displacement in firmicutes and proteobacteria

The only bacterium outside the actinomycetes known to harbor the two complete isoprenoid biosynthesis pathways is the firmicute *Listeria monocytogenes*. However, even the closely related *Listeria innocua*, of which the genome has been completely sequenced, is missing the *gcpE* and *lytB* genes in its MEP pathway. Although isoprenoid metabolism has not been studied in *Listeria*, it is unlikely that the partial MEP pathway in *L. innocua* is functional, the two enzymes it is missing catalyzing essential steps (Campos *et al.*, 2001; McAteer *et al.*, 2001). The two pathways having the same final products (IPP and DMAPP), they might have become redundant. Consequentially, one of them (the MEP pathway) might be on its way to elimination. Some other firmicutes, which use the MVA pathway for the biosynthesis of their isoprenoids (Wilding *et al.*, 2000), also harbor genes of the MEP pathway: DXS is found in *Lactococcus lactis* and *ychB* is found in *Oceanobacillus iheyensis* and *Staphylococcus aureus* (Table 3.2).

There are two possible explanations for the presence of MEP pathway enzymes in these firmicutes. A first possibility is displacement. Since the MEP pathway is most likely ancestral in bacteria, it could have been displaced by an MVA pathway acquired through LGT by the common ancestor of these firmicutes. The few MEP pathway genes that can still be seen would simply be remnants of this ancestral pathway. Another firmicute, *Listeria innocua*, with its complete MVA pathway and its MEP pathway missing two genes, might be an example of the start of such a displacement. The α-proteobacteria *Paracoccus zeaxanthificans* could be an example of such a displacement process once it is complete. This proteobacterium has been shown, through carbon tracing studies, to use solely the MVA pathway for the biosynthesis of its isoprenoids (Eisenreich *et al.*, 2002). It is the only member of the α-proteobacteria that harbors the MVA pathway, all other members of this class use the MEP pathway to synthesize their isoprenoids. *Paracoccus*' MVA pathway is also encoded by a single operon (Humbelin *et al.*, 2002), making its acquisition by LGT in a single event possible. The HMGR and HMGS genes of *Paracoccus* are likely to have been acquired from actinomycetes, as they cluster strongly with their orthologs from the latter group in phylogenetic analyses (Figure 3.3A, B).

A second explanation for the presence of MEP pathway enzymes in firmicutes would be lateral transfer of the genes encoding these proteins from other bacteria. The MEP pathway is not generally found as an operon; its genes are usually scattered around a genome. As a consequence, the pathway is unlikely to be transferred as a unit. *Listeria* are therefore unlikely to have acquired the whole pathway by LGT. The possibility that the other firmicutes might have acquired a few of the MEP pathway genes by LGT is difficult to prove or disprove, as the phylogenetic trees of most MEP enzymes show little backbone support (Figure 3.5).

Whether the MEP pathway genes found in firmicutes are remnants of an ancestral pathway displaced by the MVA pathway or have been acquired more recently, LGT played an important role in the evolution of their isoprenoid metabolism

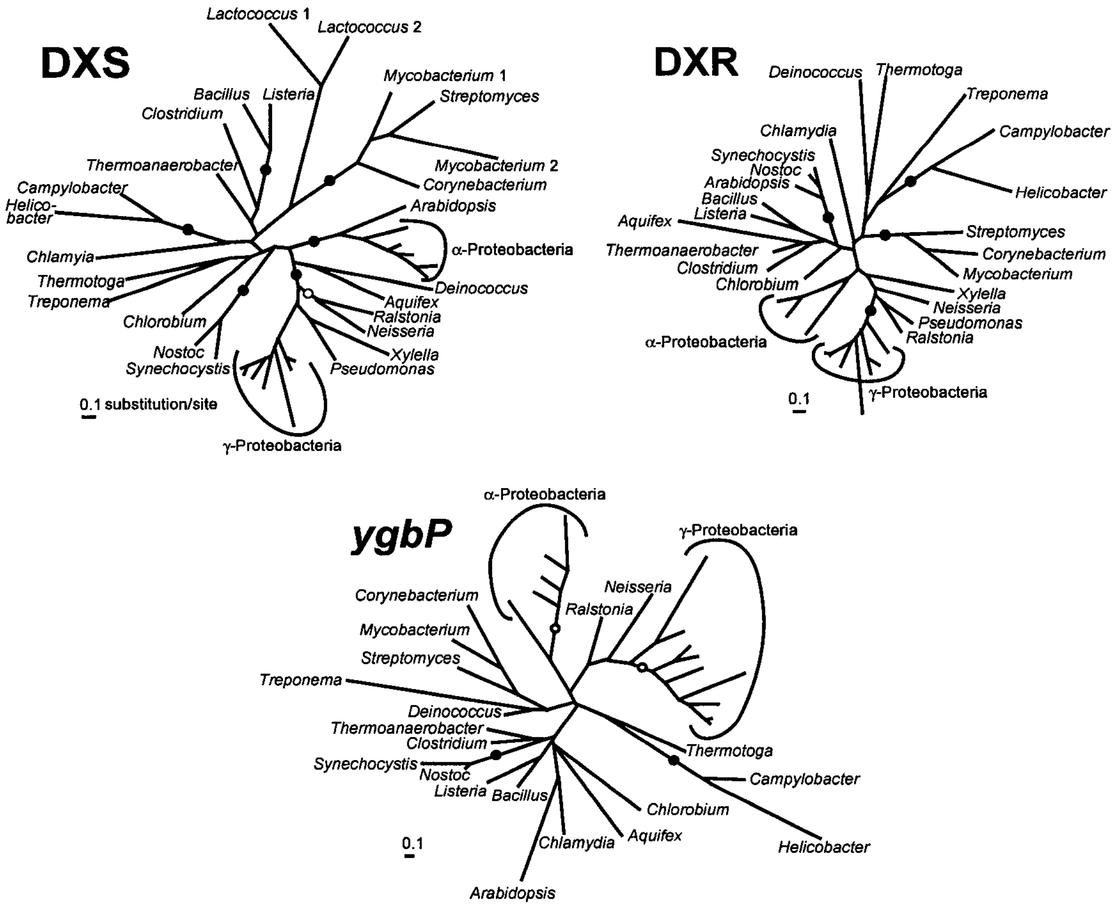


Figure 3.5 Best maximum likelihood phylogenetic trees of the MEP pathway enzymes. Black dots indicate maximum likelihood distances bootstrap support over 95% and white dots support over 80%. Archaea are highlighted in bold. DXS (1-deoxy-D-xylulose-5-phosphate synthase), DXR (1-deoxy-D-xylulose-5-phosphate reductoisomerase), *ygbP* (4-diphosphocytidyl-2-C-methyl-D- erythritol synthase), *ychB* (MEP kinase), *ygbB* (2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase), *gcpE* (4-hydroxy-3-methylbut-2-enyl diphosphate synthase, *lytB* (LytB).

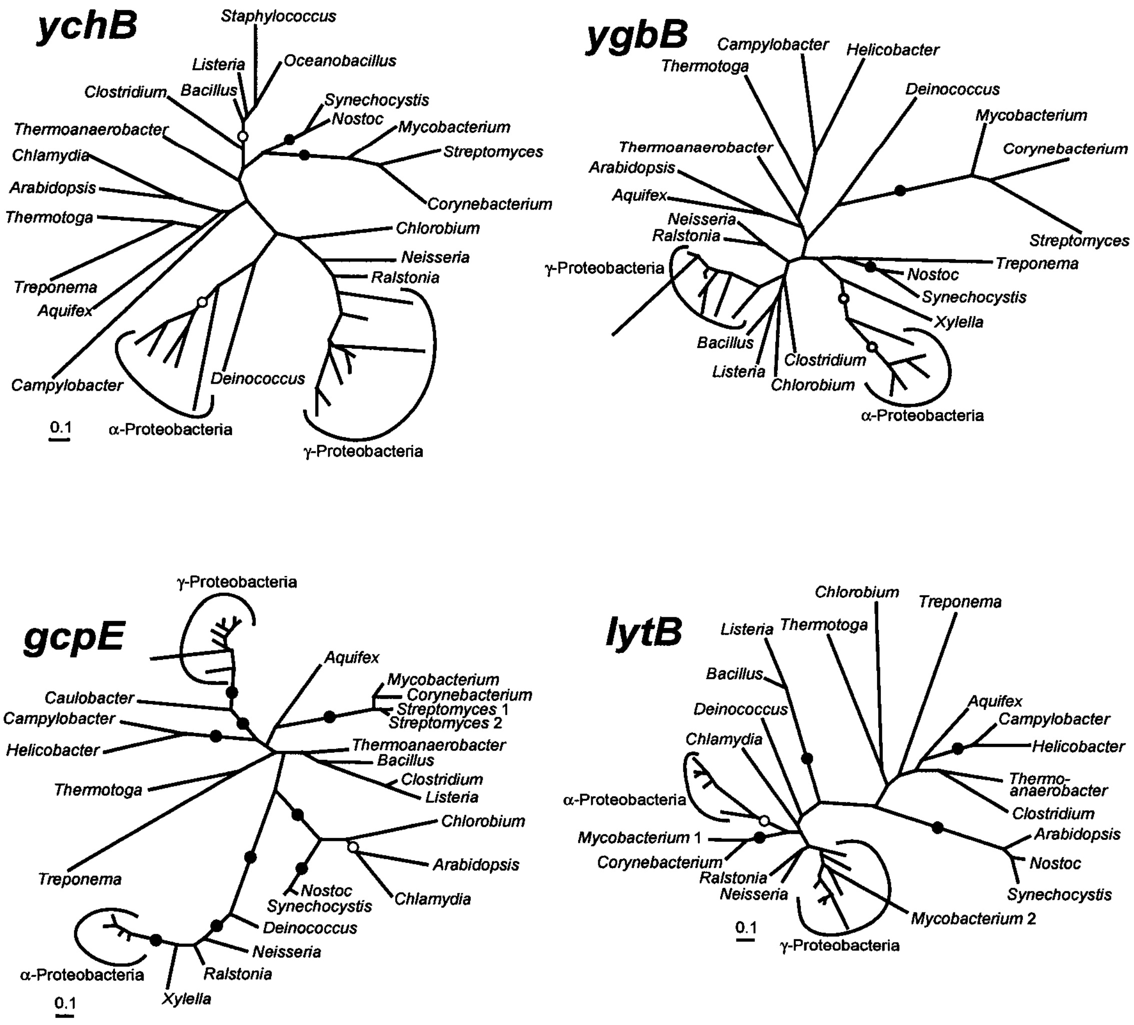


Figure 3.5 Best maximum likelihood phylogenetic trees of the MEP pathway enzymes (continued).

Endosymbiotic origin of the MEP pathway in eukaryotes

As mentioned earlier, photosynthetic eukaryotes like plants and algae possess both IPP biosynthesis pathways. The MVA pathway is believed to be ancestral to all eukaryotes, but an endosymbiotic origin has been suggested for the MEP pathway. Indeed the cyanobacterial origin of the MEP pathway found in eukaryotes is supported by a strong clustering of photosynthetic eukaryotes with cyanobacteria in the phylogenetic trees of DXR and *lytB* (Figure 3.5). Moreover the genes encoding the MEP pathway enzymes are found in the nucleus and their products are targeted to the chloroplast (Lange *et al.*, 1998). According to carbon tracing studies, unicellular green algae, such as *Scenedesmus obliquus*, *Chlamydomonas reinhardtii* and *Chlorella fusca,* (Sprenger *et al.*, 1997), unlike their photosynthetic multicellular relatives, would seem to have completely lost the MVA pathway after acquisition of the MEP pathway (Disch *et al.*, 1998; Schwender *et al.*, 1996). This type of metabolic displacement might also have occurred independently in the apicomplexa *Plasmodium falciparum*. This protist possesses an apicoplast, an organelle derived from the same ancestral cyanobacterial endosymbiont as plastids (Waller *et al.*, 1998). Like plants and algae, all the genes of the MEP pathway are found encoded in the nuclear genome of *Plasmodium* and yield active proteins (Jomaa *et al.*, 1999). Furthermore, there is no detectable activity of the MVA pathway enzymes in this protist, which suggest that the latter is absent and therefore, like unicellular green algae, displaced by the MEP pathway.

Lateral transfer of individual isoprenoid biosynthesis pathway genes

Orthologous replacement of MEP pathway genes in bacteria

I mentioned earlier the possibility that several MVA pathway-using firmicutes acquired individual MEP pathway genes by LGT. Much stronger evidence suggests that some of the genes of this pathway might also be exchanged through homologous gene replacement between different organisms harboring this pathway. Phylogenetic analysis shows that the *gcp*E gene was most likely transferred from plants to *Chlamydia* and to *Chlorobium* (in the latter case the source might also be cyanobacteria) (Figure 3.5). This gene is also likely to have experienced extensive transfer among the Proteobacteria: from the γ-Proteobacteria to the α-proteobacterium *Caulobacter crescentus* and from the α-Proteobacteria to a γ-proteobacterium, *Xyllela fastidiosa* and two β-proteobacteria, *Neisseria meningitidis* and *Ralstonia solanacearum* (Figure 3.5).

Giardia: essential genes can be replaced in eukaryotes as well

Orthologous replacement of isoprenoid biosynthesis pathway genes does not seem to be limited to prokaryotes. *Giardia lamblia* is a parasitic protozoan (diplomonad), once thought to be an early branching eukaryote that secondarily lost its mitochondria (Roger, 1999). Like most other non-photosynthetic eukaryotes, it uses the MVA pathway to synthesize its isoprenoids. I obtained its HMGR gene by PCR, using exact match primers designed using partial sequences obtained from the *Giardia* genome sequencing project (McArthur *et al.*, 2000). Surprisingly, the sequence of its genes revealed that it encodes a bacterial class 2 enzyme. This result is strongly supported by phylogenetic analyses

(Figure 3.3B) and by protein sequence motifs shared by *G. lamblia* and bacterial HMGRs (Figure 3.4). Such "bacterial" genes in eukaryotic nuclear genomes are often taken to be of "mitochondrial origin". Since all other known eukaryotic HMGR genes are of class 1, the *G. lamblia* gene would have had to be transferred from an α-proteobacterial symbiont that gave rise to mitochondria after diplomonads separated from the rest of eukaryotes. This constraint makes us favor the idea that *G. lamlia* acquired this class 2 HMGR independently, from a bacterial source. The ancestor of all eukaryotes, like the extant representatives of this domain, most certainly possessed a full MVA pathway, including a class 1 HMGR. It seems that *G. lamblia* lost its ancestral class 1 HMGR, as it has not so far been found by the complete genome sequencing project (McArthur *et al.*, 2000). The intestinal environment in which this diplomonad lives makes it easy to acquire genes from bacteria, and *G. lamblia* can harbour bacterial endosymbionts (Adam, 1991). The operation of a gene transfer ratchet could explain the replacement of nuclear genes by genes originating from the bacterial endosymbionts, even in the absence of selection (Doolittle, 1998). Selective pressure could further facilitate this process (for example, acquisition of resistance to naturally-occurring statins by replacement of an HMGR class 1 gene by a class 2 gene).

Acquisition of MVA pathway genes by bacteria

In some instances, only one or a few genes of the MVA pathway are present in an organism. This is the case for *Vibrio cholerae*, a human pathogen that can live in estuarine and marine environments (Colwell and Huq, 1994). The complete genome sequence data suggest that *V. cholerae* has a class 1 HMGR: phylogenetic analysis strongly supports this claim (Figure 3.3B). More specifically, *V. cholerae* always groups within archaea in the HMGR phylogenetic trees, with all three tree reconstruction methods used. In addition, a four amino-acid insertion present in archaeal HMGRs and absent from all bacterial class 2 and eukaryotic class 1 enzymes is found in *V. cholerae* HMGR (Figure 3.4). All seven genes of the MEP pathway are found in *V. cholerae*, as well as in all other completed γ-proteobacteria genomes (Table 3.2). Also, all γ-proteobacteria in which isoprenoid biosynthesis has been examined by biochemical studies use the MEP pathway (Rosa Putra *et al.*, 1998). Thus, the presence of these genes in *V. cholerae* probably indicates that it uses the MEP pathway for its primary biosynthetic needs. The HMGR gene would have been subsequently acquired by an ancestor of *V. cholerae* from an archaeon.

Pseudomonas mevalonii is another γ-proteobacterium where HMGR is the only gene of the mevalonate pathway present. However, the origin and role of this gene in *P. mevalonii* are presumably very different from *V. cholerae*. The former harbors an operon encoding for HMGR and an HMG-CoA lyase, the two enzymes being used in the reverse direction for mevalonate degradation (Gill *et al.*, 1985). This small operon is not found in any other *Pseudomonas* or γ-proteobacteria. Phylogenetic analysis indicates that *P. mevalonii* HMGR is a typical class 2 bacterial enzyme, as opposed to the class 1 archaeal HMGR found in *V. cholerae*. However, the fact that HMGR is absent from the vast majority of γ-proteobacteria and that it is found in an operon in *P. mevalonii* is highly suggestive of a lateral transfer of HMGR from other prokaryotes to this proteobacterium.

Individual transfer of MVA pathway genes also occur in bacteria that possess this pathway in its entirety and use it anabolically. Carbon tracing studies indicated that the green non-sulfur bacterium *Chloroflexus aurantiacus* synthesizes its isoprenoids through the MVA pathway (Hefter *et al.*, 1993). I confirmed the presence of HMGR gene in this bacterium by PCR amplification with degenerate universal HMGR primers. Phylogenetic analysis indicated that this gene encodes for a bacterial class 2 enzyme (Figure 3.3B). Phylogeny of other MVA pathway genes obtained from the partial genome sequence suggests that *C. aurantiacus* cobbled together its MVA pathway from multiple sources. It clearly acquired its HMGS from archaea (Figure 3.3A) and its PPMD gene clusters strongly with archaeal sequences and could have originated from Thermoplasmatales or extremely halophilic archaea (Figure 3.3D). Alternatively, *Chloroflexus* could be the source of the PPMD found in these archaea. This second alternative seems less likely, as *Chloroflexus* PPMD gene clusters strongly in-between the Halobacteriales and Thermoplasmatales homologs in phylogenetic analyses (as opposed to occupying a basal position if it was the ancestral sequence).

Acquisition of bacterial and eukaryotic MVA pathway genes by archaea

As mentioned earlier, although archaea share the enzymes catalyzing the early steps of the MVA pathway with eukaryotes and bacteria, the last two enzymes of this standard bacterial/eukaryotic pathway (PMK and PPMD) are missing in most archaea (Smit and Mushegian, 2000). Regardless of what enzymes catalyze those two steps in the majority of archaea, some representatives of that domain seem to have acquired the bacterial or eukaryotic versions of PMK and PPMD.

Our database survey revealed that both *Sulfolobus* species of which the genome was completely sequenced (*S. tokadaii* and *S. solfataricus*) harbour a protein with a high degree of similarity to the PMK found in fungi and bacteria. Phylogenetic analysis of MVK and PMK (which are homologous enzymes) cluster this *Sulfolobus* protein with PMKs from the fungi *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Figure 3.3C). A protein showing high similarity to PPMD from eukaryotes is encoded adjacent to this putative PMK in the genomes of *S. tokadaii* and *S. solfataricus*. In phylogenetic analysis, this protein clusters very strongly with eukaryotic homologs (Figure 3.3D). This suggests that this PPMD homolog was acquired from eukaryotes by an ancestral *Sulfolobus*. The same is likely to hold for PMK, as this enzyme is not found in any other archaea and the *Sulfolobus* homologs show some affinity to PMKs from fungi. The presence of these two enzymes in *S. tokadaii* and *S. solfataricus* means that *Sulfolobus* is the only genus of archaea known to possess all five enzymes of the standard mevalonate pathway.

However, *Sulfolobus* are not the only archaea to harbour a PPMD. Genes encoding for homologs of this enzyme were found in the complete or partial genome sequences of three Halobacteriales (*Halobacterium sp.* NRC-1, *Haloferax volcanii* and *Haloarcula marismortui*) (Ng *et al.*, 2000). Our database survey also detected this enzyme in the three partially or completely sequenced Thermoplasmatales genomes (*Thermoplasma acidophilum*, *Thermoplasma volcanium* and *Ferroplasma acidarmanus*). However, phylogenetic analysis indicates that the PPMDs found in the Halobacteriales and the Thermoplasmatales are likely to have a different origin than the *Sulfolobus* enzyme. Indeed, these PPMD genes cluster with bacterial homologs and are distinct from eukaryotic sequences (Figure 3.3D).

Conclusion

There are many and complex stories here, but a general picture emerges. Isoprenoid biosynthesis is an essential function that can be carried out by two sets of analogous genes. In Bacteria, one of them (the MVA pathway) has acquired mobility by being encoded in only one or two operons, and in some cases has displaced the ancestral (MEP) pathway. In other cases, it has been maintained along with the MEP pathway, bringing additional functionality (i.e. synthesis of secondary metabolites). Individual genes of both pathways have also been exchanged through homologous replacement or acquired as stable units, over short and long evolutionary distances. The absence of the MEP pathway in Archaea suggests some fundamental incompatibility, which is not genetic since individual genes float freely between domains and bacteria can tolerate either or both pathways. This incompatibility helps define the Archaea as a group and characterizes their metabolism.

Section II: Origins and evolution of isoprenoid lipid biosynthesis in archaea

Archaeal membrane lipids have several interesting characteristics that distinguish them from their bacterial and eukaryotic counterparts: 1) isoprenoid, not fatty acid, side chains 2) ether, not ester, links joining these side chains to the glycerol phosphate backbone 3) the *sn*-2,3, not *sn*-1,2, stereochemistry of this backbone. Archaeal lipids side chains, like all other isoprenoids, are assembled from two universal precursors: isopentenyl diphosphate (IPP) and its isomer dimethyllalyl diphosphate (DMAPP) (Figure 3.6). Archaea produce IPP through the MVA pathway, as discussed in section 1 of this chapter. DMAPP, which is also required to synthesize polyisoprenoids, originates from the structural rearrangement of IPP by an isopentenyl diphosphate isomerase (IDI) of either of two functionally analogous but non-homologous types (type 1 and 2). The type 2 IDI (IDI2) was only recently discovered by Kaneda *et al.* (Kaneda *et al.*, 2001), who have shown IDI2 to be part of a cluster with genes of the MVA pathway in the actinomycete *Kitasatospora griseola* (also known as *Streptomyces griseolosporeus*) and identified its presence in several other bacteria and archaea. The type 1 IDI (IDI1), traditionally found only in eukaryotes and bacteria, was discovered in the genome sequence of the extremely halophilic archaeon *Halobacterium sp.* NRC-1 (Ng *et al.*, 2000).

Figure 3.6 Pathways for the synthesis of isopentenyl diphosphate and its assembly in isoprenoid ether lipids side chains. A boxed "A" indicates that an enzyme is present in all archaea and a boxed "S" indicates that it is present in only a subset of them. Abbreviations: phosphate (P), isopentenyl diphosphate (IPP), dimethylallyl diphosphate (DMAPP), GPP (geranyl diphosphate), FPP (farnesyl diphosphate), GGPP (geranylgeranyl diphosphate), FGPP (farnesylgeranyl diphosphate), GGGP (geranylgeranylglyceryl phosphate), DGGGP (digeranylgeranylglyceryl phosphate).

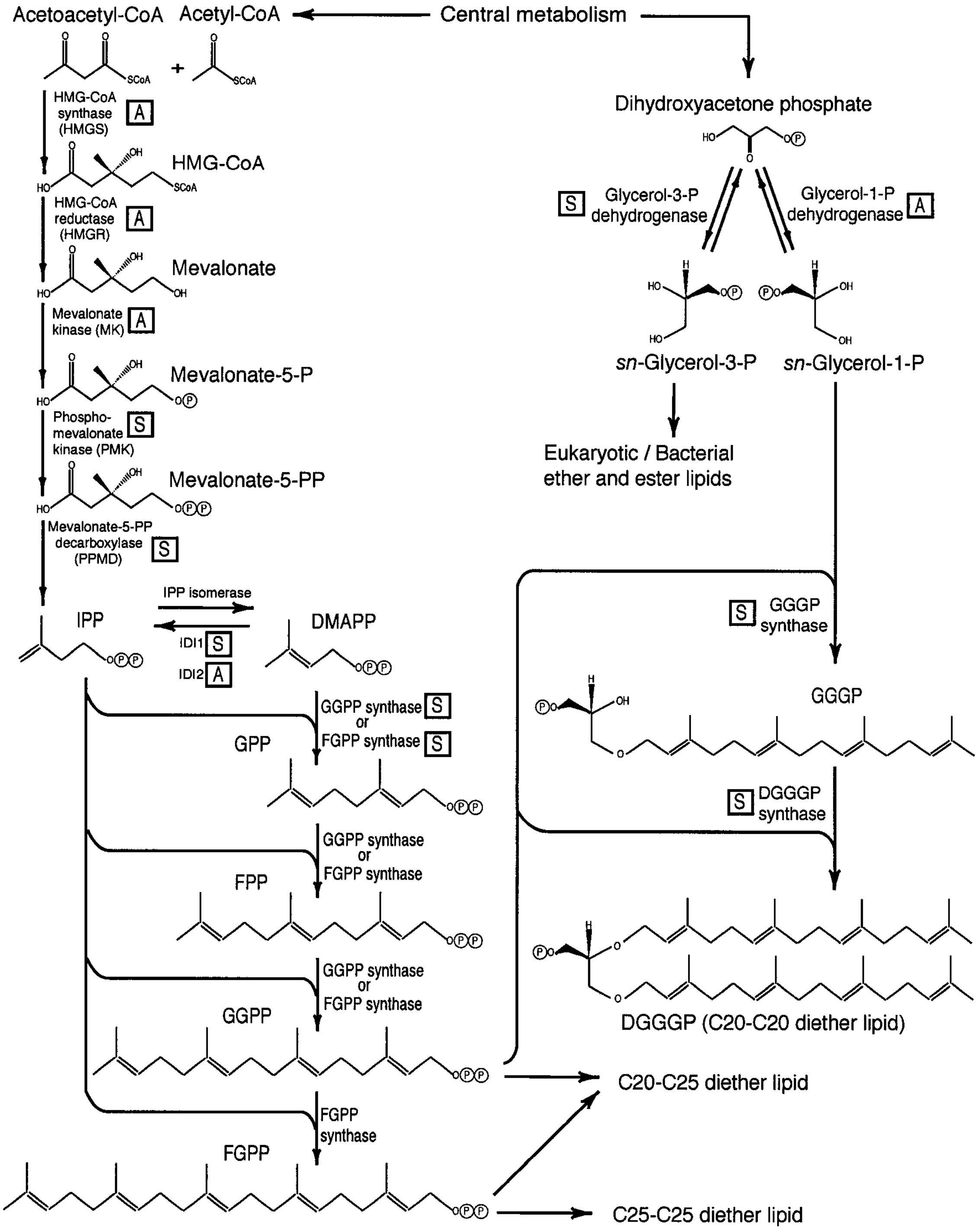


Figure 3.6 Pathways for the synthesis of isopentenyl diphosphate and its assembly in isoprenoid ether lipids side chains.

Archaeal lipid side chains (for acyclic lipids) are composed of 20 or 25 carbons (C20 or C25). Such a length is reached through the sequential condensation of IPP to a growing allylic polyisoprenoid diphosphate (Figure 3.6). The first molecule in the chain is the IPP isomer DMAPP, to which IPP molecules are successively added to obtain GPP (10 carbons), FPP (15 carbons), GGPP (20 carbons) and FGPP (25 carbons) (Figure 3.6). These chain elongation reactions are catalyzed by isoprenyl diphosphate synthases, such as GGPP and FGPP synthases, which can differ from one another by the allylic substrate they accept to start the elongation process and the chain length of their product(s) (Kellogg and Poulter, 1997). In archaea, GGPP synthase can elongate DMAPP to obtain both FPP and GGPP, the latter being the isoprenyl forming the side-chains of C20-C20 diether lipids. Archaea harboring C20-C25 or C25-C25 diether lipids require an FGGP synthase, which can either elongate directly from DMAPP {i.e. *Aeropyrum pernix*, see (Tachibana *et al.*, 2000)} or from longer allylic substrates like GGPP {*Natronobacterium*, see (Tachibana, 1994)}.

The sequence of the enzyme responsible for linking the first side-chain to the glycerol backbone of C20-C20 diether lipids was recently obtained from *Methanobacterium thermoautotrophicum* (Soderberg *et al.*, 2001). This enzyme, termed geranylgeranylglyceryl phosphate (GGGP) synthase, was identified in six other archaeal species by similarity searches (Soderberg *et al.*, 2001). It strongly favors *sn*-glycerol-1-phosphate as a substrate, and therefore plays a role in defining the stereoconfiguration of archaeal lipids. The enzyme responsible for adding the second C20 isoprenoid side-chain on *Methanobacterium* diether lipids, DGGGP synthase, was identified in a separate cellular fraction but its amino acid sequence has yet to be determined (Zhang and Poulter, 1993).

For a long time after the unique stereoconfiguration of archaeal lipids was determined, there have been questions about the precursor used to synthesize the glycerol phosphate backbone. Nishihara *et al.* (Nishihara and Koga, 1995) identified *sn*-glycerol-1-phosphate (G1P) dehydrogenase to be responsible for the synthesis of the *sn*-glycerol-1-phosphate phospholipid backbone from dihydroxyacetone phosphate (DHAP) in *Methanobacterium thermoautotrophicum.* They subsequently demonstrated that the activity specific to this enzyme occurred in five other archaeal species and that the gene encoding for this enzyme was present in all complete archaeal genome sequences available at the time (Nishihara *et al.*, 1999).

To better understand their origin and evolution, phylogenetic analyses were performed on all characterized enzymes involved in any steps of archaeal isoprenoid lipid biosynthesis, from the synthesis of the isoprenoid precursor DMAPP to the linkage between the glycerol phosphate backbone and the first isoprenoid side-chain.

Materials and Methods

Archaeal strains and DNA extraction

The following strains of extremely halophilic archaea were used in this study: *Halobacterium halobium* JCM9120, *Halobacterium salinarum* ATCC19700, *Halococcus morrhuae* NRC16008, *Haloferax denitrificans* ATCC35960, *Haloferax mediterranei* ATCC33500, *Halorhabdus utahensis* DSM12940, *Halorubrum dustributum* JCM9100, *Haloterrigena turkmenicus* VKM B-1734, *Natrialba asiatica* 172P1, *Natrinema versiforme* XF10, *Natronobacterium gregoryi* NCMB2189, *Natronomonas pharaonis*

DSM2160, *Natronorubrum sp.* Tenzan-10 and *Natrinema* sp. XA3-1. Genomic DNA was extracted from these strains using the protocol from Wilson (1994).

PCR (polymerase chain reaction) amplification, cloning and sequencing

The amplification of IDI1, IDI2, GGGPS and GD genes was done in two steps. First, several degenerate primers were designed from available database sequences for the amplification of each gene from the genomic DNA of all available strains of halophilic archaea. Sequences obtained for each gene were subsequently aligned to design better degenerate primers for all five genes (Table 3.4). Each gene was amplified from two independent PCR reactions. Amplifications were carried out in a final volume of 25μl containing 1-5 ng of template DNA, 1 X PCR buffer, 2.5 mM MgCl2, 0.2 mM dNTPs, 1.0 mM of each primer, and 0.5-1 U of Platinum Taq High Fidelity DNA polymerase (INVITROGEN). The reactions were performed with an initial denaturation at 94°C for 1 min., 35 cycles with a denaturation at 94°C for 30 sec., primer annealing at 48-52°C for 30 sec., and primer extension at 72°C for 1 min. PCR products were gel purified with the MinElute kit (QIAGEN) and cloned in TopoTA (INVITROGEN). Two clones were sequenced from both strands for each PCR product using a LiCor 4000L automated sequencer.

Table 3.4 Primers used for the amplification of isoprenoid lipid biosynthesis genes.

