8

Molecular Sequences and the Early History of Life

RADHEY S. GUPTA

The evolutionary history of life, spanning a period of more than 3.5 billion years (Giga annum or Ga) constitutes one of the most fascinating problems in the life sciences.[1–4] This chapter will critically examine our understanding of a number of aspects of early evolutionary history. The topics covered are critical issues in Bacterial phylogeny, lateral gene transfer (LGT) and its influence on evolutionary relationships, the relationship of Archaea to Bacteria, and the origin of eukaryotes.

Bacterial Phylogeny: Some Critical Issues

The Bacteria make up the vast majority of prokaryotes. Hence, discerning the evolutionary relationships among them constitutes a major part of understanding prokaryotic phylogeny. On the basis of branching in the 16S rRNA trees, about 25 main groups or phyla within Bacteria are recognized at this time.[5] Although Bacteria have been divided into phyla on the basis of 16S rRNA, the criteria as to what actually constitutes a phylum remain to be defined.[5,6] In the beginning, when the sequence database was limited, the main phyla could be clearly distinguished in phylogenetic trees on the basis of long, "naked" internal branches that separated them. However, the explosive increase in sequence database entries in recent years has filled most of these naked branches and, as a result, distinguishing between phyla has become increasingly difficult and imprecise.[5,6] In the absence of objective criteria for the main divisions, it is unclear at present how many phyla exist within Bacteria and how to distinguish them from their subdivisions. On the basis of 16S rRNA, the proteobacterial phylum is presently divided into five subdivisions, named α, β, γ, δ, and ε.[3,7] Some of these subdivisions, for example, α, β, and γ,

contain from several hundreds to thousands of species, and they are much larger than most other main groups within Bacteria. The members of these subdivisions are also clearly distinguished from each other and other bacterial divisions, both in phylogenetic trees and by distinctive signature sequences.[7] Lacking objective criteria, it is unclear why these major groups have been assigned subdivision status, whereas many poorly characterized taxa consisting of only a few species are recognized as distinct divisions.

Any understanding of bacterial phylogeny requires that one can determine how main groups are related to each other and their branching order from a common ancestor. Unfortunately, phylogenetic trees based on rRNA have not been able to resolve these relationships and, thus, have yielded contradictory results.[3,5,8,9] This has led to the notion that this important problem is unsolvable, and even to the erroneous assumption that most, if not all, main groups within Bacteria may have branched off from a common ancestor at about the same time.[5,9] Over the last few years we have developed a means for understanding bacterial phylogeny.

Signature Approach for Determining Bacterial Phylogeny

Our new approach is grounded on conserved inserts and deletions (referred to as indels or signature sequences) in protein sequences for deducing phylogeny. On the basis of the presence or absence of shared conserved indels, different species can be divided into distinct groups, and their specific evolutionary relationships can be revealed. The rationale for our approach is that when a conserved indel of defined length and sequence is present in the same position in a given gene or protein from all members from one or more groups of bacteria, but not in the other groups, the simplest and most parsimonious explanation is that the indel was introduced only once, in a common ancestor of the group of species that possess this characteristic. All evolutionary useful signatures need be flanked on either side by conserved regions to ensure their reliability.

The signatures that we have identified are of two main kinds. Many of them are group specific: they are uniquely present in protein homologs from particular phyla or subdivisions (referred to as groups) of Bacteria. One example of a group-specific signature is provided in figure 8.1, where a conserved insert of 18–21 aa is present in the DNA polymerase I (Pol I) from various cyanobacteria, but not in any other groups of bacteria. Cyanobacteria-specific signatures are also present in many other proteins including DNA helicase II, ADP-glucose pryophosphorylase, FtsH protease, phytoene synthase, EF-Tu, Sec A, ribosomal S1 protein, IMP-dehydrogenase, and the major sigma factor 70 (table 8.1).[10] Similar to cyanobacteria, a large number of signatures that are distinctive for the chlamydiae have also been identified.[11] Group-specific signatures have also been identified for most other major groups within Bacteria including Proteobacteria, Aquificales, Firmicutes, Actinobacteria, Deinococcus-Thermus, Spirochetes, Cytophaga-Flavobacteria-Bacteriodetes-Green sulfur bacteria (CFBG)[4,7] (Gupta, R.S., unpublished results). These signatures provide a means for identifying different bacterial groups in clear molecular terms and for assignment of species to these groups.

A second category of signatures comprises those in which a conserved indel is commonly present in several groups of bacteria, but is absent in other groups. These signatures, which I will refer to as main line signatures, have been introduced at important

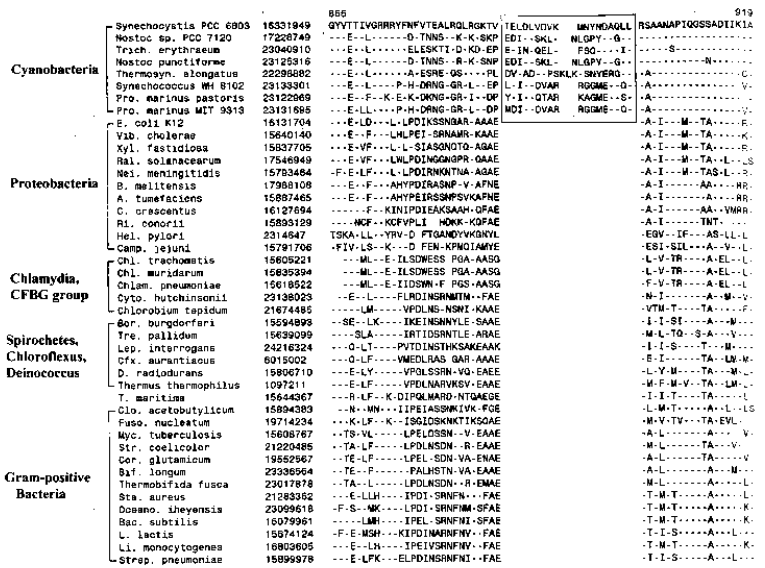

Figure 8.1. Sequence alignment of DNA polymerase I showing a large insert (boxed) that is specific for cyanobacteria. Dashes in the alignment indicate identity with the amino acid on the top line. Polymerase I is present in all bacterial genomes, and sequence information for only representative species is presented.

evolutionary branch points and thus are very useful in understanding the branching order and interrelationships among different groups.[4] If an indel was introduced in an ancestral lineage at a critical branch point (i.e., in the main trunk of the tree), then it is expected that all species diverging from this ancestor at later times should contain the signature, whereas all other species originating from the branches that existed before introduction of the signature should be lacking the indel.[6,12] Thus, on the basis of different main-line signatures that have been introduced in the main evolutionary trunk at different stages, the order of divergence of different groups can be established.

Two examples of main line signatures are shown. In the Hsp70 (DnaK) protein found in all bacteria (fig. 8.2), a 21–23-aa insert is present in various Proteobacteria, Chlamydiae, CFBG, Aquifex, Spirochetes, Cyanobacteria, and Deinococcus-Thermus groups of bacteria, but it is not found in any of the species from the Thermotoga, Clostridia-Fusobacteria, Actinobacteria, and Firmicute groups. This indel is also absent in various Archaea, and on the basis of the established rooting of the prokaryotic tree between Archaea and Bacteria,[13,14] this observation indicates that the groups lacking this indel are ancestral and that this indel constitutes an insert in the later branching groups. The ancestral nature of the groups lacking this indel is also supported by other lines of evidence discussed in earlier work[4,15] Another example of a main-line signature found in the RNA polymerase β-subunit (RpoB), is shown in figure 8.3. RpoB is a core compo-

Table 8.1. Sequenced bacterial genomes

Proteobacteria (γ-subdivision)
Escherichia coli K12
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
Escherichia coli CFT073
Buchnera sp. APS
Buchnera aphidicola
Buchnera aphidicola Sg
Pasteurella mutocida
Pseudomonas aeruginosa
Pseudomonas putida KT 2400
Pseudomonas syringae
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Xylella fastidiosa
Xyella fastidiosa Temecula
Haemophilus influenzae
Yersinia pestsis C092
Yersina pestsis KIM
Salmonella typhimurium LT2
Salmonella typhi
Xanthomonas citri
Xanthomonas campestris
Xyellela fastidiosa
Shewanella oneidensis
Shigella flexneri 2a
Wiggelsworthia brevipalpis
Coxiella burnetii

Proteobacteria (α-subdivision)
Rickettsia prowazekii
Caulobacter crescentus
Mesorhizobium loti
Bradyrhizobium japonicum
Agrobacterium tumefaciens-Dupont
Agrobacterium tumefaciens-Cereon
Rickettsia conorii
Sinorhizobium loti
Brucella melitensis
Brucella suis
Rhodopseudomonas palustris

Proteobacteria (β-subdivision)
Neisseria meningitidis MC58
Neisseria meningitidis Z2491
Ralstonia solanacearum

Proteobacteria (δ, ϵ-subdivision)
Helicobacter pylori 26695
Helicobacter pylori J99
Campylobacter jejuni

Aquifex
Aquifex aeolicus

Chlamydia-CFBG
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae J138
Chlamydophila pneumoniae AR39
Chlorobium tepidum
Bacteroides thetaiotamicron

Spirochetes
Borrelia burgdorferi
Treponema pallidum
Leptospira interrogans

Cyanobacteria
Synechocystis sp. PCC6803
Nostoc sp. PCC7120
Thermosynechococcus elongatus

Clostridia-Thermotoga
Thermotoga maritima
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani E88
Fusobacterium nucleatum
Thermoanaerobacter tengcongensis

