

Stabilization against Hyperthermal Denaturation through Increased CG Content Can Explain the Discrepancy between Whole Genome and 16S rRNA Analyses[†]

T. E. Meyer*‡ and A. K. Bansal§

Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, Arizona 85721, and
Department of Computer Science, Kent State University, Kent, Ohio 44242

Received February 15, 2005; Revised Manuscript Received June 3, 2005

ABSTRACT: Based largely upon analysis of ribosomal RNA, a third domain of life, called archaea, had been proposed in addition to bacteria and eukaryotes. However, quantitative analysis of 73 whole genomes shows only a two-domain division of life: into eukaryotes and prokaryotes. Thousands of orthologous genes in archaea and bacteria show an essentially unimodal distribution of sequence identities. Thus, whole genome analyses indicate that archaea are a phylum of bacteria rather than a separate domain of life. In contrast, archaeal rRNA and that of hyperthermophilic bacteria differ from the rRNA of mesophilic bacteria. Thus, there is a bimodal distribution of rRNA sequence identities which differ by 12%. This discrepancy in rRNA and gene content based analyses of whole genomes is likely due to a 15% elevated C:G content of the rRNA of archaea and hyperthermophilic bacteria. The elevated C:G content is consistent with stabilization against thermal denaturation caused by additional hydrogen bonding (3 bonds) in C:G pairs compared to A:U pairs (2 bonds). Based upon this premise, there is no reliable way to correct rRNA for such differences in base composition and it is not possible to quantitatively compare hyperthermophiles with mesophiles by the rRNA method. Furthermore, quantitative study of whole genomes shows that the extent of change in both bacterial and archaeal genes, including rRNA, has reached a limit. Thus, direct sequence comparisons work with closely related genomes, but it is not possible to differentiate the most divergent prokaryotic species, which are currently designated as separate phyla. We believe that the differences in characteristics of archaeal species is based primarily upon selection of genes and pathways compatible with the extreme environmental lifestyle, i.e., hyperthermophily.

The use of protein sequences for evolutionary analysis was pioneered by Zuckerkandl and Pauling (1) using hemoglobin. Cytochrome *c*, which is more widespread and conservative than are the globins, has also been used extensively (2). However, neither protein is universally distributed among prokaryotes. Consequently, ribosomal RNA, due to its widespread occurrence, its larger size, and even more conservative sequence, became the molecule of choice for evolutionary analysis (3) with the current database numbering more than 16 000 species (4). These studies have had a particularly significant impact on current bacterial classification (5) with one of the major conclusions being that there is a new division of life, called the archaeabacteria (6), which was eventually elevated to a higher rank, called the archaea, that was reported to be equal in stature to eukaryotes and bacteria (7). This is known as the three-domain hypothesis.

The methods of sequence analysis used to arrive at these conclusions are generally accepted as reliable, although unrelated approaches have led to differing results. The availability of whole genome sequences now permits quantitative analysis of thousands of genes for comparison with

RNA. Thus, we recently developed methods for analysis of whole genomes and concluded that there are only two domains of life (8). We herein examine possible reasons for the discrepancy.

When measuring the extent of change in one species compared with another, it is important to determine the maximum possible amount of change in the property being studied, for example, in sequence identities. It is thought that “slowly evolving” genes such as rRNA allow one to sample a greater taxonomic range than with the average protein, which appears to change at a more rapid rate, but that has never been proven. It is furthermore implied that rRNA should eventually reach the same limit to change as the average protein if given enough time. However, there are only 4 different bases and 20 amino acids and, based upon average composition alone, the maximum amount of change in nucleic acids should be significantly less than in proteins. This alone would prevent rRNA from reaching the same limit to change as a protein. Thus, they are not strictly comparable. Proteins also vary in their apparent rate of change, and there should be an explanation for this phenomenon in addition to variable compositions because such variation is generally too small to account for the effect. The most likely explanation for variable rates of change is that there are many bases and amino acids that are required to maintain structure and function. Thus, it is not possible to reach the compositional limit to change. The functional limit is expected to

[†] T.E.M. was supported by a grant from the National Institutes of Health, GM21277.

* Corresponding author. Phone: 520-621-5256. Fax: 520-621-6603.
E-mail: temeyer@u.arizona.edu.

‡ University of Arizona.

§ Kent State University. Fax: 330-672-7824. E-mail: arvind@cs.kent.edu.

vary from one gene to another, depending on how many bases or residues are required to remain constant for maintenance of structure and function, accounting for the illusion of rapidly and slowly evolving genes. There are many proteins, similar to rRNA, that have a relatively small limit to change as we have demonstrated in our previous paper (8). For those species comparisons in which the proteins have apparently reached a functional limit to change, it is reasonable to assume that the rRNA should also have reached such a functional limit to change. The only actual advantage of a slowly evolving gene appears to be in the improved capability for alignment of the most divergent sequences, which in effect will extend the range somewhat. Thus, one aspect of our study is to further examine whether genes and proteins have reached a functional limit to change in the most divergent species and at what taxonomic level that limit is reached.

Whole genome analysis, which was initiated with the complete sequence of *Haemophilus influenzae* (9), now includes over 140 published species and has further transformed our thinking about bacterial evolution in that we now know that gene transfer, duplication, and loss are relatively common (10). As a consequence, evolutionary analyses based upon single genes or combinations thereof including rRNA have been cast in doubt (11). The errors resulting from single gene analyses should be minimized by utilization of the thousands of genes identified from complete genome sequences. The major advantage of whole genome analysis is that errors due to misidentification of orthologues, to misalignment, to the order in which sequences are aligned, to localized gene transfer, and to lineage-specific variation in rates of change are reduced through utilization of all of the genes rather than a subset.

