Submit to *Gene* (a draft)

Evolutionary Tree Reconstructed Based on Oligonucleotide Frequencies And Conserved Words in 16s Ribosomal RNA

Abstract

Evolutionary distance is defined by oligonucleotide (n-bases) frequency difference of two sequences. Phylogenetic tree is reconstructed using a set of 16S (18S) rRNA sequences and the definition of distance. The quality of trees generally improves with increasing n and reaches a plateau of best fit at n=7 or 8. So, the 7-mer or 8-mer frequencies provides a basis to describe rRNA evolution. Then, a group of (612 in total) 7-mers are deduced which are correlate well with evolution Representative conservative words for Bacteria and Archaea in 16S rRNA sequences have been found which are evolution-related oligonucleotides and located on nearly same sites of sequences for a vast range of organisms (in a kingdom). The structural meaning of these conservative words is discussed briefly.

1. **Introduction**

The investigation of oligo-nucleotide correlation in a DNA sequence is an important approach to the understanding of genetic language (Luo et al, 1998; Lobzin & Chechetkin, 2000). Trifonov et al (1986) has compiled a dictionary of oligo-nucleotide words from the statistical analysis of nucleotide frequencies in DNA sequences. In principle the oligo-nucleotide correlations in a DNA sequence can be studied by use of joint probabilities p_{abc} , p_{abcd}, etc or in general, p_σ where $\sigma = abc\ldots$) is an oligonucleotide of some length. For a sequence with p_σ much larger than the expected frequency in a random sequence, σ is a preferred word in this sequence. Oppositely, if p_σ is much smaller than the , then σ is suppressed or even forbidden word. In earlier sequence analyses many dinucleotide and trinucleotide preferred words and their evolutionary relations have been found (Burgel et al, 1992; Luo & Ji, 1997; Karlin et al, 1997). The forbidden and preferred words of six to eight nucleotides in some genomes have also been indicated by many authors (Hao & Lee, 1999; Blattuer et al, 1996).

To understand the meaning of a preferred or a forbidden word, apart from the concrete knowledge on the biological function of the DNA fragment, one should make a comparison of a group of evolution-related sequences and search for its occurrence commonly in different sequences. The strategy usually used is to make the multi-alignment of sequences and to find the conservative fragments. However, for a group of remote homologue sequences the multi-alignment is difficult to be used since there is no statistically significant sequence-similarity between them. Another problem related to multi-alignment is its large computationally intense if the sequence length is large enough.

However, one may investigate the problem from another point of view. If a group of preferred words is evolutionary conservative then they may play a role in reconstruction of the evolutionary tree.. So, one may study the relation between evolutionary tree and word frequency and try to find some evolution – related words by observing the tree reconstruction.

In this paper, instead of multi-alignment we shall suggest a new definition of sequence distance which is based on the calculation of n-tupe frequency difference, the difference of p_σ between two sequences. We call it n-distance. It measures the difference in the frequency distributions of oligonucleotides of n bases long in two sequences. It can be used for remote homologue sequences as well. The computation of an n-distance is far less intense than that in the alignment method. It neither requires sequence alignment nor depends on other sequences in the sequence set. In particular, it is independent on the size of the set of sequences being compared. We shall prove for ribosomal RNA, a clock-like gene, the partition tree deduced from multi-comparison of a group of evolution-related sequences based on this definition agrees well with biological evolutionary tree. So, we shall be able to investigate the relation between tree topology and n and deduce some information on the evolutionary meaning of oligo-nucleotide correlations. One interesting conclusion is: the branching of evolutionary tree is mainly controlled by oligo ($n \geq 7$) nucleotide frequencies in clock gene rRNA. To reconstruct evolutionary relation through oligo nucleotide frequencies -- this is the first motivation of the present study. The second motivation is related to the early evolution of ribosomal RNA. As is well known, the universal phylogenetic tree based on rRNA is a valid representation of organismal genealogy (Woese, 2000). The most interesting is its deepest branchings which has extended back to an era when cells were more primitive than today. Chief among molecular components at that time was the primitive translation apparatus, especially its RNA component. Though the horizontal gene flow (Deckert et al, 1998; Nelson et al, 1999) had severely jumbled the evolutionary histories, the universal tree based on ribosomal RNA still retained the clear vestige of the ground structure. One may assume that in rRNA sequence there exist some functional sites which are highly conservative in evolution. We shall demonstrate how to search for these conservative sites. After the oligonucleotide evolutionary tree has been reconstructed we are able to find these conservative oligonucleotides.