Deinococcus-Thermus
Deinococcus radiodurans

Actinobacteria
Mycobacterium tuberculosis H37
Mycobacterium tuberculosis 1551
Mycobacterium leprae
Corynebacterium glutamicum
Corynebacterium efficiens
Streptomyces coelicolor
Bifidobacterium longum
Tropheryma whipplei Twist
Tropheryma whipplei TW08/27

Firmicutes
Bacillus subtilis
Bacillus halodurans
Bacillus antharics
Oceanabacillus iheyensis
Staphylococcus aureus N315
Staphylococcus aureus MW2
Staphylococcus epidermidis
Staphylococcus aureus Mu50
Streptococcus pyogenes
Streptococcus pyogenes S315
Streptococcus pyogenes S8232
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus agalactiae 2603
Streptococcus agalactiae NEM316
Streptococcus mutans UA159
Mycoplasma genitalilum
Mycoplasma pneumoniae
Mycoplasma pulmonis
Mycoplasma penetrans
Ureaplasma urealyticus
Lactococcus lactis
Lactobacillus plantarum
Listeria innocua
Listeria monocytogenes

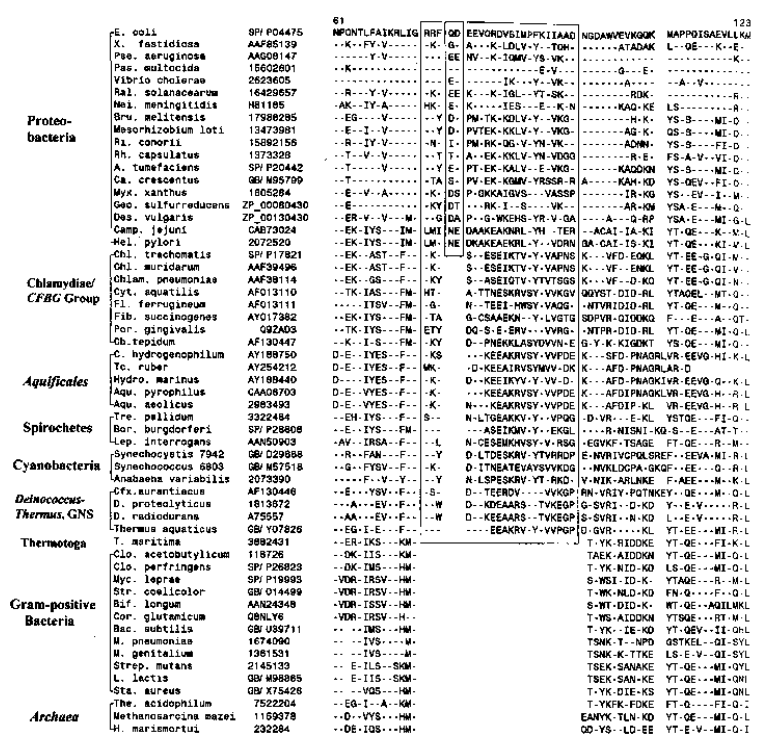

Figure 8.2. Partial alignment of Hsp70 sequences showing two main line signatures. The large insert (box 1) is a distinctive characteristic of various gram-negative bacteria (as defined by the presence of an outer membrane), and it is not found in any gram-positive or monoderm bacteria. This insert is also not present in any Archaea. The smaller, 2-aa insert (box 2) is a distinctive characteristic of Proteobacteria.

nent of the RNA polymerase found in all bacterial genomes. The signature in this case consists of a large indel of between 90 and 133 aa that is commonly present in various proteobacteria, Aquifex, and the Chlamydia–CFBG groups, but not in any other groups of bacteria.[16,17] This indel is also not present in the RpoB homologs from Archaea, indicating that it constitutes an insert in the latter branching groups. On the basis of its specific presence only in Proteobacteria, Aquificales, and the Chlamydia–CFBG groups, this insert was likely introduced in a common ancestor of these groups after branching of the other groups (fig. 8.4).

We have described a large number of other main-line signatures that are helpful in determining Bacterial evolutionary relationships. Our analyses of these signatures, as shown here for Hsp70 and RpoB proteins, indicate that they have been introduced at

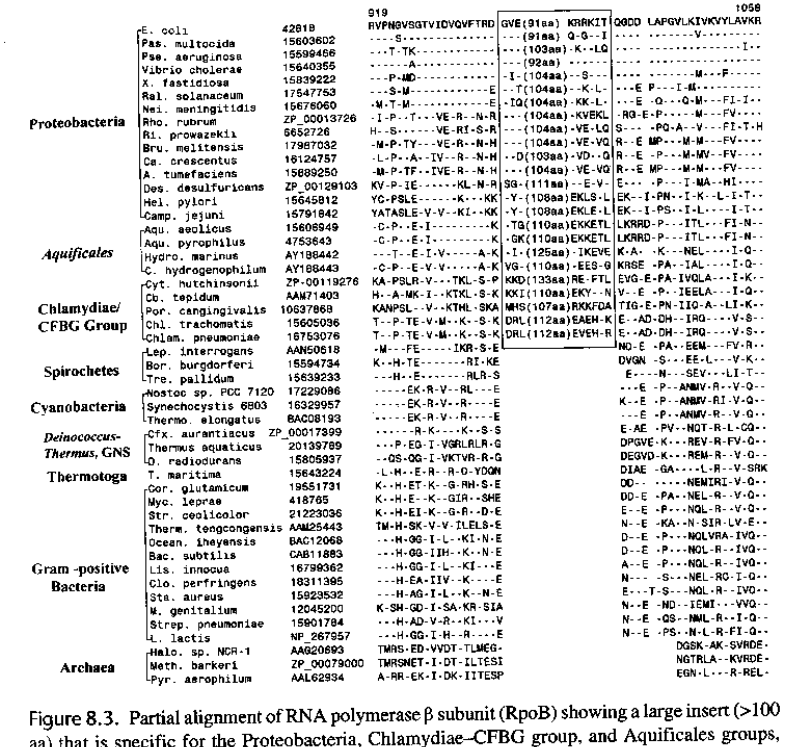

Figure 8.3. Partial alignment of RNA polymerase β subunit (RpoB) showing a large insert (>100 aa) that is specific for the Proteobacteria, Chlamydiae–CFBG group, and Aquificales groups, but that is not found in any other bacteria. The absence of this insert in archaeal homologs provides evidence that the groups lacking this insert are ancestral.

specific stages in bacterial evolution, as depicted in figure 8.4. On the basis of the presence or absence of these signatures, all main groups within Bacteria can be clearly distinguished, and it is also possible to logically deduce that they have branched off from a common ancestor in the order shown in figure 8.4.[6,12,17,18]

Testing the Indel Model on Bacterial Genomes

The completed bacterial genomes provide an objective means to test the reliability of the deduced branching order based on our signature sequence approach. At the end of April 2003 sequences for 100 bacterial genomes were available in the public domain (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/complete.html). The main groups to which these species belong are indicated in table 8.2. The branching order of these groups as shown in figure 8.4 makes very specific predictions as to which of the different indels should be present or absent in different species. According to the indel model,

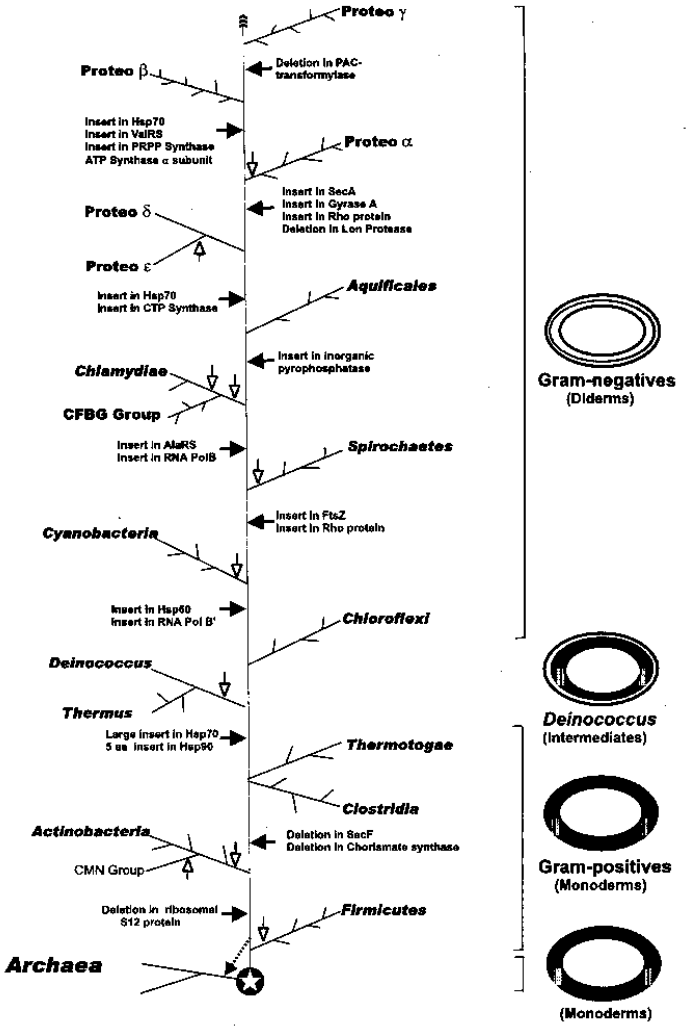

Figure 8.4. Evolutionary model based on signature sequences indicating the branching order of the main bacterial groups. The filled arrows depict the stages at which the different main-line signatures indicated in table 8.3 have been introduced. These signatures are expected to be present in bacterial groups that have diverged at a later time (i.e., those lying above the indicated insertion points), but they should be absent in the earlier branching groups. The unfilled arrows denote the positions of many group-specific signatures (not shown here). The cell structures of different groups of bacteria are indicated on the right. The dotted arrow at the bottom indicates the possible derivation of Archaea from gram-positive bacteria.