Primer	**Sequence (5' to 3')**
Isopentenyl diphosphate isomerase type 1 (IDI1)	
Halobacteriales forward F1	GGCTTCCAGTASACCCNCC
Halobacteriales reverse R1	TGGGACACCTNCTGGGAYGG
Isopentenyl diphosphate isomerase type 2 (IDI2)	
Halobacteriales forward F1	GCGATCTCGAACCANGGRCA
Halobacteriales reverse R1	ATCGACTCAATGACNGGNGG
Geranylgeranylglyceryl phosphate synthase (GGGPS)	
Halobacteriales forward F1	GTGCCCCTCTAYCAGGARCC
Halobacteriales forward F2	CACGTCACGAARTGNGAYCC
Halobacteriales reverse R1	CTGGATGCCRCCRCCRTAGAA
Halobacteriales reverse R2	CCCACCACGACSGCGTCSGC
Glycerol dehydrogenase (GD)	
Halobacteriales forward F1	CCGTCCACGTAKGTNCARGG
Halobacteriales forward F2	CTGGCGACNTCCTTYGARGC
Halobacteriales reverse R1	ATATTGACCTTYTCNCCRTG
Halobacteriales reverse R2	GGCTCATCGTGDATNGTYTC

Sequence retrieval, multiple sequence alignment and phylogenetic analysis

All genes sequences were retrieved from the NCBI website (http://www.ncbi.nlm.nih.gov/) through similarity searches performed using BLASTP (http://www.ncbi.nlm.nih.gov/BLAST/). A biochemically characterized archaeal homolog of each enzyme was used as the query. Functional homology of significant hits (e-value lower than $1X10^{-5}$) was inferred if amino acid motifs important to the functionality of the particular enzyme were conserved. Preliminary sequence data from the unfinished genomes of *Haloferax volcanii* (http://wit-scranton.mbi.scranton.edu/Haloferax/) and *Haloarcula marismortui* (http://zdna2.umbi.umd.edu/cgi-bin/blast/blast.pl) were obtained from their respective websites. The retrieved amino acid sequences and the new sequences from this study were aligned using CLUSTALW (Thompson *et al.*, 1994). The alignment was subsequently edited manually to remove ambiguous characters and regions corresponding to the primers sequence in novel sequences. 5' and 3' regions absent from the novel sequences (that are only partial) were kept in the alignment. The number of sites used in the different protein sequence alignments were as follows (the first number is the length of the novel partial sequences and the second number is the full length of the alignment): IDI1 (87/139), IDI2 (78/243), GGPPS (198/198) GGGPS (131/177), GD/G1PD (146/268). Maximum likelihood phylogenetic analyses were performed using PROML with the JTT amino acid substitution matrix, a rate heterogeneity model with gamma-distributed rates over four categories with the α parameter estimated using TREE-PUZZLE, global rearrangements and randomized input order of sequences (10 jumbles). Bootstrap support values represent a consensus (obtained using CONSENSE) of 100 Fitch-Margoliash distance trees (obtained using

PUZZLEBOOT and FITCH) from pseudo-replicates (obtained using SEQBOOT) of the original alignment. The settings of PUZZLEBOOT were the same as those used for PROML, except that no global rearrangements and randomized input order of sequences are available in this program. PROML, CONSENSE, FITCH and SEQBOOT are from the PHYLIP package version 3.6a (http://evolution.genetics.washington.edu/phylip.html). TREE-PUZZLE and PUZZLEBLOOT can be obtained from the programs website (http://www.tree-puzzle.de).

Phylogenetic analyses were also carried at the DNA level for the taxa belonging to the order Halobacteriales. These analyses were performed with PAUP* 4.04b (Swofford, 1998) applying the heuristic-search option and using the TBR branch-swapping algorithm. Maximum likelihood was used as the tree reconstruction method, with the nucleotide substitution model, gamma rates parameter α, proportion of invariable sites and nucleotide frequencies determined independently for each gene using MODELTEST (Posada and Crandall, 1998). The confidence of each node was determined by building a consensus tree of 100 bootstrap replicates.

Results

Biosynthesis of an isoprenoid building block: biosynthesis of DMAPP from IPP

The presence of type 2 isopentenyl diphosphate isomerase is a universal feature of archaea

This alternative IPP isomerase, first discovered by these workers in the mevalonate gene cluster of *Kitasatospora griseola*, does not share detectable sequence similarity with the bacterial/eukaryotic IDI1. In addition to archaea and *Kitasatospora*, IDI2 is found in various bacteria (although only those that do not possess IDI1) and in the trypanosomatid *Leishmania major*. An IPP isomerase is essential to all organisms possessing only the mevalonate pathway to synthesize their isoprenoids, as they need the enzyme to make the essential precursor DMAPP in addition to IPP.

The presence of two types of isopentenyl diphosphate isomerase is an ancestral feature of extremely halophilic archaea

Although IDI2 is universally found in archaea, type 1 isopentenyl diphosphate isomerase (IDI1) was originally thought to be restricted to eukaryotes and bacteria. Our database survey indicated that the IDI found in the genome of *Halobacterium sp*. NRC-1 (Ng *et al.*, 2000) was homologous to the type 1 isomerase of eukaryotes and bacteria. This makes *Halobacterium*, which like all other archaea harbours a type 2 IDI, the only organism known to harbour both types of isopentenyl diphosphate isomerase.

To confirm that the presence of two types of IDIs was a general trait of the Halobacteriales (the archaeal order representing extreme halophiles) and not simply a specific feature of *Halobacterium*, the genes encoding these two enzymes were amplified from representatives of several genera of this order. Table 3.5 reports the results of this survey, in which IDI1 was obtained from seven species, and IDI2 from three. IDI1 has also been identified in the three haloarchaea for which a genome sequencing project is under way, while IDI2 has been detected in two of the three (Table 3.5).

The fact that these genes are short and present relatively few highly conserved regions makes the design of efficient primers difficult. Failure to consistently amplify these genes from the DNA of all strains on which PCR was performed should therefore not be interpreted as the absence of the gene in question. The high numbers of strains from which each of these genes was nonetheless amplified does suggest that they are a universal feature of Halobacteriales, or at least present in the vast majority of species in this group. In phylogenetic analyses of the two IDI genes (Figure 3.7A, B), Halobacteriales form monophyletic clusters, suggesting that they are ancestrally present in this order. The distribution and phylogeny of IDI1 suggest that it was acquired by LGT from bacteria. This enzyme is found in several groups of bacteria and eukaryotes, but limited to extreme halophiles in archaea. Also, haloarchaeal IDI1s cluster with bacterial homologs in phylogenetic analysis, albeit weakly (Figure 3.7A).

Table 3.5 Isoprenoid and lipid biosynthesis enzymes found in Archaea

ARCHAEA	ENZYME										REFERENCE
	HMGS	HMGR	MVK	PMK	PPMD	IDI1	IDI2	IPPS	GGGPS	G1PD	
HALOBACTERIALES	■	■	■	▨	B	B	■	G	D	■	*Halobacterium sp.* NRC-1 (complete genome)
Haloarcula	■	■	■		B	B			D	■	*Haloarcula marismortui* (partial genome)
Halococcus					B				D		This study
Haloferax	■	■	■		B	B	■		D	■	*Haloferax volcanii* (partial genome)
Halorhabdus					B	B					This study
Haloterrigena						B	■		D		This study
Halorubrum						B	■		D		This study
Natrialba						B			D		This study
Natrinema						B					This study
Natronobacterium					B	B	■		■		This study
Natronomonas						B					This study
Natronorubrum					B						This study
METHANOCOCCALES	■	■	■	▨	▨	■	■	G	■	■	*M. jannaschii* (complete genome)
METHANOBACTERIALES	■	■	■	▨	▨	■	■	G	■	■	*M. thermoautotrophicus* (complete genome)
METHANOPYRALES	■	■	■	▨	▨	▨	■	G	■	■	*Methanopyrus kandleri* (complete genome)
METHANOSARCINALES	■	■	■	▨	▨	▨	■	G	■	■	*Methanosarcina* (3 complete genomes)
THERMOCOCCALES	■	■	■	▨	▨	▨	■	G	■	■	*Pyrococcus* (3 complete genomes)
ARCHAEOGLOBALES	■	B	■	▨	▨	■	■	G	D	■	*Archaeoglobus fulgidus* (complete genome)
THERMOPLASMATALES	■	B	■	▨	B	■	■	G	■	■	*Thermoplasma* (2 complete genomes)
SULFOLOBALES	■	■	■	E	E	▨	■	G	■	■	*Sulfolobus* (2 complete genomes)
DESULFUROCOCCALES	■	■	■	▨	▨	▨	■	F	■	■	*Aeropyrum pernix* (complete genome)
THERMOPROTEALES	■	■	■	▨	▨	▨	■	G	■	■	*Pyrobaculum aerophilum* (complete genome)

HMGS	3-hydroxy-3-methylglutaryl CoA synthase
HMGR	3-hydroxy-3-methylglutaryl CoA reductase
MK	mevalonate kinase
PMK	phosphomevalonate kinase
PPMD	diphosphomevalonate decarboxylase
IDI1 and IDI2	isopentenyl diphosphate isomerase (IDI1 and IDI2 are two analogues)
IPPS	short-chain isoprenyl diphosphate synthase
GGGPS	geranylgeranylglyceryl phosphate synthase
G1PD	glycerol-1-phosphate dehydrogenase

Gray = absent or too divergent to be detected
Black = present
White = no information
E = Eukaryotic type
B = Bacterial type
D = Divergent from main archaeal type, but origin uncertain
F = farnesylgeranyl diphosphate (FGPP) synthase
G = geranylgeranyl diphosphate synthase (GGPP) synthase

Figure 3.7 Best maximum likelihood phylogenetic trees for various enzymes involved in the biosynthesis of isoprenoid lipids in archaea. Trees are arbitrarily rooted and only relevant nodes with support values over 50% are displayed. Archaeal taxa are highlighted in bold. A) Isopentenyl diphosphate isomerase type 1 (IDI1). B) Isopentenyl diphosphate isomerase type 2. C) Short (10 to 25 carbons) and medium (30 to 50 carbons) chain prenyltransferases, all archaeal short-chain prenyltransferases are geranylgeranyl diphsophate (GGPP, 20 carbons) synthases, except for the *Aeropyrum pernix* homolog, which is a farnesylgeranyl diphsophate (FGPP, 25 carbons) synthase, the residues found at positions 74 and 77 in the amino acid sequence of each taxa (positions based on *Sulfolobus acidocaldarius* GGPP synthase) are indicated in parenthesis beside the taxa names. These residues are believed to be important in limiting the length of the isoprenyl diphosphate produced (see text). D) Geranylgeranylglyceryl phosphate (GGGP) synthase. E) *sn*-Glycerol-1-phosphate dehydrogenase (G1PD) rooted with Glycerol dehydrogenase (GD).

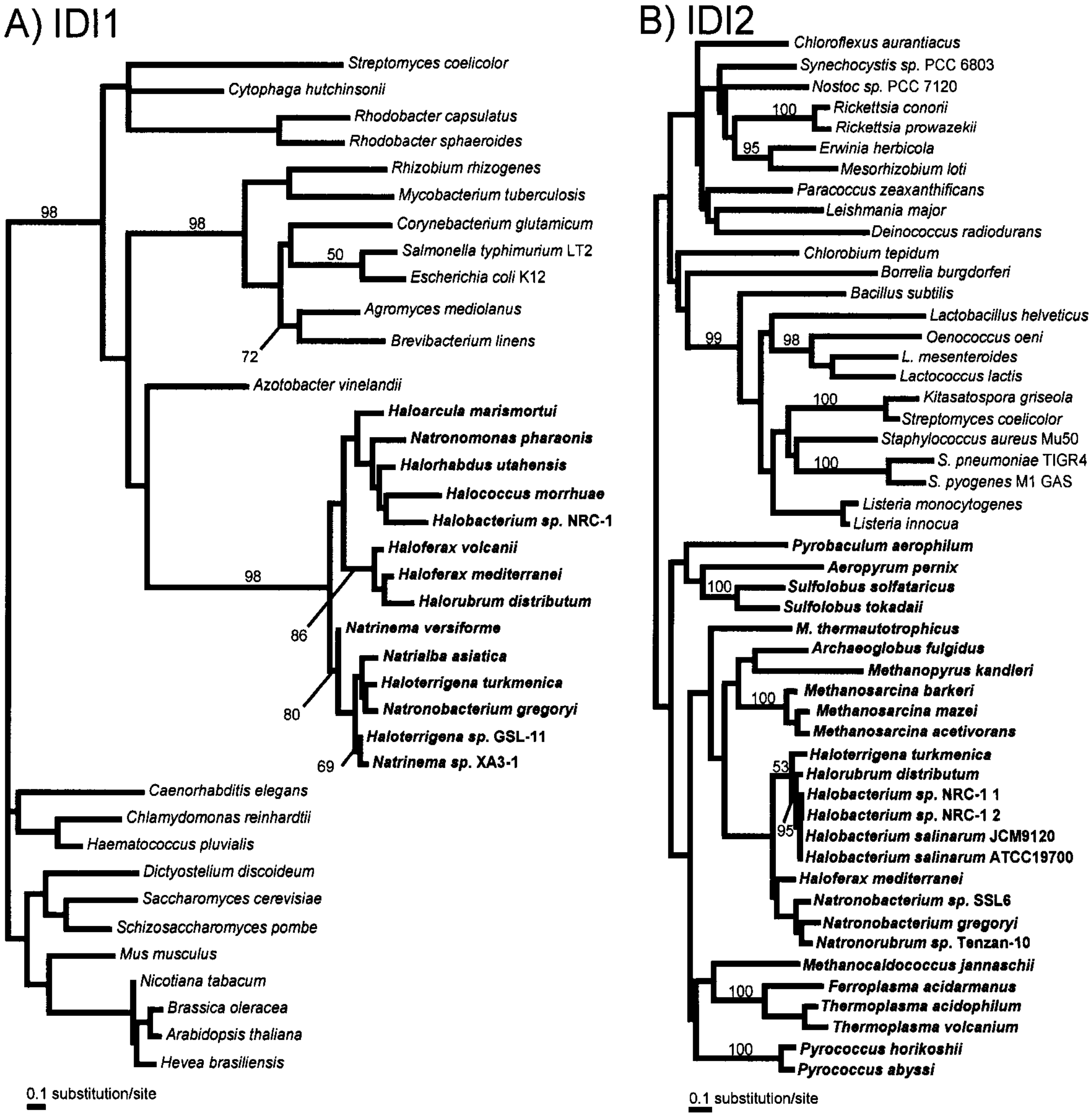
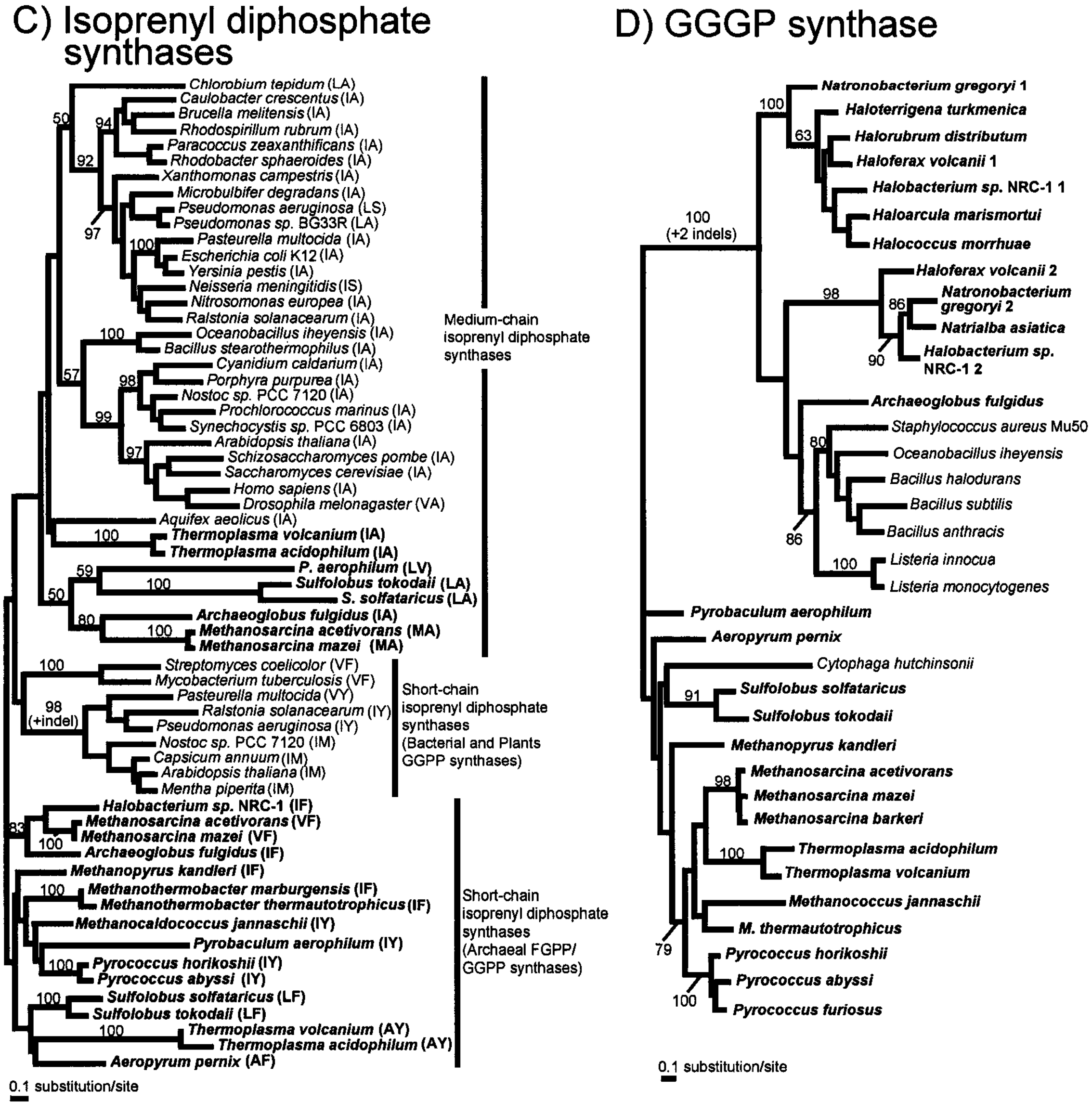


Figure 3.7 Best maximum likelihood phylogenetic trees for various enzymes involved in the biosynthesis of isoprenoid lipids in archaea.

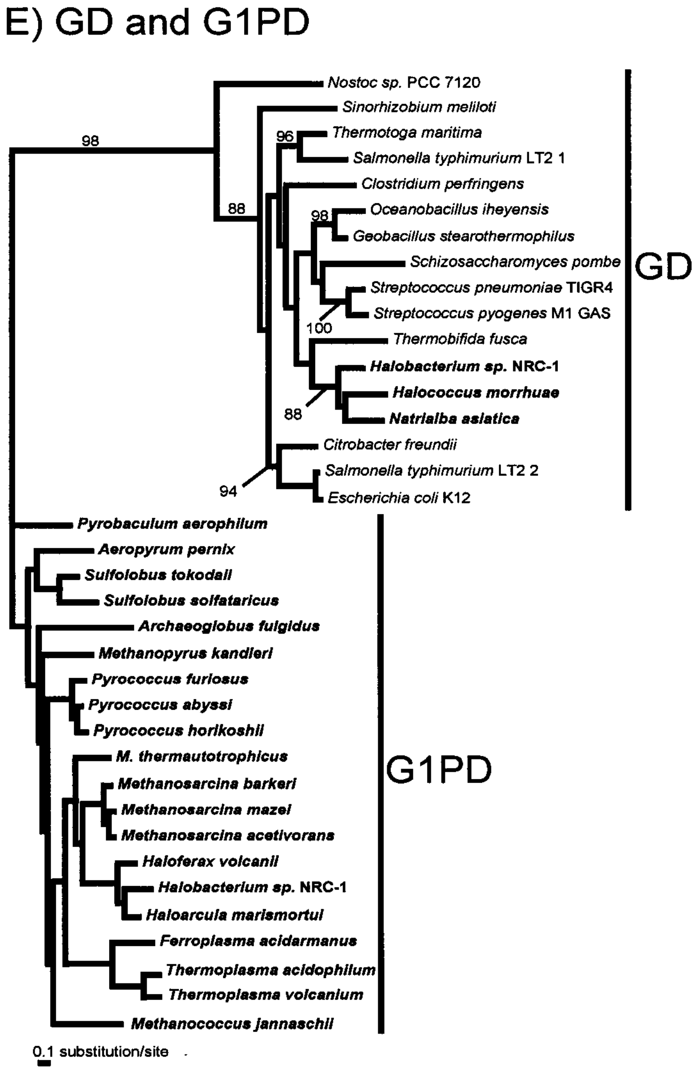


C) Isoprenyl diphosphate synthases

D) GGGP synthase

Chlorobium tepidum (LA)
Caulobacter crescentus (IA)
Brucella melitensis (IA)
Rhodospirillum rubrum (IA)
Paracoccus zeaxanthificans (IA)
Rhodobacter sphaeroides (IA)
Xanthomonas campestris (IA)
Microbulbifer degradans (IA)
Pseudomonas aeruginosa (LS)
Pseudomonas sp. BG33R (LA)
Pasteurella multocida (IA)
Escherichia coli K12 (IA)
Yersinia pestis (IA)
Neisseria meningitidis (IS)
Nitrosomonas europea (IA)
Ralstonia solanacearum (IA)
Oceanobacillus iheyensis (IA)
Bacillus stearothermophilus (IA)
Cyanidium caldarium (IA)
Porphyra purpurea (IA)
Nostoc sp. PCC 7120 (IA)
Prochlorococcus marinus (IA)
Synechocystis sp. PCC 6803 (IA)
Arabidopsis thaliana (IA)
Schizosaccharomyces pombe (IA)
Saccharomyces cerevisiae (IA)
Homo sapiens (IA)
Drosophila melonagaster (VA)
Aquifex aeolicus (IA)
Thermoplasma volcanium (IA)
Thermoplasma acidophilum (IA)
P. aerophilum (LV)
Sulfolobus tokodaii (LA)
S. solfataricus (LA)
Archaeoglobus fulgidus (IA)
Methanosarcina acetivorans (MA)
Methanosarcina mazei (MA)
Medium-chain isoprenyl diphosphate synthases
Streptomyces coelicolor (VF)
Mycobacterium tuberculosis (VF)
Pasteurella multocida (VY)
Ralstonia solanacearum (IY)
Pseudomonas aeruginosa (IY)
Nostoc sp. PCC 7120 (IM)
Capsicum annuum (IM)
Arabidopsis thaliana (IM)
Mentha piperita (IM)
Short-chain isoprenyl diphosphate synthases (Bacterial and Plants GGPP synthases)
Halobacterium sp. NRC-1 (IF)
Methanosarcina acetivorans (VF)
Methanosarcina mazei (VF)
Archaeoglobus fulgidus (IF)
Methanopyrus kandleri (IF)
Methanothermobacter marburgensis (IF)
Methanothermobacter thermautotrophicus (IF)
Methanocaldococcus jannaschii (IY)
Pyrobaculum aerophilum (IY)
Pyrococcus horikoshii (IY)
Pyrococcus abyssi (IY)
Sulfolobus solfataricus (LF)
Sulfolobus tokodaii (LF)
Thermoplasma volcanium (AY)
Thermoplasma acidophilum (AY)
Aeropyrum pernix (AF)
Short-chain isoprenyl diphosphate synthases (Archaeal FGPP/ GGPP synthases)
50
94
92
97
100
100
57
98
99
97
100
59
100
50
80
100
100
98 (+indel)
83
100
100
100
100
100
0.1 substitution/site
Natronobacterium gregoryi 1
Haloterrigena turkmenica
Halorubrum distributum
Haloferax volcanii 1
Halobacterium sp. NRC-1 1
Haloarcula marismortui
Halococcus morrhuae
Haloferax volcanii 2
Natronobacterium gregoryi 2
Natrialba asiatica
Halobacterium sp. NRC-1 2
Archaeoglobus fulgidus
Staphylococcus aureus Mu50
Oceanobacillus iheyensis
Bacillus halodurans
Bacillus subtilis
Bacillus anthracis
Listeria innocua
Listeria monocytogenes
Pyrobaculum aerophilum
Aeropyrum pernix
Cytophaga hutchinsonii
Sulfolobus solfataricus
Sulfolobus tokodaii
Methanopyrus kandleri
Methanosarcina acetivorans
Methanosarcina mazei
Methanosarcina barkeri
Thermoplasma acidophilum
Thermoplasma volcanium
Methanococcus jannaschii
M. thermautotrophicus
Pyrococcus horikoshii
Pyrococcus abyssi
Pyrococcus furiosus
100
63
100 (+2 indels)
98
86
90
80
86
100
91
98
100
79
100
0.1 substitution/site

Figure 3.7 Best maximum likelihood phylogenetic trees for various enzymes involved in the biosynthesis of isoprenoid lipids in archaea (continued).



E) GD and G1PD

Figure 3.7 Best maximum likelihood phylogenetic trees for various enzymes involved in the biosynthesis of isoprenoid lipids in archaea (continued).

For its part, IDI2 does not seem to have been involved in inter-domain LGT. It is present in all archaea, which form a monophyletic cluster that includes the Halobacteriales orthologs. Despite the apparent absence of inter-domain lateral transfer, there does seem to have been exchange of this gene between species of Halobacteriales. Phylogenies of the different isoprenoid lipid biosynthesis genes amplified from a variety of Halobacteriales are mostly unresolved, as these genes are short and only partial sequences were obtained for most of them. However, there are two clades of Haloarchaea found in most of these phylogenies as well as in a phylogeny of the small ribosomal subunit (SSU) rRNA gene: one includes mostly neutrophilic halophiles (subgroup I) and the other contains mostly haloalkaliphiles (subgroup II) (Figure 3.8). This division does not hold for IDI2, as *Haloterrigena turkmenica*, clustering strongly with the subgroup II in the SSU and IDI1 phylogenies, clusters as strongly with subgroup I in the IDI2 phylogeny (Figure 3.8). This suggests LGT between genera of Halobacteriales for the IDI2 gene.

Figure 3.8 Best maximum likelihood phylogenetic trees for various enzymes involved in the biosynthesis of isoprenoid lipids in extremely halophilic archaea. The trees shown are from an analysis at the amino acid level. The bootstrap support values displayed before the dash corresponds to the consensus of 100 Fitch-Margoliash maximum likelihood distance trees (amino acid analysis) and the value displayed after the dash corresponds to the consensus of 100 maximum likelihood trees (DNA analysis). If (--) is displayed instead of a bootstrap value, it means that the bootstrap consensus tree disagreed with the best tree or was not resolved at that node. The top tree was constructed from an alignment of small ribosomal subunit rRNA genes (SSU). Since this gene is not protein coding, only the DNA analysis bootstrap value is presented. The nucleotide substitution model used for the DNA analysis of each gene is indicated under its respective tree (I= invariant sites, G= gamma distributed rates).

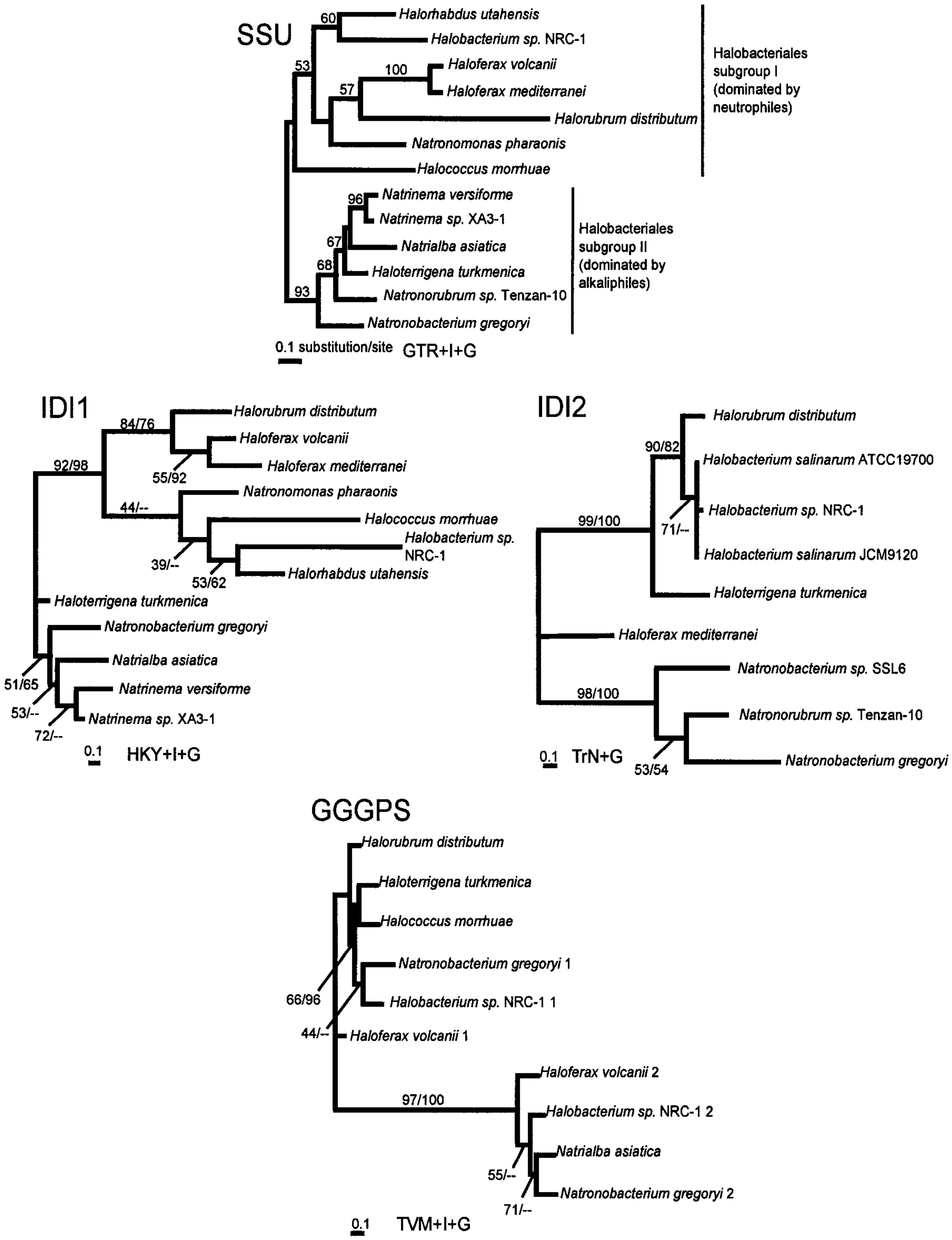


Figure 3.8 Best maximum likelihood phylogenetic tree for various enzymes involved in the biosynthesis of isoprenoid lipids in extremely halophilic archaea.

Biosynthesis of isoprenoid side-chains: isoprenyl diphosphate synthases

Functional plasticity of isoprenyl diphosphate synthases

Mutational studies on the GGPP synthase of *Sulfolobus acidocaldarius* demonstrated the extent of the plasticity of isoprenyl diphosphate synthases regarding the length of the isoprenoid products. A single amino acid substitution changed this archaeon GGPP synthase into an enzyme synthesizing FGPP as its main product and able to produce small amounts of hexaprenyl diphosphate (30 carbons) (Ohnuma *et al.*, 1997). With two or three amino acid substitutions, using DMAPP, FPP or GPP as the allylic substrate, the main product could reach lengths of 35 or 40 carbons (heptaprenyl and octaprenyl diphosphates) and secondary products reaching lengths of up to 65 to 120 carbons were also obtained (Ohnuma *et al.*, 1998).

A naturally occurring example of such chain elongation plasticity is found in *Aeropyrum pernix*. This archaeon harbors only C25-C25 diether lipids, a rare feature among archaea, most of which have only C20-C20 lipids (De Rosa and Gambacorta, 1988). This hyperthermophile does not possess a GGPP synthase, but a single FGPP synthase (Tachibana *et al.*, 2000). The latter was most likely derived from an ancestral archaeal GGPP synthase, as suggested by a phylogenetic analysis performed by Tachibana *et al.* (Tachibana *et al.*, 2000). Our phylogeny of isoprenyl diphosphate synthases also supports this claim. In the best maximum likelihood tree, *A. pernix* FGPP synthase is found grouping among archaeal GGPP synthases (Figure 3.7C). Beside *A. pernix* FGPP synthase, *S. acidocaldarius* GGPP synthase and *M. thermoautotrophicus* GGPP synthase, which have been biochemically characterized in detail (Chen and

Poulter, 1993; Ohnuma *et al.*, 1996; Tachibana *et al.*, 2000), assessment of the chain length specificity of isoprenyl diphosphate synthases is based on the amino acid residue found at position 74 (amino acid positions of *S. acidocaldarius* GGPP synthase). GGPP synthases harbour an isoleucine or leucine residue at this position, as opposed to the alanine found in the *Aeropyrum* enzyme (Figure 3.7C). Bulkier residues at this position have been shown to "floor" the hydrophobic pocket that contains the isoprenoid chain during the elongation process, effectively limiting the length of the isoprenoid products (Ohnuma *et al.*, 1998). An exception to this is *Thermoplasma*, which, like *Aeropyrum*, harbors an alanine at position 74. However, the *Thermoplasma* enzyme is almost certainly a GGPP synthase, as this archaeon is known to only contain C20-C20 diether lipids (Shimada *et al.*, 2002). This suggests that other residues beside the one found at position 74 could limit the length of the isoprenoid chain produced.