In this work, to construct the phylogenetic tree based on computing the n-distances, thirty–five 16s rRNA (for archaeons and bacteria) and 18s rRNA (for eukaryotes) sequences have been selected . Three methods for tree construction are employed , namely, the neighbor-joining (NJ) method, the unweighed pair-group mean arithmetic (UPGMA) method (see book by Li, 1998) and the fuzzy clustering (FC) method.(Luo and Ji, 1995). For simplicity, we call a tree constructed from *n*-distances an *n*-tree. Ideally the benchmark against which on can test the quality of an *n*-tree could be a tree extracted from the Tree of Life (Olsen et al, 1994; Cavalier-Smith, 1993; Paterson & Sogin, 1993) by removing all except the 35 organisms considered in this paper (called the life tree). It is not entirely fair to judge the usefulness of an n-distance by how close the resultant n- tree is to the life tree, because how organisms are grouped on a tree depends on the size and population of the tree. To gain an estimation of this size effect we construct a 35-organism tree using standard distance based on multiple alignment of the 35 rRNA sequences or these sequences in three kingdoms separately (called the alignment tree).

The requirement of a theoretically deduced tree consistent with life tree is a very strong constraint. It is a sensitive test on any evolutionary model. We find the $n \geq 7$ tree consistent well with the Tree of Life. It means the possible existence of some preferred words with length near or larger than 7, the frequency of which correlates with evolution, though they lie hidden in the strong background of noises. We shall find these words through following steps. At first, from calculation of the correlation between n-distance and the distance defined by single n-long- nucleotide frequency (called single-word-distance) we can find all n-long-oligonucleotides the frequency difference of which has meaningful correlation with evolutionary distance. They are called evolution – related oligonucleotides. In this work we have found 612 evolution – related oligonucleotides with n=7. The next step is to confirm their occurrence and determine the location of these oligonucleotides. When an evolution – related oligonucleotide occurs at the same or nearly same sites in rRNA sequences of different organisms we call it a conserved word in these species. By use of BLAST algorithm (Altschul et al, 1990) we have found many conserved words. The most interesting is some conserved words occurring in a vast of species, being special for a kingdom of species. We shall investigate the biological meaning of these words through inspection of the secondary and tertiary structure of ribosomal RNA. The full explanation of these words seems difficult and has not been given in this paper. Further works on this line are waited for.

Methods

Database. The 35 organisms – 9 archaeons, 19 bacteria and 7 eukaryotes – studied in constructing evolutionary tree and the accession number of their 16S/18S rRNA sequences are listed in Table 1. To facilitate the possibility of future comparative studies based on other genes we have chosen representative species whose genomes either have been completely sequenced or will soon be so. In the table each archaeon is coded by an upper-case Roman alphabet, each bacterium by a lower-case alphabet, each eukaryote by a non-alphabet symbol.

For studying the conserved words an extended set of 16S rRNA sequences was selected for the purpose of testing . Since the conserved words we are interested in are mainly in prokaryotes the test set includes 61 organisms – 20 archaeons and 41 bacteria which are selected by reference to Olsen, Woese and Overbeek (1994) . The names of organisms in test set are listed in Table 2.

All sequence data are taken from GenBank. The life tree (shown in Figure 1) is obtained from existing consensus, alignment-based trees by removing from them all organisms or species not included Table 1. Its three-kingdom topology is from Woese (1987). Its Archaea and Bacteria branches are reconstructed from the prokaryotic tree of Olsen, Woese and Overbeek (1994). The Eukarya branch is from Cavalier-Smith (1993) and Paterson & Sogin (1993). The rRNA alignment tree (the 35- sequence tree) is deduced based on multiple alignment of 35 sequences in table 1 by use of software package OMIGA 1.13 (Calvet, 1998). If the three kingdoms were assumed and 3 set of sequences were aligned separately, the 3-kingdom alignment tree is deduced.

The *n*- distance. Denote the probability of base *a* (*a* =A,G,C or T) occurring in a sequence by p_a , and the joint probability of base *a* and *b* occurring sequentially in the sequence by p_{ab} . In general, if $\sigma = abc\ldots$ is an oligonucleotide n bases long, denote the joint probabilities of the bases in σ, or relative frequency of σ, occurring in the sequence by p_σ . In the calculation of joint probabilities all sequences are assumed to be circular. For any *n* we always have $\sum_\sigma p_\sigma = 1$, where the summation over σ is over the set $\{\sigma\}$ of the 4^n oligonucleotides of length n. So long as n is much less than the sequence length N, with increasing n the set $\{\sigma\}$ is an increasingly fine-grained characterization of a sequence. Given two sequences Σ and Σ' with sets of joint probabilities { p_σ} and {p_σ '}, respectively, define a distance, called an *n*- distance , between the two sequences based on the difference in the two sets of joint probabilities as follows