Table 8.2. Predicted versus observed distribution of indels in 100 bacterial genomes

Protein	Signature Description	No. Genomes with Protein	No. Genomes with Indels Expected/ Found	No. Genomes Lacking the Indel Expected/ Found	Exceptions Observed
Rib. S12 protein	13 aa *Firmicute* insert	100	25/25	75/75	0
Hsp70/DnaK	21–23 aa G+/G- insert	100	60/60	40/40	0
Hsp90	5 aa G+/G- insert	52	11/11	41/41	0
Chorismate Synthase	15–17 aa deletion after *Actinobacteria*	89	29/29	60/60	0[a]
SecF protein	3–4 aa deletion after *Actinobacteria*	81	15/17	56/54	2[b]
Hsp60/GroEL	1 aa insert after *Deinococcus*	98	65/66	33/32	1[c]
RNA Polymerase β^{1}- subunit	>150 aa after *Deinococcus*	100	59/59	41/41	0
FtsZ protein	1 aa insert after cyanobacteria	91	51/51	40/40	0
Rho protein	2 aa insert before spirochetes	83	56/57	27/26	1[d]
Ala-tRNA Synth.	4 aa after spirochetes	100	53/53	47/47	0
RNA Polymerase β- subunit	90–120 aa insert after spirochetes	100	53/53	47/47	0
Inorganic pyro-phosphatase	2 aa insert common to *Aquifex* and proteo.	71	45/45	26/26	0
Hsp70/DnaK	2 aa Proteo insert	100	45/45	55/55	0
CTP Synthetase	10 aa Proteo Indel	92	45/45	47/47	0
Lon protease	1 aa deletion in αβγ-proteobacteria	70	41/43	29/27	2[e]
Rho Protein	3 aa αβγ-Proteo indel	83	42/43	41/40	1[f]
DNA Gyrase A subunit	26–34 aa insert in αβγ-proteobacteria	100	42/42	58/58	0
SecA protein	7 aa αβγ-Proteo indel	100	42/42	58/58	0
HSP70/DnaK	4 aa βγ-Proteo insert	100	31/34	69/66	3[g]
ATP Synthase α-subunit	11 aa insert in βγ-proteobacteria	92	31/32	61/60	1[h]
Val-tRNA Synth.	37 aa βγ-Proteo insert	100	31/31	69/69	0
PRPP synthetase	1 aa βγ-Proteo insert	94	31/31	63/63	0
PAC-formyltransferase	2 aa γ-Proteo deletion	83	55/55	28/28	0

Abbreviations in the proteins names are: PAC, 5'- phoshoribosyl-5-aminoimidazole -4 carboxamide formytransferase; PRPP, phosphoribosyl pyrophosphate;

[a]Smaller inserts also present in this region *in D. radiodurans, A. aeolicus, Cb. tepidum,* and *C. tetani.*

[b]*T. whipplei* contains the insert that is not expected.

[c]*M. penetrans* constitutes an exception.

[d]*T. maritima* contains the insert that is not expected.

[e]*B. japonicum* and *P. putida* are exceptions.

[f]*B. thetaiotamicron* is the exception.

[g]*B. longum* and *T. whipplei* are exceptions.

[h]*Cb. tepidum* also contains this insert.

once a main-line signature has been introduced in an ancestral lineage, all species from the latter branching groups should contain the indel, whereas all species from groups that branched off before the introduction of the signature should be lacking the indel.[4] If the deduced branching order is reliable, then the observed distribution of these indels in different genomes should be close to that predicted by the model. However, if such indels could arise independently, or if the genes harboring them were subjected to frequent lateral transfer, then their presence or absence in different species would not follow the prediction of the model. Thus, the reliability of the deduced branching order can be objectively assessed by determining how closely the distribution of these signatures in different genomes follows the predictions of the model.

Results of these analyses, examining the presence or absence of different main line signatures in various bacterial genomes, are presented in table 8.3. The number of genomes in which these proteins have been found, and the number of species, we would expect to contain or lack these indels based on their postulated insertion positions, are also indicated in table 8.3. For example, for the main-line indels in RpoB and AlaRS, the model predicts that 53 of the 100 species from various groups that branched off after the insertion points of these indels (i.e., groups on the top in figure 8.4) should contain the indel, whereas all 47 species from groups that diverged before the introduction of the indels (those below the insertion point) should not possess it. Similarly, the large, 21–23-aa insert in the Hsp70 protein should be present in 60 of the 100 species branching after the Thermotoga-Clostridia clade (figure 8.4), but it should not be found in the remaining 40 species from groups branching earlier. The last few columns in table 8.3 summarize the results obtained for different indels and the number of exceptions or contradictions that were observed. The results of these studies are strikingly clear (table 8.3), as the presence or absence of various signatures in different genomes was found to be

Table 8.3. Statistical significance of cyanobacterial signatures

Protein	Signature	Presence of Indel in Cyanobacteria and Plastids	Presence of Indel in other Bacteria	χ^2 Probability
DNA Helicase II (UvrD)	6 aa. insert	10/10	0/>100	$<10^{-15}$
	7 aa insert	10/10	0/>100	$<10^{-15}$
	28 aa insert	10/10	0/>100	$<10^{-15}$
DNA Pol I	18–21 aa insert	8/8	0/>100	$<10^{-15}$
ADP-Glucose Pyrophosphorylase	14 aa insert	17/17	0/40	$<10^{-15}$
FtsH protease	3 aa insert	8/8	0/>90	$<10^{-15}$
Phytoene Synthase	11–13 aa insert	13/13	0/40	$<10^{-15}$
Elongation factor-Tu	5 aa insert	15/15	0/>100	$<10^{-15}$
Ribosomal S1 protein	2 aa and 7 aa deletions	15/15	0/>60	$<10^{-15}$
SecA protein	2 aa insert	15/15	0/>80	$<10^{-15}$
IMP dehydrogenase	2 aa deletion and 6 aa insert	9/9	0/>80	$<10^{-15}$
Major sigma factor-70	1 aa deletion	13/13	0/>80	$<10^{-15}$

Data for these signatures taken from reference 35. The χ^2 probability for the random occurrence of these indels in different bacteria was calculated using two degrees of freedom.

almost exactly as predicted by the model. In 2,079 observations examining the presence or absence of these indels in 100 genomes, only 11 exceptions or ambiguities were observed. These exceptions could be the results of LGTs or other nonspecific mechanisms. The ability of the indel model (fig. 8.4) to predict with such remarkable accuracy (>99%) the presence or absence of various indels in different genomes provides compelling evidence of its reliability and predictive power.

The relationships depicted in figure 8.4, with a few notable exceptions (to be discussed later), are generally in accordance with phylogenetic trees based on different genes and proteins.[4,8,10,16,19] Most published phylogenetic trees show groups such as Thermotoga, Deinococcus-Thermus, Cyanobacteria, and green nonsulfur bacteria to be deep branching, whereas other groups such as Proteobacteria and Chlamydiae–CFBG clade are late-branching lineages. The Spirochetes generally branch in middle in proximity of the Chlamydiae–CFBG and cyanobacterial taxa. In contrast to these groups, the branching of gram-positive bacteria (Firmicutes, Actinobacteria, Clostridia) and Aquificales is found to be highly variable in different trees.[5,7,8,16]

Despite its uncertain phylogenetic position, the deep branching of Aquifex, as seen in the rRNA trees, has become a cornerstone of our present understanding of bacterial phylogeny.[5,20] However, this view is not supported by signature sequences in various proteins. These studies instead provide strong and consistent evidence that the order Aquificales has diverged late in bacterial evolution, branching in between Chlamydiae–CFBG and the δ,ϵ-proteobacterial groups (fig. 8.4).[6,12,17] We have recently obtained evidence that this inference is not limited to Aquifex but also applies to various other species belonging to the order Aquificales (viz. *Calderobacterium hydrogenophilum*, *Hydrogenobacter marinus*, and *Thermocrinis ruber*).[21]

The branching order of different groups is also consistent with the major structural differences seen within Bacteria (fig. 8.4). Bacteria can be divided into two distinct groups, depending on whether they are bounded by one membrane (monoderms) or two different membranes separated by a periplasmic compartment (diderms).[4,22] These two groups roughly correspond to the gram-positive and gram-negative bacteria. The signature sequences support this important structural distinction and indicate that of these two groups, the monoderm or gram-positive bacteria are ancestral (fig. 8.4).The deduced branching order also places the Deinococcus-Thermus group in an intermediate position between the gram-positive and gram-negative bacteria. This placement is in accordance with the unique structural characteristics of Deinococcus species, which contain a thick peptidoglycan layer and show positive Gram staining, but they are surrounded by both inner and outer cell membranes, similar to various gram-negative bacteria.[23] The branching of cyanobacteria after the Deinococcus-Thermus group is also of interest because their cell walls are intermediate, in terms of thickness of the peptidoglycan layer and degree of cross-linking, between gram-positive and gram-negative bacteria.[24] These observations are indicative that Deinococcus-Thermus and Cyanobacteria are evolutionary intermediates in the transition from gram-positive bacteria to gram-negative bacteria. The genotype–phenotype correspondence is a central theme of biology, and the picture of bacterial phylogeny that is emerging based on signature sequences is now showing a good correspondence between these two important aspects.[6,12,17]

Signature sequences allow us to finally establish objective criteria for distinguishing the main groups within Bacteria and the major subdivisions within them. In the scheme shown in figure 8.4, all suggested main groups are required to meet two different criteria.