Diversity of isoprenyl diphosphate synthases in archaea

Similarity searches for archaeal isoprenyl diphosphate synthases also recovered enzymes predicted to synthesize medium-chain isoprenoids (30 to 50 carbons). A characteristic distinguishing short and medium-chain isoprenyl diphosphate synthases is the residue found in position 77. Short-chain specific enzymes almost always display a small residue at this position (alanine, serine or valine) while enzymes producing medium-chain products usually harbor a bulkier residue (phenylalanine or tyrosine). This residue, like position 74, is thought to be oriented toward the channel in which the isoprenoid chain is elongated, limiting the length of the product (Ohnuma *et al.*, 1998). Several archaeal homologs divergent from the clade containing characterized short-chain isoprenyl diphosphate synthases all display a phenylalanine or tyrosine residue at position 77 (Figure 3.7C) and are therefore most likely medium-chain isoprenyl phosphate synthases, possibly involved in the biosynthesis of respiratory quinones (Kellogg and Poulter, 1997). These medium-chain enzymes were detected in only some of the archaea that have their genomes completely sequenced (Table 3.5). Consequently, the archaea in which this medium-chain isoprenyl diphosphate synthase cannot be detected must use a distantly related homologous prenyltransferase or a functionally analogous enzyme to synthesize their essential isoprenoids.

Biosynthesis of the glycerol phosphate backbone and its association with side-chains: sn-glycerol-1-phosphate dehydrogenase and geranylgeranylglyceryl phosphate synthase

Evolution of the stereoconfiguration of archaeal lipids

Two enzymes involved in the biosynthesis of archaeal lipids show stereospecificity. *sn*- glycerol-1-phosphate (G1P) dehydrogenase introduces stereospecificity into archaeal lipids by specifically synthesizing glycerol phosphate with the *sn*-1 stereoconfiguration from dihydroxyacetone phosphate (DHAP) (Nishihara and Koga, 1995). Geranylgeranylglyceryl phosphate (GGGP) synthase strongly favors this glycerol phosphate stereoisomer when attaching the first isoprenoid side-chain, yielding 3-*O*-geranylgeranyl-*sn*-glycero-1-phosphate. The attachment of the second side-chain is also stereospecific, as DGGGP synthase will only recognize the *sn*-1 monoether as opposed to an *sn*-3 monoether substrate, yielding only archaeol (2,3-*O*-geranylgeranyl-*sn*-glycero-1-phosphate) as a product (Zhang and Poulter, 1993). CDP-archaeol (the cytidylated version of archaeol) is thought to be a major precursor of phospholipids (Morii *et al.*, 2000). The enzyme responsible for its synthesis from CTP and archaeol, CDP-archaeol synthase, does not recognize the stereochemical structure of the glycerol phosphate backbone or the linkage between glycerol and the isoprenoid side chains (ester or ether linkage). It therefore seems that the specific stereoconfiguration of archaeal lipids is established by G1P dehydrogenase and the GGGP and DGGGP synthases.

Origin of *sn*-glycerol-1-phosphate dehydrogenase

G1P dehydrogenase and glycerol dehydrogenase catalyze similar reactions, the substrate and product being phosphorylated in one case and not in the other (dihydroxyacetone vs. dihydroxyacetone phosphate as substrate and glycerol vs. *sn*-glycerol-1-phosphate as product). These enzymes are also homologous, sharing about 20-25% identity in their amino acid sequences. They are both part of the NAD-dependent dehydrogenase superfamily, which also includes G3P dehydrogenases. Although the latter enzyme is functionally equivalent to G1P dehydrogenase, with the exception of its stereospecificity, they share little sequence similarity.

Our database survey confirms earlier claims, based on a more limited sampling, that G1P dehydrogenase is solely found in archaea (Nishihara *et al.*, 1999). Phylogenetic analysis also presents G1P dehydrogenases as a monophyletic cluster among the larger NAD-dependent dehydrogenase superfamily (Figure 3.7E). This suggests that G1P dehydrogenase is an archaeal invention derived from an enzyme of the NAD-dependent dehydrogenase superfamily, most likely glycerol dehydrogenase, which shares similar sequence, substrate and product. Interestingly, GD itself is absent from all archaea with the exception of Halobacteriales, which seem to have acquired the enzyme from Bacteria (Figure 3.7E).

Presence of two homologous types of GGGP synthases in archaea

The recently described sequence of the gene encoding *Methanobacterium thermoautotrophicum* GGGP synthase was the first obtained for this protein (Soderberg *et al.*, 2001). This enzyme is selective not only for the *sn*-glycerol-1-phosphate acceptor, but also for the isoprenoid side-chain added, strongly favoring GGPP over shorter or longer chains (Zhang and Poulter, 1993). No phylogenetic analysis had been performed for this gene prior to this study.

Our similarity searches reveal the presence of homologs of this enzyme in all archaea, even *A. pernix*, which, as mentioned earlier, is known to produce only C25-C25 diether lipids. Although *M. thermoautotrophicum* GGGP synthase shows very little activity with FGPP as a prenyl donor, the *A. pernix* homolog might be able to use this longer substrate, as it would have little need for a GGPP-specific enzyme (Soderberg *et al.*, 2001). Phylogenetic analysis shows the *A. pernix* homolog branching among other archaeal enzymes which are likely to display GGGP synthase activity, as the major fraction of the lipids of these archaea are C20-C20 diether lipids. This outlines that GGGP synthase homologs might harbor a functional plasticity similar to other prenyltransferases; a few mutations could be sufficient to alter the length of the prenyl donor used by the enzyme.

A surprising finding from phylogenetic analysis is the existence of two distinct but homologous types of GGGP synthase. *Halobacterium sp.* NRC-1 and *Archaeoglobus* *fulgidus* both possess very divergent enzymes that cluster with homologs from various species of the bacterial order Bacillales. This separate cluster is not only supported by a high bootstrap value (100%, see Figure 3.7D), but also by two insertions shared only by its members (data not shown). To investigate if this divergent enzyme is a specific feature of *Halobacterium* or a more general characteristic of the Halobacteriales, various genera of extremely halophilic archaea were surveyed for its presence. We obtained this divergent GGGP synthase from eight genera of Halobacteriales (Table 3.5). In a few species of haloarchaea, including *Halobacterium*, two paralogs of this divergent enzyme were present (Figure 3.7D). The role of these two paralogs is unclear, but it is possible that one of them encodes a farnesylgeranylglyceryl phosphate synthase, as some Halobacteriales (among others *Haloterrigena, Halococcus, Natronobacterium, Natrinema, Natrialba and Natronomonas*) produce C20-C25 diether lipids (Kamekura and Kates, 1999). However, some Halobacteriales that only have C20-C20 lipids (*Haloferax* and *Halobacterium*) also exhibit two paralogs.

Given its presence in several Halobacteriales genera, a divergent type of GGGP synthase is most likely an ancestral characteristic of this order. The situation could be similar for the Archaeoglobales, but a survey of other species in addition to *Archaeoglobus fulgidus* would be required to confirm this. The evolutionary origin of this divergent type of GGGP synthase is unclear. The distribution of GGGP synthase homologs, ubiquitous in archaea but only found in the order Bacillales and a *Cytophaga* among bacteria, suggest that this enzyme is an archaeal invention, which was later acquired by bacteria through LGT. The Halobacteriales, *Archaeoglobus* and bacillales homologs are closely related to each other and distinct from other archaeal homologs and the *Cytophaga* gene. The bacillales GGGP synthase homologs would therefore have originated from the Archaeoglobales or the Halobacteriales, while the *Cytophaga* homolog would descend from the enzyme of some other archaeal group. Since no ether lipids of the *sn*-2,3 stereoconfiguration are found in bacteria, the enzymes present in the bacillales and *Cytophaga* were probably co-opted for a different function. All bacillales for which sequence information is available harbour the GGGP synthase homolog, which is always encoded upstream of a DNA helicase and an NAD-dependent DNA ligase. The homolog found in *Staphylococcus aureus* S20 (termed *pcrB*) has been identified as part of a chromosomal operon that include the *pcrA* gene, which encodes a helicase required for cell viability and the replication of plasmid pT180 (Iordanescu, 1993). The ubiquity of this gene in Bacillales and its linkage with *pcrA* points to an important function, possibly in DNA replication.

Lateral transfer of GGGP synthase homologs is also likely to have happened among archaea, as *Archaeoglobus* and Halobacteriales are never found as each other closest relatives in phylogenetic analyses using known molecular markers (Matte-Tailliez *et al.*, 2002). Halobacteriales are usually most closely related to the orders Methanosarcinales and Methanomicrobiales (Matte-Tailliez *et al.*, 2002), which is not the case in the GGGP synthase phylogeny (Figure 3.7D). This implies that the divergent enzymes found in Halobacteriales and *Archaeoglobus* are not simply the result of an increase in the evolutionary rate of a particular archaeal clade, as then the GGGP synthase from *Methanosarcina* would be divergent as well. Most likely, the GGGP synthase gene of either Halobacteriales or *Archaeoglobus* diverged and was later transferred to the other archaeal group.

Conclusion

Like all the enzymes of the MVA pathway of bacteria/eukaryotes that have been laterally transferred to archaea (PMK and PPMD in Sulfolobales, PPMD in Halobacteriales and Thermoplasmatales), IDI1 is found in all representatives of the order by which it was acquired. This makes it likely that all LGT events so far identified to be responsible for the presence of isoprenoid lipid biosynthesis enzymes in particular orders of archaea have occurred prior to the diversification of these groups.

These ancestral transfer events are not the only thing that affected the evolution of isoprenoid lipid biosynthesis in archaea. Further LGT events within (IDI2) and between (GGGP synthase) archaeal orders must also have influenced the lipid metabolism of certain archaea. Also, the prenyltransferases involved in the elongation of the isoprenoid side chain (isoprenyl diphosphate synthases) and its association to the glycerol phosphate backbone (GGGP synthase) both seem to have experienced specificity switches during archaeal evolution, an ancestral archaeal GGPP synthase giving rise to FGPP synthase found in *Aeropyrum pernix* and an ancestral archaeal GGGP synthase possibly giving rise to an FGGP synthase in the same archaeon.

Most of the isoprenoid lipid biosynthesis enzymes seem to have been present before the divergence of the domain archaea. Two of them, however, are likely to be archaeal inventions: GGGP synthase and G1P dehydrogenase. Besides the GGGP synthase homologs found in the bacterial order Bacillales and a *Cytophaga* (likely to have been co-opted for a different function), these two enzymes are solely found in archaea and catalyze stereospecific reactions that have never been observed in the bacterial domain. The type of enzyme the GGGP synthases originated from is difficult to identify, as they do not share significant similarity with any other prenyltransferases. However, G1P dehydrogenases are clearly part of the NAD-dependent dehydrogenase superfamily, and most likely derived from glycerol dehydrogenase, with which they share sequence similarity and similar substrate and product.

Much has yet to be discovered concerning the genetics of isoprenoid lipid biosynthesis in archaea. Key enzymes like the archaeal analogs catalyzing the last two steps of the mevalonate pathway, those responsible for the hydrogenation of the isoprenoid side chains and the synthesis of cyclic tetraethers, are still uncharacterized. However, the information currently available about genes involved in biosynthesis of isoprenoid lipids tell us that this apparatus evolved through the co-option of ancestral enzymes for novel functions (GGGP synthase, G1P dehydrogenase), tinkering with specificity (GGPP/FGPP synthases), orthologous displacement (HMGR), evolution of archaeal specific analogs (archaeal analogs of PMK and PPMD), integration of components from eukaryotes and bacteria (PMK, PPMD and IDI1), rapid divergence (GGGP synthase) and LGT within (IDI2 in Halobacteriales) and between (GGGP synthase) archaeal orders.

Section III: Bacterial origin for the HMG-CoA reductase of the archaeal orders Archaeoglobales and the Thermoplasmatales

The MVA pathway gene encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR or HMG-CoA reductase, E.C.1.1.1.34) provides an ideal system with which to study LGT in detail, for several reasons. First, the existence of two homologous classes of the enzyme {eukaryotic/archaeal class 1 and bacterial class 2 (Bochar *et al.*, 1999)} makes it easy to detect events of LGT between bacteria and archaea and identify the donor and the recipient of the transfer. Indeed, identification of the class by sequence is unambiguous as the two classes share less than 20% amino acid identity and enzymes within a class are ~40% identical. One very clear-cut case of LGT was indeed identified from the complete genome sequence of *Archaeoglobus fulgidus*. This hyperthermophilic archaeon harbours a bacterial class 2 HMGR instead of the class 1 enzyme that is found in other archaea. In some bacteria, most eukaryotes and all archaea, HMGR is an essential enzyme, and thus cannot be simply lost but only displaced. Displacement is the acquisition of a functionally homologous gene followed by the loss of the ancestral gene, so an enzyme is always present to fulfil the specific function.

Not only is *A. fulgidus* HMGR of bacterial origin, but it also shows striking amino acid identity (61%) with the HMGR of the soil proteobacterium *Pseudomonas mevalonii.* These two homologous enzymes likely play very different metabolic roles in their respective cells. In all archaea (including *A. fulgidus*), HMGR catalyzes the first committed step of the mevalonate pathway (reductive deacylation of HMG-CoA to mevalonate). This central biosynthetic pathway leads to isopentenyl pyrophosphate, which is, with its isomer dimethylallyl pyrophosphate, the universal precursor for the synthesis of isoprenoids (Qureshi and Porter, 1981). Isoprenoids are quantitatively very important in archaea. The latter membrane lipids, which are a unique and characteristic taxonomic feature of this domain, are formed of glycerol (or more complex polyols) ether linked to isoprenoid alcohols (De Rosa *et al.*, 1986). In contrast, as mentioned in the previous section, the proteobacterium *P. mevalonii* does not use its HMGR in the biosynthesis of isoprenoids, but to biodegrade mevalonate by oxidative acylation. It can use this compound as sole source of carbon and energy (Scher and Rodwell, 1989). This is thus a case where an essential anabolic archaeal enzyme was displaced by a biodegradative bacterial homologue sharing little amino acid sequence identity (less than 20% between HMGR classes).

I decided to investigate this metabolically important case of inter-domain LGT more thoroughly. A survey of close relatives of *A. fulgidus* and of other archaea that live in similar environments was performed, searching for other events of LGT of the HMGR gene. Among the archaea surveyed were the Archaeoglobales (*Archaeoglobus veneficus*, *Archaeoglobus lithotrophicus*, *Archaeoglobus profundus* and *Ferroglobus placidus*) and the Thermococcales (*Thermococcus litoralis*, *Thermococcus stetteri* and *Thermococcus celer*). Although they occupy very different ecological niches, two species of acidophilic Thermoplasmatales (*Thermoplasma volcanium* and *Thermoplasma acidophilum*) were also inspected, to get representation of all major euryarchaeal groups. Some of the few bacteria known to possess biochemical activity requiring HMGR were also surveyed. Small ribosomal subunit (SSU) genes were sequenced from all species from which an

HMGR gene sequence had been obtained. The large number of SSU sequences available for archaea and its widespread use as a phylogenetic marker make this gene ideal to build a reference phylogeny. HMGR phylogenies were compared to SSU and other phylogenetic markers to identify cases of LGT and try to reconstruct the evolutionary history of the HMGR gene in archaea.

Materials and Methods

Genomic DNA Extraction

For archaea, total genomic DNA was extracted from frozen cell pellets following the protocol of Charbonnier *et al.* (1995). For bacteria, genomic DNA was isolated from fresh cell cultures as described in Wilson (1998). The archaeal strains used are listed in Table 3.6.

PCR primer design

The amplification of 600-950 bp of the HMGR gene from genomic DNA of hyperthermophilic archaea and bacteria was carried out in two steps: (1) amplification with class-specific degenerate primers designed with the help of an alignment of all available HMGR gene sequences (2) amplification with degenerate primers designed from the partial fragments previously amplified (5' end) and *A. fulgidus* complete HMGR gene sequence (3' end). This second round of amplification was only performed for archaea that displayed a class 2 HMGR to obtain more complete sequences for

Table 3.6 Provenance and description of known (2001) archaeal HMG-CoA reductase genes.

		Small ribosomal subunit		HMG-CoA reductase		
Organism	**Strain**	**Length (bp)**	**Accession[a]**	**Length (bp)**	**Accession[a]**	**Class**
Archaeoglobales						
Archaeoglobus fulgidus	VC-16	1492	AE000965	1311	028538	2
Archaeoglobus profundus	AV18	1500[b]	**AJ299219**	912[b]	**AJ299205**	2
Archaeoglobus lithotrophicus	TF2	1495[b]	**AJ299218**	597[b]	**AJ299203**	2
Archaeoglobus veneficus	SNP6	1488[b]	**Y10011**	885[b]	**AJ299204**	2
Ferroglobus placidus	AED II 12 DO	1492[b]	**AJ299217**	885[b]	**AJ299206**	2
Thermoplasmatales						
Thermoplasma acidophilum	DSM 1728	1471	M38637	786[b]	**AJ299207**	2
Thermoplasma volcanium	DSS1	1411[b]	**AJ299215**	831[b]	**AJ299208**	2
Other euryarchaeotes						
Pyrococcus horikoshii	OT3	1495	AP000001	1023	E71191	1
Pyrococcus abyssii	Orsay	1503	AJ248283	1233	G75150	1
Thermococcus celer	DSM2476	1486	M21529	443[b]	**AJ299209**	1
Thermococcus stetteri	K3	1463	Z75240	443[b]	**AJ299211**	1
Thermococcus litoralis	type strain	1368	Z70252	648[b]	**AJ299210**	1
Methanosarcina mazei	Gö1	1475	-	1263	-	1
Methanococcus jannaschii	DSM 2661	1475	U67473	1215	Q58116	1
M. thermoautotrophicum	delta H	1479	AE000930	1191	026662	1
Haloferax volcanii	WFD11	1472	K00421	1209	Q59468	1
Halobacterium sp.	NRC-1	1472	AE005128	1224	AAG20075	1
Crenarchaeotes						
Aeropyrum pernix	K1	1423	AP000062	1263	E72573	1
Sulfolobus solfataricus	P2	-	-	1227	008424	1
Pyrobaculum aerophilum	DSM 7523	2210	L07510	1203	-	1
Pyrodictium occultum	PL-19	1497	M21087	441 b	AJ299213	1

[a] Accession numbers in bold indicates sequences from this study

[b] These represent partial sequences

phylogenetic analysis (Archaeoglobales and Thermoplasmatales). The HMGR primer sequences can be found in Table 3.7.

Part of the small ribosomal RNA subunit (SSU) genes (1468-1500 bp) was amplified by PCR from all Archaeoglobales and Thermoplasmatales species. Different combinations of primers designed with the help of SSU sequence alignments from "The European Small Subunit Ribosomal RNA database" (Van de Peer *et al.*, 2000) were used. For the Archaeoglobales, the internally transcribed spacer (ITS) between the small and large ribosomal subunits genes (SSU and LSU) and the 5' end of LSU were also amplified by PCR. The sequences of the primers used to amplify the SSU genes and the ITS can be found in Table 3.8.

DNA amplification, cloning and sequencing

Each gene was amplified from two independent PCR reactions. Amplifications were carried out in a final volume of 25μl containing 1-5 ng of template DNA, 1X PCR buffer, 2.5 mM MgCl2, 0.2 mM dNTPs, 1.0 mM of each primer, and 0.5-1 U of Platinum Taq High Fidelity DNA polymerase (INVITROGEN). The reactions were performed with an initial denaturation at 95°C for 1 min., 30 cycles with a denaturation at 95°C for 30 sec., primer annealing at 48-52°C for 30 sec., and primer extension at 72°C for 1 min (2 min for SSU genes). PCR products were gel purified with the MinElute kit (QIAGEN) and cloned in TOPO TA (INVITROGEN). The vector was then transformed into chemically competent TOP10 *E. coli* cells (INVITROGEN) and plated on LB-kanamycin agar plates containing X-Gal for blue-white screening (INVITROGEN, TOPO-TA cloning kit, manufacturers recommendations). White colonies were then picked and used

Table 3.7 Primers used for the amplification of the HMG-CoA reductase gene.

Primer	**Sequence (5' to 3')**
HMG-CoA reductase (HMGR)	
Universal class 1 forward F2	CCATTGGCTACNACNGARGG
Universal class 1 forward F3	GACGCCATGGGNATGAAYATG
Universal class 1 reverse R2	GTTCCACCGCCNACNGTNCC
Universal class 1 reverse R3	GCTGCTAGCGANAGYTCNCC
Universal class 2 forward F4	TTAGCTACCGARGARCCNTC
Universal class 2 forward F5	CATGATGCCATGGGNGCNAA
Universal class 2 reverse R4	CCGTTCATGATNCCYTTRTT
Universal class 2 reverse R5	CATGTGGCCTCTYTGDATNCC
Archaeoglobales class 2 forward LF1	GAGAACGTCATHGGRACYTT
Archaeoglobales class 2 forward LF5	ATGATAGGVCAGATWCAGGT
Thermoplasmatales class 2 forward LF4	AACAGCATGTGYGAATACGT
Archaeoglobales/Thermoplasmatales class 2 reverse LR1	CATGTGCCCYCTYTGDATNCC
Archaeoglobales/Thermoplasmatales class 2 reverse LR2	TAAAGCTGCRAARTTYTGNGC
Giardia lamblia forward F1	ATCCAGGCTCTTGATACAATG
Giardia lamblia reverse R1	GTGCTCCATCTCCAGGCTCTT

Table 3.8 Primers for the amplification of SSU and ITS1 in Archaeoglobales and Thermoplasmatales.

Primer	Sequence (5' to 3')
Small ribosomal subunit (SSU)	
Bacteria forward 27F	AGAGTTTGATCMTGGCTCAG
Bacteria reverse RTU8	AAGGAGGTGATCCANCCRCA
Archaeoglobales/Thermoplasmatales forward AF4	CTGGTTGATCCTGCCAG
Archaeoglobales/Thermoplasmatales forward AF42	GACTAAGCCATGCRAGTCA
Archaeoglobales/Thermoplasmatales forward AF99	ACGGCTCAGTAACACGTGGA
Archaeoglobales/Thermoplasmatales reverse AR1470	TCCCCTACGGMTACCTTGTTA
Archaeoglobales/Thermoplasmatales reverse AR1491	GGAGG TGATCCAKCCGCAG
Archaeoglobales/Thermoplasmatales reverse AR1492	AAGGAGG TGATCCAGCC
Internally transcribed spacer (SSU-**ITS1**-LSU)	
Archaeoglobales forward F1	CTGCAACTCGCCCTCGTGAAC
Archaeoglobales forward F2	GAGGGCTGYAACTCGCCCTC
Archaeoglobales reverse R1	ACGGTCTAAACCCAGCTCACG
Archaeoglobales reverse R2	CTTGGCACGYCCTTCGTCGGC

to inoculate 25μl of PCR mixture for confirmation of the insert size. This mixture contained 1X PCR buffer, 2.5 mM MgCl2, 0.2 mM dNTPs, 1.0 mM of each primer (M13 forward and M13 reverse), and 0.5-1 U of Taq DNA polymerase (INVITROGEN). For each PCR product, two positive clones (vectors with inserts of the right size) were sequenced from both strands using a LiCor 4000L automated sequencer.

Southern transfers and hybridization

To confirm their origin, the HMGR gene fragments sequenced were digoxigenin (DIG)-dUTP labeled and used as probes for Southern hybridization on the genomic DNA from which they had been amplified. DNA probe labeling was done by random primed labeling (DIG DNA labeling kit, Roche Molecular Biochemical). Genomic DNA was transferred by capillary transfer to a positively charged nylon membrane (Brown, 1993) to which the labeled probe was then hybridized (DIG application manual, Roche Molecular Biochemical). Detection of the probe was done using anti-DIG antibody conjugated to alkaline phosphatase followed by reaction of an ultra-sensitive chemiluminescent substrate with the alkaline phosphatase (CDP-*Star*, Roche Molecular Biochemical). The chemiluminescent signal was captured using BioMax single emulsion films (Kodak).

DNA sequence analysis and assembly

Traces obtained from automated DNA sequencing were visualized using Sequencher 4.1.2 (Gene Codes Corporation) and base calls were corrected manually. This program also allowed the assembly of several sequencing reads from multiple clones, making possible the correction of DNA polymerase errors, whether they occurred at the amplification or the sequencing stage.

Gene alignment construction

The gene alignments were constructed and edited as describe in Chapter 2 with the following modifications. Regions corresponding to the PCR primers in novel sequences were removed from the alignments during the manual editing in MacClade (Maddison and Maddison, 1989). 5' and 3' regions absent from the novel sequences (that are only partial) were kept in the alignment. DNA sequences of protein coding genes were obtained from Genbank along with amino acid sequences and were similarly aligned and edited.

Phylogenetic analysis

For phylogenetic analysis of HMGR, sequences were retrieved from the NCBI web site (http://www.ncbi.org). Preliminary sequence data from the unfinished genomes of *Staphylococcus aureus*, *Enterococcus faecalis*, *Streptococcus pyogenes* and *Streptococcus pneumoniae* was obtained from The Institute for Genomic Research web site (http://www.tigr.org). *Methanosarcina mazei* (http://www.g2l.bio.uni-goettingen.de/), *Pyrobaculum aerophilum* (http://www.tree.caltech.edu/bes.html) and *Halobacterium sp*. NRC-1 (http://zdna.micro.umass.edu/haloweb/) HMGR sequences were provided by their respective genome sequencing projects. The retrieved amino acid sequences and the new sequences from this study were aligned using CLUSTALW (Thompson *et al.*, 1994). The alignment was subsequently edited manually to remove gaps and ambiguous characters. The number of sites used in the HMGR alignments were as follows: 203 for class 1 and class 2, 253 for Archaeoglobales and Thermoplasmatales, 335 for archaeal class 1.

To build a reference archaeal phylogeny and look at the relationship between Thermoplasmatales and Archaeoglobales, different phylogenetic markers were chosen. This choice was based on the availability of sequences from most archaeal groups. The selected molecules were as follows: small and large ribosomal subunit (SSU and LSU), archaeal release factor (aRF1), elongation factor 1 and 2 (ef1/EF-Tu and ef2/EF-G) and RNA polymerase subunit β and β' (rpoB and rpoC). Markers that are protein-coding genes were aligned at the amino acid level, all available archaeal sequences were included and the trees were rooted with crenarchaeotes. For SSU, an alignment of representative archaeal sequences was obtained from "The European Small Subunit Ribosomal RNA database" (Van de Peer *et al.*, 2000). New sequences were added to this alignment and local adjustments were performed manually. The alignment was also edited to exclude gaps and ambiguous characters. A total of 1124 sites were used in the archaeal SSU alignment, which included sequences from 66 taxa, covering all euryarchaeal groups and including representative crenarchaeotes. A subset alignment containing only representatives of the Archaeoglobales and Thermoplasmatales was also assembled (1372 sites used). The LSU alignment was obtained (De Rijk *et al.*, 1995) and edited (2218 sites used) by the same procedure. For comparison with the archaeal class 1 HMGR phylogeny, for which only 9 taxa were available, the SSU and LSU alignments were reduced to include only the same 9 taxa, plus *A. fulgidus* and *T. acidophilum*.

To evaluate the influence of taxon sampling on the SSU phylogeny, a random sampling analysis was performed. In this random taxon sampling analysis, the SSU alignment was broken down in 10 different groups composed of 4 to 8 taxa each, based on recognized phylogenetic groups: Thermoplasmatales, Archaeoglobales, Thermococcales, Methanosarcinales, Methanomicrobiales, Methanococcales, Methanobacteriales, Halobacteriales, Methanopyrales (only one sequence in this group, which is from *Methanopyrus kandleri*) and Crenarchaeota (8 representative crenarchaeotes were chosen). A program based on a strategy used by Van de Peer *et al.* (1994) was written to randomly choose one taxon from each of these groups and then produce a Neighbor-joining (NJ) Jukes-Cantor distance tree (10 taxa, rooted with the crenarchaeal taxon). The procedure was repeated 1000 times, giving a set of SSU trees, each including one randomly chosen taxon from each archaeal group. A consensus tree was built based on the frequency of occurrence of different groups of taxa in the 1000 SSU trees (Van de Peer *et al.*, 1994).

DNA sequences were used for phylogenetic analysis of the genes coding for HMGR, SSU and LSU. All DNA analyses of the HMGR gene included only the first and second codon positions, as a large GC composition bias could be detected at the third position for some taxa (Chi-square = 35.166077, df = 12 and P-value < 0.0005). DNA analyses were performed with PAUP* 4.04b (Swofford, 1998) applying the heuristic-search option and using the TBR branch-swapping algorithm. Logdet distance trees were reconstructed taking into account the proportion of invariable sites. Maximum likelihood analyses were performed under the TIM+I+G substitution model as selected using MODELTEST (Posada and Crandall, 1998). Parsimony analyses were also performed with PAUP*, with ACCTRAN character-trait optimization selected. Quartet-puzzling maximum likelihood and bootstrapped quartet-puzzling maximum likelihood distance trees were reconstructed with TREE-PUZZLE and PUZZLEBOOT, respectively. TREE-PUZZLE settings were as follows: HKY85 nucleotide substitution model, rate heterogeneity model with gamma distributed rates over eight categories plus one invariable category and the α parameter and amino acid frequencies estimated from the data. For the large euryarchaeal SSU dataset (66 taxa), each archaeal group mentioned above was constrained as one moveable group for phylogenetic tree-reconstruction. Confidence of nodes in DNA analyses was estimated by 1000 bootstrap replicates (PAUP* and PUZZLEBOOT) or 1000 quartet-puzzling steps (TREE-PUZZLE).

Phylogenetic analyses were performed on amino acid sequences for HMGR, ef1, ef2, rpoB, rpoC and aRF1. For distance analyses, Fitch-Margoliash trees of PAM-based distances were made using PROTDIST and FITCH from the PHYLIP package ver.3.572 (Felsenstein, 1993). Protein parsimony analyses were performed using PAUP*, again with ACCTRAN character-trait optimization selected. Maximum likelihood analyses were carried using PROTML (quick-add search with 2000 replicates, JTT-F amino acid substitution model) from the MOLPHY package ver.2.3 (Adachi, 1996) and TREE-PUZZLE 4.0 (JTT-F amino acid substitution model, rate heterogeneity model with gamma distributed rates over eight categories plus one invariable category and the α parameter and amino acid frequencies estimated from the data). PUZZLEBOOT was used with the same settings as TREE-PUZZLE to bootstrap quartet-puzzling maximum likelihood distance trees. Confidence of nodes in amino acid analyses was estimated by 1000 bootstrap replicates (PROTDIST and PUZZLEBOOT), 1000 quartet-puzzling steps

(TREE-PUZZLE) or RELL values (PROTML). The bootstrap replicates of PROTDIST were generated using SEQBOOT and compiled in a consensus tree with CONSENSE.

When required, different tree topologies were compared using the Kishino-Hasegawa (KH) test, which was performed in TREE-PUZZLE. The trees used for input were obtained using PROTML (quick-add search with 2000 replicates, JTT-F amino acid substitution model) for amino acid sequences and by a parsimony heuristic search for the most parsimonious trees in PAUP* (ACCTRAN character-trait optimization selected) for DNA sequences. When investigating the possibility of the Thermoplasmatales and Archaeoglobales being sister taxa, the 100 most likely or most parsimonious trees obtained when these taxa are constrained to form a clade were compared with the 100 trees obtained without this constraint.