$$E_n(\Sigma,\Sigma') = \sum_\sigma \left| p_\sigma - p_\sigma' \right|$$

$$(n = 1,2,\ldots) \qquad (1)$$

In the following when there is no confusion the arguments of E_n will be suppressed. An n-distance is well defined for sequences that of different lengths and are not aligned. By repeated application of relations such as

$$\left| p_\sigma - p'_\sigma \right| = \left| \sum_a (p_{\sigma a} - p'_{\sigma a}) \right| \le \sum_a \left| p_{\sigma a} - p'_{\sigma a} \right| \qquad (2)$$

where σ is any *n*-nucleotide and σa is an (*n*+1)- nucleotide, it can be deduced that

$$E_{n+1} \geq E_n \qquad n = 1,2,... \tag{3}$$

The limit of the increasing series {E_n } (*n*=1,2,…) is 2 (if neither sequence is a subsequence of the other, otherwise the limit is less than 2).

Given a set of organisms labeled by *i,j,k,…*, we can use the *n*-distance to obtain a distance matrix *D* for the set by having the matrix element D_{ij} equal to the E_n (Σ_i, Σ_j), where Σ_i is the sequence representing the organism *i*. By definition D has vanishing diagonal elements. An *n*-distance matrix will have insufficient differentiating power if n is either too small or too large. It is so when n is so small that the characterization of the sequences is too coarse grained. When n is too large E_n becomes binary – E_n =0 if the sequences are identical and E_n = 2 if the sequences are different – and loses its resolving power.

Now if the two sequences are aligned so that the aligned sequence length is *L*, then

$$E_1 = \frac{1}{L}\sum_a \left| n_{\bar{a}\to a} - n_{a\to\bar{a}} \right| \leq \frac{1}{L}\sum_a (n_{\bar{a}\to a} + n_{a\to\bar{a}}) \equiv 2M < 2 \tag{4}$$

where *M* is the total number of single mutations divided by *L*, namely the fraction of positions in which the two sequences differ, $n_{a\to\bar{a}}$ is the number of incidents where base *a* in the first sequence is either changed to another base or is missing in the second sequence, and $n_{\bar{a}\to a}$ is the number of incidents where either a blank or a base that is not *a* in the first sequence is changed to base *a* in second sequence. The parameter *M* is actually the conventional definition of evolutionary distance in the alignment approach. Because *M* neglects part of the effect of multiple mutation and because mutations can reduce as well as increase the difference between two sequences, the parameter *M* is actually a lower limit of evolutionary distance. For any two sequences there is always some *n* such that $E_n > M$. Where an *n*-distance emphasizes the role of nucleotide correlation in evolution, *M* basically counts single-base mutations. For long sequences an *n*- distance is insensitive to minor misalignments between two sequences.

There is practical reason for considering *n*-distances as alternatives for *M*. The computation time for an *n*-distances grows linearly with sequence length whereas, owing to the need for sequence alignments, that for *M* grows exponentially with sequence length.

The *n*- tree . For each n, $2 \leq n \leq 9$, we compute a distance matrix D for the 35 organisms in Tab 1, where the element D_{ij} is the n-distance E_n (Σ_i, Σ_j) computed using eq. (1) between the 16S/18S rRNA sequences Σ_i and Σ_j of the i-th and j-th organisms. In the computation no alignment is made of the sequences. Dendograms, or n-trees, are then constructed from the distance matrix using the UPGMA method, NJ method and FC method respectively. The former two are well-known methods (Li, 1998). For tree construction and plotting the software package PHYLIP version 3.5c has been used (Felsenstein, 1988). The FC method does not directly use the distances to construct a distance tree, rather it first converts the distances to a set of equivalence relations which are then used to construct a tree by partition. Given a distance matrix D one construct a similarity matrix $S = 1 - (1/2)\,D$. Because D is symmetric with vanishing diagonal elements, S is symmetric and reflexive ($S_{jj} = 1$). The element $S_{ij} = 1 - (1/2)\,D_{ij}$ therefore measures the closeness of the two objects i and j. Using the method of fuzzy clustering, one can compute from S a fuzzy equivalence matrix from which one can construct a partition tree based on alpha-cut technique (Luo & Ji, 1995). The NJ tree (n=7), the UPGMA tree (n=8) and the FC tree (n=8) are shown in Figure 2,3 and 4 respectively. Their comparisons with life tree are given in Table 3.