There is some debate on the best way to compare whole genome data (12–16). There are two major approaches for use of whole genome data to study genome evolution: gene content based analysis (8) and genome rearrangement-based analysis (17). Genome rearrangement appears to be relatively common, it is far more complex, and it includes the study of gene acquisition and loss as well as statistical analysis of gene and gene-group duplications (17). These analyses include recognition of multiple gene domain fusions/fissions. While rearrangements are important to understand the characteristics and proteomics of the genomes and pathways, these methods are still evolving, and are largely handicapped due to the lack of knowledge of the correct annotation in the absence of experimental verification of functionality of the genes.

We have developed two methods for the analysis of gene content (8): normalized gene content (NGC) and median identity of orthologues (MIO). These methods differ from previous analyses in that the NGC method corrects for the influence of genome size on gene content (17) and the MIO method accounts for the possible limit to change of sequence identities inherent in the structure/function relationship (18). We previously applied the NGC and MIO methods to 35 available species of archaea, bacteria, and eukaryotes (8). The quantitative results showed that only eukaryotes are significantly different by these criteria (8). NGC indicates that eukaryotes share significantly fewer genes with archaea as well as with bacteria. However, archaea and bacteria have essentially the same overall gene content. MIO shows that

archaeal genes have reached approximately the same limit to change as have the bacterial genes and cross-comparison of the two showed only an insignificantly lower distribution of sequence identities as for the individual domains. These results favor the bipartite division of life into eukaryotes and prokaryotes. However, it is important to demonstrate that the size of the database does not bias the results, especially in terms of the limit to change in genes and proteins. If the genes have reached a limit to change, then doubling the size of the database should not affect the outcome, but if they are still capable of further change, the limit will increase.

To examine the apparent discrepancy between the three-domain theory suggested by 16S rRNA analysis (6, 7) and the identification of only two domains derived by quantitative whole genome analysis (8), we compared the 16S rRNA sequences for the same set of species as used for whole genome analysis. In addition, the size of the database was doubled from 37 to 75 species (57 bacteria, 16 archaea, and 2 eukaryotes). The rationale for reanalyzing the rRNAs was (1) to take into account the most recent results obtained by genome sequencing, (2) to reduce any possibility of error in the previous databases collected over years by different groups often using oligonucleotide libraries involving essentially incomplete sequences, (3) to reduce the possibility of variation of evolutionary trees when new species are added, and (4) to use the same comparison techniques both for whole genome and for rRNAs. Thus, the topology of evolutionary trees based upon 16S rRNA analysis can change when new species are added to the database, presumably due in part to differing choices of parameters, to the resulting alignments, and to the order of alignment.

Our quantitative results confirmed the discrepancy (see Table 1); gene-content analysis showed significant overlap between archaea and bacteria, but 16S rRNA based analysis showed separation between mesophilic bacteria and archaea. However, hyperthermophilic bacteria also appeared to be different from the mesophiles based upon 16S rRNA. Our analysis showed that differences in rRNA base composition caused by elevated C:G content due to a hyperthermophilic lifestyle of a majority of the archaea plus a few bacteria are a likely cause of the discrepancy although there could be other contributing factors. This base compositional bias in thermophilic rRNA has been well documented (19–22) but was thought to result in localized artifacts only (23, 24). However, we maintain that the higher C:G content in hyperthermophilic species, particularly in the archaeabacteria, strongly affects their placement in rRNA trees.

METHODS

Whole genome data of main chromosomes and the corresponding plasmids for 75 species (57 bacteria, 16 archaea, 2 eukaryotes) were extracted from Genbank. Since some of the associated plasmids have a large number of genes, we included both main chromosomes and the corresponding plasmids for pairwise genome comparisons to identify orthologues. Some of the plasmids are approaching the size of chromosomes, and it is likely that there has been some gene transfer from chromosome to plasmid, both of which justify their inclusion in the analysis.

Orthologues are those homologues that share the same functional role and common descent, whereas paralogues

Table 1: Genes in Chromosome(s) and Sequenced Plasmids Archived in Genbank, Chromosomal C:G Content, rRNA C:G Content, and Optimum Growth Temperature^a