Results

Classification of sequenced archaeal HMGR genes

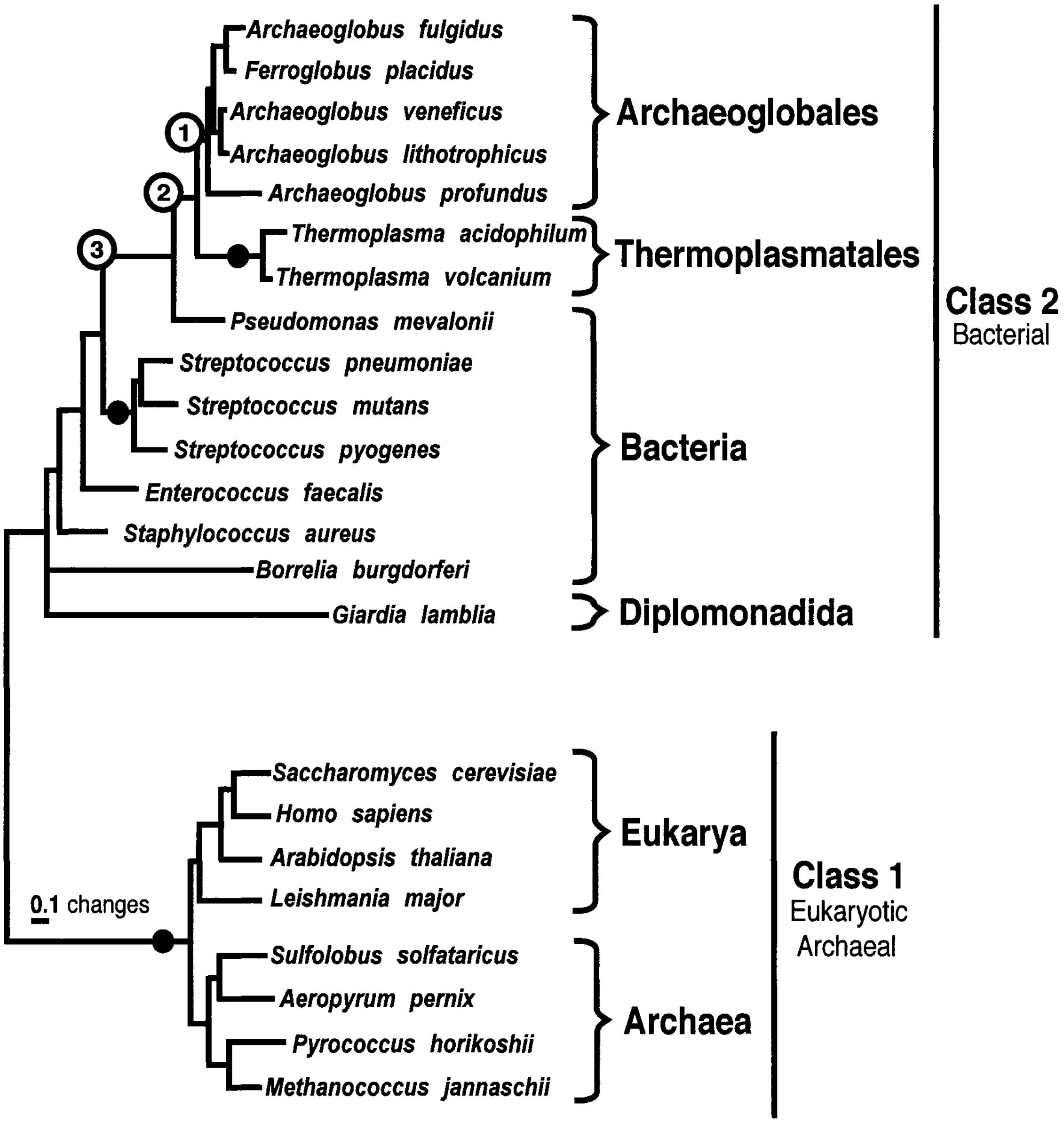
Of the different archaea sampled, those sharing an environment with *A. fulgidus* without being its phylogenetic neighbors (*Thermococcus celer*, *Thermococcus stetteri* and *Thermococcus litoralis* and *Pyrodictium occultum)* were found to have the standard archaeal class 1 HMGR (Table 3.6). All close relatives of *A. fulgidus* (*A. profundus*, *A. lithotrophicus*, *A.veneficus* and *F. placidus*) and the two Thermoplasmatales sampled (*T. volcanium* and *T. acidophilum*) harboured an unusual class 2 bacterial HMGR (Table 3.6).

Pairwise sequence comparison with SIM (Huang and Miller, 1991) revealed an unusually high similarity between the HMGRs of the Thermoplasmatales, Archaeoglobales and *P. mevalonii.* The amino acid sequence identity between HMGRs from different species of the three lineages ranges from 53 to 66%, with an average of 56.5% (sd 4.3). Comparatively, the average amino acid identity is 41.1% (sd 9.1) among class 2 HMGRs and 18.9% (sd 2.3) between classes 1 and 2 (only one representative of a given genus included in the calculations in both cases). The amino acid identity between different members of a genus for class 2 HMGRs averaged 70.8% (sd 1.6) for *Streptococcus*, 79.1% (sd 9.7) for *Archaeoglobus* and 80.3% for *Thermoplasma.*

Phylogenetic analysis of Thermoplasmatales and Archaeoglobales HMGRs

Phylogenetic trees based on amino acid sequences of all class 2 HMGR rooted with representative class 1 enzymes were constructed with a variety of methods. The best maximum likelihood distance tree is shown in Figure 3.9. As expected, class 1 and class 2 genes formed well-separated clades. However, Archaeoglobales and Thermoplasmatales clustered with bacteria rather than other archaea in the HMGR tree. In fact, the HMGRs from these two archaeal groups are much more similar to the *P. mevalonii* enzyme than any other bacterial homologues. The clustering of these three genes was observed in all phylogenetic reconstructions. *P. mevalonii* also occupied a basal position relative to the Archaeoglobales and Thermoplasmatales in all analyses in which only class 2 sequences were included (not shown). When the HMGR tree was rooted with class 1 enzymes, *P. mevalonii* still grouped with Archaeoglobales and Thermoplasmatales but the branching order was unresolved in distance and quartet-puzzling maximum-likelihood analyses (Figure 3.9).

Figure 3.9 Best maximum likelihood distance tree for the amino acid HMGR dataset as determined by the Kishino-Hasegawa test. The tree is rooted by representative class 1 HMGRs from archaea and eukaryotes. Support values (%) of critical nodes are displayed in boxes: (PROT) amino acid analysis, (R) tree rooted with class 1 HMGRs, (U) unrooted tree of class 2 HMGRs, (DT) Logdet distances bootstrap, (MP) Maximum parsimony bootstrap, (ML) Maximum likelihood RELL values, (PZ) Quartet-puzzling maximum likelihood support values (PB) Quartet-puzzling maximum likelihood distances bootstrap. (--) The node is unresolved or not recovered. (**) The node is supported with a 100% confidence level. (●) The node is recovered by all tree-reconstruction methods with a confidence of 90% or more.



①

	DT	MP	ML	PZ	PB
PROT(U)	98	89	85	64	62
PROT(R)	74	91	94	--	71

②

	DT	MP	ML	PZ	PB
PROT(U)	71	58	97	89	89
PROT(R)	--	67	97	--	73

③

	DT	MP	ML	PZ	PB
PROT(R)	89	97	**	85	98
PROT(U)	**	93	**	96	**

Figure 3.9 Best maximum likelihood distance tree for the amino acid HMGR dataset as determined by the Kishino-Hasegawa test.

The Kishino-Hasegawa test was used to compare all 124 maximum likelihood trees obtained by PROTML (quick-add search with 2000 replicates). The best tree showed the topology presented in Figure 3.9. However, of the 124 trees compared, 10 did not show *P. mevalonii* as basal to the two archaeal clades and were not significantly worse than the best tree. These alternative trees either placed *P. mevalonii* between the Thermoplasmatales and the Archaeoglobales clades (5 trees) or with either one of them (in 3 trees with the Thermoplasmatales and in 2 trees with the Archaeoglobales). The Thermoplasmatales clade was present in all trees and the Archaeoglobales were monophyletic in all but 8 of the 124 best trees. In all of these 8 cases, it is *A. profundus*, a notably long branch among Archaeoglobales, that is probably attracted to the Thermoplasmatales or *P. mevalonii*.

Are Thermoplasmatales and Archaeoglobales sister taxa?

No previous phylogenetic studies have shown Thermoplasmatales and Archaeoglobales to be sister taxa. To test the possibility, however, we examined trees constructed with different phylogenetic markers (Table 3.9). For each marker, the Kishino-Hasegawa test was used to compare the 100 most likely trees obtained by a PROTML quick-add search with another 100 trees obtained by a similar search with Thermoplasmatales and Archaeoglobales constrained to be sister taxa. None of the markers gave Thermoplasmatales and Archaeoglobales as sister taxa in its best tree. However, several trees that placed these groups as monophyletic were not significantly worse than the best tree for most markers (Table 3.9). Thermoplasmatales and

Table 3.9 Comparison of the most likely or parsimonious archaeal trees of different phylogenetic markers by the Kishino-Hasegawa (KH) test.

Gene	No. of sites used[a]	No. of taxa	KH best tree (log likelihood)	Trees not signicficantly worse than best tree with TP and AG as sister taxa (log likelihood)[b]	
ef1	379aa	11	-4663.07	3	(-4676.49 to -4683.40)
ef2	684aa	16	-11954.78	4	(-11971.65 to -11997.06)
rpoB	976aa	12	-13708.58	12	(-13732.23 to -13750.49)
rpoC	1080aa	12	-15351.32	34	(-15360.47 to -15396.61)
aRF1	335aa	11	-5396.24	21	(-5404.18 to -5425.82)
16S	1124nu	66	-13514.43	61	(-13519.24 to -13538.03)
23S	2218nu	14	-15218.4	1	(-15237.79)

[a] aa = amino acids nu = nucleotides

[b] TP = Thermoplasmatales, AG = Archaeoglobales

Archaeoglobales were found as adjacent but distinct branches in the best trees of LSU and aRF1. No other marker has indicated these clades to be adjacent, even in their best tree. With the large taxon sampling of SSU and well-defined clades for archaea, it was possible to perform a random taxon sampling analysis, in which single taxa are randomly selected from each of several pre-defined clades and trees constructed and results averaged over 1000 repetitions {described in (Van de Peer *et al.*, 1994)}. None of the 1000 trees showed the Thermoplasmatales and the Archaeoglobales as sister taxa. The only nodes recovered in a significant number of trees were the one grouping the Methanosarcinales and the Methanomicrobiales (1000 trees), the node grouping these two clades with Halobacteriales (969 trees), and the position of the Methanopyrales as the deepest euryarchaeal branch (878 trees) (Figure 3.10A).

An interesting characteristic was found in one of the phylogenetic markers used, RNA polymerase subunit β, encoded by *rpoB*. The gene is split into two smaller genes encoding independent subunits (B' and B") in all euryarchaeal taxa from which it has been sequenced so far except Thermococcales and Thermoplasmatales (Schleper *et al.*, 1995). Crenarchaeal *rpoB* genes are not split. This gene is thus split only in Archaeoglobales and not Thermoplasmatales. This split can occur without altering the RNA polymerase function, as it as been introduced experimentally in *E. coli rpoB* without removing the activity of the encoded polymerase (Severinov *et al.*, 1996).

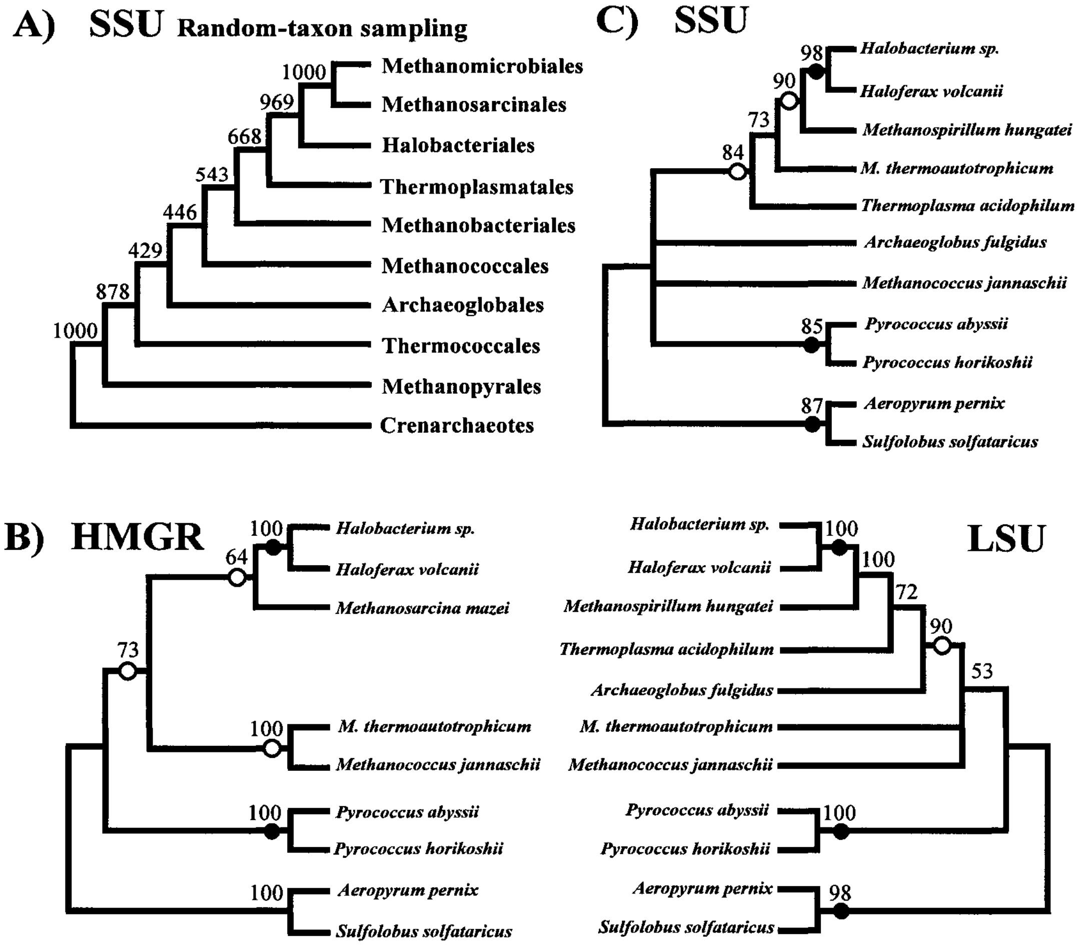


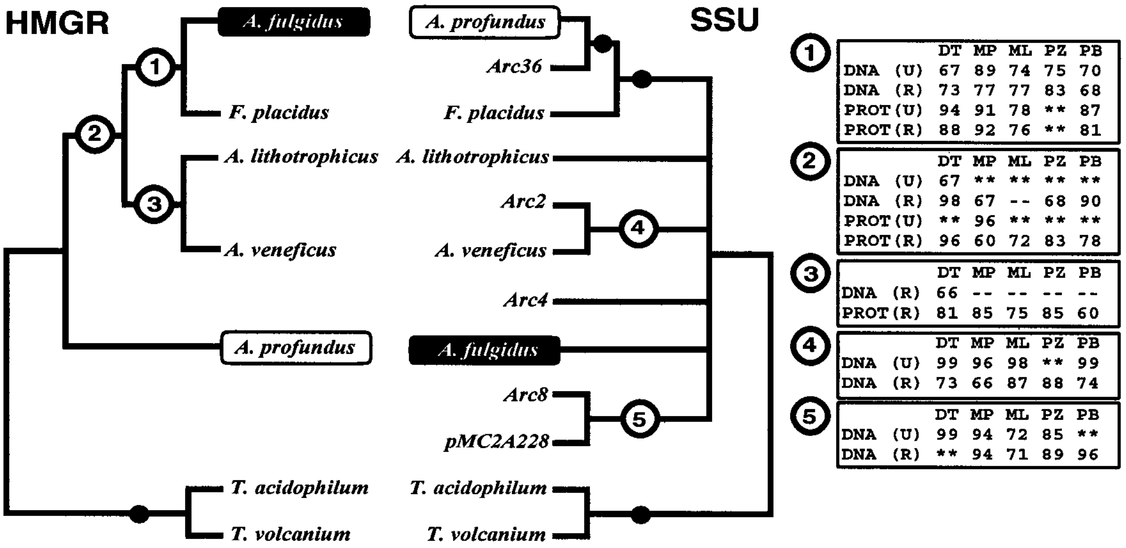
Figure 3.10 Phylogeny of the euryarchaeotes according to different genes. HMGR, SSU and LSU trees are the best maximum likelihood trees obtained in TREE-PUZZLE. (○) The node is recovered by all tree-reconstruction methods used (distance, maximum parsimony, maximum likelihood, quartet-puzzling maximum likelihood and quartet-puzzling maximum likelihood distances). (●) The node is recovered by all tree-reconstruction methods used with a confidence of 90% or more. A) Representation of the frequency of occurrence of different groupings in the random taxon-sampling analysis on a total of 1000 trees. B) Comparison of the LSU and class 1 HMGR phylogenies. C) SSU phylogeny.

Phylogenetic relationships among Archaeoglobales

The phylogeny of the five different cultured Archaeoglobales species was determined using the SSU gene, also including sequences from six uncultured taxa (Figure 3.11). Both unrooted trees and trees rooted with the Thermoplasmatales gave a highly supported clustering of *A. profundus* and *F. placidus* with all tree-reconstruction methods. *A. lithotrophicus* and *A. veneficus* were found to cluster together only with some tree-building methods and even then, only weakly. When only the five cultured Archaeoglobales species are included, regardless of the tree-reconstruction method used or whether the tree is rooted with Thermoplasmatales or not, the clustering of *A. profundus* and *F. placidus* is maintained and *A. lithotrophicus* and *A. veneficus* cluster together.

For the HMGR gene, rooted and unrooted phylogenies on amino acid datasets gave the same topology with all tree-reconstruction methods (Figure 3.11). DNA phylogenies, when unrooted, also gave the topology presented in Figure 3.11. However, when DNA phylogenies were rooted with Thermoplasmatales, the *A. veneficus*/*A. lithotrophicus* clade was unresolved with maximum likelihood and the node placing *A. profundus* as a basal branch was completely unresolved with most tree-reconstruction methods.

When the HMGR phylogeny was compared to that obtained for the SSU gene, they matched quite well, except for *A. fulgidus* and *A. profundus*, which switched position (Figure 3.11). The Kishino-Hasegawa test was used to verify if this topology of the Archaeoglobales HMGR tree was significantly better than other topologies that would better agree with the SSU tree. This comparison of the 15 possible trees for a dataset of



①

	DT	MP	ML	PZ	PB
DNA (U)	67	89	74	75	70
DNA (R)	73	77	77	83	68
PROT(U)	94	91	78	**	87
PROT(R)	88	92	76	**	81

②

	DT	MP	ML	PZ	PB
DNA (U)	67	**	**	**	**
DNA (R)	98	67	--	68	90
PROT(U)	**	96	**	**	**
PROT(R)	96	60	72	83	78

③

	DT	MP	ML	PZ	PB
DNA (R)	66	--	--	--	--
PROT(R)	81	85	75	85	60

④

	DT	MP	ML	PZ	PB
DNA (U)	99	96	98	**	99
DNA (R)	73	66	87	88	74

⑤

	DT	MP	ML	PZ	PB
DNA (U)	99	94	72	85	**
DNA (R)	**	94	71	89	96

Figure 3.11 Comparison of the HMGR and SSU phylogenies for the members of the Archaeoglobales. Both trees are maximum likelihood trees obtained in TREE-PUZZLE and are rooted with the Thermoplasmatales (distances are not represented in branch lengths). SSU sequences for uncultured members of the Archaeoglobales were included in the phylogeny: pMC2A228 is from Takai and Horikoshii (1999), all other sequences are from Reysenbach *et al.* (1999). Support values (%) of critical nodes are displayed in boxes: (PROT) amino acid analysis, (DNA) DNA analysis, (R) tree rooted with class 1 HMGRs, (U) unrooted tree of class 2 HMGRs, (DT) Logdet distances bootstrap, (MP) Maximum parsimony bootstrap, (ML) Maximum likelihood RELL values, (PZ) Quartet-puzzling maximum likelihood support values (PB) Quartet-puzzling maximum likelihood distances bootstrap. (--) The node is unresolved or not recovered. (**) The node is supported with a 100% confidence level. (●) The node is recovered by all tree-reconstruction methods with a confidence of 90% or more.

five Archaeoglobales HMGRs amino acid sequences confirmed the topology presented in Figure 3.11 as the best tree. However, this topology was not significantly better than the alternative topology matching the SSU tree. Also, *A. profundus* seemed to be a long branch, three times as long as the second longest branch (0.35462 Vs 0.12993 branch length) in an unrooted amino acid maximum likelihood phylogeny of HMGR.

Comparison of phylogenies of the HMGR gene and the SSU and LSU phylogenetic markers

To look for differences in their evolutionary histories, a phylogenetic tree of class 1 HMGRs was compared with trees of the commonly used phylogenetic markers SSU and LSU. Only taxa for which all three genes were available were included. Since no LSU sequence was available for *Methanosarcina mazei*, we substituted a sequence from *Methanospirillum hungatei*, another representative of the large methanogen cluster that include *M. mazei*. LSU and SSU tree topologies were found to be slightly different from each other (Figure 3.10B and Figure3.10C). For the latter, however, the branching order of the groups varied substantially depending on the reconstruction method used, compared to a generally more stable topology with LSU (data not shown). Comparative analyses have shown that LSU seems less prone to variation due to changes in method and dataset and is generally a more reliable phylogenetic marker (De Rijk *et al.*, 1995). LSU was thus chosen over SSU as a reference to which the archaeal HMGR phylogeny is to be compared. Other than the fact that the HMGR tree is missing *A. fulgidus* and *T. acidophilum* (class 2 HMGRs were excluded from the analyses), its topology was found to be very similar to the LSU tree (Figure 3.10B).

Discussion

The common ancestor of archaea and eukaryotes almost certainly had a class 1 HMGR, since it is found in all but two archaeal groups (Thermoplasmatales and Archaeoglobales) as well as in all eukaryotes known to use the mevalonate pathway except for the diplomonad *Giardia lamblia* (which has a bacterial-like HMGR) (see previous sections). Knowing that class 1 HMGR is ancestral to archaea, it would be interesting to know how the Thermoplasmatales and Archaeoglobales acquired a bacterial-like class 2 enzyme. Two general evolutionary scenarios are discussed here.

Acquisition of the bacterial-like HMGR in an ancestral euryarchaeote

In this LGT scenario, the euryarchaeote ancestral to the Thermoplasmatales, the Archaeoglobales and all later-branching lineages, acquired a bacterial HMGR in addition to its endogenous class 1 enzyme. This acquisition would have been followed by differential losses in subsequently diverging lineages. The Thermoplasmatales and Archaeoglobales would have kept the class 2 enzyme and lost the class 1 homologue. All other lineages descending from the original recipient of the transfer would have kept only the class 1 enzyme, or lost it and re-acquired it later (Figures 3.12A and 3.12B).

Under a differential loss scenario, since Thermoplasmatales and Archaeoglobales occupy distinct intermediate branches in the rooted tree of the euryarchaeotes, several events would be needed to account for the presence of a different class of HMGR in these two lineages and subsequent ones (Figure 3.12A). Also, for differential loss to be possible, there must have been two HMGRs encoded in the genome of the ancestral euryarchaeote. This situation would have had to be maintained long enough for at least two lineages to diverge (Thermoplasmatales and Archaeoglobales). Since there is no known case of the presence of both classes of HMGR in any single genome, it is unlikely that the two enzymes would have persisted together on such a large time scale.

Displacement of the ancestral enzyme by a class 2 homologue followed by a re-acquisition of a class 1 HMGR after divergence of the Archaeoglobales and Thermoplasmatales also seems unlikely (Figure 3.12B). The LSU phylogeny of euryarchaeotes (assumed here to model the organismal phylogeny) matches the class 1 HMGR phylogeny very well (Figure 3.10B). This is incompatible with the displacement/re-acquisition scenario postulated above. The most likely donor for a re-acquisition of a class 1 HMGR being another archaeon, the recipient would show an affinity for other archaeal lineages, thus leading to a topology fairly different from the one obtained with LSU.

If Archaeoglobales and Thermoplasmatales are neither sister taxa nor adjacent branches in the euryarchaeal tree, the possibilities mentioned above become even less appealing. Although the possibility that Archaeoglobales and Thermoplasmatales are sister taxa cannot be formally excluded, the evidence presented here weights heavily against it. None of the best maximum likelihood trees for the seven different phylogenetic markers used here places these two archaeal orders as sister taxa. Furthermore, in a random taxon sampling analysis performed on the SSU gene including representatives of all archaeal orders, none of the 1000 trees obtained placed Archaeoglobales and Thermoplasmatales as sister taxa. Also supporting the results obtained with the phylogenetic markers is the fact that the rpoB gene is split in two smaller genes in Archaeoglobales and not in Thermoplasmatales. This situation would

Figure 3.12 Possible scenarios for the acquisition of a bacterial-like HMGR by the Thermoplasmatales and the Archaeoglobales. The trees correspond to the class 1 HMGR tree with the addition of the Thermoplasmatales and Archaeoglobales at the position were they are found in the LSU phylogeny of euryarchaeotes. Representations of the most parsimonious sequence of events for each scenario are mapped on the trees. A black line represents the presence of a class 1 HMGR and a gray line the presence of a class 2 HMGR. A) Acquisition of a class 2 HMGR by an ancestral euryarchaeote and differential loss in subsequent lineages. B) Displacement of the class 1 HMGR by a class 2 enzyme in an ancestral euryarchaeote followed by a re-acquisition of the class 1 enzyme in the ancestor(s) of some of the subsequent lineages. C) Displacement of the ancestral class 1 HMGR by a class 2 enzyme in either Thermoplasmatales or Archaeoglobales followed by a transfer to the other lineage. D) Independent displacements of the ancestral class 1 HMGR by a class 2 enzyme in Thermoplasmatales and Archaeoglobales.

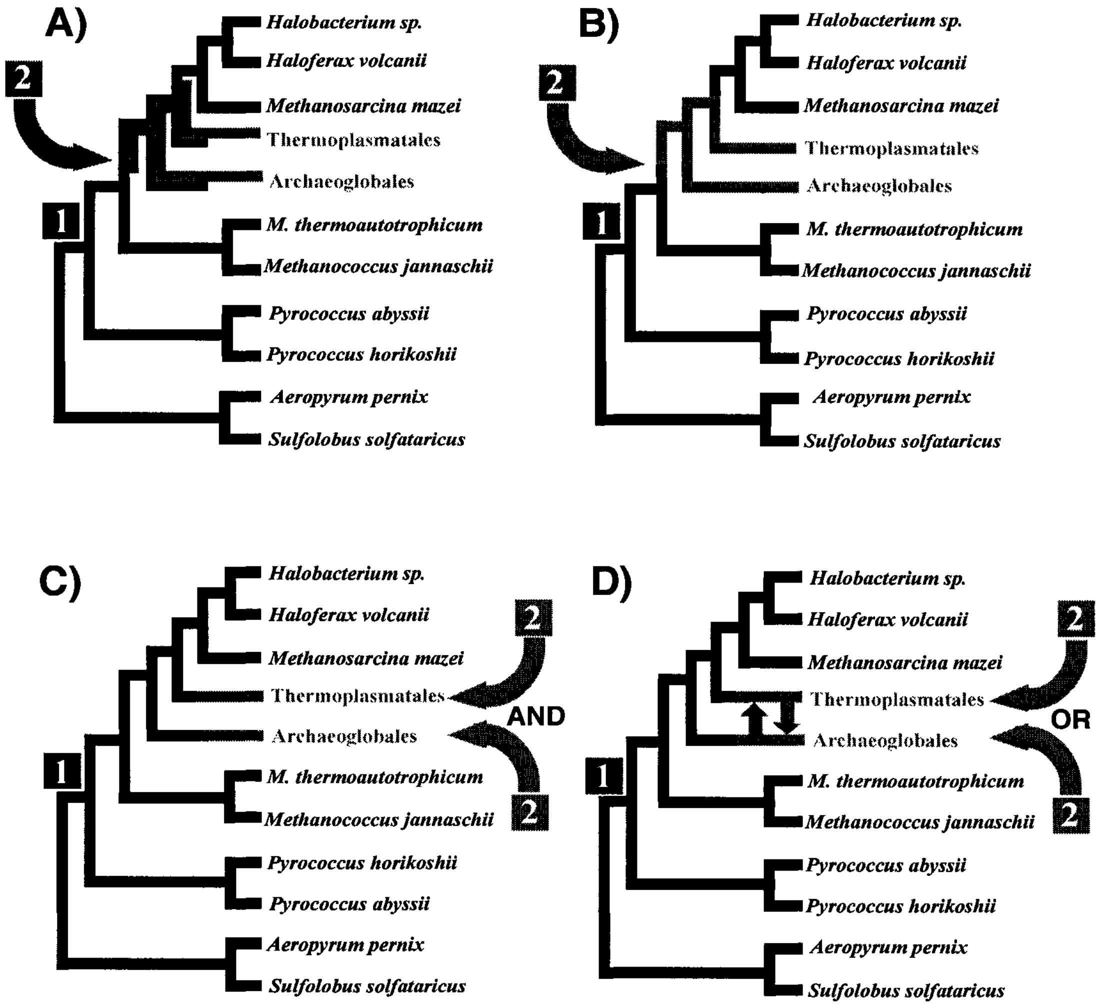


Figure 3.12 Possible scenarios for the acquisition of a bacterial-like HMGR by the Thermoplasmatales and the Archaeoglobales.

require an unlikely evolutionary event to have occurred twice if these two archaeal orders are to be sister taxa. It is however likely that Archaeoglobales and Thermoplasmatales comprise consecutive branches within the euryarchaeal tree. Two of the markers, LSU and aRF1, did indeed place *T. acidophilum* and *A. fulgidus* as adjacent but distinct branches in their trees, regardless of the reconstruction method used.

Overall, if the bacterial-like class 2 HMGR displayed by the Thermoplasmatales and the Archaeoglobales was acquired by an ancestral euryarchaeote, the evolutionary history of this gene would be very complex. An unparsimonious scenario of differential losses and/or gene displacements followed by re-acquisition would have to be used to explain the presence of a class 2 HMGR only in two euryarchaeal lineages (Thermoplasmatales and Archaeoglobales).

Direct acquisition of the bacterial-like HMGR by the Thermoplasmatales and Archaeoglobales

The alternative possibility to the acquisition of a class 2 HMGR in an ancestral euryarchaeote is LGT directly to the Thermoplasmatales, the Archaeoglobales or both. Very few events would be required to explain the presence of bacterial-like HMGRs in the Thermoplasmatales and Archaeoglobales under this scenario. This is two displacement events if they are distinct lineages, as the analysis presented earlier suggested (Figures 3.12C and 3.12D). Acquisition of a class 2 HMGR would have either proceeded independently for each lineage (Figure 3.12C) or would have occurred in one of the lineages, the new gene being spread to the other by a second LGT event (Figure 3.12D). Independent acquisition by the two lineages seems to us less likely. This is because the high similarity of Thermoplasmatales and Archaeoglobales HMGRs (60.7% average amino acid identity, sd 3.1) would require the displacement of the same gene occurring in two different clades to have involved very similar donors. Furthermore, if the Thermoplasmatales and Archaeoglobales enzymes did not originate from the same bacterial donor, the *P. mevalonii* homologue would then be more similar to the HMGR from one of the lineages and thus cluster with either one of them. On the other hand, if either the Archaeoglobales or the Thermoplasmatales lineage acquired the enzyme first and then transferred it to the other lineage, it is expected that they would be related to the exclusion of all other lineages. A sister group relationship is indeed what is observed in the HMGR phylogeny (Figure 3.9), supporting a single inter-domain event, followed by an exchange between the archaeal clades. However, a small number of HMGR trees with high likelihood, 10 out of 124 tested, did not display *P. mevalonii* as basal and were not significantly worse than the best tree (Kishino-Hasegawa test). This means that we cannot eliminate two separate inter-domain LGT events as a possibility, even if less likely than an inter-domain event followed by a second event between archaea. Both possibilities nonetheless correspond to a common origin for the HMGR found in Thermoplasmatales and Archaeoglobales, involving either a single donor or two closely related donors. The exact identity of the donor of the bacterial HMGR remains unknown. It is unlikely to be *P. mevalonii* itself, an aerobic mesophile that does not share a habitat with either the hyperthermophilic Archaeoglobales or the thermoacidophilic Thermoplasmatales and that most likely acquired its own HMGR by LGT (see section 1 of this chapter).

More Exchange at the Species Level for the Archaeoglobales HMGR Gene?

The HMGR trees of Archaeoglobales, both DNA and amino acid, differ from the SSU tree for the position of *A. profundus* and *A. fulgidus* (Figure 3.11). Duplication of all DNA isolation, PCR amplification and DNA sequencing for each gene sequence obtained greatly reduces the risk of this inconsistency being caused by a mix-up of the HMGR/SSU sequences from different *Archaeoglobus* species. Furthermore, the sequences for HMGR and SSU from the *A. fulgidus* genome were confirmed here by PCR amplification of these genes from genomic DNA of the same strain as the one used in the genome sequencing project (VC-16). The inconsistency between the HMGR and SSU trees of Archaeoglobales is thus either due to a gene exchange of SSU or HMGR among Archaeoglobales or a methodological problem with phylogenetic tree reconstruction. The branching position difference of *A. fulgidus* and *A. profundus* between the SSU and HMGR trees seems well supported by the different tree reconstruction methods (Figure 3.11). However, the HMGR phylogeny does not seem reliable because of an insignificant difference between the best HMGR tree for Archaeoglobales and an alternative topology that agrees with the SSU tree (as shown by the Kishino-Hasegawa test). Therefore, we cannot say with certainty whether or not LGT of HMGR has occurred between different Archaeoglobales species.

Gene exchange between closely related species is usually difficult to establish because of the lack of variability among sequences. Its frequency should however be higher than between distant organisms, since less barriers and more vectors are present (Ochman *et al.*, 2000).