Evolution–related oligonucleotides with *n* = 7. For a set of 35 sequences, the distance matrix includes 35×34/2=595 different elements. In what follows these elements will be treated as being independent. If in eq.(1) the summation on the right-hand-side is removed, and a single term , that of the oligonucleotide σ, is retained, then a “single-word-distance” based on σ , $D_{sw}(\sigma)$, is obtained. Now define a correlation coefficient between D and $D_{sw}(\sigma)$ as

$$Cor\,(\sigma) = \mathrm{Cov}\,(D\,, D_{sw}(\sigma)\,)\,/\quad(\mathrm{Var}\,(D)\ \mathrm{Var}\,(D_{sw}(\sigma)))^{1/2} \qquad (5)$$

For a sampling size of 595, a value for *Cor* that is substantially greater than the threshold value (at 99% C.L.) of 0.11 may be considered as indicating special significance, in this case in the evolution process. Among all n=7 σ ’s 612 are found to have $Cor \geq 0.30$. Taking the latter value arbitrarily to be the cut-off value, we call these σ ’s n=7 evolution-related-oligonucleotides (ERO7s). Some examples of ERO7s with $Cor \geq 0.58$, together with their normalized occurrence frequencies in the 16S/18S rRNA sequences of the 35 organisms, are shown in Table 4.

Conserved words in three kingdoms

From set of ERO7s we find conserved words (CWs) in the three kingdoms by a procedure involving three steps: 1) Identify EROs longer than 7 bases; 2) Identify those EROs as candidate CWs whose relative positions in a large number of organisms (in Table 1) are approximately the same; 3) Identify as CWs those candidate CWs that also appear at approximately the same position in a larger number of organisms in the extended set of organisms in Table 2.

1. Searching for all EROs. At first we match 612 n=7 EROs in each sequence of Table 1. Note that some words are partly overlapped and they should be melted each other, forming a longer word. Then we collect all obtained words with length equal or larger than 8 in one database. For example, a part of 16SRNA sequence of *E.coli*, from 750 to 820 , are shown in the following

```
ctgacgctcaggtgcgaaagcgtggggagcaaacaggattagataccctggtagtccacgccgtaaacgat
-*-------------------------------****--*-**----*--*--------------------
```

where * means the starting site of an ERO7. There are two EROs, one of length 7, tgacgct, and another melted ERO of length 24, caggattagataccctggtagtcc, that are matched in the segment of 16S rRNA.

2. Matching the EROs (n≥8) in 35 rRNA sequences by use of BLAST program (Altschul et al, 1990). Observing the sites in sequences where the ERO occur we retain those matched words (conserved words) that they occur at the same or nearly same sites in rRNA sequences of different organisms . For example, the marching of word gcggtgaatacgt is as follows :

bact-q *Deinococcus Radiopugnan*

Query: 1 gcggtgaatacgt 13

Sbjct: 1316 gcggtgaatacgt 1328

bact-p *Flavobacterium heparinum*

Query: 1 gcggtgaatacgt 13

Sbjct: 1360 gcggtgaatacgt 1372

Query: 1 gcggtgaa 8

Sbjct: 683 gcggtgaa 690

The site (1316-1328) is near (1360-1372) and word gcggtgaatacgt is conserved in bact-q and bact-p. Another word gcggtgaa matches bact-p in two places but bact-q only in one place. So the word gcggtgaa as a part of gcggtgaatacgt should be retained but the word gcggtgaa (683-690) should be omitted.

3. Checking conservative words in test set. If the wrong matching and the inserting /deleting in some site are not permitted the matching is called a full one. There have been found many conservative words in above approach but full matched are few in a kingdom. Only considering full matching , after checking them in enlarged set of organisms we have obtained all conservative words in Bacteria and Archaea. The representative conservative words are listed in Table 4.

3. Results and Discussions

1) We have investigated how the *n*-tree changes with *n* for n< 10. We find that the quality of *n*-trees improves with increasing *n* when *n*≤6 , and reaches a plateau at *n*=7 or 8. The trend is general, irrelative with *n*-tree construction. The *n*=7 or 8 tree is the best one that is most similar to alignment tree. The overall pattern is: recognizable Archaea from *n*=2, formation of Eukarya as a separate group from *n*=4, and formation of the three kingdoms from *n*=7. However, for FC tree, the three kingdoms are recognizable even on the 2-tree. The detailed comparisons of best trees are given in Table 2. The number of moves needed to bring the branching pattern into agreement with the life tree is given in last column. It shows that the three best n-trees agree well with life tree. Their quality is comparable with alignment tree. For the set of 35 organisms chosen for this work, the branching pattern of the bacterial group is the key for ranking the trees. As seen from table 2, The n=7 NJ tree is the best *n*-tree through the comparison of all its details with life tree and alignment tree. Another interesting finding is the placing of the two thermotogales *A. Aeolicus* and *T. Maritima*, which display a persistent tendency to be "half-bacterial and half- archaeal" (Achenbach-Richter et al, 1987; Burggraf, 1992).