First, all such groups should be clearly distinguishable from the other main groups based on group-specific signature sequences. Second, their branching position should be distinct from all other identified main groups. On the basis of these criteria, the Proteobacterial phylum has been divided into four main groups (δ, ε-, α-, β-, and γ-), each of which is distinct from the others and also branches in a different position.[6,7,12] In contrast, a number of other groups such as Chlamydiae and the CFBG, which branch in the same position based on signature sequences, have not been assigned separate main group statuses, even though Chlamydiae are clearly distinguishable from the CFBG group by a large number of signatures.[11] We have suggested that distinct groups of species that branch in the same position should be tentatively recognized as subdivisions of a given main group or phylum. Using this criterion, the δ- and ε-proteobacteria, whose branching order cannot be distinguished at present,[7] are also placed in the same group. The division of Bacteria into different main groups or their subdivisions as proposed here is based strictly on genealogical considerations, and it reflects both the degrees of similarities among different groups and their hierarchal order, which is the most logical and natural way for understanding evolutionary relationships.[25]

Lateral Gene Transfer—Its Prevalence and Effects

Analyses of whole genome sequences have led to the widespread notions that LGT among prokaryotic organisms is rampant and that it poses a serious problem for discerning evolutionary relationships.[5,9,26–29] A survey of the recent literature gives the impression that LGT among prokaryotes is so pervasive that it has almost completely obliterated any phylogenetic signal resulting from vertical descent (i.e., Darwinian evolution). According to this view, LGT and homology-dependent recombination are the major mechanisms of prokaryotic evolution, and any observed similarities within a group of species is mainly a consequence of selective LGTs.[9,26] Before offering an alternative view, we first briefly examine the evidence that has led to this belief.

The occurrence of LGT has been inferred from a variety of observations. One involves the unexpected branching of species in phylogenetic trees.[30,31] The identification of LGT by this means is based on the assumption that phylogenetic relationships among the groups under consideration are well understood. The vast majority of LGTs that have been identified on this basis involves unexpected branching of an archaeal species within the Bacteria or a bacterial species within Archaea.[30,32,33] Although the Archaea and the Bacteria are presently recognized as two distinct domains, their evolutionary relationship is not well understood (see next section).[34,35] In addition to deduced LGT between Archaea and Bacteria, there have also been several reports describing the branching of individual genes or proteins in unexpected phyletic positions.[11,26,30] The number of such cases is rather limited, however, and does not support the view that LGT between different groups is rampant and occurs indiscriminately.

The major stimulus for the belief in the widespread occurrence of LGT within prokaryotes, particularly Bacteria, has come from analyses of genome sequences using a variety of indirect approaches. They include reliance on atypical base composition or codon usage, use of BLAST searches to ascertain species relatedness, and the presence or absence of genes in closely related genomes.[9,27,29,32,33,36] One of the most influential studies of this nature was by Lawerence and Ochman, who, on the basis of atypical GC content

and pattern of codon usage inferred that about 17.6% of the open reading frames in *Escherihia coli* and *Salmonella* have undergone LGT, after the divergence of these species.[29] Because these species are estimated to have diverged only about 100 million years ago, the inferred high rate of LGT from this study indicated that over long evolutionary periods, LGTs can completely abrogate the evolutionary relationships among prokaryotic organisms.[29] Likewise, several authors have inferred massive LGTs between certain bacterial lineages (Thermotoga, Aquifex, Deinococcus) and Archaea, using BLAST searches to identify closest relatives.[27,32,33,36]

However, many recent studies point out the fallacies of using these approaches to infer the incidence of LGT.[31,37–40] In a detailed study examining the use of BLAST searches to identify closest relatives, Koski and Golding[39] determined that the genes that appear to be most related by BLAST are often not each others' closest relatives, phylogenetically. The extent to which this occurs depends on the availability of closest relatives in the database. Other studies provide evidence that atypical base composition or codon usage are also not reliable indicators of LGTs.[31,37,38,40] Many genes previously classified as laterally transferred using these criteria are, in fact, native.[37] In view of the incongruent results obtained using these approaches for different genes, none of these methods are reliable indicators of LGTs, and much caution needs to be exercised in interpreting the results of such studies.[31,37–41] Alarmed with the growing bandwagon of rampant LGT across different group boundaries, which began with his earlier studies, Ochman[42] has recently warned that such is not the case: "Whereas LGT has certainly been a significant factor in the rapid adaptation and speciation of many bacterial lineages, the overall stability of the genome is, in fact, what allows one to assess the role of lateral gene transfer."

There is little doubt at present that LGT constitutes an important evolutionary mechanism and that it may have affected different genes to various extents. For certain genes whose acquisitions confer selective advantage (e.g., those involved in antibiotic resistance or virulence), one expects that they would be readily transferred from one species to another. However, the extent to which LGT has affected numerous others genes involved in various essential functions remains to be determined. The main challenge before us is to determine whether such genes have also been subjects of rampant LGTs, or whether a significant number of them have undergone either minimal or no LGTs and whether they could provide a stable core of well-preserved molecular fossils, on the basis of which the early evolutionary history can be reliably deduced. I will examine this question here mainly in the context of bacterial evolution, which has been studied in most detail.

To identify LGT, an essential prerequisite is to define all the groups under consideration in unequivocal terms. Without this, it is difficult to assess or quantify LGT. Second, it is also necessary to have a reliable model as to how these groups are related to each other. In the absence of such a model, it is difficult to evaluate whether an observed relationship is natural or whether it is a consequence of LGT. Within Bacteria, a number of main groups have been identified in phylogenetic trees (e.g., Deinococcus-Thermus, Cyanobacteria, low-G+C gram-positive, high-G+C gram-positive, Spirochetes, Chlamydiae-CFBG, and Proteobacteria).[3,5,8] These groups can now be clearly distinguished on the basis of signature sequences in different proteins. Table 8.1 lists a number of signatures that have been identified for the Cyanobacteria.[10] The distribution of these signatures in cyanobacteria and other groups of bacteria is also indicated. We could

now ask the question of whether any LGT for these genes has occurred between cyanobacteria and other bacteria. On the basis of a Darwinian model of evolution, these signatures were introduced in a common ancestor of the cyanobacteria at the time when this lineage evolved. The model predicts that these signatures should be present in various cyanobacterial species but not elsewhere. However, if these genes were either subjects of frequent LGTs, or if such indels could independently arise in various species, then their distribution should differ greatly from that predicted by the Darwinian model. From the observed distribution of these indels in various bacteria, it is clear that these indels are highly specific for cyanobacteria (many of them are also shared by plastid homologs that have originated from cyanobacteria via endosymbiosis),[2,43,44] but they are not found in any other bacteria, even though most of these proteins are present in all such organisms. The statistical significance of these results can be determined by means of a simple χ^2 test, with two degrees of freedom. The χ^2 probability that the observed distribution of these indels in different groups can result from random occurrence or LGTs is virtually nil ($<10^{-15}$). On the basis of other identified signatures, similar strong arguments can be made for most of the other major groups within Bacteria including Proteobacteria, Deinococcus-Thermus, Chlamydiae, Spirochetes, Aquificales, Actinobacteria, and Firmicutes.

If the major groups within Bacteria have not undergone extensive LGT, then one could inquire next whether the interrelationships between these groups have been affected by LGTs. As discussed earlier, at present there is no reliable model as to how different groups within Bacteria are related to each other or how they branched off from a common ancestor.[5,9] In the absence of a reliable model, the branching pattern of these groups in the 16S rRNA trees has been assumed as a working model.[5,8] However, there is no stable or consistent branching pattern of these groups in the rRNA trees, and there are numerous disagreements between these trees and those based on other genes and proteins.[5,8,19]

Although most of these differences have been attributed to LGTs, in the absence of any reliable model, it is difficult to determine that this is actually the case. In our work, based on signature sequences in different proteins, we have proposed a very specific model of how these groups are related and of their branching order from a common ancestor. This model, which makes very definite predictions regarding the presence or absence of indels in different species, allows one to objectively determine whether the genes in question have been affected by LGT. As discussed earlier, the excellent correspondence between the predicted and observed distribution of different main-line indels in 100 completed bacterial genomes (table 8.3) provides strong evidence that the genes containing these indels have not been affected to any significant extent by LGT. The χ^2 probability that the observed distribution of these indels could be caused by random occurrence or LGT is virtually nil ($<10–^{15}$ in all cases), indicating that these indels were introduced only once in the common ancestors of different groups (as indicated in fig. 8.4) and then passed on to other species by vertical descent.

Concerns have been privately expressed that these inferences are based on a small number of indels that may be biased toward portraying the indicated relationship. We emphasize in response that, during our extensive work on signature sequences, we have not detected any other indels that challenge the consistent picture developed here. Furthermore, in contrast to those phylogenetic trees based on a single gene or protein, all inferences drawn here are based on a large number of different and widely distributed

proteins. Our inferences are not restricted to a particular group or family of genes or proteins. The proteins in question are responsible for various essential functions including transcription, translation, replication, DNA repair, protein folding, cell division, metabolic enzymes, and cell wall biosynthesis. The fact that all of these signatures yield a highly consistent picture, and that they are also in accordance with cell structural characteristics, strongly indicates that the model presented here is reliable.