Chapter 4: Frequent intraspecific heterogeneity of ribosomal RNA operons among extremely halophilic archaea

This chapter includes work intended to be published as Y. Boucher, C.J. Douady, A.K. Sharma, M. Kamekura and W.F. Doolittle (2003) Frequent intraspecific heterogeneity of ribosomal RNA operons among extremely halophilic archaea. *Journal of Bacteriology.*

Introduction

The gene coding for the small ribosomal subunit (SSU) has been microbiology's "gold standard" for identification and classification of microorganisms over the last decade (Ludwig and Schleifer, 1999). More recently, its role has also been extended to the evaluation of microbial diversity by its direct amplification from environmental DNA followed by sequencing or denaturing gradient gel electrophoresis (Muyzer, 1999). It is also by far the most common probe for the detection of specific strains in the environment by fluorescent *in-situ* hybridization (Amann *et al.*, 2001).

Several major assumptions have been made by molecular sytematists in considering rRNA genes as ideal phylogenetic markers: 1) The intraspecific (within a single organism) variability between multiple rRNA gene copies is generally assumed to be low (Clayton *et al.*, 1995); 2) This gene is rarely (if ever) affected by LGT (including recombination) (Woese, 2000); 3) It has a conserved function across all domains of life; 4) The gene has slow and fast evolving regions, which should allow phylogenetic analysis of both distantly and closely related taxa. In the last few years, it has become obvious that some of these assumptions might not hold true for all prokaryotes. Low levels of intraspecific variability of the SSU gene seem to be very common in prokaryotes (Clayton *et al.*, 1995), and even more important variability can be present in the internal transcribed spacer (ITS) found between the SSU and LSU genes (Cilia *et al.*, 1996). Even these low levels of divergence can undermine attempts to evaluate prokaryotic diversity at the molecular level, especially when using methods as sensitive as DGGE (Dahllof *et al.*, 2000). Also, when using SSU gene sequence as a criterion to assign strains to a particular species, low level heterogeneity (~1.0-2.0%) can lead to misidentification (Ninet *et al.*, 1996).

High levels of heterogeneity between multiple SSU gene copies found within a single cell have also been identified in numerous organisms. Important intraspecific SSU variability has been reported for some eukaryotes, such as the apicomplexan *Plasmodium berghei* {harbouring SSU gene copies divergent at 5.0% of their nucleotide positions (Gunderson *et al.*, 1987)} and the metazoan *Dugesia mediterranea* {8.0% divergence, (Carranza *et al.*, 1996)}. Among prokayotes, divergence levels over 5.0% have been found in two different instances: the thermophilic actinomycete *Thermomonospora chromogena* and the extremely halophilic archaeal genus *Haloarcula*. However, it seems that these two cases of intraspecific SSU variability have different origins. *T. chromogena* harbours six rRNA operons, one of which differs from the others at 6.0% of the nucleotide positions in the SSU gene and at 10.0% in the LSU gene (Yap *et al.*, 1999). This divergent operon was found to be a mosaic of the other operons of *T. chromogena* and an rRNA operon from the thermophilic actinomycete species

Thermobispora bispora. This mosaic operon was only found in one species, *T. chromogena*, and was hypothesized by the authors to have been originally acquired by LGT from *T. bispora* and subsequently recombined with the other operons of its new host by gene conversion, on its way to be homogenized (Yap *et al.*, 1999).

Haloarcula's case seems fairly different. Intraspecific rRNA heterogeneity was first detected in *Haloarcula marismortui*, which displayed two rRNA operons, divergent at 5% of the positions in their SSU genes and 1.3% in their LSU gene (Mylvaganam and Dennis, 1992). The processing of the rRNA product was shown to differ between the two operons, one going through the canonical archaeal processing and the other displaying a unique processing (Dennis *et al.*, 1998). Later on, similar levels of divergence were identified between the two to four SSU genes found within numerous species of *Haloarcula* (Gemmell *et al.*, 1998). This demonstrates that the phenomenon is not an isolated event affecting only one species (as was probably the case for *T. chromogena*), but rather an evolutionarily stable characteristic. Both of the *H. marismortui* rRNA operons were shown to be expressed when it is grown in rich media in the laboratory (Amann *et al.*, 2000), suggesting that they are both functional. The recent identification of SSU genes that are 7% divergent in the Halobacteriales *Halosimplex carlsbadense* brings the possibility that this level of heterogeneity might be common in extremely halophilic archaea (Vreeland *et al.*, 2002). This prompted us to look in other Halobacteriales genera for high levels of intraspecific rRNA genes heterogeneity. We detected the presence of heterogeneous rRNA operons in strain XA3-1, which most likely belong to the genus *Natrinema*, a group distantly related to both *Halosimplex* and *Haloarcula* according to SSU gene phylogenies.

The rRNA operons of strain XA3-1 were obtained through methods that do not involve PCR amplification. Most studies rely on PCR to obtain the SSU gene(s) of a species, which can lead to the formation of chimeric molecules if multiple heterogeneous rRNA operons are found in the organism in question. A clear example of this is the novel haloarchaeal species *H. carlsbadense*, thought to harbour three rRNA operons from PCR-amplification and sequencing of its SSU genes for classification purposes (Vreeland *et al.*, 2002). We obtained the rRNA operons of *H. carlsbadense* through a combination of PCR-independent and chimera-limiting PCR methods, clearly showing that only two operons are present and the third one originally identified is in fact a chimera of these two.

Beside *H. marismortui*, no haloarchaeon possessing more than one rRNA operon has had all of its copies completely sequenced. With the addition of *H. carlsbadense* and *Natrinema sp.* XA3-1, all heterogeneous rRNA operons of representatives of three divergent groups of Halobacteriales are available to understand the origins and evolutionary importance of intraspecific rRNA variability.

Materials and Methods

Genomic DNA extraction

Genomic DNA samples of *Natrinema versiforme* and *Natrinema sp.* XA3-1 were both obtained from M. Kamekura (Noda Institute for Scientific Research, Noda, Japan). *Halosimplex carlsbadense* genomic DNA was isolated from a cultured strain following the protocol from Wilson (1994).

Genomic DNA digests

The genomic DNAs of *Natrinema versiforme* and *Natrinema sp.* XA3-1 were completely digested with *Cla* I, while *Halosimplex carlsbadense* gDNA was digested using both *Not* I and *Sca* I. These digests were performed overnight at 37°C following the manufacturer's recommendations (New England Biolabs). The restriction endonucleases used were chosen for an average length of digestion products between 5 and 10 kb and a low probability of cutting within the rRNA operons of the extreme halophiles under investigation. To determine which restriction endonucleases are unlikely to cut within these rRNA operons, we obtained the sequence of one rRNA operon from each species. These sequences were obtained from PCR amplified fragments (using primers F1 and R1, see Figure 4.1) cloned in a plasmid vector (Topo-XL, invitrogen). One clone was sequenced for each strain using the primers described in Table 4.1. The resulting sequences were analyzed to identify enzymes that did not cut within them. Also, the selected enzymes were required not to cut within other Halobacteriales rRNA operons of known sequence: *Haloferax volcanii* (http://wit-scranton.mbi.scranton.edu/haloferax/), *Haloarcula marismortui* (http://zdna2.umbi.umd.edu/cgi-bin/blast/blast.pl), *Natrialba magadii* (Lodwick *et al.*, 1991) and *Halobacterium sp.* NRC-1 (Ng *et al.*, 2000).

Table 4.1 Primers for the amplification and sequencing of rDNA operons in Halobacteriales.

Primer[1]	Position[2]	Gene	Sequence
F1	1-17	SSU	ATTCCGGTTGATCCTGC
F2	463-480	SSU	CCGCGGTAATACCGGCAG
F3	822-839	SSU	CCGCCTGGGAAGTACGTC
F4X	1136-1155	SSU	GCAACGGTAGGTCAGCATGC
F4	1140-1156	SSU	CGGTAGGTCAGTATGCC
F5	1458-1473	SSU	GGCTGGATCACCTCCT
F6	2050-2066	LSU	GGACGTGCCAAGCTGCG
F7X	2435-2454	LSU	TACTCCTCGAGACCGATAGC
F7	2460-2479	LSU	AGTAGTGTGAACGAACGCTG
F8	2863-2879	LSU	GGTGAAAGGCCCATCGA
F9	3189-3207	LSU	CAGCTTACCGGCCGAGGTT
F10	3722-3739	LSU	ACGTTAGGGAATTCGGCA
F11	4050-4069	LSU	CCAGTGCGGAGTCTGGAGAC
F12	4527-4545	LSU	CGGTTCCCTCCATCCTGCC
R1	4834-4856	LSU	CGCGCACACCCCGAGTCTATCGA
R2X	4485-4501	LSU	CGATATGTACTCTTGCG
R2	4485-4501	LSU	CGATATGTGCTCTTGCG
R3	4078-4094	LSU	ATAGGGTCTTCGCTTCC
R4X	3722-3741	LSU	ATTGCCGAATTCCCTAACGT
R4	3722-3741	LSU	CTTGCCGAATTCCCTAACGT
R5	3189-3207	LSU	AACCTCGGCCGGTAAGCTG
R6	2863-2879	LSU	TCGATGGGCCTTTCACC
R7X	2524-2543	LSU	CAGCGTTCGCTCGCGCTACT
R7	2519-2537	LSU	TCGCTCGATCGCCAACTGA
R8	2050-2066	LSU	CGCAGCTTGGCACGTCC
R9	1427-1447	SSU	CTACGGCTACCTTGTTACGAC
R10X	1241-1261	SSU	CCTCAATCCGAACTACGACC
R10	1173-1190	SSU	CCATTGTAGCCCGCGTGT
R11	822-839	SSU	GACGTACTTCCCAGGCGG
R12	499-516	SSU	ACGCTTTAGGCCCAATAA

[1] F (forward), R (reverse), X (Specific to *Halosimplex*)

[2] Correspond to the nucleotide position in *Halobacterium sp.* NRC-1 rDNA operon

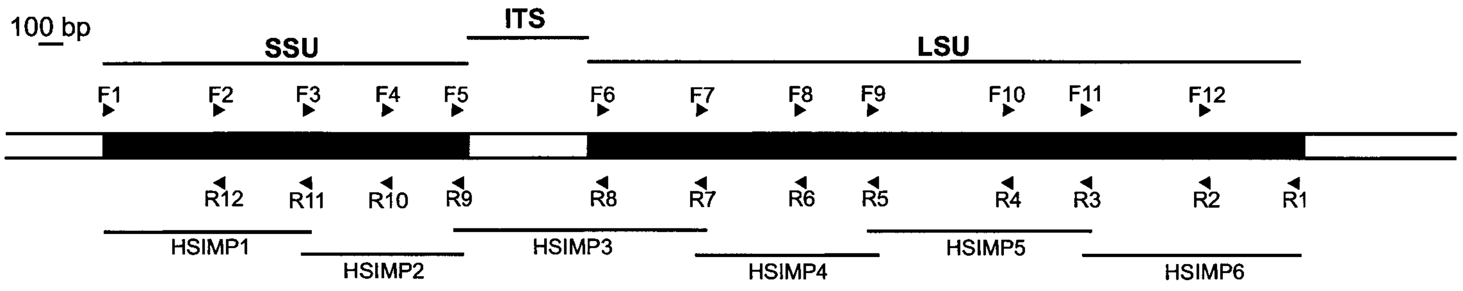


Figure 4.1 Schematic diagram of the rRNA operon sequencing strategy and the PCR amplification of one of *Halosimplex carlsbadense* two rRNA operons. Cloned rRNA operons were sequenced using primers matching regions conserved in all Halobacteriales SSU and LSU genes. For each operon, both DNA strands were sequenced, with one sequencing primer every ~500 bp on each strand. ITS represents the internally transcribed spacer found between the SSU and LSU genes. The fragments labeled HSIMP 1 to 6 represent the PCR fragments used to obtain the complete sequence of *Halosimplex carlsbadense* second rRNA operon (see materials and methods).

Determination of the number of rRNA operons by Southern hybridization

Digested gDNA samples were loaded onto a 0.8% agarose gel, which was run for 16h at 60 volts in 1X TAE buffer (4°C). Transfer of the electrophoresis gel to a positively charged nylon membrane was performed as described in chapter 3, section 3. SSU rRNA probes were directly amplified from genomic DNA of the species to which they were hybridized. Probes were amplified with universal Halobacteriales primers (F1 and R9 in Table 4.1), generated from an alignment of all Halobacteriales SSU gene sequences available in public databases. Amplified probes were gel-purified (MinElute, QIAGEN) and subsequently labeled with digoxigenin (DIG) dUTP (as described in chapter 3, section 3). Each probe was individually hybridized to the membrane along with labeled DNA ladder (GeneRuler 10 kb, MBI Fermentas). The hybridization of the probes was detected by a CDP-star chemiluminescence system (Roche).

Genomic DNA library construction

Southern hybridization revealed the location of the rRNA operons relative to the DNA ladder (between 5 and 10 kb). For both *Natrinema sp.* XA3-1 and *Halosimplex carlsbadense*, genomic DNA was re-digested and run on an agarose gel under conditions identical to those used for Southern hybridization. gDNA was then extracted from the regions of the gel corresponding to DNA fragments containing an rRNA operon (MinElute, QIAGEN). Before purified DNA fragments could be cloned, protruding 5' or 3' overhangs resulting from digestion had to be blunt-ended and dephosphorylated in preparation for blunt end cloning (manufacturers instructions, TOPO-Zeroblunt, Invitrogen). The resulting products were ligated into the Topo-Zeroblunt plasmid vector

(Invitrogen) and transformed in chemically competent TOP10 *E. coli* cells (Invitrogen), which were plated on kanamycin media to select for positive transformants.

Since the average genome size for our species was approximately 2.0 Mb, 2000 clones carrying an average of 8 kb of exogenous DNA (depending on where the band containing the rRNA operon was extracted from on the gel) should provide an 8-fold genome coverage. Such coverage was assessed to be sufficient to recovere all rRNA operons copies, as the gDNA libraries are in fact sub-libraries enriched for rRNA genes.

Library screening

The libraries were transferred to positively charged nylon membranes (Roche) and screened directly for the presence of rRNA genes by hybridizing with a DIG-dUTP labeled probe (PCR amplified SSU gene of the species from which the gDNA in the library originates). Positive clones were confirmed by end-sequencing (using the M13 forward and M13 reverse primers) and direct sequencing (using a universal Halobacteriales SSU primer, F1 in Table 4.1). Confirmed clones were completely sequenced (see below).

PCR amplification of rRNA genes not obtained in genomic libraries

All four rRNA operons of *Natrinema sp.* XA3-1 and one of the two operons of *H. carlsbadense* were obtained after repeated screening of the libraries. A PCR strategy aimed at minimizing the formation of chimeras was used to obtain the second rRNA operon of *H. carlsbadense.* PCR reactions were carried out under standard conditions with the following modifications: 5X more of each primer, using a *Pyrococcus furiosus*

DNA polymerase (PFU turbo, Stratagene), increased extension time (5X) and fewer (25) PCR cycles. These modifications result in the following PCR conditions: a final volume of 50 µl containing 1-5 ng of template DNA, 1 X PCR buffer, 1 ul of 10mM dNTPs, 5 µl of each 10 µM primer and 1 µl of PFU Turbo DNA polymerase (Stratagene). The initial denaturation took place at 95°C for 2 minutes, followed by 25 cycles with a denaturation at 95°C for 30 seconds, primer annealing at 55°C for 30 seconds, and primer extension at 72°C for 5 minutes. The operon was amplified in six fragments of ~500bp overlapping at their extremities, as chimera formation is less likely for shorter products (primers used are identified in Figure 4.1). After the cloning of each fragment (TOPO-ZeroBlunt, Invitrogen), a mixed population of clones (from each of the two operons) was obtained. Sixteen clones were sequenced for each fragments and subsequently compared to the sequence of the complete rRNA operon obtained from the library. This allowed detection of chimeric fragments (which were eliminated) and the determination of fragments belonging to the missing operon. The fragments were then assembled together to obtain the final sequence of the second *H. carlsbadense* operon.

DNA sequencing, analysis and assembly

Positive clones from the library found to carry an entire rRNA operon were sequenced using MegaBase technology and BigDye chemistry. Multiple primers were used for sequencing the entire rRNA operon, giving overlapping reads to obtain reliable sequence (Figure 4.1). The sequencing primers target conserved regions of rRNA operons, as determined by the examination of an alignment of all available Halobacteriales rRNA genes. Sequences of these primers can be found in Table 4.1.

Sequencher 4.1.2 (Gene Codes Corporation) was used to analyze sequence chromatograms and assemble sequence fragments as described in chapter 3, section 3.

Phylogenetic analysis

Phylogenetic analyses were performed with PAUP* 4.04b (Swofford, 1998) applying the heuristic-search option and using the TBR branch-swapping algorithm. Maximum likelihood and maximum likelihood distances were used as the tree reconstruction methods, with the nucleotide substitution model, gamma rates parameter α, proportion of invariable sites and nucleotide frequencies determined independently for each gene using MODELTEST (Posada and Crandall, 1998). The confidence of each node was determined by building a consensus tree of 100 bootstrap replicates (using minimum evolution or maximum likelihood distances or "full" maximum likelihood).

Results

Detection of heterogeneity in the rRNA operons of Halobacteriales

Genomic DNA was extracted from a pure culture of *Natrinema sp.* XA3-1 and its SSU gene amplified by PCR, cloned and sequenced. The multiple clones sequenced showed high sequence divergence (at ~5.0% of nucleotide positions). As a control, the same procedure was applied to the Halobacteriales species *Natrinema versiforme*, whose SSU gene sequence was the most similar (among sequences present in GenBank) to the sequences obtained from XA3-1. Heterogeneity was also detected at high levels in that species.

Halosimplex carlsbadense SSU genes were amplified by PCR as part of its taxonomic identification by Vreeland *et al.* (2003) (Vreeland *et al.*, 2002). Three divergent genes were identified, one of which was different from the other two at about 7% of the nucleotide positions, the highest level of divergence so far reported among multiple copies of the SSU gene within one prokaryotic organism. However, upon close examination, one of these gene copies seemed to be a chimera of the other two. The 5' end of this "chimeric" gene was identical to the homologous region in the second copy and its 3' end identical to the third copy.

Determination of the number of rRNA operons in *Natrinema versiforme*, *Natrinema sp.* XA3-1 and *Halosimplex carlsbadense*

The number of rRNA operons found in each of the three species where heterogeneity was detected was determined by Southern hybridization (Figure 4.2). *N. versiforme* and *Natrinema sp.* XA3-1 were both found to have four rRNA operons. In contradiction with previous claims that *H. carlsbadense* harbours three rRNA operons {based on the PCR amplification of three divergent SSU genes from genomic DNA, see (Vreeland *et al.*, 2002)}, we found that this haloarchaeon had only two operons. This result was obtained when its genomic DNA was digested to completion with a combination of NotI and ScaI (Figure 4.2), as well as NotI/ClaI and ScaI/ClaI double digests (results not shown).

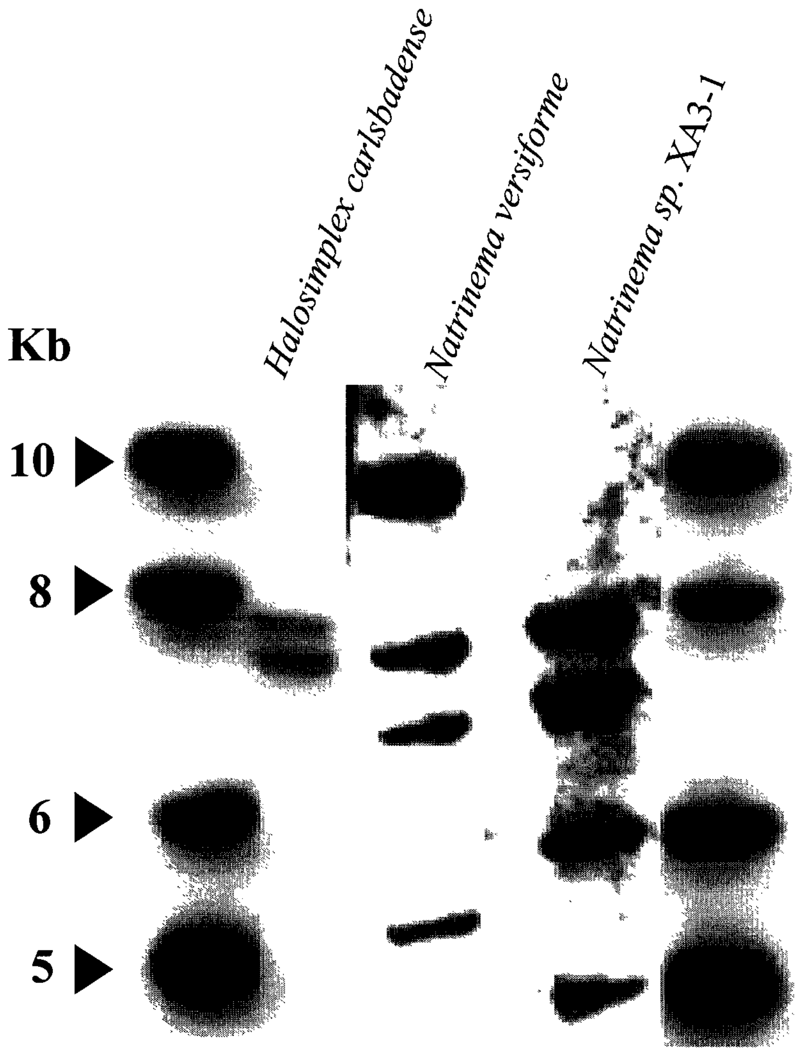


Figure 4.2 Southern hybridization of SSU probes to complete genomic DNA restriction endonuclease digest. The restriction endonucleases used were ClaI for *Natrinema sp.* XA3-1 and *Natrinema versiforme* and both NotI and ScaI for *Halosimplex carlsbadense*. The SSU probes used were directly amplified from genomic DNA of the species to which they were hybridized.

Degree of divergence of the rRNA operons of Halobacteriales containing multiple heterogeneous copies

Sequence comparisons of the multiple rRNA operons found in *Natrinema sp.* XA3-1 and *H. carlsbadense* confirms the high degree of divergence detected between their PCR amplified SSU genes (Table 4.2). Three of the four SSU genes of *Natrinema sp.* XA3-1 (from operons B, C and D) are almost identical, the other (from operon A) being different at 5.0% of nucleotide positions. The divergence levels observed for the LSU genes follow a different pattern, as the genes of all operons, including operon A, diverge similarly from each other with values going from 0.9 to 1.9%. The LSU gene from operon B is the most divergent, differing from the others at 1.2 to 1.9% of nucleotide positions. The SSU-LSU intergenic spacers (ITS1) of all rRNA operons of *Natrinema sp.* XA3-1 are identical in sequence, including a tRNA^{ala}.

The two SSU genes of *H. carlsbadense* differ at 6.8% of their nucleotide positions. Similarly to *Natrinema sp.* XA3-1, the LSU genes display a lower divergence (2.6%). The ITS1 are extremely divergent, showing significant similarity only in the first 33bp of their 5' end (which are identical) and in the last 135 bp (39 of the last 135 positions are divergent). Operon B ITS1 is 109 bp shorter than its operon A counterpart (348 bp vs. 457 bp) and does not contain a functional tRNA^{ala} (one can clearly be identified in the ITS1 of operon A).

Table 4.2 Details of nucleotide substitution and levels of divergence between the rDNA operons of Halobacteriales species harbouring multiple heterogeneous copies

Species	Operons compared	SSU					LSU				
		Substitutions				%ID	Substitutions				%ID
		all	TV	TI	Indels		all	TV	TI	Indels	
Haloarcula marismortui	A and B	74	20	54	0	95.0	39	11	28	0	98.7
Halosimplex carlsbadense	A and B	100	39	56	5	93.2	77	30	46	1	97.4
Natrinema sp. XA3-1	A and B	73	15	53	5	95.1	53	22	31	0	98.1
	A and C	74	15	54	5	95.0	30	12	18	0	98.9
	A and D	73	15	53	5	95.1	25	10	15	0	99.1
	B and C	1	0	1	0	99.9	35	19	16	0	98.8
	B and D	2	0	2	0	99.9	54	22	32	0	98.1
	C and D	1	0	1	0	99.9	32	13	19	0	98.9

TV = transversions
TI = transitions
SSU = small ribosomal subunit gene
LSU = large ribosomal subunit gene
ITS1 = internal transcribed spacer (SSU-ITS1-LSU)
Indels = insertion/deletion
%ID = percentage of nucleotide identity

Location of variable positions of heterogeneous rRNA genes on the secondary structure of their rRNA product

Nucleotide substitutions between the multiple SSU and LSU genes found in *Halosimplex*, *Natrinema* and *Haloarcula* were mapped to the secondary structure of their rRNA products. *Haloarcula* is included here, as it is the only genus of haloarchaea beside those being studied here known to display intraspecific heterogeneity in its rRNA genes. Most substitutions between the heterogeneous rRNA genes found in any of those haloarchaea are either compensatory mutations occurring in stems or are located in loop regions. This suggests that these substitutions would have little or no effect on the overall secondary structure of the ribosomal RNA. Furthermore, all of the ribosomal protein genes of *Haloarcula marismortui* have been cloned and sequenced (Scholzen and Arndt, 1992), all appearing to be single copy genes. This implies that the same protein should have the capacity to bind to rRNA from either of the two heterogeneous genes to form functional ribosomes. When grown in rich media, *H. marismortui* has indeed been shown (using FISH probes specific to each of the two SSU genes) to have a mixed ribosome population (Amann *et al.*, 2000).

Figure 4.3 presents the distribution of nucleotide sequence differences in pairwise comparisons of SSU and LSU genes. Substitutions between heterogeneous SSU genes are mostly found in hypervariable regions, which is where most interspecies divergence also occurs. Regions corresponding to helices 21, 22 and 26 (5' domain) in SSU secondary structure are variable in all three haloarchaeal genera displaying intraspecific rRNA gene heterogeneity. The region corresponding to helices 7, 8 and 9 (5' domain) is highly variable in both *Halosimplex* and *Natrinema* but not in *Haloarcula*.

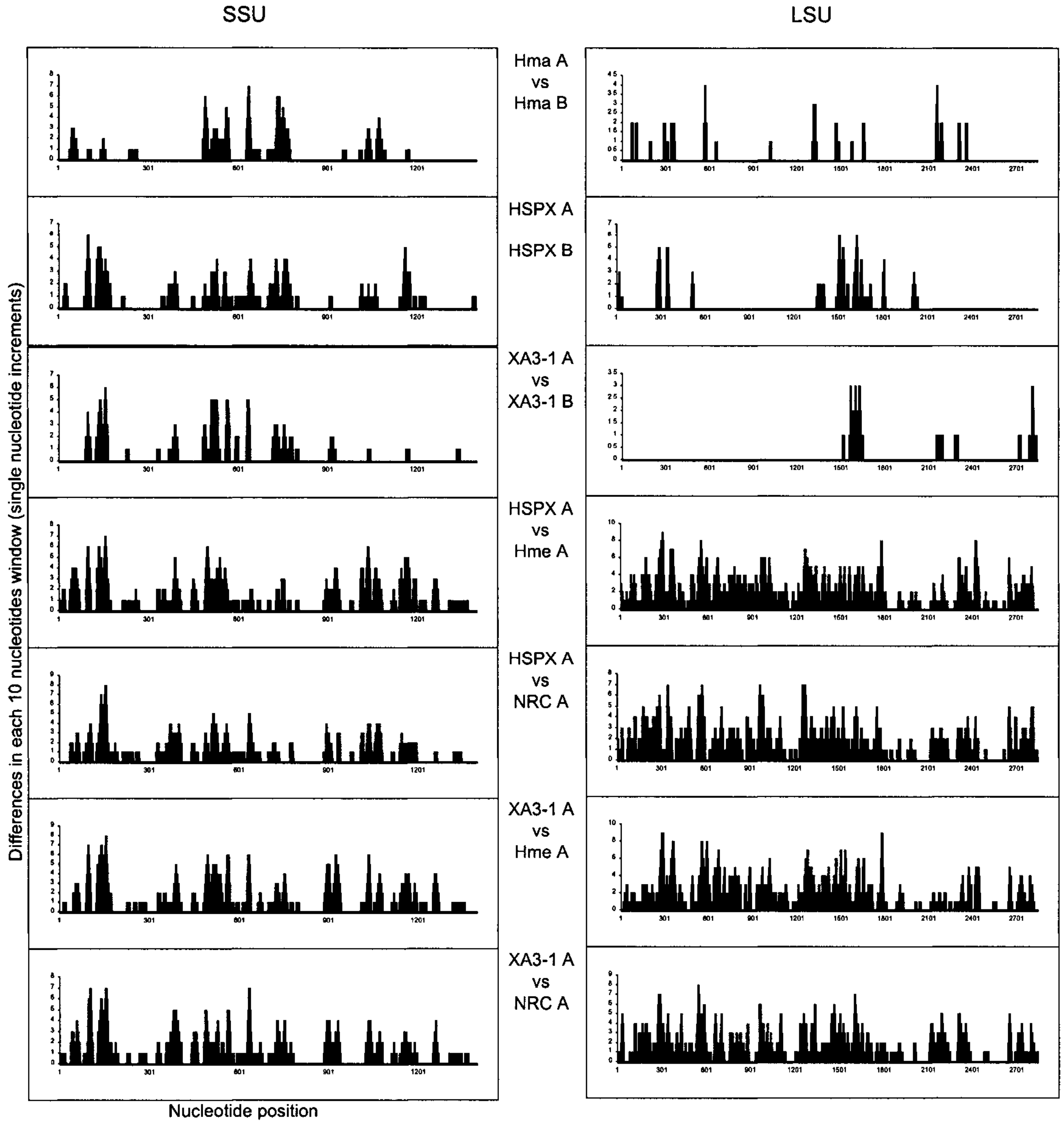
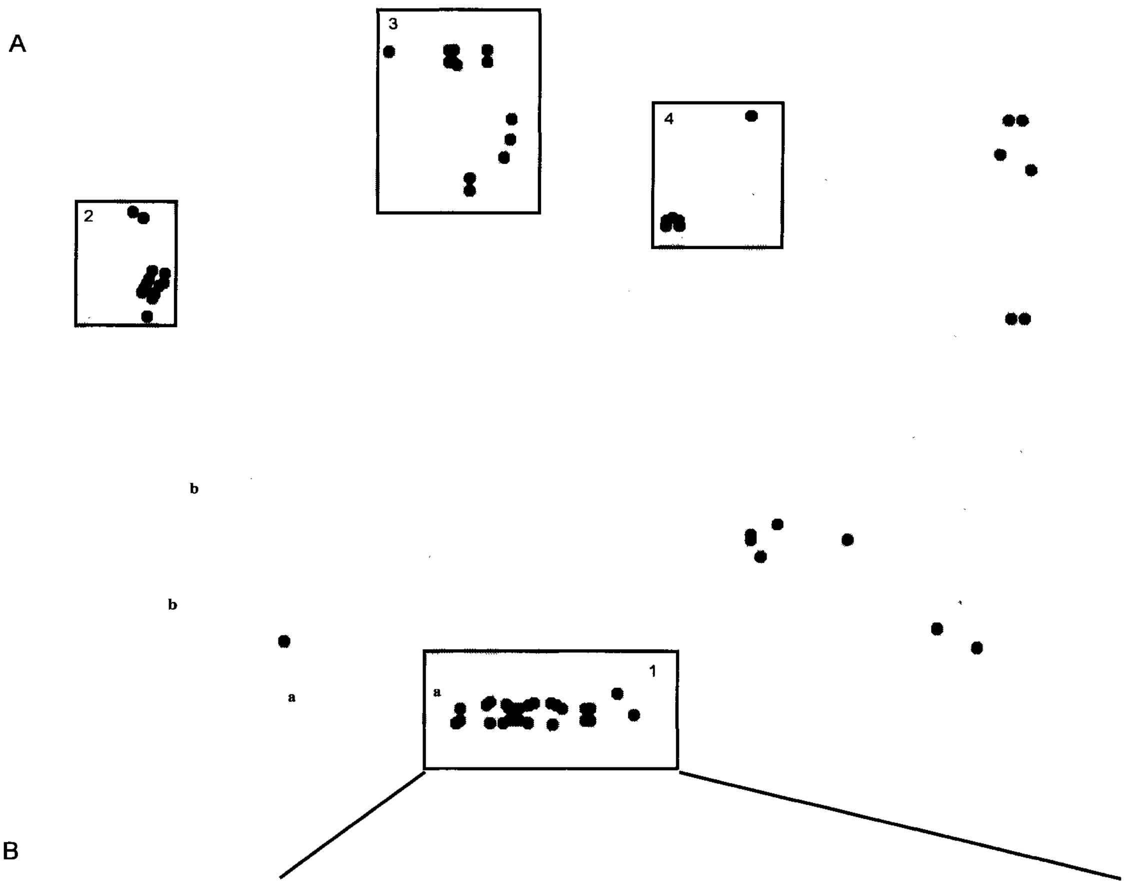


Figure 4.3 Distribution of nucleotide differences in pairwise comparison of SSU and LSU genes. The number of nucleotide sequence differences was determined using a 10 nucleotides wide sliding window (single nucleotide increments). Species abbreviations: Hma (*Haloarcula marismortui*), HSPX (*Halosimplex carlsbadense*), XA3-1 (*Natrinema sp.* XA3-1), Hme (*Haloferax mediterranei*) and NRC (*Halobacterium sp.* NRC-1).