	genes				
	chr	plas	C:G(DNA)	C:G(RNA)	T _{opt}
Archaea (16 Species)					
<i>Aeropyrum pernix</i> K1 DSM11879 ^T	1841 (1)	0 (0)	56	65	95
<i>Archaeoglobus fulgidus</i> VC16 DSM4304 ^T	2420 (1)	0 (0)	49	63	83
<i>Halobacterium</i> sp. NRC1	2075 (1)	547 (2)	67	57	40
<i>Methanosa</i> cina acetivorans C2A DSM2834 ^T	4540 (1)	0 (0)	43	55	35
<i>Methanosa</i> cina mazei Go1	3371 (1)	0 (0)	42	54	35
<i>Methanopyrus</i> kandleri AV19 DSM6324 ^T	1687 (1)	0 (0)	62	69	105
<i>Methanocaldococcus jannaschii</i> JAL1 DSM2661 ^T	1786 (1)	0 (0)	31	64	85
<i>Methanothermobacter thermoautotrophicus</i> ΔH ^T	1873 (1)	0 (0)	50	57	65
<i>Pyrobaculum aerophilum</i> IM2 DSM7523 ^T	2605 (1)	0 (0)	51	65	100
<i>Pyrococcus abyssi</i> GE5	1896 (1)	0 (0)	45	66	96
<i>Pyrococcus furiosus</i> VC1 DSM3638 ^T	2125 (1)	0 (1)	41	63	100
<i>Pyrococcus horikoshii</i> OT3 DSM12428 ^T	1956 (1)	0 (0)	42	63	98
<i>Sulfolobus solfataricus</i> P2 DSM1617	2977 (1)	0 (0)	36	63	80
<i>Sulfolobus tokodaii</i> 7 DSM16993 ^T	2825 (1)	0 (0)	33	64	80
<i>Thermoplasma acidophilum</i> DSM1728 ^T	1482 (1)	0 (0)	46	53	59
<i>Thermoplasma volcanium</i> GSS1 DSM4299 ^T	1499 (1)	0 (0)	38	53	60
Proteobacteria (25 Species)					
<i>Agrobacterium tumefaciens</i> C58 _UWASH	4751 (2)	741 (2)	62	54	27
<i>Brucella melitensis</i> 16M	3198 (2)	0 (0)	57	54	37
<i>Buchnera aphidicola</i> APS	564 (1)	11 (2)	26	48	20
<i>Buchnera aphidicola</i> SG	546 (1)	0 (0)	30	48	
<i>Caulobacter crescentus</i> CB15	3737 (1)	0 (0)	65	55	23
<i>Campylobacter jejuni</i> NCTC11168	1629 (1)	0 (0)	30	48	37
<i>Escherichia coli</i> K12	4242 (1)	0 (0)	51	54	37
<i>Escherichia coli</i> O157: H7	5253 (1)	88 (2)	50	54	37
<i>Haemophilus influenzae</i> Rd	1657 (1)	0 (0)	38	51	36
<i>Helicobacter pylori</i> 26695	1576 (1)	0 (0)	39	49	37
<i>Mesorhizobium loti</i> MAFF303099	6743 (1)	529 (2)	63	54	26
<i>Neisseria meningitidis</i> MC58	2079 (1)	0 (0)	52	52	36
<i>Pasteurella multocida</i> PM70	2015 (1)	0 (0)	38	51	37
<i>Pseudomonas aeruginosa</i> PAO1	5567 (1)	0 (0)	67	53	37
<i>Ralstonia solanacearum</i> GMI1000	3440 (1)	1676 (1)	67	54	32
<i>Rickettsia conorii</i> Malisch 7	1374 (1)	0 (0)	33	49	35
<i>Rickettsia prowazekii</i> Madrid E	835 (1)	0 (0)	29	48	35
<i>Salmonella enterica</i> serovar cholerasuis	4445 (1)	221 (2)	52	54	37
<i>Salmonella typhimurium</i> LT2	4425 (1)	102 (1)	53	54	37
<i>Sinorhizobium meliloti</i> 1021	3341 (1)	2864 (2)	63	54	26
<i>Vibrio cholerae</i> N16961	3835 (2)	0 (0)	47	53	37
<i>Xanthomonas axonopodis</i> Citri 306	4312 (1)	115 (2)	65	54	26
<i>Xanthomonas campestris</i> ATCC33913	4181 (1)	0 (0)	65	54	26
<i>Xylella fastidiosa</i> 9a5c	2766 (1)	66 (2)	53	53	26
<i>Yersinia pestis</i> CO92	3885 (1)	182 (3)	48	53	29
Other Bacteria (32 Species)					
<i>Aquifex aeolicus</i> VF5	1529 (1)	31 (1)	43	65	95
<i>Bacillus halodurans</i> C125	4066 (1)	0 (0)	44	54	30
<i>Bacillus subtilis</i> 168	4105 (1)	0 (0)	43	54	39
<i>Borrelia burgdorferi</i> B31DSM4680 ^T	851 (1)	788(21)	29	45	37
<i>Chlamydia muridarum</i> Nigg	911 (1)	0 (0)	40	50	37
<i>Chlamydia trachomatis</i> D	895 (1)	0 (0)	41	50	37
<i>Chlamydophila pneumoniae</i> CWL029	1052 (1)	0 (0)	41	48	37
<i>Chlorobium tepidum</i> TLS DSM12025 ^T	2252 (1)	0 (0)	57	52	48
<i>Clostridium acetobutylicum</i> ATCC824 ^T	3672 (1)	172 (1)	28	50	37
<i>Clostridium perfringens</i> 13	2660 (1)	63 (1)	29	52	45
<i>Corynebacterium glutamicum</i> ATCC13032 ^T	3057 (1)	0 (0)	54	54	40
<i>Deinococcus radiodurans</i> R1 DSM20539 ^T	2997 (2)	185 (2)	66	57	37
<i>Fusobacterium nucleatum</i> ATCC25586 ^T	2067 (1)	0 (0)	28	47	37
<i>Lactococcus lactis</i> IL1403	2321 (1)	0 (0)	35	50	26
<i>Listeria innocua</i> Clip11262	2968 (1)	75 (1)	37	53	34
<i>Listeria monocytogenes</i> EGD	2846 (1)	0 (0)	39	53	34
<i>Mycobacterium leprae</i> TN	1605 (1)	0 (0)	58	54	37
<i>Mycobacterium tuberculosis</i> H37Rv	3917 (1)	0 (0)	66	57	37
<i>Mycoplasma genitalium</i> G37	484 (1)	0 (0)	32	46	37
<i>Mycoplasma pneumoniae</i> M129	689 (1)	0 (0)	40	46	37
<i>Mycoplasma pulmonis</i> UAB CTIP	782 (1)	0 (0)	27	46	37
<i>Nostoc</i> sp. PCC7120	5366 (1)	689 (6)	41	50	26
<i>Staphylococcus aureus</i> MW2	2632 (1)	0 (0)	34	51	34
<i>Streptococcus pneumoniae</i> R6	2043 (1)	0 (0)	40	51	37
<i>Streptococcus pyogenes</i> SF370	1697 (1)	0 (0)	38	51	37
<i>Streptomyces coelicolor</i> A3 (2)	7769 (1)	385 (2)	72	58	26