2) Our result shows that the n-distance is a good definition of evolutionary distance. The n-mer frequencies , as a set of evolutionary parameters, provide a reasonable basis to reproduce an evolutionary tree . The n-mer frequency reflect the oligo-nucleotide correlation in a DNA sequence . The frequency difference of n-mers between two DNA sequences describes their evolutionary distance. The definition surmount the limitation of conventional definition of evolutionary distance based on the number of single mutation. It emphasizes the role of nucleotide correlation in molecular evolution and especially suits for definition about distance for remote homologues. Our method is different from oligonucleotide catalog - an old method of using oligonucleotides to characterize a sequence (Fox et al, 1977; Woese and Fox, 1977)). Before the time when rRNAs could be completely sequenced they were characterized by their oligonucleotide catalogs. But in oligonucleotide catalog only a partial catalog of oligomers occurring in a sequence is actually generated through cleaving by nucleases. It does not give information on the frequencies of occurrence of the oligomers, so, no distance between two rRNAs can be defined in that method through frequency difference of oligomers. The success of our approach to reconstruct evolutionary tree indicates the conservation of oligo-nucleotide frequency. We have found a group of preferred words with length near 7, the frequency of which correlates with evolution. There are 612 in total 7-mers with correlation coefficient *Cor*≥0.3, including 126 with correlation coefficient *Cor* ≥0.5, 175 with 0.40≤ *Cor* <0.5, and 311 with 0.30≤ *Cor* <0.40. They are classified into 6 categories mainly, namely, 1, mainly existed in archae and eukaryotes; 2, mainly existed in eukaryotes; 3, mainly existed in archae and eubacteria; 4, mainly existed in eubacteria; 5, mainly existed in archae ; 6, mainly existed in eukaryotes and eubacteria. The last two categories occur only as *Cor* <0.4 . To save space only top four 7-mers with *Cor* >0.58 and next three from 7-mers with *Cor* =0.58 are listed in Table 4. They belong to the first and the second class respectively. The above result means there exist some 7-mers with frequency common to all organisms in one or two kingdoms. The occurrence frequency of these 7-mers in one organism is about ten times of the stochastic value (1/16384). They are highly conservative in evolution. So, the bifurcation of evolutionary tree is mainly determined and can be described by the occurrence of these oligomers.

3). From the comparison of locations of EROs in different organisms we have found a lot of conservative words. Each conservative word is located on nearly same sites of sequence for a vast range of organisms. They may change their location for different 16S rRNAs only smaller than 100 bases. Some representative conservative words of bacteria and archaea are summarized in Table 4. The conservation of 16S rRNA sequences has been investigated by many authors (Gutell et al, 1994). However, to our knowledge, the fully matched conserved word which is conserved in such a large range as a kingdom is firstly indicated by us. The word ggattagataccc in *E.coli* is located on end loop near H24 (Brimacombe, 1995). It is an active center responsible for subunit association of the ribosome molecule (Levin, 1995). The word is highly conservative in two kingdoms – Archaea and Bacteria - of species. It transcends the era of earliest branching of universal phylogenetic tree. So, the conservation of the word perhaps means the subunit association as the first important event in the evolution of primitive translation apparatus. Note that in *E.coli* the H24 is a P site tRNA footprint and H24(791) and H24(793) are IF-3 (initiation factor) footprint (Mueller and Brimacombe, 1997). The word aacgagcg in *E.coli* is located on a helix H35. This is a 8-bases long word and also conservative in Archaea and Bacteria. Interestingly, its expansion, a 32-bases long word, tgttgggttaagtcccgcaacgagcgcaaccc, is conservative in the kingdom Bacteria. This mean probably the expansion occurring in the bifurcation of Bacteria from universal tree. Another word aaactcaaa conservative in Bacteria is located between two helices, H27 and H2, while H2(912) and H2(912-915) are mutation sites causing resistance to streptomysin and footprint sites for streptomysin (Mueller et al, 1997). The structural information indicated above are gained by reference to 16S rRNA of *E.coli*. Since no relevant structural datum on 16S rRNA has been obtained for Archaea at present, we can't analyze the structural meaning about conservative word for Archaea temporarily . Though the concrete explanation on the meaning of these conserved words has not been given one may reasonably assume that these words are closely related to the basic structure and function of bacterial ribosome , related to the early evolution of the primitive translational apparatus. The archaeal phlogenetic tree in its root is divided into two major lineages. *Crenarchaeota* is one of the two branches (Woese, 1991). We have found some highly conservative words in this kingdom (including 9 organisms in total) that are also listed in Table 5 for reference in further study.