Intraspecific divergence between LSU genes is much lower than between SSU genes and is very patchily distributed across the length of the gene (Figure 4.3). The region containing helices 18, 19 and 20 (domain 1) is the only one that displays intraspecific heterogeneity in all three haloarchaeal species, other variable regions being mostly found in only one of the three species. Figure 4.3 displays, among others, a sequence comparison of the LSU genes from *Natrinema sp.* XA3-1's rRNA operons A and B (LSU-A and LSU-B). This mapping of substitutions distribution is not representative of the overall LSU intraspecific heterogeneity found in this species. In contrast with the SSU gene, the bulk of the heterogeneity between the LSU genes of XA3-1 is not contained in one highly divergent copy (SSU-A for the small subunit gene). Each of the four copies diverges from the others in specific regions (Figure 4.4).

Figure 4.4 A) Location of variable positions of *Natrinema sp.* XA3-1 LSU genes on the secondary structure of their rRNA product. Conserved positions are represented by gray dots while variable ones are indicated by black dots. Numbered boxes represent different region of the LSU genes where heterogeneity between the multiple copies can be observed: 1) Divergence of LSU-B at 27 positions between nucleotides 249 and 370; 2) Divergence of LSU-B and C from LSU-A and D at 13 positions between nucleotides 1113 and 1284; 3) Divergence of LSU-D at 13 positions between nucleotides 1523 and 1652; 4) Divergence of LSU-C at 6 positions between nucleotides 1786 and 1856. Variable positions that are not found within a box mostly represent divergence in LSU-A. B) Alignment of all available Halobacteriales LSU genes from positions 200 to 400, which corresponds to box 1 in the secondary structure diagram above. At any given position, nucleotides identical to the one found in the first row are indicated by a dot.



B

```
                                       210       220       230       240       250       260       270       280       290       300
                                        .         .         .         .         .         .         .         .         .         .
Natrinema sp. XA3-1 LSU-A       ATGTCGTTAGTAAGGGCGACTAAACGCGACACAGTCCAAACCGAAGCCTTCACGCAATGTGGTGTTCGGGCTGACTCTCAGCAGCAAAACGGCGACGCGA
Natrinema sp. XA3-1 LSU-B       ................................................C....................A.....GA...CTCT.GG..A.T....AA..
Natrinema sp. XA3-1 LSU-C       ....................................................................................................
Natrinema sp. XA3-1 LSU-D       ....................................................................................................
Natrialba magadii               ............CT.....AGG.......T.....................GG................A.....GA...CTCT.CG..ACT....A...
Halococcus morrhuae             .........C...C.....G.G.........................T.A.GG................A.............C.GG.T..T.TGT.T..
Halobacterium sp. NRC-1         .C.C.........T.....A.G......G......................GG................A......T...T.GTTTG.C..TTCGT.T..
Haloarcula marismortui LSU-A    .............CC....G.G.......T....C.............C................CA.....AC......T....CG.C..T.TCGA...
Haloarcula marismortui LSU-B    .............CC....G.G.......T....C.............C................CA.....AC......T....CG.C..T.TT.A...
Haloferax volcanii              .............C.....G.G.......T.....................GG............A...A..AC.CA.A.CGGA.CTGGAATTC..A...
Halosimplex carlsbadense LSU-A  ..A....C.....CC....G.G.......T....C......T....G.C.................T....C.G.A....T.GT.GG...CCTTCT.T..
Halosimplex carlsbadense LSU-B  ..A....C.....CC....G.G.......T....C......T....G.C.................T....C.G.GA...A.GTTCG....ATCTT....

                                       310       320       330       340       350       360       370       380       390       400
                                        .         .         .         .         .         .         .         .         .         .
Natrinema sp. XA3-1 LSU-A       AGTCTCTTGGAATAGAGCACGAAACAGGGTGACAGTCCCGTACCGTCGACAAGTACGCTGTGCGTCAGTCCCAGAGTATCGGGGGTTGGATATCCCTCGT
Natrinema sp. XA3-1 LSU-B       .................T....G...................TT......G....A.AGAC.A.....CT..............................
Natrinema sp. XA3-1 LSU-C       ....................................................................................................
Natrinema sp. XA3-1 LSU-D       ....................................................................................................
Natrialba magadii               ............C.........G....................T......G....A.AGAC.A.....CT..............................
Halococcus morrhuae             ......C.....CG....GT..G.........A.AC......A.ACA...C......G..........CT..............................
Halobacterium sp. NRC-1         ......C..A..CG....G...T............C......T.ACG...C......AC.........CT........G...........A.........
Haloarcula marismortui LSU-A    ............C.....GT..T...........AC.......TCGA...C......AC.....GT...G........G....................C
Haloarcula marismortui LSU-B    ..................GT..T...........AC.......T.AA...C.............GT...G........G....................C
Haloferax volcanii              ......C.....CG....GT..T.........A.AC......GT..GA.GG.A-.GATCCC.G.GTT..T.....A..G.....................
Halosimplex carlsbadense LSU-A  ......C.....CG....G...C.........A..C......A..GAAG.T....T.AC...A.C.G..T..G.....G...........A.........
Halosimplex carlsbadense LSU-B  ......C.....CG....G...C.........A..C.........AG.T.T.T....AC...A.C.G..T..G.....G...........A.........
```

Figure 4.4 A) Location of variable positions of Natrinema sp. XA3-1 LSU genes on the secondary structure of their rRNA product. B) Alignment of all available Halobacteriales LSU genes from positions 200 to 400.

Recombination between *Natrinema sp.* XA3-1 and *Natrialba magadii* LSU genes

XA3-1's LSU genes display heterogeneity between paralogs in several distinct regions (Figure 4.4A). Between positions 270 and 370 (corresponding to box 1 in Figure 4.4A), XA3-1's LSU-B gene is virtually identical to its *Natrialba* homolog (differs at only three positions), while it diverges strongly from the three other XA3-1 LSU genes (differs at 26 positions) (Figure 4.4B). None of the other regions where one or two of XA3-1's LSU genes differs from the others had such strong similarity to the LSU gene(s) of another species of Halobacteriales.

The surprisingly high similarity between positions 270 and 370 of XA3-1 LSU-B gene and the equivalent region in its *N. magadii* homolog is most easily explained by a homologous recombination event between the two genes. This recombination event is made even more likely by the fact that it is limited to a specific structural region of the LSU molecule (helices 18, 19 and 20, domain 1). Most of the nucleotide substitutions caused by this recombination, when found in stem regions, are complementary to each other, therefore conserving the rRNA secondary structure (Figure 4.4A). The maintenance of the secondary structure is important, given the complexity of the ribosome and the numerous interaction of the LSU with proteins and other RNA components. To our knowledge, this is the first example of a recombination event between rRNA genes in archaea. Also, the event took place over one of the largest phylogenetic distances reported for recombination of rRNA genes (between two genera diverging at 5% of the positions in their SSU genes). Other regions that are heterogeneous between the LSU genes of XA3-1 might also be the result of recombination. This could only be verified, however, if strong similarity was found between these regions and their equivalent in other organisms LSU genes. The incapacity to find a donor for these regions, if they indeed originated through recombination, probably lies in the very limited sampling of the LSU gene among Halobacteriales (which is available from only seven species).

Phylogenetic distribution of organisms harbouring intracellular heterogeneity in their rRNA genes

A phylogeny of the order Halobacteriales based on the SSU gene is presented in Figure 4.5. With the addition of data from this study, rRNA heterogeneity is now known to happen in three divergent lineages: *Halosimplex*, *Haloarcula* and *Natrinema*. These lineages, according to the SSU gene, do not share specific phylogenetic affinity for each other and are spread across Halobacteriales diversity (Figure 4.5A).

Within *Haloarcula*, all species that have been investigated harbor rRNA heterogeneity (Gemmell *et al.*, 1998). *Halosimplex carlsbadense* is the only cultured species of its genus. We present here two species of *Natrinema*, *Natrinema sp.* XA3-1 and *N. versiforme*, displaying strong intraspecific rRNA heterogeneity. A phylogeny of the SSU genes from representatives of the haloarchaeal clade to which *Natrinema* belongs is presented in Figure 4.5B, which is a subset of the tree presented in Figure 4.5A (indicated by the black triangle). The most divergent of the 1 four SSU genes from *Natrinema sp.* XA3- (SSU-A) clusters outside of the clade formed by its other three genes and SSU genes from various *Natrinema* species (which include, among others, *N. versiforme*, along with several uncharacterized Halobacteriaceae). More specifically, the

XA3-1 SSU-A gene clusters very strongly with a homolog from *Haloterrigena sp.* arg-4 (Ihara *et al.*, 1999). In fact, the XA3-1 SSU-A gene differs only at 27 positions from its arg-4 homolog (only 1.8% of the nucleotide positions), comparatively to a 4.9-5% divergence of SSU-A from XA3-1's other SSU genes.

When phylogenetic trees of the SSU and LSU genes of haloarchaea are compared to each other (using the same taxon sampling), the overall structure is found to be very similar (Figure 4.6). The LSU tree, resulting from a dataset with more informative positions, seems to be slightly more resolved. The only conflict between the two trees resides in the branching order of the SSU and LSU genes from *Natrinema sp.* XA3-1. In the SSU tree, the SSU-A gene is by far the most divergent and strongly branches in a basal position in relation to the other three copies. In the LSU tree, on the other hand, the most divergent copy is LSU-B, LSU-A grouping strongly with the other two copies.

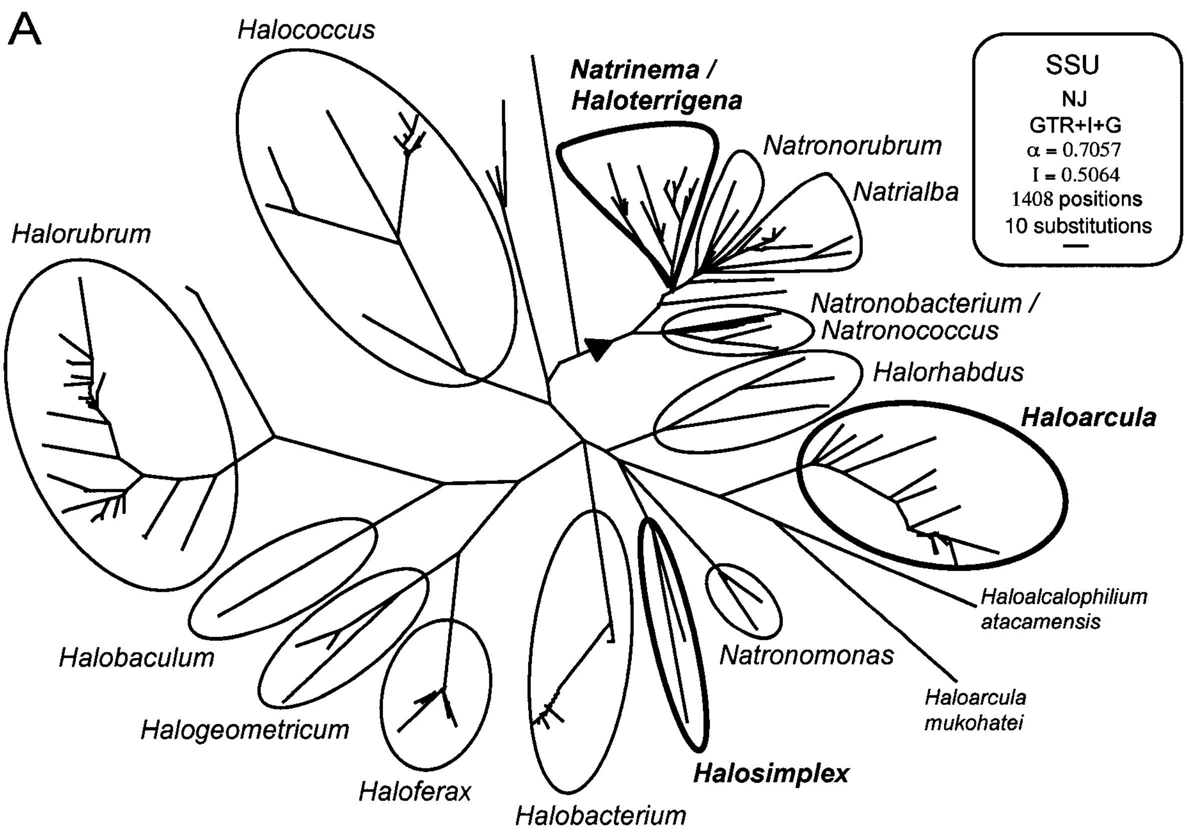


Figure 4.5 A) Best maximum likelihood distance tree of the SSU gene for the archaeal order Halobacteriales and B) Best maximum likelihood tree of the Halobacteriales subgroup indicated by the black triangle in A. The evolutionary models and parameters used for the phylogenetic analyses are presented in boxes on the right-hand side of the trees (NJ: neighbor-joining, ML: maximum likelihood, GTR: general time reversible, I: proportion of invariable sites, G: gamma distributed among-site rate variation, α: gamma rate shape parameter alpha, Γ: number of gamma distribution rate categories). The genera in which single organisms contain highly heterogeneous rRNA operons are highlighted in bold. Bootstrap values were obtained using a distance maximum likelihood tree reconstruction method (only values >50% are displayed).

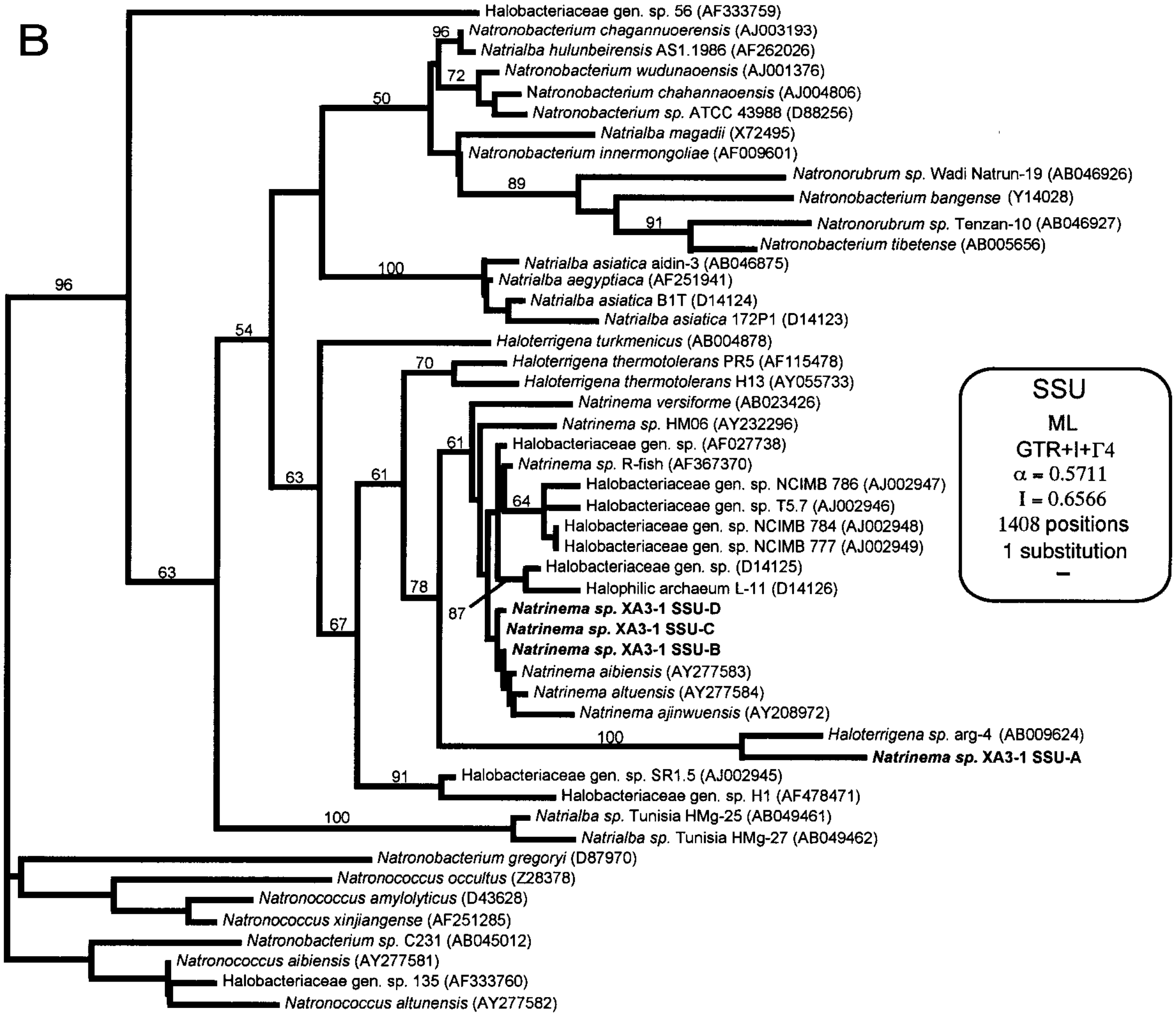

B
Halobacteriaceae gen. sp. 56 (AF333759)
96
Natronobacterium chagannuoerensis (AJ003193)
Natrialba hulunbeirensis AS1.1986 (AF262026)
72
Natronobacterium wudunaoensis (AJ001376)
Natronobacterium chahannaoensis (AJ004806)
Natronobacterium sp. ATCC 43988 (D88256)
50
Natrialba magadii (X72495)
Natronobacterium innermongoliae (AF009601)
89
Natronorubrum sp. Wadi Natrun-19 (AB046926)
Natronobacterium bangense (Y14028)
91
Natronorubrum sp. Tenzan-10 (AB046927)
Natronobacterium tibetense (AB005656)
100
Natrialba asiatica aidin-3 (AB046875)
Natrialba aegyptiaca (AF251941)
Natrialba asiatica B1T (D14124)
Natrialba asiatica 172P1 (D14123)
96
54
Haloterrigena turkmenicus (AB004878)
70
Haloterrigena thermotolerans PR5 (AF115478)
Haloterrigena thermotolerans H13 (AY055733)
Natrinema versiforme (AB023426)
Natrinema sp. HM06 (AY232296)
61
Halobacteriaceae gen. sp. (AF027738)
Natrinema sp. R-fish (AF367370)
63
61
Halobacteriaceae gen. sp. NCIMB 786 (AJ002947)
64
Halobacteriaceae gen. sp. T5.7 (AJ002946)
Halobacteriaceae gen. sp. NCIMB 784 (AJ002948)
Halobacteriaceae gen. sp. NCIMB 777 (AJ002949)
Halobacteriaceae gen. sp. (D14125)
63
78
Halophilic archaeum L-11 (D14126)
87
Natrinema sp. XA3-1 SSU-D
67
Natrinema sp. XA3-1 SSU-C
Natrinema sp. XA3-1 SSU-B
Natrinema aibiensis (AY277583)
Natrinema altuensis (AY277584)
Natrinema ajinwuensis (AY208972)
100
Haloterrigena sp. arg-4 (AB009624)
Natrinema sp. XA3-1 SSU-A
91
Halobacteriaceae gen. sp. SR1.5 (AJ002945)
Halobacteriaceae gen. sp. H1 (AF478471)
100
Natrialba sp. Tunisia HMg-25 (AB049461)
Natrialba sp. Tunisia HMg-27 (AB049462)
Natronobacterium gregoryi (D87970)
Natronococcus occultus (Z28378)
Natronococcus amylolyticus (D43628)
Natronococcus xinjiangense (AF251285)
Natronobacterium sp. C231 (AB045012)
Natronococcus aibiensis (AY277581)
Halobacteriaceae gen. sp. 135 (AF333760)
Natronococcus altunensis (AY277582)
SSU
ML
GTR+I+Γ4
α = 0.5711
I = 0.6566
1408 positions
1 substitution

Figure 4.5 B) Best maximum likelihood tree of the Halobacteriales subgroup indicated by the black triangle in A (continued).

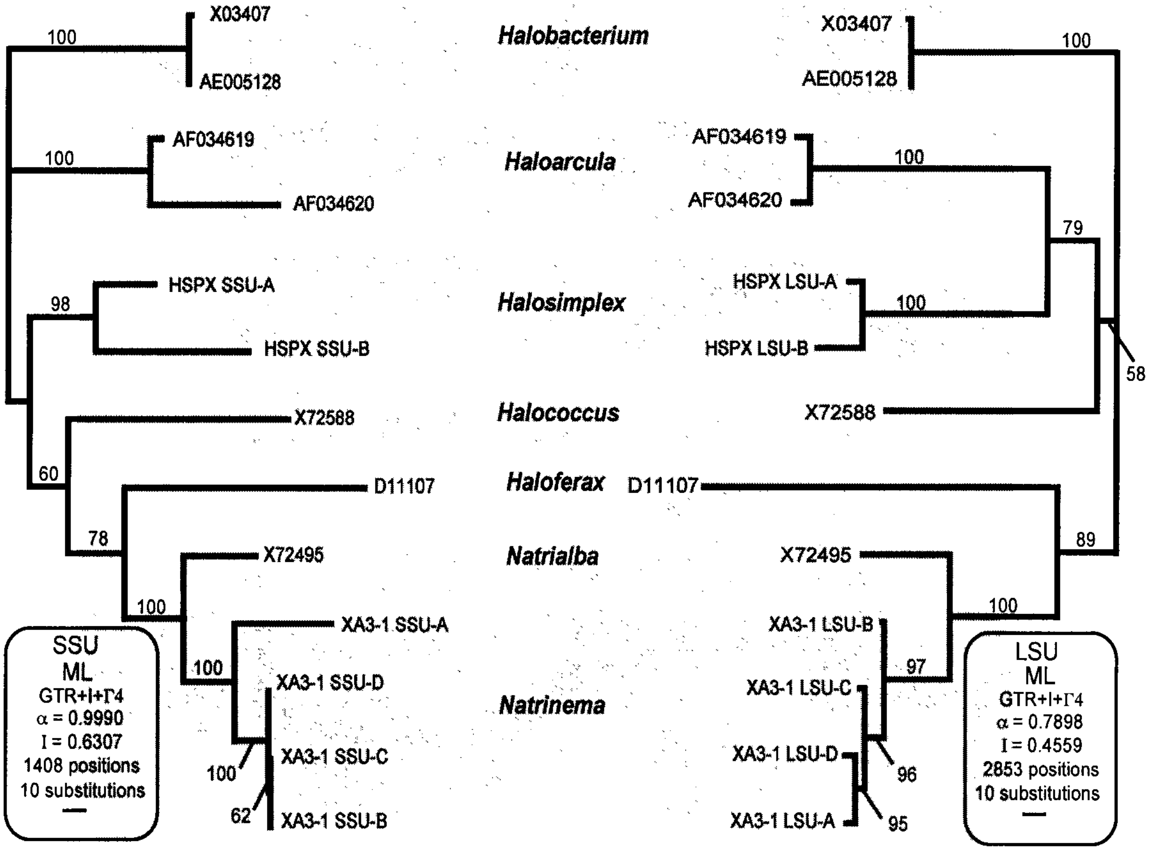


Figure 4.6 Comparison of best maximum likelihood trees for the SSU and LSU genes of all genera of Halobacteriales for which an both gene sequences were available. The evolutionary models and parameters used for the phylogenetic analyses are presented in boxes at the base of the trees (ML: maximum likelihood, GTR: general time reversible, I: proportion of invariable sites, α: gamma rate shape parameter alpha, Γ: number of gamma distribution rate categories). Bootstrap values represent the consensus of maximum likelihood trees from 100 pseudo-replicates.

Discussion

The high intraspecific heterogeneity observed here in three different genera of Halobacteriales clearly poses a problem to the use of SSU and LSU as a reliable classification tool for this archaeal order. The degree of intraspecific variability observed here (5-7% in SSU) is far greater than the average between-species divergence (generally around 1.5-2.0, although it can be as low as 0% or over 10%) and is more in the range of the divergence usually found between genera in the Halobacteriales order (~5-10%). One example of taxonomic confusion caused by intraspecific SSU heterogeneity is the status of the genera *Haloterrigena* and *Natrinema*. As discussed earlier, high similarity is observed between one of *Natrinema sp.* XA3-1 SSU genes and one SSU gene sequence PCR-amplified from *Haloterrigena sp.* arg-4. This suggests that this "*Haloterrigena*" strain could in fact be a *Natrinema* harbouring multiple heterogeneous rRNA operons, only one of which has been obtained by PCR. This could also be the case for other species of *Haloterrigena*, which are close neighbors of the *Natrinema* in phylogenetic analyses (Figure 4.5B). Furthermore, as Halobacteriales SSU and LSU genes are almost always sequenced from PCR-amplified DNA fragments, many more genera than the three identified here could harbour intraspecific heterogeneity. Excluding the taxa discussed earlier (*Natrinema sp.* XA3-1, *N. versiforme*, *H. carlbadense* and most species of the genera *Haloarcula*), *Halobacterium sp.* NRC-1 is the only haloarchaeon for which reliable information about potential rRNA genes intraspecific heterogeneity is available, and its complete genome sequence has shown that it harbours only one rRNA operon.

The similarity observed in a 100 bp stretch between one of *Natrinema sp.* XA3-1 LSU genes and *Natrialba magadii* LSU gene, which must be due to an interspecies recombination event, suggests an origin for intraspecific rRNA heterogeneity. As the other heterogeneous regions of XA3-1 LSU genes are also restrained to 100-200 bp stretches and found in only one or two of its rRNA operons, it is likely that they also originated by recombination (although this cannot be confirmed unless the source of the recombined fragment is identified). The source of recombined fragments is often very difficult to identify for several reasons. Recombination usually occurs in hypervariable regions, which means that the sequence of the recombined fragment will rapidly diverge from its source sequence, making it hard to detect by eye or using software (which will be unable to find a statistically significant match with the source). The limited number of rRNA gene sequences available in databases for Halobacteriales, compared to the diversity found in nature, also makes the search for a recombination donor difficult (especially for the LSU gene, of which less then a dozen sequences are available).

It is clear that the four LSU genes of *Natrinema sp.* XA3-1 did not evolve in coordination with their SSU and ITS1 partners. First, all LSU gene copies diverge from each other by roughly similar numbers of substitutions (0.9 to 1.9% of divergent nucleotide positions). In contrast, only one of XA3-1 SSU genes is really divergent from the others (at 5.0% of the positions), the other three diverging from each other at only one or two positions. ITS1, which is usually the most divergent part of rRNA operons (Perez Luz *et al.*, 1998), is here completely identical in all four operons. Obviously, gene conversion must operate between the four copies, as the fast-evolving ITS1 region is kept homogeneous. There are two possible explanations (which are not mutually exclusive) for such inequality in the evolutionary rate of different parts of XA3-1 rRNA operons: Gene conversion does not happen uniformly across the length of the operons and/or interspecies homologous recombination is frequent. Recombination between rRNA genes of different strains or species could indeed occur at a much higher levels than it was originally suspected. In addition to the between-genera case presented here, such recombination has been observed in several bacterial lineages at the species or subspecies levels (Anton *et al.*, 1999; Parker, 2001; Pereira *et al.*, 2001; Smith *et al.*, 1999). The highly evolutionarily conserved stretches of DNA found in rRNA genes, often presented as an advantage to trace phylogeny of organisms over large evolutionary distances, could also facilitate homologous recombination between divergent organisms (Woese, 2000).

The presence of rRNA intraspecific heterogeneity is unlikely to be an ancestral feature of haloarchaea. This phenomenon has been detected in only three lineages so far and it seems unlikely that heterogeneity would be lost in all other and only conserved in those. More importantly, rRNA genes are still monophyletic within each of those three groups (*Haloarcula*, *Halosimplex* and *Haloterrigena/Natrinema*). This monophyly suggests that the heterogeneity originated in the respective ancestors of each of those clades. This is especially likely for *Haloarcula*, a genus for which all species have been shown to display rRNA heterogeneity (Gemmell *et al.*, 1998). What would be the source of this heterogeneity? A divergent rRNA operon acquired by LGT or a duplication of the rRNA operon in the ancestor of each lineage followed by divergence of the paralogs? It is difficult to distinguish between those possibilities, which are not necessarily mutually exclusive. The identification of clear traces of a between-genera recombination event in the LSU of *Natrinema sp.* XA3-1 suggests that the latter process played a role either in the origin or maintenance of this heterogeneity.

Could the intraspecific heterogeneity of rRNA operons be maintained by evolutionary pressure? Such a link between functionality and intraspecific rRNA divergence has been observed in the apicomplexa *Plasmodium berghei*. The two types of SSU genes (which differ at 5.0% of their nucleotide positions) are preferentially expressed in different stages of the life cycle of this eukaryotic parasite (Gunderson *et al.*, 1987). It has been suggested that in extremely halophilic archaea, selective advantage could be gained from the differential expression of divergent rRNA operons dependent on the salt concentration in the environment (Dennis *et al.*, 1998). Indeed, salinity has a significant influence on most biochemical reactions and is very variable in the environments occupied by halophiles, being subjected to constant fluctuations caused by solubilization-precipitation and dilution-evaporation (Dennis and Shimmin, 1997). There is even some experimental evidence of salinity-dependence for rRNA expression. Indeed, the promoters used for the expression of the unique rRNA operon of *Halobacterium cutirubrum* vary according to the salt concentration in which the organism is grown (Dennis, 1999). Although both rRNA operons of *H. marismortui* have been shown to be expressed under standard laboratory growth conditions (Amann *et al.*, 2000), looking at their differential expression under variable growth conditions has yet to be attempted.

Intraspecific heterogeneity, as we just discussed, causes problems at the level of rRNA data analysis {see also (Clayton *et al.*, 1995)}. More importantly, it can also cause artifacts at the data acquisition level. Indeed, PCR-amplification is known to be susceptible to the formation of chimeric products if the template DNA contains multiple divergent copies of the target gene (Qiu *et al.*, 2001). A certain rate of chimera formation is known to occur when the SSU gene is amplified directly from environmental DNA samples, which usually contain a great diversity of the target gene (von Wintzingerode *et al.*, 1997). Intraspecific heterogeneity means that this type of artifact can also occur for PCR-amplification of rRNA genes using DNA extracted from a pure culture of an organism. Although PCR-conditions can be modified to reduce the risk of chimera formation, it can never be completely excluded. The vast majority of the SSU gene sequences of Halobacteriales available in the database have been amplified by PCR without knowledge of the possibility of intraspecific heterogeneity occurring in other lineages than the genera *Haloarcula*. Several of the sequences found in the database could therefore be chimeric, warranting caution in their use for phylogenetic analysis.

The presence of intraspecific heterogeneity, at least partly caused by homologous recombination of rRNA genes, can have a significant impact on the acquisition of true (non-chimeric) gene sequences and their use in identification and classification of organisms through phylogenetic analysis. Therefore, efforts towards development of non-PCR methods of acquiring rRNA genes and the evaluation of the frequency of intraspecific heterogeneity and homologous recombination among prokaryotes are needed to ensure the reliable use of these genes as molecular markers in microbiological disciplines.

Conclusion

As I have already discussed individual aspects of LGT in the different chapters of this thesis, I will not discuss them at length about them here. I will rather describe what I think are the major questions yet unanswered concerning LGT, as well as the approaches that should be taken toward their elucidation.

How much can we know about the tree of life?

LGT, just as spontaneous generation did hundreds of years ago, puts into question our classification scheme of prokaryotes. The holy grail of prokaryotic systematists has always been to trace the evolutionary history of major groups of prokaryotes, from their origin, through their separation from other lineages to their differentiation into multiple minor groups. This was first dreamed of by Darwin when he set up the foundations of the theory of evolution by natural selection. The feasibility of a major classification of organisms was first questioned by the possibility of spontaneous generation, which would mean that not one, but several trees of life do exist (Wallace, 1872). After Thomas Huxley pushed back the occurrence of spontaneous generation to ancient times (Farley, 1972a), it was thought possible to classify microbes based on their morphological, and later, physiological (Buchanan, 1925), characteristics. But it eventually appeared, as Stanier and Van Niel voiced it, that such characteristics were too variable to be used to construct a natural (phylogenetic) scheme of classification (Stanier and van Niel, 1941). The discovery of molecular markers that could be used to uncover the phylogenetic relationships between organisms, in a more reliable way than physiological characteristics do, rekindled hopes for an organismal tree of life (Woese, 1987). However, it became rapidly evident that, although reasonable results could be obtained to classify relatively closely related organisms, methodological artifacts and lack of phylogenetic signal in genes could prevent us from discovering relationships between major groups (Gribaldo and Philippe, 2002). LGT came to further complicate the picture, as it limits the number of genes usable for the purpose of reconstructing distant phylogenetic relationships, since only markers having a relatively linear (vertical) evolution would be useful in this endeavor (Doolittle, 1999b). The presence of LGT also means that the information retrieved by these markers has to be interpreted with care, because it represents the history of a core of genes, not of a whole genome. This means that the best that can be achieved without making philosophical assumptions are gene trees, rather than organismal trees (Doolittle, 1999b).