Table 1. (Continued)

	genes				
	chr	plas	C:G(DNA)	C:G(RNA)	T_{opt}
Other Bacteria (32 Species) (Continued)					
<i>Synechocystis PCC6803</i>	3167 (1)	0 (0)	48	53	29
<i>Thermoanaerobacter tengcongensis</i> MB4	2588 (1)	0 (0)	38	59	75
<i>Thermosynechococcus elongatus</i> BP1	2475 (1)	0 (0)	54	53	55
<i>Thermotoga maritima</i> MSB8 DSM3109 ^T	1858 (1)	0 (0)	46	63	80
<i>Treponema pallidum</i> Nichols	1036 (1)	0 (0)	53	53	37
<i>Ureaplasma urealyticum</i> ATCC700970	614 (1)	0 (0)	26	46	37
Eukaryotes (2 Species)					
<i>Caenorhabditis elegans</i>	17085 (6)				
<i>Saccharomyces cerevisiae</i>			6333 (16) (also includes 28 mitochondrial genes)		

^a The number of archived chromosomes and plasmids are within parentheses.

result from duplication or gene transfer and do not necessarily share the same function. Computational methods for identification of orthologues attempt to maximize the number of unique or best homologues (above a certain threshold above other homologues) using approximate string-matching and graph-matching techniques while at the same time minimizing the numbers of paralogues or false positives.

To computationally identify orthologues, the amino acids in the protein coding regions were compared using Goldie 2.0, a software library (25). This involves progressive pairwise genome comparisons (8, 17) first using BLAST (26) with an expect-value of 10^{-3} to prune the search space by removing the dissimilar gene pairs, then using the local Smith-Waterman alignment (27) to identify the matches between pairs of similar genes, and finally modeling the pair of whole genomes as a bipartite graph-matching problem (25) to identify the best matching gene pairs. Computationally, orthologues were determined by identifying the best match between two genes (above a certain threshold) and deleting the remaining similarity edges from the orthologue pairs. Paralogues were pruned by identifying the low similarity value between a group of genes which could not be clearly separated. We have shown previously that changing the expectation value (*E*-value) from 10^{-3} to 10^{-5} would only affect the outcome by approximately 3% (8). Our scheme is effective when compared with those based solely upon lower expectation value of BLAST comparisons since we use the best possible match among all gene pairs using bipartite graph matching. Statistical analysis was used to identify the mean and median identity of orthologues.

Approximate sequence matching does not always predict the same functionality. However, in the absence of cost-prohibitive experimental verification of function, sequence-based comparisons are good indicators. At any cutoff, there will always be some paralogues present. Inevitably, some orthologues also will be lost, thus resulting in a bias toward higher identity in the remaining sequences. We believe that our choice of *E*-value is the best compromise for reducing the magnitude of the computational overhead without the loss of significant orthologues and without biasing the result for evolutionarily distant organisms (8). It is apparent from the distribution of sequence identities we obtained that only a fraction of orthologues were lost and that they would not account for more than a percentage point or two in the mean of the distribution. On the other hand, confining analysis to the forty or so universal genes would not provide significant

advantage over single gene analysis. Consequently, to carry out whole genome analysis, it was necessary to include all orthologues that were identified. For each species pair, a somewhat different set of genes were compared. This too could introduce some error into the analysis, but we expect that randomly selected large sets of orthologues should contain similar distributions of rapidly and slowly evolving genes.

Ribosomal RNA data were extracted from various sources: Genbank, EBI rRNA database, and LLNL 16S rRNA database. The data were verified by matching against one another, aligning the rRNAs from similar species, and blasting against genome data in Genbank. This was necessary to correct a small number of annotation errors in some submitted genomes in Genbank. Ribosomal RNAs from different genomes were also compared pairwise using the Smith-Waterman local alignment, and the identity value was counted for each ribosomal RNA pair. Multiple sequence alignment software was not used in order to be consistent with the method used in whole genome comparison and to avoid inherent imprecision present in multiple sequence alignment techniques.

Figure 1 was drawn according to our previous paper (8). That is, the orthologues in species 1 that are present in any other species were expressed as a percentage of the total numbers of genes in species 1. These values were then plotted against the total numbers of genes present in species 2. Positive and negative deviations from the straight line would reflect greater or lesser degrees of relatedness. For Figure 2, sequences of orthologues were aligned, percentage identities were determined, and the numbers of pairwise comparisons that gave particular values of sequence identity were plotted against sequence identity. Because there are only 16 archaea and 57 bacteria in the study, the frequency is much smaller for the archaea.

In our study (see Figure 1), we used *Deinococcus radiodurans* as a control, since it has no close relatives in the dataset, but is typical of all species of prokaryotes in that it shares the same overlap in gene content with the archaea as with the bacteria.