Acknowledgement This work was supported by grants from the National Science Foundation 90103030 to L.F.L and the National Science Council NSC87-2112-M-008-002 to H.C.L.

References

Achenbach-Richter, L., Gupta, R., Setter, K.O. & Woese, C.R. 1987. Were the original eubacteria thermophiles? Syst. Appl. Microbial. 9 , 34-39 .

Altschul, S.F.,et al. 1990. Basic local alignment aearch tool. J. Mol. Biol.,215, 403-410.

Blattuer, F.R. 1996. The complete genome sequence of E. coli. K-12. Science 277, 1453-1462.

Burge, C., Campbell, A.M. & Karlin, S. 1992. Over and under-representation of short oligonucleotide in DNA sequences. Proc. Natl. Acad. Sci. USA 89 , 1358-1362 .

Brimacombe,R. 1995. The structure of ribosomal RNA. Eur. J. Biochem. 230, 365-383.

Burggraf, S., Olsen, G.J., Stetter, K.O. & Woese, C.R. 1992. A phylogenetic analysis of Aquifex pyrophilus. Syst. Appl. Microbial. 15 , 352-356 .

Calvet, J.P. 1998. Comprehensive sequence analysis: OMIGA 1.1. Science 282 1057-1058 .

Cavalier-Smith, T. 1993 Kingdom protozoa and its 18 phyla. Microbiol. Rev. 57 , 953-994

Deckert, G. et al. 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature 392 , 335-358 .

Felsenstein, J 1998. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22 521-565 (1988). For the software package PHYLIP see the website [evolution.genetics.washington.edu/phylip/software.pars.html#PHYLP].

Fox, G.E., Peckman, K.J. & Woese, C.R. 1977. Comparative cataloging of 16S rRNA: molecular approach to prokaryotic systematics. Int. J. Syst. Bacteriol. 27 , 44-57 ; Fox, G.E., et el. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc. Natl. Acad. Sci. USA 74 , 4537-4541 .

Gutell, R., Larson, N., Woese,C.R. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. 58, 10-26.

Hao B.L. , Lee, H.C. 1999. Private communication.

Karlin, S., Mrazek, J. and Campbell, A.M. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriology 179 , 3899-3913.

Lake, J.A., Jain, R. & Rivera, M.C. 1998. Mix and match in the tree of life. Science 280 , 2027-2028.

Lewin, B. 1995. Gene 5 . Ch 9. Oxford Univ. Press

Li, W.H. 1997. Molecular Evolution. (Sinauer Associates)

Lobzin, V.V., Chechetkin, V.R. 2000. Order and correlation in genomic DNA sequences. Physics Uspekhi. 43, 55-78.

Luo, L.F., Ji, F.M. & Li, H. 1995. Fuzzy classification of nucleotide sequences and bacterial evolution. Bull. Math. Biol. 57 , 527-537 .

Luo, L.F. & Ji, F.M. 1997. The preferential mode analysis of DNA sequence. J. Theor. Biol. 188, 343-353.

Luo, L.F., et al. 1998. Statistical correlation of nucleotides in a DNA sequence. Phys. Rev. E 58 , 861-871 .

Mueller,F., Brimacombe,R. 1997. A new model for the 3-D folding of E.coli 16S ribosomal RNA. 1. J. Mol. Biol. 271, 524-544.

Mueller,F., et al 1997. A new model for the 3-D folding of E.coli 16S ribosomal RNA. 3. J. Mol. Biol. 271, 566-587.

Nelson, K.E. et al. 1999. Evidence for horizontal gene transfer between archaea and bacteria from genome sequence of T. maritima. Science 399 , 323-329

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

Patterson, D.J., & M.L. Sogin. 1993. Eukaryote origins and protistan diversity. In: The Origin and Evolution of Prokaryotic and Eukaryotic Cells. Eds. Hartman, H., and K. Matsuno. (World Scientific Pub.) pp. 13-46. See also the “Tree of Life" website: [phylogeny.arizona.edu/tree/eukaryotes/crown eukaryotes.html].