The use of molecular sequences for deducing evolutionary relationships is analogous to evaluating fossil evidence to piece together the evolutionary history of extinct species. Paleontologists give greater credence to well-preserved fossils and much less to those that have been disintegrated. Similarly, molecular sequences vary greatly in terms of their degree of conservation and usefulness for evolutionary studies. During the course of evolution, although some genes and proteins have undergone extensive changes, others are less affected by such factors. Given the long evolutionary history of prokaryotes, it should not be surprising to find many examples of genes that have been affected by LGT. However, it is wrong to infer that all genes have been similarly affected. This would amount to throwing the baby with the bath. As evolutionary scientists, our focus should not be limited to finding examples of genes that have been affected by LGTs, but to identify and discover genes that are minimally affected by LGTs, on which reliable evolutionary models can be developed and validated. Such models should enable us to reliably identify the LGT events as well as to understand why they have occurred.

On the basis of this rationale, our work on signature sequences, still in its initial phases, has focused on documenting highly conserved molecular signatures that are minimally affected by LGT or gene duplication and that can be interpreted with minimal ambiguity. Our model was first developed at a time when fewer than ten bacterial genomes were available,[4] and it is now being rigorously tested for its various predictions using sequence data for >100 bacterial genome. From the outset, our model made specific predictions as to which of these indels should be present or absent in various bacteria for which no sequence information was available. At present, when sequence information for >100 bacterial genomes is available, it is gratifying to note that the predictions made by this model are borne out with such high degree of accuracy (>99.0%) in over 2,000 observations. These results strongly validate our evolutionary model and indicate that the genes and proteins on which it is based provide a stable core that is minimally affected by factors such as LGTs.

Origin of Archaea and Their Relationship to Bacteria

Unlike the evolutionary relationships within Bacteria which are beginning to be understood, the relationship between Archaea and Bacteria (and also the relationships within Archaea) have proven much more difficult to resolve.[4,28,34,35,41,45–50] That the Archaea were distinct from Bacteria (or eubacteria) was first proposed by Woese and his coworkers on the basis of pronounced differences in their 16S rRNA oligonucleotide catalogues and their discrete branching in the 16S rRNA trees.[3,46] The distinction between these two groups was also supported by a number of other characteristics including lack of muramic acid in archaeal cell walls, membrane lipids that contain ether-linked isoprenoid side chains rather than the diacyl esters found in bacteria, distinctive RNA polymerase subunits structures, differences in sensitivity profile to various toxins and antibiotics,

and so forth. Subsequent studies based on duplicated pairs of gene sequences indicated that the root of the universal tree lay between Archaea and Bacteria, with eukaryotic homologs derived from the same branch as Archaea.[13,14] The inference from these studies that Archaea are the closest relatives of Eukarya has had a profound influence on their acceptance as a separate domain. However, it is now firmly established that all eukaryotic cells have received major gene contributions from both Bacteria and Archaea.[4,35,51,52] The ancestral eukaryotic cell, therefore, is not a direct descendent of the archaeal lineage. On the basis of fossil evidence, the eukaryotic organisms have also evolved much later (>2 Ga) than the prokaryotes. Therefore, if Archaea and Bacteria were the only organisms that existed for much of the early history of life, it is important for us take a closer look at them to see how they differ from each other and how such differences possibly arose.

The majority of the genes that indicate Archaea to be different from Bacteria are for the information transfer processes, such as those responsible for DNA replication, transcription, and protein synthesis.[53] Of these, the DNA replication machinery appears to be most different between the two domains. Archaea do not contain typical bacterial DNA polymerases (PolI, Polβ), helicase, or most other proteins (e.g., DnaA, SSB, and DnaG) involved in different stages of DNA replication. Although proteins that carry out analogous functions are present in Archaea, they do not show significant sequence similarity to the bacterial counterparts. In terms of transcription, the core subunits of the RNA polymerase (α, β, and β') are the same in Bacteria and Archaea, but the archaeal enzyme also contains several smaller subunits not present in bacteria.[53] Archaea also contain a variety of transcription factors not found in bacteria.[53] The translation machinery is generally quite similar between Bacteria and Archaea: All rRNAs, most r-proteins, the major elongation factors, various amino acid-charging enzymes and tRNAs, and so forth. are common to both these groups of prokaryotes. The vast majority of r-proteins in Archaea are also arranged in operons similar to that seen in Bacteria. However, Archaea differ from Bacteria in having a small number of unique r-proteins as well as many translation initiation factors.[53]

Apart from these differences and dissimilarities in their cell envelope biosynthetic enzymes,[54] Archaea and Bacteria are extensively similar.[36] Most of the metabolic pathways, which make up the vast majority of any organism's gene repertoire, are common between Archaea and Bacteria. In terms of their cell structures, *Archaea* are indistinguishable from gram-positive bacteria.[4,34] Within prokaryotes, only these two groups of organisms are bounded by a single unit lipid membrane, and they generally contain a thick sacculus of varying chemical composition. Some Archaea also show positive Gram staining, and a few of them (e.g., Thermoplasma), similar to certain gram-positive bacteria (e.g., mycoplasma), are unique in not containing any cell wall.[4] The similarity between Archaea and Bacteria extends to numerous other characteristics including their cellular size, which is much smaller (<100–1000-fold) than that of eukaryotic cells; absence of nucleus; cytoskeleton; histones; spliceosomal introns; circular organization of their genomes; organization of genes into operons; presence of 70 S ribosomes; and so forth.[47] Koonin et al.[36] have reported that about 63% of the genes in *Methanococcus janaschii* are also found in other bacteria, whereas only 5% of them are uniquely shared with Eukarya.[36] Although about one-third the total genes in this Archaea are unique (i.e., no similarity seen to any other organisms), the same is generally true for most other prokaryotic genomes.

The similarity between Archaea and gram-positive bacteria, as noted above, is not limited to their cell structures. In phylogenetic trees based on a number of different proteins, archaeal species show polyphyletic branching within gram-positive bacteria.[4,19,55] If one considers only prokaryotic homologs, then phylogenetic trees for the majority of proteins indicate that the Archaea are more closely related to gram-positive bacteria (i.e., monoderm bacteria, which include Thermotoga) than to gram-negative bacteria (R.S. Gupta, unpublished results).[4,19] Strong evidence of a closer relationship between Archaea and gram-positive bacteria, as compared with gram-negative bacteria, is also provided by several prominent signature sequences (e.g., 21–23-aa indel in Hsp70 [fig. 8.2] and 26-aa indel in GS I), which are commonly and uniquely shared by these two groups of prokaryotes.[4,15]

The question can now be asked of how Archaea and Bacteria are related to each other? Because the majority of the genes that indicate Archaea to be distinct from Bacteria are for the information transfer processes, and because these processes are of fundamental importance, it has been assumed that differences in these regards arose in the universal ancestor before separation of these two domains.[3] According to both Woese and Kandler, these two primary domains, as well as the eukaryotic cells, evolved from a precellular community that contained different types of genes that define these lineages by a process leading to the fixation of specific subsets of genes in the ancestors of these domains.[28,50,56] To account for the presence of different genes, it is postulated that the universal ancestor was not a unique organism but a loose community of precellular entities that evolved independently and also in an interdependent manner. These precellular entities did not have stable genealogy or chromosome, and they also lacked a typical cell membrane, thus allowing unrestricted LGTs among them.[28,50,56] These proposals thus postulate that all differences between Archaea and Bacteria originated at a precellular stage by non-Darwinian means, but they suggest no rationale as to how or why the observed differences between these two groups of prokaryotes arose or evolved.[28,50,56] Cavalier-Smith has suggested the possibility that the Archaea have evolved from gram-positive bacteria as an adaptation to hyperthermophily or hyperacidity,[48] but his proposal fails to explain how various differences in the information transfer genes that distinguish Archaea from Bacteria arose.

Our work offers an alternate proposal as to how Archaea and Bacteria may be related: Archaea are related to gram-positive bacteria, as seen by the striking similarities in their cell structures and by a large number of gene phylogenies.[4,34,35] An important distinctive characteristic of all Archaea is that they are resistant to a wide variety of antibiotics that are primarily produced by gram-positive bacteria.[4] Further, the majority of the genes that indicate Archaea to be distinct from Bacteria are related to either information transfer processes or to synthesis of cell wall and membrane lipids, and they provide the main cellular targets for these antibiotics.[4,35,57] These observations are of central importance for understanding the origin of Archaea.[4,34,35] If the differences that characterize Archaea evolved at a precellular stage before the formation of Bacteria, then it is difficult to understand how Archaea developed resistance to most antibiotics that are produced by gram-positive bacteria. Further, it seems too much of a coincidence that most of the genes that indicate Archaea to be distinct from Bacteria provide the main targets for these antibiotics.

To account for these observations, I have suggested that the earliest groups of prokaryotes that evolved were related to the gram-positive bacteria. The characteristics that

distinguish Archaea from Bacteria, rather than evolving independently at a precellular stage, evolved from gram-positive bacteria in response to antibiotic selection pressure.[4,35] In one plausible scenario, after a certain group of gram-positive bacteria developed the ability to produce different types of antibiotics to survive in this strongly selective environment, some sensitive bacteria underwent extensive changes in genes that provided the targets for these antibiotics. The changes leading to resistance were of different kinds including mutations, insertions and deletions, nonhomologous recombinations, and replacement of the target genes with nonorthologous genes.[58,59] A prolonged and successive selection in this environment led to the eventual development of a resistant strain that had undergone extensive changes in many genes that were targets of these antibiotics, and this strain represented the common ancestor of present-day Archaea.[4,35] The evolution of Archaea in response to antibiotic selection also provides a plausible explanation for their adaptation to harsher environments such as high temperature, high salts, high acidity, and so forth,[3] which could have been a defensive strategy on their part to find niches that are "hostile" to antibiotic-producing organisms.[4,35] Thus, this proposal can logically explain the evolution of most of the distinguishing characteristics of Archaea from known groups of bacteria by normal evolutionary mechanisms without attributing such differences to the unusual properties of the universal ancestor. Because differences between Archaea and Bacteria evolved at a very early stage in prokaryotic history (fig. 8.4), the Archaea appear distinct from Bacteria in phylogenetic trees or other studies based on such characteristics.