RESULTS AND DISCUSSION

Based upon the NGC method (8), comparison of 75 genomes confirmed that archaea share essentially the same overall set of orthologous genes as bacteria, but eukaryotes

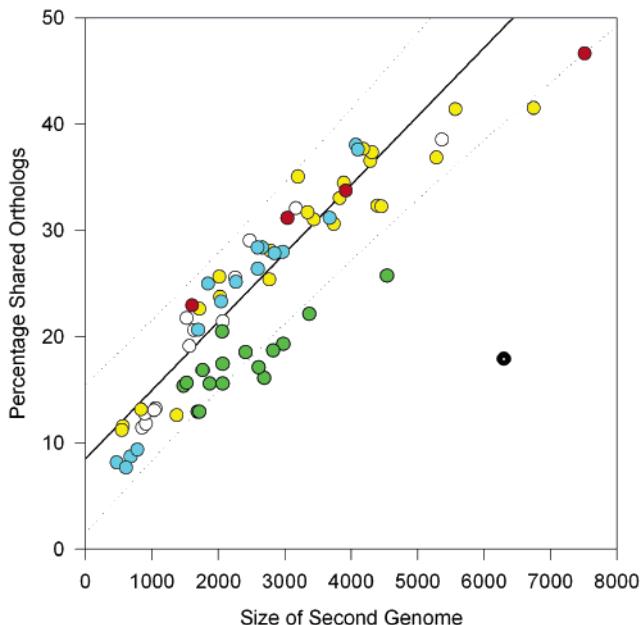


FIGURE 1: Normalized gene content of *Deinococcus radiodurans* with respect to other bacteria. Archaeabacteria are colored green, proteobacteria yellow, firmicutes blue, actinobacteria red, and the remainder represented as open circles. Yeast is shown in black. The bacterial data were fit to a straight line (slope 0.0058, intercept 10, standard deviation 3.3). The dotted lines are at two standard deviations.

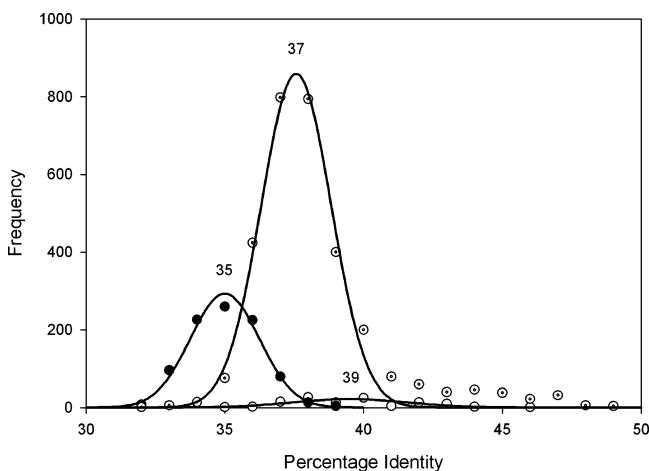


FIGURE 2: Median identity of orthologues. The archaea (open circles) were plotted separately from the bacteria (dotted circles) and the cross-comparison of archaea with bacteria (filled circles). The mean of the median identity of orthologues for archaea is 39.5% with standard deviation 2.0%. The mean for bacteria is 37.6% with standard deviation 1.3%. The mean for the cross-comparison is 35.0% with standard deviation 1.2%. The data for archaea and bacteria at identities greater than 50% are not shown.

have significantly fewer orthologues in common with either archaea or bacteria (see Figure 1). If archaea actually belonged to a separate domain of life, they should have significantly fewer genes in common with bacteria as well as with eukaryotes, but this is not the case. The archaea do have somewhat fewer genes in common with the other bacteria, but on average by a marginal 1.8 standard deviations in the example shown. In contrast, yeast has significantly fewer genes in common with prokaryotes by 9 and worm by 24 standard deviations. Thus, we now have stronger quantitative evidence based upon gene content that archaea are much more like bacteria than they are to eukaryotes at

the gene-content level. That is, on a whole genome basis, the archaea are virtually indistinguishable from the bacteria in terms of gene content. While this result is not novel (see ref 8), it was important to demonstrate consistency and to show that the data were not biased by the size of the database.

Using the MIO method with 35 genomes (8), we previously showed that there is an apparent limit to change in proteins of 36.9% identity. Using twice as many genomes, involving 16 archaea and 57 bacteria, we clearly establish that the limit does not significantly change with the size of the database (now 37.1%, SD 1.9%) and that archaeal proteins (which we will label A) have reached virtually the same limit to change as have the bacterial proteins (B), i.e., 39.5% A/A vs 37.6% B/B median identity. Cross-comparison of the archaeal and bacterial genes results in a median of 35.1% A/B, which overlaps extensively with the other two curves: bacteria against bacteria and archaea against archaea, resulting in an overall distribution that is essentially unimodal as illustrated in Figure 2. That is, the difference in the two distributions is at most 4% or about one to two standard deviations. This is in fact insignificantly small (by nearly an order of magnitude) in comparison to the relatively large differences in rRNA as shown below. To verify that the small variation in MIO between archaea and bacteria is insignificant, we compared 25 genomes in the “proteobacterial phylum” against the remaining 32 bacterial genomes. We found that the proteobacterial proteins had reached a limit of 40% median identity among themselves and 37% with the remaining species, which is not unlike the archaeal distribution. This indicates that the orthologous genes approach a functional limit to change at the currently recognized taxonomic level of phylum as necessitated by the structure/function relationship, which is represented by the mean of the distribution, i.e., 37%. Because sequence differences in orthologues approach a limit to change at or before the level of phylum, they do not allow inferences about relationships above that taxonomic rank whether comparing archaea with bacteria or the most divergent bacteria with one another. Only those comparisons at two standard deviations or more above the mean or 41% identity approach significance. Thus, it is surprising that rRNA sequence comparisons, which should follow the same rules that regulate the structure/function relationship in proteins, suggest a third domain of life (7), which is supposed to be two taxonomic levels above the phylum.