Can we get a reliable picture of prokaryotic evolution through gene trees alone? This depends on how many genes have significant phylogenetic signal, low levels of paralogy, and limited amounts of lateral transfer. Also, certain genes will have a tractable evolutionary history only at certain phylogenetic levels, depending on the factors mentioned above and their rate of evolution. A systematic study of how individual genes have been affected by these factors would allow an estimation of how much of prokaryotic evolution we can reliably reconstruct. Refinement of automated methods for ortholog identification, multiple gene alignment, gene alignment editing, tree reconstruction and calculation of statistical support would allow for a thorough phylogenetic analysis of a maximum number of genes (Sicheritz-Ponten and Andersson, 2001). Phylogenetic analysis could also be made more reliable by getting a more representative sampling of natural diversity. Such sampling is unlikely to be achieved through complete genome sequencing alone. However, techniques such as environmental DNA libraries sequencing and draft genome sequencing have a high enough processivity to allow for a more elaborate sampling of gene diversity.

What is a prokaryotic species?

The idea of some coherent unit of evolution gave birth to the biological species concept, first expressed by Ernst Mayr for sexual species (Mayr, 1982). Since then, several of species concepts have been put forward, both for sexual and clonal organisms (Rossello-Mora and Amann, 2001). LGT however, makes prokaryotes neither clonal nor sexual. It has also become obvious that various groups of prokaryotes cannot be described using the same basic criteria. Polyphasic approaches to define prokaryotic species, involving phylogenetic information from the SSU gene, DNA-DNA hybridization and phenotypic analysis, have been proposed to delineate what is part of a particular species and what is not (Gillis *et al.*, 2001). However, the range of each of these characteristics varies greatly among different "species" of prokaryotes. For example, a species is generally defined by 70% DNA-DNA hybridization, but in many bacterial families, members of a single species can share DNA hybridization values of 40-100% (Gillis *et al.*, 2001). Also, SSU divergence is not always correlated with genome content variation (Clayton *et al.*, 1995; Fox *et al.*, 1992). Many factors could be at the source of this dissociation: variation in the SSU evolutionary rate, gene acquisition by LGT, recombination occurring in the SSU gene, intraspecific heterogeneity of SSU. Although we are aware of the existence of these phenomena, we do not know much about their range or prevalence, even in model organisms. A more diverse (covering a large phylogenetic diversity) and thorough (involving many closely related organisms) sampling would help us determine important factors such as the frequency of intraspecific heterogeneity and homologous recombination of rRNA genes as well as the variability in genome content. It is not only the use of massive sequencing by large genomic facilities that can allow us to accumulate the vast amount of information required, but also more directed strategies, focusing on a specific question. For example, genome content variability can be tackled through the use of techniques such as subtractive-suppressive hybridization, long-walk PCR or lambda libraries of genomic DNA, which require only moderate sequencing capacities (Nesbo *et al.*, 2002). Questions about rRNA gene evolution can be investigated through methods targeting these genes, such as the construction of rDNA libraries from environmental DNA based on the LSU-specific recognition sequence of some intron-homing endonucleases (Lucas *et al.*, 2001).

The importance of homologous recombination in creating genetic diversity

It was long thought that the vast majority of the genetic diversity found in prokaryotes came from point mutations (Dykhuizen and Green, 1991). Studies in the last decade revealed that in several species of bacteria, homologous recombination was responsible for as many if not more nucleotide substitutions than point mutations (Feil *et al.*, 1999; Feil *et al.*, 2000; Feil *et al.*, 2001; Guttman and Dykhuizen, 1994). The importance of homologous recombination versus point mutation also seems to vary from organism to organism, such as the relative importance of these two diversifying forces varies a great deal across species (Smith *et al.*, 1993). The sampling, however, is still very limited, and overwhelmingly comprised of pathogenic bacteria, most of them low G+C Gram-positive cocci. The reason as to why point mutation seems to be prevalent in some organisms while homologous recombination dominates in others has yet to be linked to particular characteristics of the organisms in question. What is required here is for the method used to calculate those recombination/point mutation ratios (multi-locus sequence typing) to be applied to a greater variety of prokaryotes, more importantly ones with different environments and physiological characteristics.

What really happens in the environment?

Vectors for lateral transfer, such as viruses, transposons and mobile plasmids, were identified a long time ago. However, little is known of their actual importance in the environment. It is hypothesized that each species of prokaryote might have about ten specific types of viruses, but very few bacteriophages have been described so far relative to this potential diversity (Rohwer, 2003). If these predictions about the abundance of phages hold true, transduction could be a major vector by which genes are transmitted between prokaryotes. Another phenomenon that has been little studied in the environment is the natural competence of prokaryotes. Competence of prokaryotes in nature could be much more prevalent than we think. Some experiments have showed that *E. coli* (which does not naturally take up DNA under laboratory conditions) could be competent in freshwater (Baur *et al.*, 1996). This suggests that several organisms, normally impervious to free DNA in the laboratory, could be competent under some conditions in nature. The frequency of free DNA elements such as gene cassettes in their circular form, which can easily be integrated and expressed in integrons, is also completely unknown (Holmes *et al.*, 2003). If such elements were common in the environment, they would play a major role in gene-flux between prokaryotes. Determination of the frequency of these different vectors for LGT in nature is required for a more complete picture of the role of lateral gene transfer in prokaryotic evolution.

References

Adachi, J.M., Hasegawa M. (1996) MOLPHY: Programs for Molecular Phylogenetics. Tokyo: Institute of Statistical Mathematics.

Adam, R.D. (1991) The biology of *Giardia* spp. *Microbiol Rev* **55**: 706-732.

Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F., and Trust, T.J. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* **397**: 176-180.

Alm, R.A., and Trust, T.J. (1999) Analysis of the genetic diversity of Helicobacter pylori: the tale of two genomes. *J Mol Med* **77**: 834-846.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., 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.

Amabile-Cuevas, C.F., and Chicurel, M.E. (1992) Bacterial plasmids and gene flux. *Cell* **70**: 189-199.

Amann, G., Stetter, K.O., Llobet-Brossa, E., Amann, R., and Anton, J. (2000) Direct proof for the presence and expression of two 5% different 16S rRNA genes in individual cells of Haloarcula marismortui. *Extremophiles* **4**: 373-376.

Amann, R., Fuchs, B.M., and Behrens, S. (2001) The identification of microorganisms by fluorescence in situ hybridisation. *Curr Opin Biotechnol* **12**: 231-236.

Ambler, R.P., Daniel, M., Hermoso, J., Meyer, T.E., Bartsch, R.G., and Kamen, M.D. (1979a) Cytochrome c2 sequence variation among the recognised species of purple nonsulphur photosynthetic bacteria. *Nature* **278**: 659-660.

Ambler, R.P., Daniel, M., Meyer, T.E., Bartsch, R.G., and Kamen, M.D. (1979b) The amino acid sequence of cytochrome c' from the purple sulphur bacterium Chromatium vinosum. *Biochem J* **177**: 819-823.

Andersson, E.S. (1966) Possible Importance of Transfer Factors in Bacterial Evolution. *Nature* **209**: 637-638.

Andersson, J.O., and Roger, A.J. (2002) Evolutionary analyses of the small subunit of glutamate synthase: gene order conservation, gene fusions, and prokaryote-to-eukaryote lateral gene transfers. *Eukaryot Cell* **1**: 304-310.

Anton, A.I., Martinez-Murcia, A.J., and Rodriguez-Valera, F. (1999) Intraspecific diversity of the 23S rRNA gene and the spacer region downstream in Escherichia coli. *J Bacteriol* **181**: 2703-2709.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus Type III. *Journal of experimental medicine* **79**: 137-158.

Bamford, D.H. (2003) Do viruses form lineages across different domains of life? *Res Microbiol* **154**: 231-236.

Bandelt, H.J., and Dress, A.W. (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* **1**: 242-252.

Banthorpe, D.V., Charlwood, B.V., and Francis, M.J. (1972) The biosynthesis of monoterpenes. *Chem Rev* **72**: 115-155.

Barton, L.L., and Tomei, F.A. (1995) Characteristics and Activities of Sulfate-reducing Bacteria. In *Sulfate-Reducing Bacteria.* Barton, L.L. (ed). New York: Plenum Press, pp. 1-32.

Baur, B., Hanselmann, K., Schlimme, W., and Jenni, B. (1996) Genetic transformation in freshwater: Escherichia coli is able to develop natural competence. *Appl Environ Microbiol* **62**: 3673-3678.

Baymann, F., Brugna, M., Muhlenhoff, U., and Nitschke, W. (2001) Daddy, where did (PS)I come from? *Biochim Biophys Acta* **1507**: 291-310.

Bedard, C., and Knowles, R. (1989) Physiology, biochemistry, and specific inhibitors of CH4, NH4+, and CO oxidation by methanotrophs and nitrifiers. *Microbiol Rev* **53**: 68-84.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**: 1520-1523.

Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., Jovanovich, S.B., Gates, C.M., Feldman, R.A., Spudich, J.L., Spudich, E.N., and DeLong, E.F. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.

Beytia, E.D., and Porter, J.W. (1976) Biochemistry of polyisoprenoid biosynthesis. *Annu Rev Biochem* **45**: 113-142.

Bishop, P., and Premakumar, R. (1992) Alternative Nitrogen Fixation Systems. In *Biological Nitrogen Fixation.* Stacey, G., Burris, R. and Evans, H. (eds). London: Chapman and Hall, pp. 736-762.

Blankenship, R.E. (1992) Origin and early evolution of photosynthesis. *Photosynth Res* **33**: 91-111.

Bochar, D.A., Stauffacher, C.V., and Rodwell, V.W. (1999) Sequence comparisons reveal two classes of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Mol Genet Metab* **66**: 122-127.

Bond, J.P., and Francklyn, C. (2000) Proteobacterial histidine-biosynthetic pathways are paraphyletic. *J Mol Evol* **50**: 339-347.

Boucher, Y., and Doolittle, W.F. (2000) The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol Microbiol* **37**: 703-716.

Britton, G., Goodwin, T.W., Lockley, W.J.S., Mundy, A.P., and Patel, N.J. (1979) Stereochemistry of Cyclization in Carotenoid Biosynthesis: Use of ^{13}C-Labelling to Elucidate the Stereochemical Behaviour of the C-1 Methyl Substituents during Zeaxanthin Biosynthesis in a *Flavobacterium*. *J Chem Soc Chem Commun*: 27-28.

Brochier, C., Bapteste, E., Moreira, D., and Philippe, H. (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18**: 1-5.

Brock, T.D. (1990) *The emergence of bacterial genetics.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Brown, T. (1993) Analysis of DNA sequences by blotting and hydridization. In *Current protocols in molecular biology.* Vol. 1. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.A., Struhl, K. and Smith, J.A. (eds). New York: John Wiley and Sons, pp. 2.9.1-2.9.15.

Buchanan, R.E. (1925) *General Systematic Bacteriology.* Baltimore: Waverly Press.

Campbell, A. (2003) The future of bacteriophage biology. *Nat Rev Genet* **4**: 471-477.

Campos, N., Rodriguez-Concepcion, M., Seemann, M., Rohmer, M., and Boronat, A. (2001) Identification of gcpE as a novel gene of the 2-C-methyl-D-erythritol 4-phosphate pathway for isoprenoid biosynthesis in Escherichia coli. *FEBS Lett* **488**: 170-173.

Carranza, S., Giribet, G., Ribera, C., Baguna, and Riutort, M. (1996) Evidence that two types of 18S rDNA coexist in the genome of Dugesia (Schmidtea) mediterranea (Platyhelminthes, Turbellaria, Tricladida). *Mol Biol Evol* **13**: 824-832.

Castresana, J., Lubben, M., Saraste, M., and Higgins, D.G. (1994) Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. *Embo J* **13**: 2516-2525.

Castresana, J. (2001) Comparative genomics and bioenergetics. *Biochim Biophys Acta* **1506**: 147-162.

Castro, H.F., Williams, N.H., and Ogram, A. (2000) Phylogeny of sulfate-reducing bacteria. *FEMS Microbiol Ecol* **31**: 1-9.

Charbonnier, F., Forterre, P., Erauso, G., and Prieur, D. (1995) Purification of plasmids from thermophilic and hyperthermophilic archaea. In *Archaea: a laboratory manual*. Robb, F.T. and Place, A.R. (eds). Cold Spring Harbor: Cold Spring Harbor Laboratory Press, pp. 87-90.

Chen, A., and Poulter, C.D. (1993) Purification and characterization of farnesyl diphosphate/geranylgeranyl diphosphate synthase. A thermostable bifunctional enzyme from Methanobacterium thermoautotrophicum. *J Biol Chem* **268**: 11002-11007.

Chien, Y.T., and Zinder, S.H. (1996) Cloning, functional organization, transcript studies, and phylogenetic analysis of the complete nitrogenase structural genes (nifHDK2) and associated genes in the archaeon Methanosarcina barkeri 227. *J Bacteriol* **178**: 143-148.

Chien, Y.T., Auerbuch, V., Brabban, A.D., and Zinder, S.H. (2000) Analysis of genes encoding an alternative nitrogenase in the archaeon Methanosarcina barkeri 227. *J Bacteriol* **182**: 3247-3253.

Chistoserdova, L., Vorholt, J.A., Thauer, R.K., and Lidstrom, M.E. (1998) C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science* **281**: 99-102.

Cilia, V., Lafay, B., and Christen, R. (1996) Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol Biol Evol* **13**: 451-461.

Clayton, R.A., Sutton, G., Hinkle, P.S., Jr., Bult, C., and Fields, C. (1995) Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *Int J Syst Bacteriol* **45**: 595-599.

Collins, M.D., and Jones, D. (1981) Distribution of isoprenoid quinone structural types in bacteria and their taxonomic implication. *Microbiol Rev* **45**: 316-354.

Colwell, R.R., and Huq, A. (1994) Vibrios in the environment: viable but nonculturable *Vibrio cholerae*. In *Vibrio cholerae and Cholera: Molecular to Global Perspectives*. Wachsmuth, I.K., Blake, P.A. and Olsvik, O. (eds). Washington, D.C.: American Society for Microbiology, pp. 117-133.

Conolly, J.D., and Hill, R.A. (1992) *Dictionary of Terpenoids*. New-York.

Cooling, F.B., Maloney, C.L., Nagel, E., Tabinowski, J., and Odom, J.M. (1996) Inhibition of sulfate respiration by 1,8-dihydroxyanthraquinone and other anthraquinone derivatives. *Applied and Environmental Microbiology* **62**: 2999-3004.

Cox, M.M. (2001) Historical overview: searching for replication help in all of the rec places. *Proc Natl Acad Sci U S A* **98**: 8173-8180.

Dahl, C., Kredich, N.M., Deutzmann, R., and Truper, H.G. (1993) Dissimilatory Sulfite Reductase from Archaeoglobus-Fulgidus - Physicochemical Properties of the Enzyme and Cloning, Sequencing and Analysis of the Reductase Genes. *Journal of General Microbiology* **139**: 1817-1828.

Dahl, C., and Truper, H.G. (2001) Sulfite reductase and APS reductase from Archaeoglobus fulgidus. *Methods Enzymol* **331**: 427-441.

Dahllof, I., Baillie, H., and Kjelleberg, S. (2000) rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl Environ Microbiol* **66**: 3376-3380.

Dairi, T., Motohira, Y., Kuzuyama, T., Takahashi, S., Itoh, N., and Seto, H. (2000) Cloning of the gene encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase from terpenoid antibiotic-producing Streptomyces strains [In Process Citation]. *Mol Gen Genet* **262**: 957-964.

de Koning, A.P., Brinkman, F.S., Jones, S.J., and Keeling, P.J. (2000) Lateral gene transfer and metabolic adaptation in the human parasite Trichomonas vaginalis. *Mol Biol Evol* **17**: 1769-1773.

De Rijk, P., Van de Peer, Y., Van den Broeck, I., and De Wachter, R. (1995) Evolution according to large ribosomal subunit RNA. *J Mol Evol* **41**: 366-375.

De Rosa, M., Gambacorta, A., and Nicolaus, B. (1980) Regularity of isoprenoid biosynthesis in the ether lipids of acrhaebacteria. *Phytochemistry* **19**: 791-793.

De Rosa, M., Gambacorta, A., and Gliozzi, A. (1986) Structure, biosynthesis, and physicochemical properties of archaebacterial lipids [published erratum appears in Microbiol Rev 1987 Mar;51(1):178]. *Microbiol Rev* **50**: 70-80.

De Rosa, M., and Gambacorta, A. (1988) The lipids of archaebacteria. *Prog Lipid Res* **27**: 153-175.

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.

Demaneche, S., Bertolla, F., Buret, F., Nalin, R., Sailland, A., Auriol, P., Vogel, T.M., and Simonet, P. (2001) Laboratory-scale evidence for lightning-mediated gene transfer in soil. *Appl Environ Microbiol* **67**: 3440-3444.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., Radman, M., and Matic, I. (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**: 711-721.

Dennis, P.P., and Shimmin, L.C. (1997) Evolutionary divergence and salinity-mediated selection in halophilic archaea. *Microbiol Mol Biol Rev* **61**: 90-104.

Dennis, P.P., Ziesche, S., and Mylvaganam, S. (1998) Transcription analysis of two disparate rRNA operons in the halophilic archaeon Haloarcula marismortui. *J Bacteriol* **180**: 4804-4813.

Dennis, P.P. (1999) Expression of Ribosomal RNA Operons in Halophilic Archaea. In *Microbiology and Biogeochemistry of Hypersaline Environments*. Oren, A. (ed). New York: CRC Press, pp. 319-329.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R.A., Martinez-Arias, R., Henne, A., Wiezer, A., Baumer, S., Jacobi, C., Bruggemann, H., Lienard, T., Christmann, A., Bomeke, M., Steckel, S., Bhattacharyya, A., Lykidis, A., Overbeek, R., Klenk, H.P., Gunsalus, R.P., Fritz, H.J., and Gottschalk, G. (2002) The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* **4**: 453-461.

Diruggiero, J., Dunn, D., Maeder, D.L., Holley-Shanks, R., Chatard, J., Horlacher, R., Robb, F.T., Boos, W., and Weiss, R.B. (2000) Evidence of recent lateral gene transfer among hyperthermophilic archaea. *Mol Microbiol* **38**: 684-693.

Disch, A., Schwender, J., Muller, C., Lichtenthaler, H.K., and Rohmer, M. (1998) Distribution of the mevalonate and glyceraldehyde phosphate/pyruvate pathways for isoprenoid biosynthesis in unicellular algae and the cyanobacterium Synechocystis PCC 6714. *Biochem J* **333**: 381-388.

Doolittle, W.F. (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14**: 307-311.

Doolittle, W.F. (1999a) Lateral genomics. *Trends Cell Biol* **9**: M5-8.

Doolittle, W.F. (1999b) Phylogenetic classification and the universal tree. *Science* **284**: 2124-2129.

Doolittle, W.F. (2000a) The nature of the universal ancestor and the evolution of the proteome. *Curr Opin Struct Biol* **10**: 355-358.

Doolittle, W.F. (2000b) Uprooting the tree of life. *Sci Am* **282**: 90-95.

Doolittle, W.F., Boucher, Y., Nesbo, C.L., Douady, C.J., Andersson, J.O., and Roger, A.J. (2003) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos Trans R Soc Lond B Biol Sci* **358**: 39-57; discussion 57-38.

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

Eisenreich, W., Schwarz, M., Cartayrade, A., Arigoni, D., Zenk, M.H., and Bacher, A. (1998) The deoxyxylulose phosphate pathway of terpenoid biosynthesis in plants and microorganisms. *Chem Biol* **5**: R221-233.

Eisenreich, W., Bacher, A., Berry, A., Bretzel, W., Humbelin, M., Lopez-Ulibarri, R., Mayer, A.F., and Yeliseev, A. (2002) Biosynthesis of zeaxanthin via mevalonate in Paracoccus species strain PTA-3335. A product-based retrobiosynthetic study. *J Org Chem* **67**: 871-875.

Emmerth, M., Goebel, W., Miller, S.I., and Hueck, C.J. (1999) Genomic subtraction identifies Salmonella typhimurium prophages, F-related plasmid sequences, and a novel fimbrial operon, stf, which are absent in Salmonella typhi. *J Bacteriol* **181**: 5652-5661.

Evans, H., and Burris, R. (1992) Highlights in Biological Nitrogen Fixation during the Last 50 Years. In *Biological Nitrogen Fixation*. Stacey, G., Burris, R. and Evans, H. (eds). London: Chapman and Hall, pp. 1-42.

Faguy, D.M., and Doolittle, W.F. (2000) Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. *Trends Genet* **16**: 196-197.

Farley, J. (1972a) The spontaneous generation controversy (1859-1880): British and German reactions to the problem of abiogenesis. *J Hist Biol* **5**: 285-319.

Farley, J. (1972b) The spontaneous generation controversy (1700-1860): The origin of parasitic worms. *J Hist Biol* **5**: 95-125.

Fauque, G.D. (1995) Ecology of Sulfate-reducing Bacteria. In *Sulfate-Reducing Bacteria*. Barton, L.L. (ed). New York: Plenum Press, pp. 217-241.

Feil, E.J., Maiden, M.C., Achtman, M., and Spratt, B.G. (1999) The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Mol Biol Evol* **16**: 1496-1502.

Feil, E.J., Smith, J.M., Enright, M.C., and Spratt, B.G. (2000) Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. *Genetics* **154**: 1439-1450.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., Zhou, J., and Spratt, B.G. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**: 182-187.

Felsenstein, J. (1993) PHYLIP, (Phylogeny Inference Package). Seattle: University of Washington.

Field, J., Rosenthal, B., and Samuelson, J. (2000) Early lateral transfer of genes encoding malic enzyme, acetyl-CoA synthetase and alcohol dehydrogenases from anaerobic prokaryotes to Entamoeba histolytica. *Mol Microbiol* **38**: 446-455.

Figge, R.M., Schubert, M., Brinkmann, H., and Cerff, R. (1999) Glyceraldehyde-3-phosphate dehydrogenase gene diversity in eubacteria and eukaryotes: evidence for intra- and inter-kingdom gene transfer. *Mol Biol Evol* **16**: 429-440.

Filee, J., Forterre, P., and Laurent, J. (2003) The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res Microbiol* **154**: 237-243.

Finan, T.M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorholter, F.J., Hernandez-Lucas, I., Becker, A., Cowie, A., Gouzy, J., Golding, B., and Puhler, A. (2001) The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont Sinorhizobium meliloti. *Proc Natl Acad Sci U S A* **98**: 9889-9894.

Forterre, P., Brochier, C., and Philippe, H. (2002) Evolution of the Archaea. *Theor Popul Biol* **61**: 409-422.

Foster, B.A., Thomas, S.M., Mahr, J.A., Renosto, F., Patel, H.C., and Segel, I.H. (1994) Cloning and Sequencing of Atp Sulfurylase from Penicillium-Chrysogenum - Identification of a Likely Allosteric Domain. *Journal of Biological Chemistry* **269**: 19777-19786.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P., Jr. (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**: 166-170.

Friedrich, M.W. (2002) Phylogenetic analysis reveals multiple lateral transfers of adenosine-5'-phosphosulfate reductase genes among sulfate-reducing microorganisms. *J Bacteriol* **184**: 278-289.

Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R.D., Cann, I., Graham, D.E., Grahame, D.A., Guss, A.M., Hedderich, R., Ingram-Smith, C., Kuettner, H.C., Krzycki, J.A., Leigh, J.A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J.N., Smith, K., Springer, T.A., Umayam, L.A., White, O., White, R.H., Conway de Macario, E., Ferry, J.G., Jarrell, K.F., Jing, H., Macario, A.J., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W., and Birren, B. (2002) The genome of M. acetivorans reveals extensive metabolic and physiological diversity. *Genome Res* **12**: 532-542.

Garcia-Vallve, S., Romeu, A., and Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**: 1719-1725.

Gemmell, R.T., McGenity, T.J., and Grant, W.D. (1998) Use of molecular techniques to investigate possible long-term dormancy of halobacteria in ancient halite deposits. *Ancient Biomolecules* **2**: 125-133.

Gennis, R.B., and Stewart, V. (1996) Respiration. In *Escherichia coli and Salmonella, cellular and molecular biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: ASM Press, pp. 217-261.

Gerdes, K., Moller-Jensen, J., and Bugge Jensen, R. (2000) Plasmid and chromosome partitioning: surprises from phylogeny. *Mol Microbiol* **37**: 455-466.

Gill, J.F., Jr., Beach, M.J., and Rodwell, V.W. (1985) Mevalonate utilization in Pseudomonas sp. M. Purification and characterization of an inducible 3-hydroxy-3-methylglutaryl coenzyme A reductase. *J Biol Chem* **260**: 9393-9398.

Gillis, M., Vandamme, P., de Vos, P., Swings, J., and Kersters, K. (2001) Polyphasic taxonomy. In *Bergey's Manual of Systematic Bacteriology*. Vol. 1. Garrity, G.M. (ed). New York: Springer-Verlag.

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

Goldstein, J.L., and Brown, M.S. (1990) Regulation of the mevalonate pathway. *Nature* **343**: 425-430.

Gough, D.P., Kirby, A.L., Richards, J.B., and Hemming, F.W. (1970) The characterization of undecaprenol of *Lactobacillus plantarum. Biochem J* **118**: 167-170.

Graham, D.E., and White, R.H. (2002) Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. *Nat Prod Rep* **19**: 133-147.

Gribaldo, S., Lumia, V., Creti, R., de Macario, E.C., Sanangelantoni, A., and Cammarano, P. (1999) Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. *J Bacteriol* **181**: 434-443.

Gribaldo, S., and Philippe, H. (2002) Ancient phylogenetic relationships. *Theor Popul Biol* **61**: 391-408.

Gunderson, J.H., Sogin, M.L., Wollett, G., Hollingdale, M., de la Cruz, V.F., Waters, A.P., and McCutchan, T.F. (1987) Structurally distinct, stage-specific ribosomes occur in Plasmodium. *Science* **238**: 933-937.

Guttman, D.S., and Dykhuizen, D.E. (1994) Clonal divergence in Escherichia coli as a result of recombination, not mutation. *Science* **266**: 1380-1383.

Hefter, J., Richnow, H.H., Fischer, U., Trendel, J.M., and Michaelis, W. (1993) (-)-Verrocusane-2b-ol from the phototrophic bacterium *Chloroflexus aurantiacus*: first report of a verrocusane-type diterpenoid from a prokaryote. *J Gen Microbiol* **139**: 2757-2761.

Henninger, T., Anemuller, S., Fitz-Gibbon, S., Miller, J.H., Schafer, G., and Schmidt, C.L. (1999) A novel Rieske iron-sulfur protein from the hyperthermophilic crenarchaeon Pyrobaculum aerophilum: sequencing of the gene, expression in E. coli and characterization of the protein. *J Bioenerg Biomembr* **31**: 119-128.

Herd, M., and Kocks, C. (2001) Gene fragments distinguishing an epidemic-associated strain from a virulent prototype strain of Listeria monocytogenes belong to a distinct functional subset of genes and partially cross-hybridize with other Listeria species. *Infect Immun* **69**: 3972-3979.

Herz, S., Wungsintaweekul, J., Schuhr, C.A., Hecht, S., Luttgen, H., Sagner, S., Fellermeier, M., Eisenreich, W., Zenk, M.H., Bacher, A., and Rohdich, F. (2000) Biosynthesis of terpenoids: YgbB protein converts 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate to 2C-methyl-D-erythritol 2,4-cyclodiphosphate. *Proc Natl Acad Sci U S A* **97**: 2486-2490.

Hilario, E., and Gogarten, J.P. (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems* **31**: 111-119.

Hipp, W.M., Pott, A.S., Thum-Schmitz, N., Faath, I., Dahl, C., and Truper, H.G. (1997) Towards the phylogeny of APS reductases and sirohaem sulfite reductases in sulfate-reducing and sulfur-oxidizing prokaryotes. *Microbiology* **143 (Pt 9)**: 2891-2902.

Hirsch, A.M., McKhann, H.I., Reddy, A., Liao, J., Fang, Y., and Marshall, C.R. (1995) Assessing horizontal transfer of nifHDK genes in eubacteria: nucleotide sequence of nifK from Frankia strain HFPCcI3. *Mol Biol Evol* **12**: 16-27.

Hoeffler, J.F., Tritsch, D., Grosdemange-Billiard, C., and Rohmer, M. (2002) Isoprenoid biosynthesis via the methylerythritol phosphate pathway. Mechanistic investigations of the 1-deoxy-D-xylulose 5-phosphate reductoisomerase. *Eur J Biochem* **269**: 4446-4457.

Holmes, A.J., Gillings, M.R., Nield, B.S., Mabbutt, B.C., Nevalainen, K.M., and Stokes, H.W. (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* **5**: 383-394.

Hooper, S.D., and Berg, O.G. (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol* **54**: 365-375.

Horbach, S., Sahm, H., and Welle, R. (1993) Isoprenoid biosynthesis in bacteria: two different pathways? *FEMS Microbiol Lett* **111**: 135-140.

Huang, X., and Miller, W. (1991) A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**: 337-357.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity [published erratum appears in J Bacteriol 1998 Dec;180(24):6793]. *J Bacteriol* **180**: 4765-4774.

Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**: REVIEWS0003.

Humbelin, M., Thomas, A., Lin, J., Li, J., Jore, J., and Berry, A. (2002) Genetics of isoprenoid biosynthesis in Paracoccus zeaxanthinifaciens. *Gene* **297**: 129-139.

Hurek, T., Egener, T., and Reinhold-Hurek, B. (1997) Divergence in nitrogenases of Azoarcus spp., Proteobacteria of the beta subclass. *J Bacteriol* **179**: 4172-4178.

Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K., and Nagashima, K.V. (2001) Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J Mol Evol* **52**: 333-341.

Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y., and Mukohata, Y. (1999) Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *J Mol Biol* **285**: 163-174.

Iordanescu, S. (1993) Characterization of the Staphylococcus aureus chromosomal gene pcrA, identified by mutations affecting plasmid pT181 replication. *Mol Gen Genet* **241**: 185-192.

Jain, R., Rivera, M.C., and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96**: 3801-3806.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., Soldati, D., and Beck, E. (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**: 1573-1576.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., and Stephens, R.S. (1999) Comparative genomes of Chlamydia pneumoniae and C. trachomatis. *Nat Genet* **21**: 385-389.

Kamekura, M., and Kates, M. (1999) Structural diversity of membrane lipids in members of Halobacteriaceae. *Biosci Biotechnol Biochem* **63**: 969-972.

Kaneda, K., Kuzuyama, T., Takagi, M., Hayakawa, Y., and Seto, H. (2001) An unusual isopentenyl diphosphate isomerase found in the mevalonate pathway gene cluster from Streptomyces sp. strain CL190. *Proc Natl Acad Sci U S A* **98**: 932-937.

Kates, M., and Kushwaha, N. (1978) Biochemistry of the lipids of extremely halophilic bacteria. In *Energetics and structure of Halophilic microorganisms*. Caplan, S.R. and Ginzburg, M. (eds). Amsterdam: Elsevier, pp. 461-480.

Katz, L.A. (1996) Transkingdom transfer of the phosphoglucose isomerase gene. *J Mol Evol* **43**: 453-459.

Ke, D., Boissinot, M., Huletsky, A., Picard, F.J., Frenette, J., Ouellette, M., Roy, P.H., and Bergeron, M.G. (2000) Evidence for horizontal gene transfer in evolution of elongation factor Tu in enterococci. *J Bacteriol* **182**: 6913-6920.

Kellogg, B.A., and Poulter, C.D. (1997) Chain elongation in the isoprenoid biosynthetic pathway. *Curr Opin Chem Biol* **1**: 570-578.

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.

Kessler, P.S., Blank, C., and Leigh, J.A. (1998) The nif gene operon of the methanogenic archaeon Methanococcus maripaludis. *J Bacteriol* **180**: 1504-1511.

Kessler, P.S., and Leigh, J.A. (1999) Genetics of nitrogen regulation in Methanococcus maripaludis. *Genetics* **152**: 1343-1351.