The Origin of the Eukaryotic Cell

The origin of the eukaryotic cell and many observations central to understanding this problem have been discussed in detail in earlier work.[4,52] I will present here only a brief overview of the main hypotheses for the formation of the eukaryotic cell and examine them critically in light of the available data. It is now well established that all eukaryotic cells possess a large number of genes (representing significant portions of their genomes) that exhibit greater similarity to either Archaea or Bacteria.[4,19,45,51,52,60] Thus, the original three-domain hypothesis that Archaea and Eukarya (or Eukaryotes) are sister lineages and that the ancestral eukaryotic cell directly evolved from an archaeal ancestor is no longer valid. To account for the unique genotype and phenotype of eukaryotic cells, several hypotheses have been proposed. Some authors have suggested that the eukaryotic cells evolved first and that prokaryotic organisms originated from them by gene loss and simplification processes.[61,62] However, such hypotheses are inconsistent with the fossil and geological evidence. Other hypotheses postulate the existence of hypothetical proto-eukaryotic organisms possessing various distinctive characteristics of the eukaryotes, which later engulfed either an Archaea or both an Archaea and Bacteria to give rise to eukaryotic cells.[63,64] These hypotheses defer an understanding of the origin of the eukaryotic cell to hypothetical entities for which there is no evidence. In a recent proposal, Cavalier-Smith[48] posits that a common ancestor of eukaryotic cells and archaebacteria (termed Neomura) evolved directly from gram-positive bacteria. This proposal is unique in postulating a very late divergence (about 0.85 Ga) of both Archaea and eukaryotes, but it is not supported by the observations

that bacterial genes in eukaryotes are derived from gram-negative bacteria rather than from gram-positive bacteria.[51,52]

The most widely accepted proposals for the origin of eukaryotic cells postulate it to be a chimera formed by fusion or association of distinct lineages of Bacteria and Archaea.[4,52,65–68] These proposals were put forward to account for the observations that whereas information-transfer genes of eukaryotic cells are most closely related to Archaea, their metabolic genes are primarily derived from Bacteria.[19,52,60,69] The chimeric proposals that have been suggested are of two main kinds. One key aspect that distinguishes these proposals is whether the chimeric event that led to the formation of the eukaryotic cell was the same as that which gave rise to mitochondria (mitochondria–nucleus co-origin model) or whether the two events differed from each other both in their nature and timings (nucleus first–mitochondria later models).[4,52,65–68,70] This distinction is important because mitochondria and plastids are already known to be derived from bacteria via endosymbiosis.[2,43,44]

The main impetus for the mitochondria–nucleus co-origin proposals has come from observations that a number of protist lineages that were earlier believed to lack mitochondria have now been shown to contain at least some genes or proteins that are distinctive of mitochondria.[65,71,72] This has led to the view that all eukaryotic species once harbored mitochondria and that absence of mitochondria in some of these lineages is the result of a secondary loss of the organelle. The wide acceptance of the mitochondria-early view in conjunction with the observation that several anaerobic protist lineages contain a mitochondria-related organelle, hydrogenosome (which releases gaseous hydrogen), has led to the "hydrogen hypothesis" for the formation of the eukaryotic cell.[65] According to this hypothesis, the ancestral eukaryotic cell arose as a result of symbiotic association between a hydrogen-dependent archaebacterium (a methanogen) and an α-proteobacterium, which, under anaerobic conditions, produced molecular hydrogen as a waste product. The driving force for the formation of the eukaryotic cell was dependance of the archaebacterium on molecular hydrogen produced by the bacterial symbiont.[65] This association led to the endosymbiotic capture of α-proteobacterium by the archaeal partner, leading to the formation of mitochondria as well as to various other eukaryotic cell characteristics.

Although the hydrogen hypothesis accounts for the origin of hydrogenosomes from mitochondria, it offers no explanation of how any of the main characteristics that define a eukaryotic cell (e.g., nucleus, endoplasmic reticulum [ER]) evolved. The hydrogen hypothesis is also called into question by a number of observations:[4,73] first, there exists now unequivocal evidence that all extant eukaryotic organisms have evolved from a common ancestor, thus indicating that the formation of the eukaryotic cell was a unique event occurring only once.[4] If metabolic symbiosis based on hydrogen production and use was the main driving force for the formation of the eukaryotic cell, then given the widespread association between methanogenic archaea and hydrogen-producing proteobacteria, it is difficult to envisage why this sort of syntrophy has not lead to the formation of eukaryotic-like cells on numerous independent occasions. Second, in all well-established cases of endosymbiosis (e.g., formation of mitochondria and plastids), the metabolic processes that have formed the basis of symbiosis have been retained,[2] yet eukaryotes have not retained any genes for methanogenesis, the proposed basis of their origin. Third, the α-proteobacteria, which the hydrogen hypothesis suggests as the

bacterial partner in the syntrophic association, evolved much later than cyanobacteria.[17] This implies that formation of the eukaryotic cell took place in an aerobic atmosphere, but the symbiosis between an anaerobic hydrogen-producing bacterium and a strictly anaerobic methanogenic archaebacterium would have produced an anaerobic organism, which would thus be at a great selective disadvantage. Fourth, the hydrogen hypothesis provides no explanation as to why eukaryotic genes for the information transfer processes are derived from the archaeal partner.[51,52,60] Finally, molecular sequence data indicate that thermoacidophilic archaea rather than methanogens are the closest relatives of eukaryotes.[4,74]

All alternative chimeric proposals posit that the primary fusion (or endosymbiotic) event that led to the formation of eukaryotic cell was distinct and that it preceded the formation of mitochondria. According to Margulis's most recent proposal, the ancestral eukaryotic cell was formed by association of a spirochete and a Thermoplasma, in which the Thermoplasma contributed the nucleocytoplasm, whereas the motility apparatus (e.g., microtubules) was provided by the spirochete.[67] However, there is no evidence at present that any of the eukaryotic genes, including those for motility functions, have been derived from spirochetes.[4] Recently, Jenkins et al.[75] have identified a protein showing strong sequence and biochemical similarity to tubulin in Prosthecobacter, indicating for the first time a possible origin of this key cytoskeletal protein. Lake and Rivera[66] have proposed that the nucleus is an endosymbiont that arose from the engulfment of an eocyte (or Crenarchaeota) archaea by a gram-negative bacterium. However, this model ignores the fact that the nucleus is not an endosymbiont in the same sense as mitochondria and plastids, which have retained their information transfer machinery and are specifically related to their parental lineages.[2,43] Zillig[76] suggested the possibility that the ancestral eukaryotic cell was formed by primary fusion of an archaebacterium and a eubacterium. However, the nature of the proposed fusion event was not elaborated.

A detailed chimeric proposal for the nucleus-first, mitochondria-later origin has emerged from our work on some of the best-characterized protein families in eukaryotic cells.[4,52,55,70] The Hsp70 family of proteins represents one such family. Distinct homologs of Hsp70 that are encoded for by different genes are present in mitochondria, cytosol, and the ER compartments.[4,70] The mitochondrial and hydrogenosomal homologs of Hsp70 are clearly derived from α-proteobacteria, as evidenced by phylogenetic analyses and many common signature sequences.[4] In contrast, the homologs of Hsp70 that are present in the cytosol and ER compartments (referred to as nuclear-cytosolic homologs), which are also derived from bacteria, show no relationship to the mitochondrial homologs.[4,52,70] These homologs contain a large number of signature sequences that are not present in any mitochondrial, hydrogenosomal, or prokaryotic homologs. These signatures are thus uniquely eukaryotic, and they were likely introduced in the Hsp70 gene at a very early stage in the formation of the eukaryotic cell.[4,52,70] The absence of these signatures in mitochondrial and hydrogenosomal homologs provides strong evidence that nuclear-cytosolic homologs have originated independently of these organelles. Importantly, although the cytosolic and ER Hsp70 homologs are present in all eukaryotic organisms, in the earliest branching eukaryotic lineages such as Giardia, no Hsp70 gene that qualifies as a mitochondrial homolog has been detected. Morrison et al.[77] have recently identified an Hsp70 homolog in Giardia that is distinct from the nuclear-cytosolic homologs. However, this highly divergent protein shows no specific affinity to the mitochondrial homologs.