Trifonov, E.N., Brendel. 1986. Gnomic – A dictionary of genetic code. Balaban, Philadelphia.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. 51 , 221-271.

Woese, C.R. 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. U.S.A. 97, 8392.

Woese, C.R. & Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. USA 74 , 5088-5090 .

Woese, C.R. 1991. The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. Evolution at Molecular Level, Selander, R.K. et al. (Eds.) (Sinauer Associates).

Table 1: The 35 organisms, their single-letter/symbol codes and the accession numbers of the DNA sequences of their 16S/18S rRNA genes in Genbank.

Code	Organism	Accession no.
A	Aeropyrum pernix	AB019522
B	Pyrococcus horikoshii	D45214
C	Archaeoglobus fulgidus	Y00275
D	Methanococcus jannaschii	M59126
E	Methanobacterium thermoautotrophicum	Z37156
F	Thermoproteus tenax	M35966
G	Methanothermus fervidus	M32222
H	Sulfolobus solfataricus	X03235
L	Halobacterium volcanii	D11107
a	Escherichia coli	Z83204
b	Haemophilus influenzae	M35019 M59433
d	Helicobacter pylori	U00679
e	Rickettsia prowazekii	M21789
f	Bacillus subtilis	AF058766
g	Mycoplasma genitalium	X77334
h	Mycoplasma pneumoniae	M29061
i	Mycobacterium tuberculosis	X52917
j	Synechococcus sp.	D90916 AB001339
k	Borrelia burgdorferi	X98233 U78152
m	Treponema pallidum	M88726 M34266
n	Chlamydia trachomatis	D85720
o	Chlamydia pneumoniae	L06108
p	Flavobacterium heparinum	M11657 M61766 M81326
q	Deinococcus radiopugans	Y11334
r	Herpetosiphon aurantiacus	M34117
s	Chlorobium limicola	Y08102
y	Aquifex aeolicus	AE000657
z	Thermotoga maritima	AE001703
%	Homo sapiens	M10098
!	Mus musculus (mouse)	X00686
@	Solanum tuberosum (potato)	X67238
*	Glycine max (soybean)	X02623
#	Drosophila melanogaster	M21017
$	Caenorhabditis elegans	X03680
&	Saccharomyces cerevisiae (yeast)	J01353 M27607

Table 2: Extended set of 61 prokaryotes including 20 archaea and 41 bacteria

Archaea			
H. volcanii	H. halobium	H. morrhuae	T. acidophilum
M.stadtmanae	M.formicicum	M.bryantii	M.igneus
M.thermolithotrophicus	M.aeolicus	M.maripaludis	M.vannielii
M.voltae	T. celer	S. shibatae	P. occultum
D. mobilis	T. pendens	P. islandicum	P. aerophilum
Bacteria			
A. pyrophilus	P. miotherma	G. petraea	F. nodosum
T. melanesiensis	T. commune	C. aurantiacus	T. roseum
T. thermophilus	D. radiodurans	P. hollandica	A. cylindrica
N. sp.	L. monocytogenes	K. zopfii	G. haemolysans
M.hyopneumoniae	M. sualvi	M. hominis	M.gallisepticum
N. otitidis-caviarum	M. avium	F. aquatile	F. columnare
E. brevis	C. vibrioforme	T. pallidum	S. stenostrepta
S. litoralis	S. aurantia	C .psittaci	P. staleyi
I. pallida	C. jejuni	W. succinogenes	R. rickettsii
E. risticii	W. pipientis	V. parahaemolyticus	P. vulgaris
E. carotovora			

Table 3: Comparison of first few levels of branchings of Eukarya, Archaea and Bacteria on the various trees. Organisms are represented by codes given in Table 1. The last column gives the number of moves needed to bring the branching pattern into agreement with the life tree (tree of life).