In agreement with previous reports, we found that archaeal rRNA is in fact different from most bacterial rRNA, which is shown by the clearly bimodal distribution of Figure 3. We tentatively conclude that bacterial rRNA reaches a limit to change of 76% identity and archaeal rRNA reaches a limit of 77% identity, which is essentially the same. On the other hand, cross-comparison of archaeal and mesophilic bacterial rRNA results in a virtually nonoverlapping distribution with a mean of 64% identity, indicating that they are indeed different. We have thus eliminated the uncertainty arising from different approaches to data analysis followed by many labs over the years and that due to variable data sets. There is little doubt that archaeal rRNA is orthologous and has essentially the same functional nuances as bacterial rRNA in spite of the fact that there are often multiple rRNA operons per species. It is rare that these multiple operons show significant intraspecies differences. Thus, the reason for the

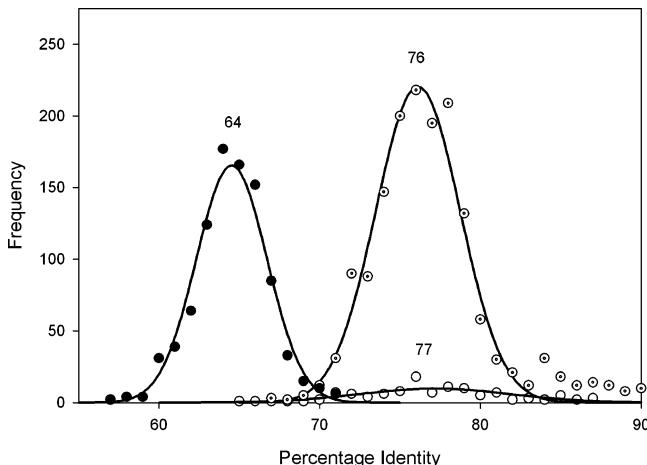


FIGURE 3: Percentage identity of 16S rRNA for archaea (open circles), bacteria (dotted circles), and the cross-comparison of archaea with bacteria (filled circles). The mean for archaea is 77.2% with standard deviation 4.7%, that for bacteria is 76.1% with standard deviation 2.6, and that for the cross-comparison is 64.5% with standard deviation 2.2. Data above 90% identity were not plotted.

discrepancy between whole genome and rRNA analyses is not immediately obvious since gene duplication, transfer, and loss can be eliminated as primary causes.

There are two ways to interpret these results: (1) the rRNA may not have reached a functional limit to change until 64% rather than 76% identity and it follows that the archaeal genomes (as opposed to the rRNA molecule alone) may actually be different from bacteria as previously reported, and (2) the rRNA may not be representative of the whole genome and there may be unappreciated environmental constraints leading to the 12% difference, such as differing base compositions of the rRNAs. The first explanation seems unlikely since thousands of genes appear to have reached a single functional limit to change. On the other hand, base compositional variation due to a hyperthermophilic life style was previously recognized to lead to artifacts in construction of trees, but it was thought to be a localized effect only (23, 24).

We considered base compositional bias to be a possible explanation for the more global discrepancy between rRNA and whole genome comparisons considering that the majority of archaea studied to date are thermophiles and the majority of bacteria are mesophiles. Because proteins contain 20 different amino acids, an imbalance in the composition of any one amino acid will not have a significant impact on sequence comparisons except in very unusual circumstances. In fact, it has been reported that archaea do have an imbalance in their amino acid compositions in favor of the hydrophobic (Ile, Val, and Leu) and charged (Glu, Lys, and Arg) amino acid residues at the expense of Gln, Asp, Asn, Ser, Thr, Cys, His, and Ala (22). This appears to be due in part to the fact that the strength of hydrophobic interactions increases with temperature. Consequently, protein stability to thermal denaturation also increases with the proportion of hydrophobic residues. However, this does not appear to have much impact on protein sequence identities as indicated in the current study by the cross-comparison of archaea and bacteria in MIO, which is only 2% lower than the overall distribution. On the other hand, nucleic acids contain only 4 different bases, and imbalances in the cytosine:guanine (C:

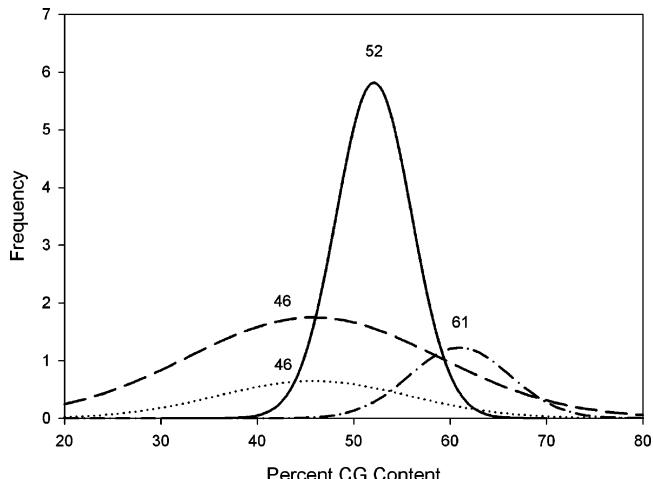


FIGURE 4: CG content of chromosomal DNA and of combined 16S plus 23S rRNA. Note that chromosomal DNA has a broader distribution than for rRNA and averages 46% for both archaea (dotted line) and bacteria (dashed line). The bacterial rRNA (solid line) has a mean of 52% with standard deviation of 3.9% and the archaea (dot-dashed line) 61% with standard deviation 5.2%.