Because ER forms the nuclear envelope, an understanding of events leading to its evolution (or proteins found in this compartment) is directly relevant to the origin of the nucleus. Phylogenetic analyses of Hsp70 and Hsp90 sequences indicate that the ER and cytosolic homologs of these proteins in different eukaryotic organisms including Giardia are the results of ancient gene duplication events.[70,78] Thus, a very early event associated with the formation of the ER (and via inference the nuclear envelope) involved duplication of genes for these proteins.[70,78]

To explain these observations, and the chimeric nature of eukaryotic nuclear-cytosolic genes, we have proposed that the ancestral eukaryotic cell evolved as a result of symbiotic association between a gram-negative bacterium (related to proteobacteria or the CFBG group) and an Archaea (fig. 8.5).[4,52,73] This symbiosis developed in an aerobic environment predominated by antibiotic-producing organisms. A combination of these two selective forces (oxygen and antibiotics sensitivities) led to the association of an antibiotic-resistant and oxygen-sensitive archaea with an oxygen-tolerant (or oxygen-using) and antibiotic-sensitive bacterium, which provided mutual protection in this environment.[4,73] The association of these two groups of prokaryotes led to the surrounding of the archaea by membrane enfolds from the bacterial partner to shield it from oxygenic environment (fig. 8.5). The cell membrane of the archaea became redundant under these conditions, and it was eventually lost. At a later stage, the membrane enfolds surrounding the archaea got separated from the bacterial membrane. This led to formation of the endomembrane system (or the ER), as well as of nucleus in the cell. Because this newly formed compartment (i.e., the ER) had to communicate (i.e., import and export proteins and other molecules) with the rest of the cell, its formation was either accompanied or preceded by duplication of the genes for the Hsp70 and Hsp90

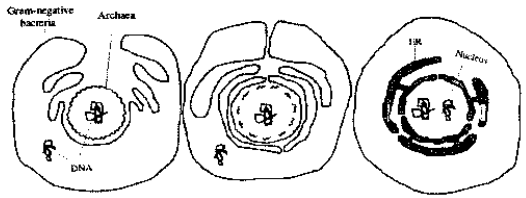

Figure 8.5. The primary fusion model for the origin of the eukaryotic cell. According to this model, the ancestral eukaryotic cell, which contained nucleus and endomembrane system, was formed before the endosymbiotic event that led to acquisition of mitochondria (not shown here). The key event leading to its formation was a long-term symbiosis between a gram-negative bacterium and an archaebacterium. This symbiosis developed in an oxygenic and antibiotic-rich environment, and its basis was sensitivity to antibiotics of the bacterial partner and the oxygen-sensitivity of the archaebacterium. As the membrane of the gram-negative bacterium surrounded the archaebacterium, its membrane (containing ether-linked lipids, shown by the wavy line) became redundant and was lost. The separation of the bacterial membrane folds surrounding the archaebacterium led to formation of the nuclear envelope and endoplasmic reticulum (ER). The resulting cell was antibiotic resistant and oxygen tolerant, and it retained the majority of the genes for the information-transfer processes, which provide main targets for antibiotics, from the archaeal partner.

chaperones, which are essential for this purpose.[70,78] The formation of this new antibiotic-resistant and oxygen-tolerant bacterium was accompanied by an assortment of genes from the two parents. During this process, most of the genes for information-transfer processes (which provide the main targets for antibiotics) were mainly retained from the archaea, whereas those for the metabolic processes were acquired from the bacterial partner.[4,73] Various eukaryotic-specific signatures were also introduced into different genes at this early stage. The transfer of all of these genes into the newly formed nuclear compartment led to integration (or primary fusion) of the original symbionts into a new type of cell, which became the prototype eukaryotic cell (fig. 8.5).[4]

Concluding Remarks

Using molecular sequence data, it is now possible to develop a reliable picture of bacterial phylogeny, where all the main groups can be clearly distinguished and their branching order can be logically inferred. This emerging picture is also consistent with the structural characteristics of prokaryotes. Our proposed model for prokaryotic evolution makes very specific predictions that are strongly corroborated by the genome sequence data. LGT, though an important evolutionary mechanism, is not a serious problem for the determination of bacterial phylogeny. The origin of the Archaea and their relationship to Bacteria remains a contentious issue. The current view is that all three domains have evolved from a precellular community containing different types of genes by an annealing process that led to stabilization of a subset of the genes in common ancestors of these domains. However, the fact that Archaea are resistant to most antibiotics produced by gram-positive bacteria, and that majority of the genes that indicate them to be distinct from Bacteria provide targets for these antibiotics, support our alternate proposal: that they could have evolved from gram-positive bacteria in response to antibiotic selection pressure. The formation of the eukaryotic cell constitutes an evolutionary discontinuity that is explained by their origin from fusion of different groups of prokaryotes. Given the unique and unusual nature of this fusion event, a clear understanding as to how the ancestral eukaryotic cell originated remains to be achieved.

References

1. J.W. Schopf, "The Evolution of the Earliest Cells," *Sci. Am.* 239 (1978): 110–120.
2. L. Margulis, *Symbiosis in Cell Evolution* (New York: W.H. Freeman, 1993).
3. C.R. Woese, "Bacterial evolution," *Microbiol. Rev.* 51 (1987): 221–271.
4. R.S. Gupta, "Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes," *Microbiol. Mol. Biol. Rev.* 62 (1998): 1435–1491.
5. W. Ludwig and H.-P. Klenk, "Overview: A Phylogenetic Backbone and Taxonomic Framework for Prokaryotic Systematics," in D.R Boone and R.W. Castenholz, eds., 2nd ed. *Bergey's Manual of Systematic Bacteriology* (Berlin, Springer, 2001), 49–65
6. R.S. Gupta and E. Griffiths, "Critical Issues in Bacterial Phylogenies," *Theor. Popul. Biol.* 61 (2002): 423–434.
7. R.S. Gupta, "The Phylogeny of Proteobacteria: Relationships to Other Eubacterial Phyla and Eukaryotes," *FEMS Microbiol. Rev.* 24 (2000): 367–402.
8. G.J. Olsen, C.R. Woese, and R. Overbeek, "The Winds of (Evolutionary) Change: Breathing New Life into Microbiology," *J. Bacteriol.* 176 (1994): 1–6.

9. W.F. Doolittle, "Phylogenetic Classification and the Universal Tree," *Science* 284 (1999): 2124–2128.
10. R.S. Gupta, M. Pereira, C. Chandrasekera, and V. Johari, "Molecular Signatures in Protein Sequences that are Distinctive of Cyanobacteria and Plastids," *Int. J. Syst. Evol. Microbiol.* 53 (2003): 1833–1842.
11. E. Griffiths and R.S. Gupta, "Protein Signatures Distinctive of Chlamydial Species: Horizontal Transfer of Cell Wall Biosynthesis Genes *glmU* from Archaebacteria to Chlamydiae, and *murA* between Chlamydiae and *Streptomyces*," *Microbiology* 148 (2002): 2541–2549.
12. R.S. Gupta, "Phylogeny of Bacteria: Are We Now Close to Understanding it?" *ASM News* 68 (2002): 284–291.
13. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary Relationship of Archaebacteria, Eubacteria, and Eukaryotes Inferred from Phylogenetic Trees of Duplicated Genes," *Proc. Natl. Acad. Sci. USA* 86 (1989): 9355–9359.
14. S.L. Baldauf, J.D. Palmer, and W.F. Doolittle, "The Root of the Universal Tree and the Origin of Eukaryotes Based on Elongation Factor Phylogeny," *Proc. Natl. Acad. Sci. USA* 93 (1996): 7749–7754.
15. R.S. Gupta and B. Singh, "Cloning of the HSP70 Gene from *Halobacterium marismortui*: Relatedness of Archaebacterial HSP70 to its Eubacterial Homologs and a Model for the Evolution of the HSP70 Gene," *J. Bacteriol.* 174 (1992): 4594–4605.
16. H.P. Klenk et al., "RNA Polymerase of *Aquifex pyrophilus*: Implications for the Evolution of the Bacterial *rpoBC* Operon and Extremely Thermophilic Bacteria," *J. Mol. Evol* 48 (1999): 528–541.
17. R.S. Gupta, "Evolutionary Relationships among Photosynthetic Bacteria," *Photosynth. Res.* 76 (2003): 173–183.
18. R.S. Gupta, "The Branching Order and Phylogenetic Placement of Species from Completed Bacterial Genomes, Based on Conserved Indels Found in Various Proteins," *Int. Microbiol.* 4 (2001): 187–202.
19. J.R. Brown and W.F. Doolittle, "*Archaea* and the Prokaryote-to-Eukaryote Transition," *Microbiol. Rev.* 61 (1997): 456–502.
20. G. Deckert, et al. "The Complete Genome of the Hyperthermophilic Bacterium *Aquifex aeolicus*," *Nature* 392 (1998): 353–358.
21. E. Griffiths and R.S. Gupta, "Signature Sequences in Widely Distributed Proteins Provide Evidence for the Late Divergence of the Order *Aquificales*," *Int. Microbiol.* 7 (2004): 41–52.
22. R.Y. Stanier, E.A. Adelberg, and J.L. Ingraham, *The Microbial World* (Engelwood Cliffs, NJ: Prentice-Hall, 1976).
23. R.G.E. Murray, "The Family Deinococcaceae," in A. Balows, H.G. Truper, M. Dworkin, W. Harder, and K.H. Schleifer eds., *The Prokaryotes* (New York: Springer, 1992), 3732–3744.
24. E. Hoiczyk and A. Hansel, "Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope," *J. Bacteriol.* 182 (2000): 1191–1199.
25. Charles Darwin, *On the Origin of Species 1859 A Facsimile of the First Edition*. (Cambridge, MA: Harvard University Press, 1964).
26. J.P. Gogarten, W.F. Doolittle, and J.G. Lawrence, "Prokaryotic Evolution in Light of Gene Transfer," *Mol. Biol. Evol.* 19 (2002): 2226–2238.
27. R. Jain, M. Rivera, and J.A. Lake, "Horizontal Gene Transfer among Genomes: The Complexity Hypothesis," *Proc. Natl. Acad. Sci. USA* 96 (1999): 3801–3806.
28. C.R. Woese, "On the Evolution of Cells," *Proc. Natl. Acad. Sci. USA* 99 (2002): 8742–8747.
29. J.G Lawrence and H. Ochman, "Molecular Archaeology of the *Escherichia coli* Genome," *Proc. Natl. Acad. Sci. USA* 95 (1998): 9413–9417.
30. M.W. Smith, D.F. Feng, and R.F. Doolittle, "Evolution by Acquisition: The Case for Horizontal Gene Transfers," *Trends Biochem. Sci.* 17 (1992): 489–493.
31. M.A. Ragan, "Detection of Lateral Gene Transfer Among Microbial Genomes," *Curr. Opin. Genet. Dev.* 11 (2001): 620–626.
32. K.E. Nelson, et al., "Evidence for Lateral Gene Transfer between Archaea and Bacteria from Genome Sequence of *Thermotoga maritima*," *Nature* 399 (1999): 323–329.