Klein, M., Friedrich, M., Roger, A.J., Hugenholtz, P., Fishbain, S., Abicht, H., Blackall, L.L., Stahl, D.A., and Wagner, M. (2001) Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes. *J Bacteriol* **183**: 6028-6035.

Kleinig, H. (1975) On the utilization in vivo of lycopene and phytoene as precursors for the formation of carotenoid glucoside ester and on the regulation of carotenoid biosynthesis in Myxococcus fulvus. *Eur J Biochem* **57**: 301-308.

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., Venter, J.C., and et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon Archaeoglobus fulgidus [published erratum appears in Nature 1998 Jul 2;394(6688):101]. *Nature* **390**: 364-370.

Kogoma, T. (1997) Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol Mol Biol Rev* **61**: 212-238.

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

Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M., and Brown, J.R. (2000) Evolution of two-component signal transduction. *Mol Biol Evol* **17**: 1956-1970.

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.

Kowalczykowski, S.C. (2000) Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem Sci* **25**: 156-165.

Kredich, N.M. (1996) Biosynthesis of Cysteine. In *Escherichia coli and Salmonella : cellular and molecular biology*. Vol. 1. Neidhardt, F.C. (ed). Washington, D.C.: ASM Press, pp. 514-527.

Kunow, J., Linder, D., Stetter, K.O., and Thauer, R.K. (1994) F420H2: quinone oxidoreductase from Archaeoglobus fulgidus. Characterization of a membrane-bound multisubunit complex containing FAD and iron-sulfur clusters. *Eur J Biochem* **223**: 503-511.

Lan, R., and Reeves, P.R. (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol* **8**: 396-401.

Lancaster, C.R. (2002) Succinate:quinone oxidoreductases: an overview. *Biochim Biophys Acta* **1553**: 1-6.

Lange, B.M., Wildung, M.R., McCaskill, D., and Croteau, R. (1998) A family of transketolases that directs isoprenoid biosynthesis via a mevalonate-independent pathway. *Proc Natl Acad Sci U S A* **95**: 2100-2104.

Lange, B.M., Rujan, T., Martin, W., and Croteau, R. (2000) Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc Natl Acad Sci U S A* **97**: 13172-13177.

Larsen, O., Lien, T., and Birkeland, N.K. (1999) Dissimilatory sulfite reductase from Archaeoglobus profundus and Desulfotomaculum thermocisternum: phylogenetic and structural implications from gene sequences. *Extremophiles* **3**: 63-70.

Lawrence, J.G., and Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**: 1843-1860.

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

Lehman, N. (2003) A case for the extreme antiquity of recombination. *Journal of molecular evolution* **56**: 770-777.

Lemos, R.S., Fernandes, A.S., Pereira, M.M., Gomes, C.M., and Teixeira, M. (2002) Quinol:fumarate oxidoreductases and succinate:quinone oxidoreductases: phylogenetic relationships, metal centres and membrane attachment. *Biochim Biophys Acta* **1553**: 158-170.

Lichtentaler, H.K. (1998) Der 1-Desoxy-D-xylulose-Biosyntheseweg pflanzlicher Isoprenoide. *Biospektrum* **4**: 49-52.

Lodwick, D., Ross, H.N.M., Walker, J.A., Almond, J.W., and Grant, W.D. (1991) Nucleotide sequence of the 16S ribosomal RNA gene from the haloalkaliphilic archaeon (archaebacterium) *Natronobacterium magadii*, and the phylogeny of the halobacteria. *Systematic and Applied Microbiology* **14**: 352-357.

Lubben, M. (1995) Cytochromes of archaeal electron transfer chains. *Biochim Biophys Acta* **1229**: 1-22.

Lucas, P., Otis, C., Mercier, J.P., Turmel, M., and Lemieux, C. (2001) Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. *Nucleic Acids Res* **29**: 960-969.

Ludwig, W., and Schleifer, K.H. (1999) Phylogeny of *Bacteria* beyond the 16S rRNA standard. *ASM news* **65**: 752-757.

Maddison, W.P., and Maddison, D.R. (1989) Interactive analysis of phylogeny and character evolution using the computer program MacClade. *Folia Primatol (Basel)* **53**: 190-202.

Maden, B.E. (2000) Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism. *Biochem J* **350 Pt 3**: 609-629.

Maeder, D.L., Weiss, R.B., Dunn, D.M., Cherry, J.L., Gonzalez, J.M., DiRuggiero, J., and Robb, F.T. (1999) Divergence of the hyperthermophilic archaea Pyrococcus furiosus and P. horikoshii inferred from complete genomic sequences. *Genetics* **152**: 1299-1305.

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F.M., and Dowson, C.G. (2000) Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol* **182**: 1016-1023.

Makarova, K.S., Ponomarev, V.A., and Koonin, E.V. (2001) Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* **2**: RESEARCH 0033.

Mathis, P. (1990) Compared structure of plant and bacterial photosynthetic reaction centers. *Biochim Biophys Acta* **1018**: 163-167.

Matic, I., Rayssiguier, C., and Radman, M. (1995) Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell* **80**: 507-515.

Matic, I., Radman, M., Taddei, F., Picard, B., Doit, C., Bingen, E., Denamur, E., and Elion, J. (1997) Highly variable mutation rates in commensal and pathogenic Escherichia coli. *Science* **277**: 1833-1834.

Matic, I., Taddei, F., and Radman, M. (2000) No genetic barriers between Salmonella enterica serovar typhimurium and Escherichia coli in SOS-induced mismatch repair-deficient cells. *J Bacteriol* **182**: 5922-5924.

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

Maynard Smith, J. (1999) The detection and measurement of recombination from sequence data. *Genetics* **153**: 1021-1027.

Mayr, E. (1982) Processes of speciation in animals. *Prog Clin Biol Res* **96**: 1-19.

Mazumdar, P.M.H. (1995) *Species and Specificity.* Cambridge: Cambridge University Press.

McArthur, A.G., Morrison, H.G., Nixon, J.E., Passamaneck, N.Q., Kim, U., Hinkle, G., Crocker, M.K., Holder, M.E., Farr, R., Reich, C.I., Olsen, G.E., Aley, S.B., Adam, R.D., Gillin, F.D., and Sogin, M.L. (2000) The Giardia genome project database. *FEMS Microbiol Lett* **189**: 271-273.

McAteer, S., Coulson, A., McLennan, N., and Masters, M. (2001) The lytB gene of Escherichia coli is essential and specifies a product needed for isoprenoid biosynthesis. *J Bacteriol* **183**: 7403-7407.

McClelland, M., Florea, L., Sanderson, K., Clifton, S.W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R.K., and Miller, W. (2000) Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **28**: 4974-4986.

Mehta, M.P., Butterfield, D.A., and Baross, J.A. (2003) Phylogenetic Diversity of Nitrogenase (nifH) Genes in Deep-Sea and Hydrothermal Vent Environments of the Juan de Fuca Ridge. *Appl Environ Microbiol* **69**: 960-970.

Michel, B. (1999) Illegitimate recombination in bacteria. In *Organization of the prokaryotic genome*. Charlebois, R. (ed). Washington, D.C.: American Society for Microbiology, pp. 129-150.

Milkman, R., and Bridges, M.M. (1993) Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons. *Genetics* **133**: 455-468.

Moldoveanu, N., and Kates, M. (1988) Biosynthetic studies of the polar lipids of *Halobacterium Cutirubrum* . Formation of isoprenyl ether intermediates. *Biochim Biophys Acta* **960**: 164-182.

Molitor, M., Dahl, C., Molitor, I., Schafer, U., Speich, N., Huber, R., Deutzmann, R., and Truper, H.G. (1998) A dissimilatory sirohaem-sulfite-reductase-type protein from the hyperthermophilic archaeon Pyrobaculum islandicum. *Microbiology* **144 (Pt 2)**: 529-541.

Morii, H., Nishihara, M., and Koga, Y. (2000) CTP:2,3-di-O-geranylgeranyl-sn-glycero-1-phosphate cytidyltransferase in the methanogenic archaeon Methanothermobacter thermoautotrophicus. *J Biol Chem* **275**: 36568-36574.

Moshier, S.E., and Chapman, D.J. (1973) Biosynthetic studies on aromatic carotenoids. Biosynthesis of chlorobactene. *Biochem J* **136**: 395-404.

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

Mylvaganam, S., and Dennis, P.P. (1992) Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium Haloarcula marismortui. *Genetics* **130**: 399-410.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.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., Fraser, C.M., and et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399**: 323-329.

Nesbo, C.L., L'Haridon, S., Stetter, K.O., and Doolittle, W.F. (2001) Phylogenetic analyses of two "archaeal" genes in thermotoga maritima reveal multiple transfers between archaea and bacteria. *Mol Biol Evol* **18**: 362-375.

Nesbo, C.L., Nelson, K.E., and Doolittle, W.F. (2002) Suppressive subtractive hybridization detects extensive genomic diversity in Thermotoga maritima. *J Bacteriol* **184**: 4475-4488.

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.W., Alam, M., Freitas, T., Hou, S., Daniels, C.J., Dennis, P.P., Omer, A.D., Ebhardt, H., Lowe, T.M., Liang, P., 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.

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

Nishida, H., Nishiyama, M., Kobashi, N., Kosuge, T., Hoshino, T., and Yamane, H. (1999) A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis. *Genome Res* **9**: 1175-1183.

Nishihara, M., and Koga, Y. (1995) sn-glycerol-1-phosphate dehydrogenase in Methanobacterium thermoautotrophicum: key enzyme in biosynthesis of the enantiomeric glycerophosphate backbone of ether phospholipids of archaebacteria. *J Biochem (Tokyo)* **117**: 933-935.

Nishihara, M., Yamazaki, T., Oshima, T., and Koga, Y. (1999) sn-glycerol-1-phosphate-forming activities in Archaea: separation of archaeal phospholipid biosynthesis and glycerol catabolism by glycerophosphate enantiomers. *J Bacteriol* **181**: 1330-1333.

Nixon, J.E., Wang, A., Field, J., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., and Samuelson, J. (2002) Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to Giardia lamblia and Entamoeba histolytica. *Eukaryot Cell* **1**: 181-190.

Ochman, H., and Jones, I.B. (2000) Evolutionary dynamics of full genome content in Escherichia coli. *Embo J* **19**: 6637-6643.

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

Ohkuma, M., Noda, S., and Kudo, T. (1999) Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. *Appl Environ Microbiol* **65**: 4926-4934.

Ohnuma, S., Hirooka, K., Hemmi, H., Ishida, C., Ohto, C., and Nishino, T. (1996) Conversion of product specificity of archaebacterial geranylgeranyl-diphosphate synthase. Identification of essential amino acid residues for chain length determination of prenyltransferase reaction. *J Biol Chem* **271**: 18831-18837.

Ohnuma, S., Hirooka, K., Ohto, C., and Nishino, T. (1997) Conversion from archaeal geranylgeranyl diphosphate synthase to farnesyl diphosphate synthase. Two amino acids before the first aspartate-rich motif solely determine eukaryotic farnesyl diphosphate synthase activity. *J Biol Chem* **272**: 5192-5198.

Ohnuma, S., Hirooka, K., Tsuruoka, N., Yano, M., Ohto, C., Nakane, H., and Nishino, T. (1998) A pathway where polyprenyl diphosphate elongates in prenyltransferase. Insight into a common mechanism of chain length determination of prenyltransferases. *J Biol Chem* **273**: 26705-26713.

Olendzenski, L., Liu, L., Zhaxybayeva, O., Murphey, R., Shin, D.G., and Gogarten, J.P. (2000) Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. *J Mol Evol* **51**: 587-599.

Olson, J.M., and Pierson, B.K. (1987) Evolution of reaction centers in photosynthetic prokaryotes. *Int Rev Cytol* **108**: 209-248.

Orihara, N., Kuzuyama, T., Takahashi, S., Furihata, K., and Seto, H. (1998) Studies on the biosynthesis of terpenoid compounds produced by actinomycetes. 3. Biosynthesis of the isoprenoid side chain of novobiocin via the non-mevalonate pathway in Streptomyces niveus. *J Antibiot (Tokyo)* **51**: 676-678.

Osborne, J.P., and Gennis, R.B. (1999) Sequence analysis of cytochrome bd oxidase suggests a revised topology for subunit I. *Biochim Biophys Acta* **1410**: 32-50.

Palmer, J.D. (1995) Rubisco rules fall; gene transfer triumphs. *Bioessays* **17**: 1005-1008.

Parker, M.A. (2001) Case of localized recombination in 23S rRNA genes from divergent bradyrhizobium lineages associated with neotropical legumes. *Appl Environ Microbiol* **67**: 2076-2082.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G., and Barrell, B.G. (2000) Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. *Nature* **404**: 502-506.

Peck, H.D., Jr. (1961) Enzimatic basis for assimilatory and dissimilatory sulfate reduction. *J Bacteriol* **82**: 933-939.

Pereira, M.M., Santana, M., and Teixeira, M. (2001) A novel scenario for the evolution of haem-copper oxygen reductases. *Biochim Biophys Acta* **1505**: 185-208.

Perez Luz, S., Rodriguez-Valera, F., Lan, R., and Reeves, P.R. (1998) Variation of the ribosomal operon 16S-23S gene spacer region in representatives of Salmonella enterica subspecies. *J Bacteriol* **180**: 2144-2151.

Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., and Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409**: 529-533.

Pesole, G., Gissi, C., Lanave, C., and Saccone, C. (1995) Glutamine synthetase gene evolution in bacteria. *Mol Biol Evol* **12**: 189-197.

Pfeifer, F., Griffig, J., and Oesterhelt, D. (1993) The fdx gene encoding the [2Fe--2S] ferredoxin of Halobacterium salinarium (H. halobium). *Mol Gen Genet* **239**: 66-71.

Posada, D., and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.

Postgate, J.R. (1952) Groth of sulfate-reducing bacteria in sulfate free media. *Research, Lond.* **5**: 189-190.

Putra, S.R., Disch, A., Bravo, J.M., and Rohmer, M. (1998) Distribution of mevalonate and glyceraldehyde 3-phosphate/pyruvate routes for isoprenoid biosynthesis in some gram-negative bacteria and mycobacteria. *FEMS Microbiol Lett* **164**: 169-175.

Qian, J., Kwon, S.W., and Parker, M.A. (2003) rRNA and nifD phylogeny of Bradyrhizobium from sites across the Pacific Basin. *FEMS Microbiol Lett* **219**: 159-165.

Qian, Q., and Keeling, P.J. (2001) Diplonemid glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and prokaryote-to-eukaryote lateral gene transfer. *Protist* **152**: 193-201.

Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., and Zhou, J. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67**: 880-887.

Qureshi, N., and Porter, J.W. (1981) Conversion of acetyl-coenzyme A to isopentenyl pyrophosphate. In *Biosynthesis of Isoprenoid Compounds*. Vol. 1. Porter, J.W.S., S.L. (ed). New York: John Wiley, pp. 47-94.

Radman, M. (1999) Enzymes of evolutionary change. *Nature* **401**: 866-867, 869.

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

Ragan, M.A. (2001b) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201**: 187-191.

Rawlings, D.E., and Tietze, E. (2001) Comparative biology of IncQ and IncQ-like plasmids. *Microbiol Mol Biol Rev* **65**: 481-496, table of contents.

Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., and Fraser, C.M. (2000) Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res* **28**: 1397-1406.

Redfield, R.J. (1993) Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. *J Hered* **84**: 400-404.

Renosto, F., Martin, R.L., Borrell, J.L., Nelson, D.C., and Segel, I.H. (1991) Atp Sulfurylase from Trophosome Tissue of Riftia-Pachyptila (Hydrothermal Vent Tube Worm). *Archives of Biochemistry and Biophysics* **290**: 66-78.

Reysenbach, A.L., Seitzinger, S., Kirshtein, J., and McLaughlin, E. (1999) Molecular constraints on a high-temperature evolution of early life. *Biol Bull* **196**: 367-371; discussion 371-362.

Roger, A.J. (1999) Reconstructing early events in eukaryotic evolution. *Am Nat* **154**: S146-S163.

Rohdich, F., Wungsintaweekul, J., Fellermeier, M., Sagner, S., Herz, S., Kis, K., Eisenreich, W., Bacher, A., and Zenk, M.H. (1999) Cytidine 5'-triphosphate-dependent biosynthesis of isoprenoids: YgbP protein of Escherichia coli catalyzes the formation of 4- diphosphocytidyl-2-C-methylerythritol. *Proc Natl Acad Sci U S A* **96**: 11758-11763.

Rohmer, M., Bouvier-Nave, P., and Ourisson, G. (1984) Distribution of haponoids triterpenes in prokaryotes. *J Gen Microbiol* **130**: 1137-1150.

Rohmer, M. (1999) The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat Prod Rep* **16**: 565-574.

Rohwer, F. (2003) Global phage diversity. *Cell* **113**: 141.

Rosa Putra, S., Disch, A., Bravo, J.M., and Rohmer, M. (1998) Distribution of mevalonate and glyceraldehyde 3-phosphate/pyruvate routes for isoprenoid biosynthesis in some Gram-negative bacteria and *Mycobacteria*. *FEMS Microbiology Letters* **164**: 169-175.

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

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. *Proc Natl Acad Sci U S A* **97**: 14668-14673.

Salamon, H., Kato-Maeda, M., Small, P.M., Drenkow, J., and Gingeras, T.R. (2000) Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. *Genome Res* **10**: 2044-2054.

Scher, D.S., and Rodwell, V.W. (1989) 3-Hydroxy-3-methylglutaryl coenzyme A lyase from Pseudomonas mevalonii. *Biochim Biophys Acta* **1003**: 321-326.

Schleper, C., Puehler, G., Holz, I., Gambacorta, A., Janekovic, D., Santarius, U., Klenk, H.P., and Zillig, W. (1995) Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. *J Bacteriol* **177**: 7050-7059.

Scholzen, T., and Arndt, E. (1992) The alpha-operon equivalent genome region in the extreme halophilic archaebacterium Haloarcula (Halobacterium) marismortui. *J Biol Chem* **267**: 12123-12130.

Schubert, W.D., Klukas, O., Saenger, W., Witt, H.T., Fromme, P., and Krauss, N. (1998) A common ancestor for oxygenic and anoxygenic photosynthetic systems: a comparison based on the structural model of photosystem I. *J Mol Biol* **280**: 297-314.

Schutz, M., Brugna, M., Lebrun, E., Baymann, F., Huber, R., Stetter, K.O., Hauska, G., Toci, R., Lemesle-Meunier, D., Tron, P., Schmidt, C., and Nitschke, W. (2000) Early evolution of cytochrome bc complexes. *J Mol Biol* **300**: 663-675.

Schwender, J., Seemann, M., Lichtenthaler, H.K., and Rohmer, M. (1996) Biosynthesis of isoprenoids (carotenoids, sterols, prenyl side-chains of chlorophylls and plastoquinone) via a novel pyruvate/glyceraldehyde 3-phosphate non-mevalonate pathway in the green alga Scenedesmus obliquus. *Biochem J* **316**: 73-80.

Schwenn, J.D. (1997) In *Sulphur metabolism in higher plants : molecular, ecophysiological and nutritional aspects*. Cram, W.J. (ed). Leiden: Backhuys, pp. xviii, 367.

Seto, H., Watanabe, H., and Furihata, K. (1996) Simultaneous operation of the mevalonate and non-mevalonate pathways in the biosynthesis of isopentenyl diphosphate in *Streptomyces aeriouvifer*. *Tetrahedron Letters* **37**: 7979-7982.

Seto, H., Orihara, N., and Furihata, K. (1998) Studies on the Biosynthesis of Terpenoids Produced by Actinomycetes. Part4. Formation of BE-40644 by the Mevalonate and Nonmevalonate Pathways. *Tetrahedron Lett* **39**: 9497-9500.

Severinov, K., Mustaev, A., Kukarin, A., Muzzin, O., Bass, I., Darst, S.A., and Goldfarb, A. (1996) Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. *J Biol Chem* **271**: 27969-27974.

Shen, Y., Buick, R., and Canfield, D.E. (2001) Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature* **410**: 77-81.

Shimada, H., Nemoto, N., Shida, Y., Oshima, T., and Yamagishi, A. (2002) Complete polar lipid composition of Thermoplasma acidophilum HO-62 determined by high-performance liquid chromatography with evaporative light-scattering detection. *J Bacteriol* **184**: 556-563.

Sicheritz-Ponten, T., and Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* **29**: 545-552.

Smit, A., and Mushegian, A. (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res* **10**: 1468-1484.

Smith, G.R. (1988) Homologous recombination in procaryotes. *Microbiol Rev* **52**: 1-28.

Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**: 4384-4388.

Smith, M.W., Feng, D.F., and Doolittle, R.F. (1992) Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci* **17**: 489-493.

Smith, N.H., Holmes, E.C., Donovan, G.M., Carpenter, G.A., and Spratt, B.G. (1999) Networks and groups within the genus Neisseria: analysis of argF, recA, rho, and 16S rRNA sequences from human Neisseria species. *Mol Biol Evol* **16**: 773-783.

Smith, R.F., and Smith, T.F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm secondary structure-dependent gap penalties for comparative protein modelling. *Protein Engineering* **5**: 35-41.

Soderberg, T., Chen, A., and Poulter, C.D. (2001) Geranylgeranylglyceryl phosphate synthase. Characterization of the recombinant enzyme from Methanobacterium thermoautotrophicum. *Biochemistry* **40**: 14847-14854.

Sonea, S. (1971) A tentative unifying view of bacteria. *Review of Canadian Biology* **30**: 3239-3244.

Sonea, S. (1988) A bacterial way of life. *Nature* **331**: 216.

Sperling, D., Kappler, U., Wynen, A., Dahl, C., and Truper, H.G. (1998) Dissimilatory ATP sulfurylase from the hyperthermophilic sulfate reducer Archaeoglobus fulgidus belongs to the group of homo-oligomeric ATP sulfurylases. *Fems Microbiology Letters* **162**: 257-264.

Sprenger, G.A., Schorken, U., Wiegert, T., Grolle, S., de Graaf, A.A., Taylor, S.V., Begley, T.P., Bringer-Meyer, S., and Sahm, H. (1997) Identification of a thiamin-dependent synthase in Escherichia coli required for the formation of the 1-deoxy-D-xylulose 5-phosphate precursor to isoprenoids, thiamin, and pyridoxol. *Proc Natl Acad Sci U S A* **94**: 12857-12862.

Stahl, D.A., Fishbain, S., Klein, M., Baker, B.J., and Wagner, M. (2002) Origins and diversification of sulfate-respiring microorganisms. *Antonie Van Leeuwenhoek* **81**: 189-195.

Stanier, R.Y., and van Niel, C.B. (1941) The Main Outlines of Bacterial Classification. *Journal of Bacteriology* **42**: 437-466.

Stetter, K.O., Lauerer, G., Thomm, M., and Neuner, A. (1987) Isolation of extremely termophilic sulfate reducers: evidence for a novel branch of archaebacteria. *Science* **236**: 822-824.

Strick, J.E. (2000) *Sparks of life.* Cambridge: Harvard Universisty Press.

Striepen, B., White, M.W., Li, C., Guerini, M.N., Malik, S.B., Logsdon, J.M., Jr., Liu, C., and Abrahamsen, M.S. (2002) Genetic complementation in apicomplexan parasites. *Proc Natl Acad Sci U S A* **99**: 6304-6309.

Suguri, S., Henze, K., Sanchez, L.B., Moore, D.V., and Muller, M. (2001) Archaebacterial relationships of the phosphoenolpyruvate carboxykinase gene reveal mosaicism of Giardia intestinalis core metabolism. *J Eukaryot Microbiol* **48**: 493-497.

Summers, W.C. (2000) History of Microbiology. In *Encyclopedia of Microbiology*. Vol. 2. Lederberg, J. (ed). New York: Academic Press, pp. 677-697.

Swofford, D.L. (1998) PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Sunderland, Massachusets: Sinauer Associates.

Syvanen, M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* **28**: 237-261.

Tachibana, A. (1994) A novel prenyltransferase, farnesylgeranyl diphosphate synthase, from the haloalkaliphilic archaeon, Natronobacterium pharaonis. *FEBS Lett* **341**: 291-294.

Tachibana, A., Yano, Y., Otani, S., Nomura, N., Sako, Y., and Taniguchi, M. (2000) Novel prenyltransferase gene encoding farnesylgeranyl diphosphate synthase from a hyperthermophilic archaeon, Aeropyrum pernix. Molecularevolution with alteration in product specificity. *Eur J Biochem* **267**: 321-328.

Takahashi, S., Kuzuyama, T., and Seto, H. (1999) Purification, characterization, and cloning of a eubacterial 3-hydroxy- 3-methylglutaryl coenzyme A reductase, a key enzyme involved in biosynthesis of terpenoids. *J Bacteriol* **181**: 1256-1263.

Takai, K., and Horikoshi, K. (1999) Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* **152**: 1285-1297.

Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S., and Horikoshi, K. (2000) Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. *Nucleic Acids Res* **28**: 4317-4331.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., and Venter, J.C. (2000) Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science* **287**: 1809-1815.

Thauer, R.K. (1998) Biochemistry of methanogenesis: a tribute to Marjory Stephenson. 1998 Marjory Stephenson Prize Lecture. *Microbiology* **144 (Pt 9)**: 2377-2406.

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

Tun-Garrido, C., Bustos, P., Gonzalez, V., and Brom, S. (2003) Conjugative Transfer of p42a from Rhizobium etli CFN42, Which Is Required for Mobilization of the Symbiotic Plasmid, Is Regulated by Quorum Sensing. *J Bacteriol* **185**: 1681-1692.

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

Van de Peer, Y., Neefs, J.M., de Rijk, P., and de Vos, P. (1994) About the order of divergence of the major bacterial taxa during evolution. *Systematic and Applied Microbiology* **17**: 32-38.

Van de Peer, Y., Baldauf, S.L., Doolittle, W.F., and Meyer, A. (2000) An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J Mol Evol* **51**: 565-576.

Van Driessche, G., Hu, W., Van de Werken, G., Selvaraj, F., McManus, J.D., Blankenship, R.E., and Van Beeumen, J.J. (1999) Auracyanin A from the thermophilic green gliding photosynthetic bacterium Chloroflexus aurantiacus represents an unusual class of small blue copper proteins. *Protein Sci* **8**: 947-957.

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

Vorholt, J.A., Chistoserdova, L., Stolyar, S.M., Thauer, R.K., and Lidstrom, M.E. (1999) Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *J Bacteriol* **181**: 5750-5757.

Vorholt, J.A. (2002) Cofactor-dependent pathways of formaldehyde oxidation in methylotrophic bacteria. *Arch Microbiol* **178**: 239-249.

Vreeland, R.H., Straight, S., Krammes, J., Dougherty, K., Rosenzweig, W.D., and Kamekura, M. (2002) Halosimplex carlsbadense gen. nov., sp. nov., a unique halophilic archaeon, with three 16S rRNA genes, that grows only in defined medium with glycerol and acetate or pyruvate. *Extremophiles* **6**: 445-452.

Wagner, M., Roger, A.J., Flax, J.L., Brusseau, G.A., and Stahl, D.A. (1998) Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration. *J Bacteriol* **180**: 2975-2982.

Wallace, A.R. (1872) Review of "The Beginnings of Life" by H. Charlton Bastian. *Nature* **6**: 299-303.

Waller, R.F., Keeling, P.J., Donald, R.G., Striepen, B., Handman, E., Lang-Unnasch, N., Cowman, A.F., Besra, G.S., Roos, D.S., and McFadden, G.I. (1998) Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum. *Proc Natl Acad Sci U S A* **95**: 12352-12357.

Watanabe, T. (1963) Infective heredity of multiple drug resistance in bacteria. *Bacteriological Reviews* **27**: 87-115.

Welch, R.A., Burland, V., Plunkett, G., 3rd, 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., 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.

Wernegreen, J.J., and Riley, M.A. (1999) Comparison of the evolutionary dynamics of symbiotic and housekeeping loci: a case for the genetic coherence of rhizobial lineages. *Mol Biol Evol* **16**: 98-113.

Widdel, F. (1986) Sulphate-reducing bacteria and their ecological Niches. In *Anaerobic Bacteria in Habitats Other Than Man*. Barnes, E.M. and Mead, G.C. (eds). Cambridge, MA: Blackwell Sci., pp. 157–184.

Widdel, F. (1988) Microbiology and ecology of sulfate- and sulfur-reducing bacteria. In *Biology of Anaerobic Microorganisms*. Zehnder, A.J.B. (ed). New York:: Wiley & Sons, pp. 469–585.

Wilding, E.I., Brown, J.R., Bryant, A.P., Chalker, A.F., Holmes, D.J., Ingraham, K.A., Iordanescu, S., So, C.Y., Rosenberg, M., and Gwynn, M.N. (2000) Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. *J Bacteriol* **182**: 4319-4327.

Wilson, K. (1994) Preparation of genomic DNA from bacteria. In *Current protocols in molecular biology*. Vol. 1. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.A., Struhl, K. and Smith, J.A. (eds). New York: John Wiley and Sons, pp. 2.4.1-2.4.5.

Winslow, C.E.A., Broadhurst, J., Buchanan, R.E., Krumwiede, C., Rogers, L.A., and Smith, G.H. (1920) The Families and Genera of the Bacteria. *Journal of Bacteriology* **5**.

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

Woese, C.R. (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* **97**: 8392-8396.

Woese, C.R., Olsen, G.J., Ibba, M., and Soll, D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* **64**: 202-236.

Wolf, Y.I., Aravind, L., Grishin, N.V., and Koonin, E.V. (1999) Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* **9**: 689-710.

Xiong, J., Inoue, K., and Bauer, C.E. (1998) Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from Heliobacillus mobilis. *Proc Natl Acad Sci U S A* **95**: 14851-14856.

Xiong, J., Fischer, W.M., Inoue, K., Nakahara, M., and Bauer, C.E. (2000) Molecular evidence for the early evolution of photosynthesis. *Science* **289**: 1724-1730.

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

Young, J. (1992) Phylogenetic Classification of Nitrogen-Fixing Organisms. In *Biological Nitrogen Fixation*. Stacey, G., Burris, R. and Evans, H. (eds). London: Chapman and Hall, pp. 43-86.

Zhang, D., and Poulter, C.D. (1993) Biosynthesis of archaebacterial ether lipids. Formation of ether linkages by prenyltransferases. *Journal of the American Chemical Society* **115**: 1270-1277.

Zuckerkandl, E., and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*. Bryson, V. and Vogel, H.J. (eds). New York: Academic Press.

Appendices

Appendix 1: List of Publications

Publications in the course of this Doctorate program

In preparation

Yan Boucher, Christophe J. Douady, Adrian K. Sharma, Masahiro Kamekura and W. Ford Doolittle. Frequent intraspecific heterogeneity of rDNA genes among extremely halophilic archaea. *Journal of Bacteriology.*

Submitted

Yan Boucher, Masahiro Kamekura and W. Ford Doolittle (2003) Origin and evolution of isoprenoid lipid biosynthesis in archaea. *Molecular Microbiology.*

In press

Yan Boucher, Christophe J. Douady, Thane R. Papke, David A. Walsh, Ellen M. Boudreau, Camilla L. Nesbø, Rebecca J. Case and W. Ford Doolittle (2003) Origin of physiological properties of major prokaryotic groups by lateral gene transfer. *Annual Review of Genetics* 37.

Published

Christophe J. Douady, Frédéric Delsuc, Yan Boucher, W. Ford Doolittle and Emmanuel J.P. Douzery (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution* Feb;20(2):248-254.

Christian Blouin, Yan Boucher, Andrew J. Roger (2003) Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic Acids Research* Jan;31(2):790-7.

W. Ford Doolittle, Yan Boucher, Camilla L. Nesbø, Christophe J. Douady, Jan Andersson and Andrew J. Roger (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philosophical Transactions of the Royal Society of London B, Series B: Biological Sciences* Dec;358:39-58.

Yan Boucher and W. Ford Doolittle (2002) Something new under the sea. *Nature* May;417(6884):27-8.

Yan Boucher, Harald Huber, Stéphane L'Haridon, Karl O. Stetter and W. Ford Doolittle (2001) Bacterial Origin for the Isoprenoid Biosynthesis Enzyme HMG-CoA Reductase of the Archaeal Orders Thermoplasmatales and Archaeoglobales. *Molecular Biology and Evolution* Jul;18(7):1378-88.

Yan Boucher, Camilla L. Nesbø and W. Ford Doolittle (2001) Microbial genomes: dealing with diversity. *Current Opinion in Microbiology* Jun;4(3):285-9.

Camilla L. Nesbø, Yan Boucher and W. Ford Doolittle. (2001) Comparative Genomics of Four Archaea: Is there a Core of Non-Transferable Proteins*? Journal of Molecular Evolution* Oct-Nov;53(4-5):340-50.

Yan Boucher and W. Ford Doolittle (2000) The Role of Lateral Gene Transfer in the Evolution of Isoprenoid Biosynthesis Pathways. *Molecular Microbiology* Aug;37(4): 703-716.