G) content conceivably could have a rather dramatic effect on apparent sequence identities. Unless the species being compared all had similar C:G contents, there could be a significant impact of base composition on DNA and rRNA sequence identities. The C:G contents of genomic DNA range from 26% to 72% for the 73 species in our study and from 22% to 77% for 764 species previously surveyed (19). Although no correlation was found between overall DNA base composition and growth temperature, there is a strong effect for rRNA where there is a dramatic increase in CG content with temperature (19). This effect is primarily seen in the hydrogen-bonded stem regions, but the single stranded regions also have an imbalance in A over C (20).

To determine if there is a compositional bias in the rRNA of our study, we combined analyses of both 16S and 23S, which separately gave a similar outcome, as shown in Figure 4. These results were compared with those for chromosomal DNA. Remarkably, the average base composition for the archaeal and bacterial chromosomal DNA in our study is the same, 46%, which is in agreement with previous studies (19, 21, 22). However, archaeal rRNA is 15% higher in C:G content than is the chromosomal DNA whereas bacterial rRNA has a higher C:G content than the DNA by only 6%. It is apparent that the difference in base composition of archaeal vs bacterial rRNA is both large and real. We conclude that differences in base composition between archaeal and bacterial rRNA are strongly correlated with the bimodal distribution of sequence identities shown in Figure 3 and are large enough to account for the discrepancy, although there could be other contributing causes. We furthermore conclude that rRNA has a functional limit to change of 76% and that the 64% distribution is an artifact of base composition. This is to be contrasted with proteins where the extent of variation is potentially more than twice as large as shown above. Thus, for a 12% difference in rRNA sequence identities, the proteins should have shown approximately 30% difference rather than the 2% that was actually observed.

Why should the C:G content of archaeal rRNA be higher than that of bacteria? The majority of archaea (with the

exception of 3 species in our database) are hyperthermophiles with optimum growth temperatures above 59 °C, which is near the melting temperature of the nucleic acids, which require stabilization against thermal denaturation (*Halobacterium* sp. is the only archaeabacterial species that has lower C:G content of its rRNA than of its chromosomal DNA; it is one of the mesophiles and already has the highest chromosomal C:G content of any of the archaea). Thus, an increase in CG content stabilizes rRNA against thermal denaturation because C:G pairs are more stable than are A:U pairs due to the presence of three rather than two hydrogen bonds. Apparently, other mechanisms are more important for stabilization of DNA such as reverse gyrase (28). The majority of bacteria in our study are mesophiles with optimum growth temperatures near 37 °C, which is much below the melting temperatures of the nucleic acids. In this case, we would expect the base composition of the rRNA to randomly vary in much the same way as the DNA. In fact, the base composition of the bacterial rRNA is not much different from that of the chromosomal DNA. However, there are three hyperthermophilic bacteria in our study, *Aquifex aeolicus*, *Thermotoga maritima*, and *Thermoanaerobacter tengcongensis*, the first two of which have been reported to be the most divergent or deeply branching of all bacteria in the universal evolutionary tree. If higher C:G content of rRNA is necessitated by the higher growth temperatures of archaea, then the three hyperthermophilic bacteria should show a similar phenomenon. This is in fact the case. The bacterial species that have optimum growth temperatures similar to the archaea also have 18%, 20%, and 22% higher than average C:G contents in their rRNA. This provides strong support for the contention that the higher C:G content of rRNA stabilizes it against thermal denaturation in both the hyperthermophilic archaea and hyperthermophilic bacteria. Furthermore, it would account for the deep branching of *Aquifex* and *Thermotoga* in the evolutionary tree based upon 16S rRNA analysis which we believe is also artifactual. As indicated above, base compositional bias in hyperthermophilic rRNA had previously been recognized, but it was thought to have localized effects only on the position of *Archaeoglobus* and *Thermus* in evolutionary trees (23, 24). A correction for the base compositional bias by selectively excluding the most variable positions for evolutionary analysis was previously proposed (23). However, those supposed corrections for base compositional bias are obviously inadequate.

In conclusion, whole genome analysis utilizing NGC and MIO shows that the archaea are bacteria living under extreme environmental conditions, and have an elevated C:G content of their rRNA, which is a consequence of adaptation to a hyperthermophilic lifestyle. This erroneously makes them appear to be a separate life form when only rRNA sequences are compared and base compositional bias is ignored. Our results on whole genome comparisons underscore the danger of using single genes or combinations of single genes to infer relationships, because of the possibilities of gene transfer, vagaries of alignment, and gene duplication, but also due to the effect that base composition can have on the apparent relationships specifically when using rRNA. However, the archaeabacteria appear to be monophyletic and should be placed in a relatively high-ranking taxon. Prior to the three-domain hypothesis, they were placed in a separate bacterial

phylum, and we consider this appropriate because of their similar normalized gene content and median identity of orthologues. It was previously argued that archaeabacteria did not deserve the status of a separate domain of life because "the phenotypic difference between the two kinds of prokaryotes is minimal as compared with the difference between, let us say, a bacterium and a plant or animal" (29). It was also argued on the basis of shared insertions and deletions as well as on the basis of cellular membrane composition that archaea are bacteria and not a separate domain of life (30).