33. L. Aravind, R.L. Tatusov, Y.I. Wolf, D.R. Walker, and E.V. Koonin, "Evidence for Massive Gene Exchange between Archaeal and Bacterial Hyperthermophiles," *Trends Genet.* 14 (1998): 442–444.
34. R.S. Gupta, "What are Archaebacteria: Life's Third Domain or Monoderm Prokaryotes Related to Gram-Positive Bacteria? A New Proposal for the Classification of Prokaryotic Organisms," *Mol. Microbiol.* 29 (1998): 695–708
35. R.S. Gupta, "The *Natural* Evolutionary Relationships among Prokaryotes," *Crit. Rev. Microbiol.* 26 (2000): 111–131.
36. E.V. Koonin, A.R. Mushegian, M.Y. Galperin, and D.R.Walker, "Comparison of Archaeal and Bacterial Genomes: Computer Analysis of Protein Sequences Predicts Novel Functions and Suggests a Chimeric Origin for the Archaea," *Mol. Microbiol.* 25 (1997): 619–637.
37. L.B. Koski, R.A. Morton, and G.B. Golding, "Codon Bias and Base Composition are Poor Indicators of Horizontally Transferred Genes," *Mol. Biol. Evol.* 18 (2001): 404–412.
38. J.A. Eisen, "Horizontal Gene Transfer among Microbial Genomes: New Insights from Complete Genome Analysis," *Curr. Opin. Genet. Dev.* 10 (2000): 606–611.
39. L.B. Koski and G.B. Golding, "The Closest BLAST Hit is Often not the Nearest Neighbor," *J. Mol. Evol.* 52 (2001): 540–542.
40. B. Wang, "Limitations of Compositional Approach to Identifying Horizontally Transferred Genes," *J. Mol. Evol.* 53 (2001): 244–250.
41. R.F. Doolittle, "Searching for the Common Ancestor," *Res. Microbiol.* 151 (2000): 85–89.
42. H. Ochman, "Lateral and Oblique Gene Transfer," *Curr. Opin. Genet. Dev.* 11 (2001): 616–619.
43. M.W. Gray, "The Endosymbiont Hypothesis Revisited," *Int. Rev. Cytol.* 141 (1992): 233–357.
44. C.W. Morden, C.F. Delwiche, M. Kuhsel, and J.D. Palmer, "Gene Phylogenies and the Endosymbiotic Origin of Plastids," *Biosystems* 28 (1992): 75–90.
45. R.S. Gupta, "Life's third domain (*Archaea*): An Established Fact or an Endangered Paradigm? A New Proposal for Classification of Organisms Based on Protein Sequences and Cell Structure," *Theor. Popul. Biol.* 54 (1998): 91–104.
46. C.R. Woese, O. Kandler, and M.L. Wheelis, "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya," *Proc. Natl. Acad. Sci. USA* 87 (1990): 4576–4579.
47. E. Mayr, "Two Empires or Three?" *Proc. Natl. Acad. Sci. USA* 95 (1998): 9720–9723.
48. T. Cavalier-Smith, "The Neomuran Origin of Archaebacteria, the Negibacterial Root of the Universal Tree and Bacterial Megaclassification," *Int. J. Syst. Evol. Microbiol.* 52 (2002): 7–76.
49. S.L. Lyons, "Thomas Kuhn is Alive and Well: the Evolutionary Relationships of Simple Life Form–A Paradigm Under Siege," *Perspect. Biol. Med.* 45 (2002): 359–376.
50. O. Kandler, "Thermophiles: The Keys to Molecular Evolution and the Origin of Life?" in J. Wiegel and W.W. Adams, eds. (Athens: Taylor and Francis, 1998), 19–31.
51. S. Ribeiro and G.B. Golding, "The Mosaic Nature of the Eukaryotic Nucleus," *Mol. Biol. Evol.* 15 (1998): 779–788.
52. R.S. Gupta and G.B. Golding, "The Origin of the Eukaryotic Cell," *Trends Biochem. Sci.* 21 (1996): 166–171.
53. G.J. Olsen and C.R. Woese, "Archaeal Genomics: An Overview," *Cell* 89 (1997): 991–994.
54. O. Kandler and H. Konig, "The Biochemistry of Archaea (Archaebacteria)," in M. Kates. D.J. Kushner, and A.T. Matheson, eds. (New York: Elsevier Science, 1993), 223–259.
55. R.S. Gupta and G.B. Golding, "Evolution of HSP70 Gene and Its Implications Regarding Relationships between Archaebacteria, Eubacteria, and Eukaryotes," *J. Mol. Evol.* 37 (1993): 573–582 .
56. C.R. Woese, "The Universal Ancestor," *Proc. Natl. Acad. Sci. USA* 95 (1998): 6854–6859.
57. J. Davies, "Inactivation of Antibiotics and the Dissemination of Resistance Genes," *Science* 264 (1994): 375–382.
58. E.V. Koonin, and M.Y. Galperin, "Prokaryotic Genomes: The Emerging Paradigm of Genome-Based Microbiology," *Curr. Opin. Genet. Dev.* 7 (1997): 757–763.

59. P. Forterre, "Displacement of Cellular Proteins by Functional Analogues from Plasmids or Viruses Could Explain Puzzling Phylogenies of Many DNA Informational Proteins," *Mol. Microbiol.* 33 (1999): 457–465.
60. M. Rivera, R. Jain, J.E. Moore, and J.A. Lake, "Genomic Evidence for Two Functionally Distinct Gene Classes," *Proc. Natl. Acad. Sci. USA* 95 (1999): 6239–6244.
61. A. Poole, D. Jeffares, and D. Penny, "Early Evolution: Prokaryotes, the New Kids on the Block," *Bioessays* 21(1999): 880–889.
62. P. Forterre and H. Philippe, "Where is the Root or the Universal Tree of Life," *Bioessays* 21 (1999): 871–879.
63. M.L. Sogin, "Early Evolution and the Origin of Eukaryotes," *Curr. Opin. Genet. Dev.* 1, (1991): 457–463.
64. H. Hartman and A. Fedorov, "The Origin of the Eukaryotic Cell: A Genomic Investigation," *Proc. Natl. Acad. Sci. USA* 99 (2002): 1420–1425.
65. W. Martin and M. Müller, "The Hydrogenosome Hypothesis for the First Eukaryote," *Nature* 392 (1998): 37–41.
66. J.A. Lake and M.C. Rivera, "Was the Nucleus the First Endosymbiont?" *Proc. Natl. Acad. Sci. USA* 91 (1994): 2880–2881.
67. L. Margulis, "Archaeal–Eubacterial Mergers in the Origin of Eukarya: Phylogenetic Classification of Life," *Proc. Natl. Acad. Sci. USA* 93 (1996): 1071–1076.
68. V.V. Emelyanov, "Mitochondrial Connection to the Origin of Eukaryotic Cell," *Eur. J. Biochem.* 270 (2003): 1599–1618.
69. G.B. Golding and R.S Gupta, "Protein-Based Phylogenies Support a Chimeric Origin for the Eukaryotic Genome," *Mol. Biol. Evol.* 12 (1995): 1–6.
70. R.S. Gupta, K. Aitken, M. Falah, and B. Singh, "Cloning of *Giardia lamblia* Heat Shock Protein HSP70 Homologs: Implications Regarding Origin of Eukaryotic Cells and of Endoplasmic Reticulum," *Proc. Natl. Acad. Sci. USA* 91 (1994): 2895–2899.
71. B.J. Soltys and R.S. Gupta, "Presence and Cellular Distribution of a 60-kDa Protein Related to Mitochondrial hsp60 in *Giardia lamblia*," *J. Parasitol.* 80 (1994): 580–590.
72. A.J. Roger, et al., "A Mitochondrial-Like Chaperonin 60 Gene in *Giardia lamblia*: Evidence that Diplomonads Once Harbored an Endosymbiont Related to the Progenitor of Mitochondria," *Proc. Natl. Acad. Sci. USA* 95 (1998): 229–234.
73. R.S.Gupta, "Origin of Eukaryotic Cells: Was Metabolic Symbiosis Based on Hydrogen the Driving Force?" *Trends Biochem. Sci.* 24 (1999): 423.
74. M.C. Rivera and J.A. Lake, "Evidence that Eukaryotes and Eocyte Prokaryotes Are Immediate Relatives," *Science* 257 (1992): 74–76.
75. C. Jenkins, et al., "Genes for the Cytoskeletal Protein Tubulin in the Bacterial Genus *Prosthecobacter*," *Proc. Natl. Acad. Sci. USA* 99 (2002): 17049–17054.
76. W. Zillig, "Comparative Biochemistry of Archaea and Bacteria," *Curr. Opin. Genet. Dev.* 1 (1991): 544–551.
77. H.G. Morrison, A.J. Roger, T.G. Nystul, F.D. Gillin, and M.L. Sogin, "*Giardia lamblia* Expresses a Proteobacterial-Like DnaK Homolog," *Mol. Biol. Evol.* 18 (2001): 530–541.
78. R.S. Gupta, "Phylogenetic Analysis of the 90 kD Heat Shock Family of Protein Sequences and an Examination of the Relationship among Animals, Plants, and Fungi Species," *Mol. Biol. Evol.* 12 (1995): 1063–1073.