Tree	Branching	pattern	No. of moves	
	Eukarya	Archaea		
Tree of Life	((@*)(&($(#(%!)))))	((H(AF))(B(D((CL)(EG)))))	-	-
Align. tree (35-sequ.)	($((#(%!))((@*)&)))	((H(AF))(L((D(BC))(EG))))	1	1
Align. tree (3-king.)	($(#((%!)((@*)&))))	((H(AF))(B(DC))(L(EG)))	1	1
n=7 NJ tree	($((#(%!))((@*)&)))	((F(AH))(L((D(BC))(EG))))	1	2
n=8 FC tree	($((#(%!))((@*)&)))	((F(AH))(L(D(C(B(EG))))))	1	2
n=8 UPGMA tree	($((#(%!))((@*)&)))	((F(AH))(L((D(BC))(EG))))	1	2
	Bacteria			
Tree of Life	((yz)(r(q(j((f(gh)i)(((no)(km)(ps))((ab)ed)))))))		-	
Align. tree (35-sequ.)	((yz)((q(fi))((j(no))(r((km)((ps)((gh)((ab)ed))))))))		3	
Align. tree (3-king.)	((yz)(r(q((fi)((j(no))((km)(((ps)(gh))((ab)ed))))))))		2	
n=7 NJ tree	((yz)(r((ij(fq))((gh)(no)(km)p)((ab)(ed)s))))		3	
n=8 FC tree	(r((gh)(s(d(p((no)((yz)(km)qj(fi)e(ab))))))))		4	
n=8 UPGMA tree	((gh)(r((yz)(d((no)(sp(km)qej(fi)(ab)))))))		5	

Table 4 Examples of Evolution – related – oligonucleotides with greater correlation coefficient and their frequencies (normalized to 3000 bases) in the 35 organisms of Table 1

oligo: aacttaa (*Cor* = 0.60)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	3
a	b	d	e	f	g	h	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: acttaaa (*Cor* = 0.60)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2
a	b	d	e	f	g	h	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: tccctgc (*Cor* = 0.60)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2
a	b	d	e	f	g	h	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: gaaactt (*Cor* = 0.59)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			2	2	2	3	2	2	2	2	2	2	2	4	4	4	2	2
a	b	d	e	f	g	h	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: ttgccaa (*Cor* = 0.58)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2
a	b	d	e	f	g	H	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: cttctta (*Cor* = 0.58)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2
a	b	d	e	f	g	H	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

oligo: gtctgtg (*Cor* = 0.58)

organism			A	B	C	D	L	E	F	G	H	%	!	@	*	$	&	#
frequency			0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2
a	b	d	e	f	g	H	i	j	k	m	n	o	z	y	p	q	r	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: Representative conserved words in the kingdoms of Bacteria and Archaea [a]

Conserved word	Site	Kingdom	Remarks [a,b]
1 ggattagataccc	785-797 [c]	Archa / Bact.	Archa. not in *M. igneus*. On end loop near H24 ; Cross link between A(794) and H23(693-696)
2 aacgagcg	1102-1109 [c]	Archa / Bact	Archa. not in *T. Acidophilum;* Bact. not in *C. Vibrioforme*; also found in *C. Elegans*. on H35
3 gacggtgag	711-719 [d]	Archa	Not in *H. Halobium*.
4 ccttgcacacac	1352-1363 [d]	Archa	Archa. not in *M.maliparudis;* Bact. only in *A.pyrophilus & A. aeolicus.*
5 aaactcaaa	907-915 [c]	Bact.	Bact. not in *M. hyopneumoniae, M. sualvi & M.hominis;* Archa. only in *D. Mobilis* . Between H27 and H2
6 tgggttaa	1086-1093 [c]	Bact.	Not in *I. Pallida, P. Staleyi, E. brevis, F. Columnare, F. Aquatile, M. hyopneumoniae, M. sualvi, M.hominis & C. Aurantiacus.* On H37 and its downstream end loop; Crosslink between loop(1090-1094) and H40(1161-1164)
7 accaccag	674-681 [d]	Archa.	Archa. all of *Crenarchaeota*; Bact. only in *C. aurantiacus;* also found in Euka except *C. elegans & S. cerevisae*.
8 gtagtcccg	759-767 [d]	Archa	Archa. all of *Crenarchaeota*; Bact. only in *A. pyrophilus & A. aeolicus.*
9 cccgtcgc	1366-1373 [d]	Archa.	Archa. all of *Crenarchaeota*; also found in Euka.

a) The conserved words listed are conservative only in Bacteria and/or in Archaea. They have not been found in Eukaryota near the sites quoted in the table unless the special indication has been given in remarks. Note that the organisms referred to are restricted to the 96 species included in Table 1 and 2. *b*) Structural information refers to 16S rRNA of *E. coli* (Mueller & Brimacombe, 1997). *c*) Sites given are in *E. coli.* Sites of conserved word in organisms other than *E. coli.* are near the sites quoted. *d*) Sites given are in *T. Tenax.* Sites of conserved word in organisms other than *T. Tenax.* are near the sites quoted.

Note: Figure 1 to Figure 4 should be taken from a preprint written by HC. The preprint I received from email is a full text in PDF format. I had not succeeded in obtaining these figures separately.

L.F. 2003-5-27