It should be emphasized that bacterial classifications as reflected, for example, in Bergey's Manual should take into account lifestyle, complexity, and metabolic pathways as well as molecular evolution. It is beyond the scope of this paper to consider the weight to be given to each characteristic used in classification; our analysis only considers molecular evolution, which many consider to be the basis for the primary division of species. It has been pointed out that the classification of the archaea not only is based upon molecular sequence comparisons but also includes other feature-based analyses and considers such properties as the unique metabolism of the methanogens and the particular adaptation of *Halobacterium* to high salt concentrations. When the individual genes and cofactors involved in these unique lifestyles are considered, archaea have many characteristics and genes in common with mesophilic bacteria. Archaea also have some unique conserved genes, but that is true of virtually any grouping of species; there are 22 such genes in archaea, but 19 in the other bacteria in our study (17). The deletion of nine ribosomal proteins in archaea may be related to their hyperthermophilic lifestyle. The deletion of conserved genes in bacterial pathogens may be due to utilization of the host's metabolic potential. Thus, the uniqueness argument is not compelling.

We believe that the differences in individual characteristics of the archaeabacteria are caused by gene selection in response to extreme environmental lifestyles such as elevated growth temperatures, that cannot be inferred by 16S rRNA or whole genome analyses alone. It follows that bacteria will likewise respond to environmental pressures in unique ways due to psychrophilic, halophilic, aerobic, anaerobic, acidiphilic, alkaliphilic, parasitic, or independent lifestyles. Whether any of these lifestyles will create sequence artifacts comparable to that seen for hyperthermophiles remains to be determined. The picture will become clearer as more bacterial genes and genomes are annotated and pathways are identified and biochemically studied experimentally in wet labs.

REFERENCES

1. Zuckerkandl, E., and Pauling, L. (1965) Molecules as documents of evolutionary history, *J. Theor. Biol.* 8, 357–366.
2. Margoliash, E., and Fitch W. M. (1968) Evolutionary variability of cytochrome c primary structures, *Ann. N.Y. Acad. Sci.* 151, 359–381.
3. Woese, C. R. (1987) Bacterial evolution, *Microbiol. Rev.* 51, 221–271.
4. Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., Chandra, S., McGarrell, D. M., Schmidt, T. M., Garrity, G. M., and Tiedje, J. M. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy, *Nucleic Acids Res.* 31, 442–443.

5. Garrity, G. M., Winters, M., Kuo, A. W., and Searles, D. B. (2002) *Bergery's Manual of Systematic Bacteriology*, 2nd ed., Springer-Verlag, New York.
6. Woese, C. R., and Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms, *Proc. Natl. Acad. Sci. U.S.A.* 74, 5088–5090.
7. Woese, C. R., Kandler, O., and Wheelis, M. L. (1990) Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya, *Proc. Natl. Acad. Sci. U.S.A.* 87, 4576–4579.
8. Bansal, A. K., and Meyer, T. E. (2002) Evolutionary analysis by whole-genome comparisons, *J. Bacteriol.* 184, 2260–2272.
9. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science* 269, 496–512.
10. Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002) Prokaryotic evolution in the light of gene transfer, *Mol. Biol. Evol.* 19, 2226–2238.
11. Doolittle, W. F. (1999) Phylogenetic classification and the universal tree, *Science* 284, 2124–2128.
12. Huynen, M. A., and Bork, P. (1998) Measuring genome evolution, *Proc. Natl. Acad. Sci. U.S.A.* 95, 5849–5856.
13. Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, D. R. (1997) 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, 619–637.
14. Fitz-Gibbon, S. T., and House, C. H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms, *Nucleic Acids Res.* 27, 4218–4222.
15. Graham, D. E., Overbeek, R., Olsen, G. J., and Woese, C. R. (2000) An archaeal genomic signature, *Proc. Natl. Acad. Sci. U.S.A.* 97, 3304–3308.
16. Teklaia, F., Lazcano, A., and Dujon, B. (1999) The genomic tree as revealed from whole proteome comparisons, *Genome Res.* 9, 550–557.
17. Bansal, A. K. (1999) An automated comparative analysis of 17 complete microbial genomes, *Bioinformatics* 15, 900–908.
18. Meyer, T. E., Cusanovich, M. A., and Kamen, M. D. (1986) Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees, *Proc. Natl. Acad. Sci. U.S.A.* 83, 217–220.
19. Galtier, N., and Lobry, J. R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes, *J. Mol. Evol.* 44, 632–636.
20. Wang, H., and Hickey, D. A. (2002) Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes, *Nucleic Acids Res.* 30, 2501–2507.
21. Hurst, L. D., and Merchant, A. R. (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes, *Proc. R. Soc. London, B* 268, 493–497.
22. Nakashima, H., Fukuchi, S., and Nishikawa, K. (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures, *J. Biochem.* 133, 507–513.
23. Weisburg, W. G., Giovannoni, S. J., and Woese, C. R. (1989) The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction, *Syst. Appl. Microbiol.* 11, 128–134.
24. Woese, C. R., Achenbach, L., Rouviere, P., and Mandelco, L. (1991) Archaeal phylogeny: Reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts, *Syst. Appl. Microbiol.* 14, 364–371.
25. Bansal, A. K., Bork, P., and Stuckey, P. J. (1998) Automated pairwise comparisons of microbial genomes, *Math. Model. Sci. Comput.* 9, 1–23.
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool, *J. Mol. Biol.* 215, 403–410.
27. Waterman, M. S. (1984) General methods for sequence comparison, *Bull. Math. Biol.* 46, 473–500.
28. Kampmann, M., and Stock, D. (2004) Reverse gyrase has heat-protective DNA chaperone activity independent of supercoiling, *Nucleic Acids Res.* 32, 3537–3545.
29. Mayr, E. (1998) Two empires or three?, *Proc. Natl. Acad. Sci. U.S.A.* 95, 9720–9723.
30. Gupta, R. S. (1998) 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, 695–707.

BI0502792